On "Generalized Finite Differences": Discretization of Electromagnetic Problems

Alain Bossavit[‡]

1 Preliminaries: Euclidean space

What we shall do in this first Section can be described as "deconstructing Euclidean space". Three-dimensional Euclidean space, denoted by E_3 here, is a relatively involved mathematical structure, made of an affine 3D space (more on this below), equipped with a metric and an orientation. By taking the Cartesian product of that with another Euclidean space, one-dimensional and meant to represent Time, one gets the mathematical framework in which most of classical physics is described. This framework is often taken for granted, and should not.

By this we do not mean to challenge the separation between space and (absolute) time, which would be getting off to a late start, by a good century. Relativity is not our concern here, because we won't deal with moving conductors, which makes it all right to adopt a privileged reference frame (the so-called laboratory frame) and a unique chronometry. The problem we perceive is with E_3 itself, too rich a structure in several respects. For one thing, orientation of space is *not* necessary. (How could it be? How could physical phenomena depend on this social convention by which we class right-handed and left-handed helices, such as shells or staircases?) And yet, properties of the cross product, or of the curl operator, so essential tools in electromagnetism, crucially depend on orientation. As for metric (i.e., the existence of a dot product, from which norms of vectors and distances between points are derived), which also seems to be involved in the two main equations, $\partial_t \mathrm{B} + \mathrm{rot}\, \mathrm{E} = 0$ (Faraday's law) and $-\partial_t \mathrm{D} + \mathrm{rot}\, \mathrm{H} = \mathrm{J}$ (Ampère's theorem), since the definition of rot depends on the metric, we shall discover that it plays no role there, actually, because a change of metric, in the description of some electromagnetic phenomenon, would change *both* rot *and* the vector fields E, B, etc., in such a way that the equations would stay unchanged. Metric is no less essential for that, but its intervention is limited to the expression of constitutive laws, that is, to what will replace in our notation the standard $\mathrm{B} = \mu \mathrm{H}$ and $\mathrm{D} = \epsilon \mathrm{E}$.[1]

Our purpose, therefore, is to separate the various layers present in the structure of E_3, to be able to use exactly what is needed, and nothing more, for each subpart of the Maxwell system of equations. That this can be done is no news: As reported by Post [76], the metric-free character of the two main Maxwell equations was independently pointed out by Cartan, as early as 1924, by Kottler [61] and by van Dantzig [31]. But the exploitation of this remark in the design of numerical schemes is a contemporary thing, which owes much to (again, working independently) Tonti [99, 98, 65] and Weiland [36, 104]. See also [93, 54, 97]. Even more recent [19, 65] is the realization that such attention to the underlying geometry would permit to soften the traditional distinctions between finite-difference, finite-element, and finite-volume approaches. In particular, it will be seen here that a common approach to error analysis applies to the three of them, which does rely on the existence of finite elements, but not on the variational methods that are often considered as foundational in finite element theory. These finite elements, moreover, are not of the Lagrange (node based) flavor. They are differential geometric objects, created long ago for other purposes, the Whitney forms [109], whose main characteristic is the interpretation they suggest of degrees of freedom (DoF) as integrals over geometric elements (edges, faces, ...) of the discretization mesh.

‡ Électricité de France, 92141 Clamart, France. Fax: 33 01 4765 4118. E-mail: Alain.Bossavit@der.edfgdf.fr.
Last update: March 25, 2002.

1 We shall most often ignore Ohm's law here, for shortness, and therefore, treat the current density J as a data. It would be straightforward to supplement the equations by the relation $\mathrm{J} = \sigma \mathrm{E} + \mathrm{J}^s$, where only the "source current" J^s is known in advance.

As a preparation to this deconstruction process, we need to recall a few notions of geometry and algebra which do not seem to get, in most curricula, the treatment they deserve. First on this agenda is the distinction between vector space and affine space.

1.1 Affine space

A *vector space*[2] on the reals is a set of objects called *vectors,* which one can (1) add together (in such a way that they form an Abelian group, the neutral element being the null vector) and (2) multiply by real numbers. No need to recall the axioms which harmonize these two groups of features. Our point is this: The three-dimensional vector space (for which our notation will be V_3) makes an awkward model of physical space,[3] unless one deals with situations with a privileged point, such as for instance a center of mass, which allows one to identify a spatial point x with the translation vector that sends this privileged point to x. Otherwise, the idea to add points, or to multiply them by a scalar, is ludicrous. On the other hand, taking the midpoint of two points, or more generally, barycenters, makes sense, and is an allowed operation in affine space, as will follow from the definition.

An *affine space* is a set on which a vector space, considered as an additive group, acts effectively, transitively and regularly. Let's elaborate.

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 a_g, such that a_1 is the identity map, and $a_{gh} = a_g a_h$. (Symbol 1 denotes the neutral element, and will later double for the group made of this unique element.) The action is *effective* if $a_g = 1$ implies $g = 1$, that is to say, if all nontrivial group elements "do something" to X. The *orbit* of x under the action is the set $\{a_g(x) : g \in G\}$ of transforms of x. Belonging to the same orbit is an equivalence relation between points. One says the action is *transitive* if all points are thus equivalent, i.e., if there is a single orbit. The *isotropy group* (or stabilizer, or little group) of x is the subgroup $G_x = \{g \in G : a_g(x) = x\}$ of elements of G which fix x. In the case of a transitive action, little groups of all points are conjugate (because $g_{xy} G_y = G_x g_{xy}$, where g_{xy} is any group element whose action takes x to y), and thus "the same" in some sense. A transitive action is *regular* (or *free*) if it has no fixed point, that is, if $G_x = 1$ for all x. If so is the case, X and G are in one-to-one correspondence, so they look very much alike. Yet they should not be identified, for they have quite distinctive structures. Hence the concept of *homogeneous space:* A set, X here, on which some group acts transitively and effectively. (A standard example is given by the two-dimensional sphere S_2 under the action of the group SO_3 of rotations around its center.) 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 neutral one. Selecting a point 0 in X (the origin) and then identifying $a_g(0)$ with g (and hence 0 in X with the neutral element of G) provides X with a group structure, but the isomorphism with G thus established is not canonical, and this group structure is most often irrelevant, just like the vector-space structure of 3D space.

Affine space is a case in point. Intuitively, take the n-dimensional vector space V_n, and forget about the origin: What remains is A_n, the affine space of dimension n. More rigorously, a vector space V, considered as an additive group, acts on itself (now considered as just a set, which we acknowledge by calling its elements *points*, instead of vectors) by the mappings[4]

[2] Most definitions will be implicit, with the defined term set, on first appearance, in *slanted* style. The same style is also used, occasionally, for emphasis.

[3] Taking $\mathbb{R}^3$, the set of triples of real numbers, with all the topological and metric properties inherited from $\mathbb{R}$, is even worse, for this implies that some basis $\{\partial_1, \partial_2, \partial_3\}$ has been selected in V_3, thanks to which a vector v writes as $v = \sum_i v^i \partial_i$, hence the identification between v and the triple $\{v^i\}$ of components (or coordinates of the point v stands for). In most situations which require mathematical modelling, no such basis imposes itself. There may exist privileged directions, as when the device to be modelled has some kind of translational invariance, but even this does not always mandate a choice of basis.

[4] We'll find it convenient to denote a map f by $x \to \text{Expr}(x)$, where Expr is the defining expression, and to

$a_v = x \to x + v$, 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 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. Reversing the viewpoint, one can say that an affine space A is a homogeneous space with respect to the action of some vector space V, considered as an additive group. (Points of A will be denoted x, y, etc., and $y - x$ will stand, by a natural notational abuse, for the vector that translates x to y.) The most common example is obtained by considering as equivalent, in some vector space V, two vectors u and v such that $v - u$ belong to some fixed vector subspace W. Each equivalence class has an obvious affine structure (W acts on it regularly by $v \to v + w$). Such a class is called an *affine subspace* of V, *parallel to W*.[5] Of course, no vector in such an affine subspace qualifies more than any other as origin, and calling its elements "points" rather than "vectors" is therefore appropriate.

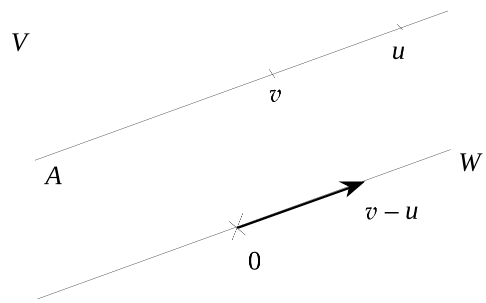

Figure 1. No point in the affine subspace A, parallel to W, can claim the role of "origin" there.

At this stage, we may introduce the *barycenter* of points x and y, with weights λ and $1 - \lambda$, as the translate $x + \lambda(y - x)$ of x by the vector $\lambda(y - x)$, and generalize to any number of points. The concepts of affine independence, dimension of the affine space, and affine subspaces follow from the similar ones about the vector space. *Barycentric coordinates*, with respect to $n + 1$ affinely independent points $\{a_0, \ldots, a_n\}$ in A_n are the weights $\lambda^i(x)$ such that $\sum_i \lambda^i(x) = 1$ and $\sum_i \lambda^i(x)(x - a_i) = 0$, which we shall feel free to write $x = \sum_i \lambda^i(x) a_i$. *Affine maps* on A_n are those that are linear with respect to the barycentric coordinates. 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 to the associate V, called the *tangent space at x*, denoted T_x. (I will call *free* vectors the elements of V, as opposed to vectors "at" some point, dubbed *bound* (or *anchored)* vectors. Be aware that this usage is not universal.) 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.[6] Note that vector fields are maps of type *POINT → BOUND_VECTOR*, actually, subject to the restriction that the value of v at x, notated $v(x)$, 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 tangent vector fields defined over a surface or a curve.

Homogeneous space is a key concept: Here is the mathematical construct by which we

link name and definition by writing $f = x \to \mathrm{Expr}(x)$. (The arrow is a "stronger link" than the equal sign in this expression.) In the same spirit, $X \to Y$ denotes the set of all maps "of type $X \to Y$", that is, maps from X to Y, not necessarily defined over all X. Points x for which f is defined form its *domain* $\mathrm{dom}(f) \subset X$, and their images form the *codomain* $\mathrm{cod}(f) \subset Y$, also called the *range* of f.

[5] Notice how the set of all affine subspaces parallel to W also constitutes an affine space under the action of V, or more pointedly—because then the action is regular—of the quotient space V/W. A "point", there, is a whole affine subspace.

[6] For a piecewise smooth manifold (see below), such a subspace may fail to exist at some points, which will not be a problem.

can best model humankind's *physical* experience of spatial homogeneity. Translating from a spatial location to another, we notice that similar experiments give similar results, hence the concept of invariance of the structure of space with respect to the group of such motions. By taking as mathematical model of space a homogeneous space relative to the action of this group (in which we recognize V_3, by observing how translations compose), we therefore acknowledge an essential *physical* property of the space we live in.

Remark. In fact, translational invariance is only approximately verified, so one should perhaps approach this basic modelling issue more cautiously: Imagine space as a seamless assembly of patches of affine space, each point covered by at least one of them, which is enough to capture the idea of *local* translational invariance of physical space. This idea gets realized with the concept of smooth manifold (see below) of dimension 3. What we shall eventually recognize as the metric-free part of the Maxwell's system (Ampère's and Faraday's laws) depends on the manifold structure only. Therefore, postulating an affine structure is a *modelling decision*, one that goes a trifle beyond what would strictly be necessary to account for the homogeneity of space, but will make some technical discussions easier when (about Whitney forms) barycentric coordinates will come to the fore. ◊

There is no notion of distance in affine space, but this doesn't mean no topology: Taking the preimages of neighborhoods of $\mathbb{R}^n$ under any one-to-one affine map gives a system of neighborhoods, hence a topology—the same for all such maps. (So we shall talk loosely of a "ball" or a "half ball" in reference to an affine one-to-one image of $B = \{\xi \in \mathbb{R}^n : \sum_i (\xi^i)^2 < 1\}$ or of $B \cap \{\xi : \xi^1 \geq 0\}$.) Continuity and differentiability thus make sense for a function f of type $A_p \to A_n$. In particular, the derivative of f at x is the linear map $\mathrm{D}f(x)$, from V_p to V_n, such that $|f(x+v) - f(x) - \mathrm{D}f(x)(v)|/|v| = o(|v|)$, if such a map exists, which does not depend on which norms $|\ |$ on V_p and V_n are used to check the property. The same symbol, $\mathrm{D}f(x)$, will be used for the *tangent map* that sends a vector v anchored at x to the vector $\mathrm{D}f(x)(v)$ anchored at $f(x)$.

1.2 Piecewise smooth manifolds

We will do without a formal treatment of manifolds. Most often, we shall just use the word as a generic term for lines, surfaces, or regions of space ($p = 1, 2, 3$, respectively), piecewise smooth (as defined in a moment), connected or not, with or without a boundary. A 0-manifold is a collection of isolated points.

For the rare cases when the general concept is evoked, suffice it to say that a p-dimensional manifold is a set M equipped with a set of maps of type $M \to \mathbb{R}^p$, called *charts*, which make M look, for all purposes, but only locally, like $\mathbb{R}^p$ (and hence, like p-dimensional affine space). *Smooth* manifolds are those for which the so-called *transition functions* $\varphi \circ \psi^{-1}$, for any pair $\{\varphi, \psi\}$ of charts, are smooth, i.e., possess derivatives of all orders. (So-called C^k manifolds obtain when continuous derivatives exist up to order k.) Then, if some property P makes sense for functions of type $\mathbb{R}^p \to X$, where X is some target space, f from M to X is reputed to have property P if all composite functions $f \circ \varphi^{-1}$, now of type $\mathbb{R}^p \to X$, have it. A manifold M *with boundary* has points where it "looks, locally, like" a closed half-space of $\mathbb{R}^p$; these points form, taken together, a (boundaryless) $(p-1)$-manifold ∂M, called the *boundary* of M. Connectedness is not required: A manifold can be in several pieces, all of the same dimension p.

In practice, our manifolds will be glued assemblies of *cells*, as follows.

First, let us define "reference cells" in $\mathbb{R}^p$, as illustrated on Fig. 2. These are bounded convex polytopes of the form

$$(1) \qquad K_p^\alpha = \{\xi \in \mathbb{R}^p : \xi^l \geq 0 \ \forall\, l = 1, \ldots, p, \ \sum_{j=1}^{p} \alpha_j^i \xi^j \leq 1 \ \forall\, i = 1, \ldots, k\},$$

where the α_j^i's form a rectangular $(k \times p)$-matrix with nonnegative entries, and no redundant rows.

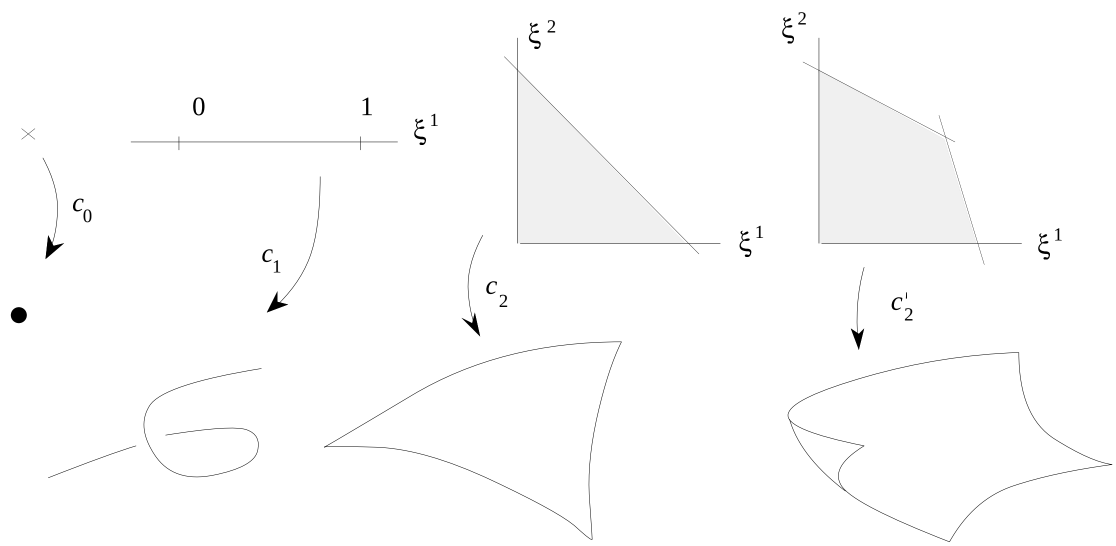

Figure 2. Some cells in A_3, of dimensions 0, 1, 2.

Now, a *p-cell* in A_n, with $0 \le p \le n$, is a smooth map c from some K_p^α into A_n, one-to-one, and such that the derivative $\mathrm{D}c(\xi)$ has rank p for all ξ in K_p^α. (These restrictions, which qualify c as an *embedding*, are meant to exclude double points, and cusps, pleats, etc., which smoothness alone is not enough to warrant.) The same symbol c will serve for the map and for the image $c(K_p^\alpha)$. The *boundary* ∂c of the cell is the image under c of the topological boundary of K_p^α, i.e., of points ξ for which at least one equality holds in (1). Remark that ∂c is an assembly of $(p-1)$-cells, which themselves intersect, if they do, along parts of their boundaries.

Thus, a 0-cell is just a point. A 1-cell, or "path", is a simple parameterized curve. The simplest 2-cell is the triangular "patch", a smooth embedding of the triangle $\{\xi : \xi^1 \ge 0, \xi^2 \ge 0, \xi^1 + \xi^2 \le 1\}$. The definition is intended to leave room for polygonal patches as well, and for three-dimensional "blobs", i.e., smooth embeddings of convex polyhedra.

We shall have use for the *open* cell corresponding to a cell c (then called a *closed* cell for contrast), defined as the restriction of c to the interior of its reference cell.

A subset M of A_n will be called a *piecewise smooth* p-manifold if (1) There exists a finite family $\mathcal{C} = \{c_i : i = 1, \ldots, m\}$ of p-cells whose union is M, (2) The open cell corresponding to c_i intersects no other cell, (3) Intersections $c_i \cap c_j$ are piecewise smooth $(p-1)$-manifolds (the recursive twist in this clause disentangles at $p = 0$), (4) The cells are properly joined at their boundaries,[7] i.e., in such a way that each point of M has a neighborhood in M homeomorphic to either a p-ball or half a p-ball.

Informally, therefore, piecewise smooth manifolds are glued assemblies of cells, obtained by topological identification of parts of their respective boundaries. (Surface S in Fig. 10, p. 12, is typical.)

Having introduced this category of objects—which we shall just call manifolds, from now on—we should, as it is the rule and almost a reflex in mathematical work, deal with maps between such objects, called *morphisms*, that preserve their relevant structures. About cells, first: A map between two images of the same reference cell which is bijective and smooth (in both directions) is called a *diffeomorphism*. Now, about our manifolds: There is a *piecewise smooth diffeomorphism* between two of them (and there too, we shall usually dispense with

[7] This is regrettably technical, but it can't be helped, if M is to be a manifold. The assembly of *three* curves with a common endpoint, for instance, is not a manifold. See also [50] for examples of 3D-spaces obtained by identification of faces of some polyhedra, which fail to be manifolds. Condition (2) forbids self-intersections, which is overly drastic and could be avoided, but will not be too restrictive in practice.

the "piecewise smooth" qualifier) if they are homeomorphic and can be chopped into two sets of cells which are, two by two, diffeomorphic.

1.3 Orientation

To get oneself oriented, in the vernacular, consists in knowing where is South, which way is uptown, etc. To orient a map, one makes its upper side face North. Pigeons, and some persons, have a sense of orientation. And so forth. *Nothing* of this kind is implied by the mathematical concept of orientation—which may explain why so simple a notion may be so puzzling to many. Not that mathematical orientation has no counterpart in everyday's life, it has, but in something else: When entering a roundabout or a circle with a car, you know whether you should turn clockwise or counterclockwise. *That* is orientation, as regards the ground's surface. Notice how it depends on customs and law. For the spatial version of it, observe what "right-handed" means, as applied to a staircase or a corkscrew.

1.3.1 Oriented spaces

Now let us give the formal definition. 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 for each frame, look at the determinant of its n vectors, as expressed in this basis, hence a $FRAME \rightarrow REAL$ function. This function is basis-dependent, but the equivalence relation defined by "$f \equiv f'$ if and only if frames f and f' have determinants of the same sign" does not depend on the chosen 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, either *direct* or *skew*, two equivalent frames getting the same label. There are two such functions, therefore two possible orientations. An *oriented vector space* is thus a pair $\{V, Or\}$, where Or is one of the two orientation classes of V. (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.) We shall find convenient to extend the notion to a vector space of dimension 0 (i.e., one reduced to the single element 0), to which also correspond, by convention, two oriented vector spaces, labelled + and −.

Remark. Once a vector space has been oriented, there are direct and skew *frames*, but there is no such thing as direct or skew *vectors*, except, one may concede, in dimension 1. A vector does not acquire new features just because the space where it belongs has been oriented! Part of the confusion around the notion of "axial" (vs. "polar") vectors stems from this semantic difficulty [12, p. 296]. As axial vectors will not be used here, the following description should be enough to deal with the issue. Let's agree that, if Or is one of the orientation classes of V, the expression $-Or$ denotes the other class. Now, form pairs $\{v, Or\}$, where v is a vector and Or any orientation class of V, and consider two pairs $\{v, Or\}$ and $\{v', Or'\}$ as equivalent when $v = -v'$ and $Or = -Or'$. *Axial vectors* are, by definition, the equivalence classes of such pairs. (*Polar* vectors is just a redundant name, inspired by a well-minded sense of equity, for vectors of V.) Notice that axial *scalars* can be defined in the same way: substitute a real number for v. Hence axial vector fields and axial functions (more likely to be dubbed "pseudo-functions" in the physical literature). The point of defining such objects is to become able to express Maxwell's equations in *non*-oriented Euclidean space, i.e., V_3 with a dot product but no specific orientation. See [13] or [14] for references and a discussion. $\Diamond$

An affine space, now, is oriented by orienting its vector associate: a *bound frame* at x in A_n, i.e., a set of n independent vectors at x, is direct [resp. skew] if these n vectors form a direct [resp. skew] frame in V_n.

Vector subspaces of a given vector space (or affine subspaces of an affine space[8]) can have

[8] An affine subspace is oriented by orienting the parallel vector subspace. A point, which is an affine subspace

their own orientation. Orienting a line, in particular, means selecting a vector parallel to it, called a *director* vector for the line, which specifies the "forward" direction along it.

Such orientations of different subspaces are a priori unrelated. Orienting 3D space by the corkscrew rule, for instance, does not imply any orientation in a given plane. This remark may hurt common sense, for we are used to think of the standard orientation of space and of, say, a horizontal plane, as somehow related. And they are, indeed, but only because we think of vertical lines as oriented, bottom up. This is the convention known as *Ampère's rule*. To explain what happens there, suppose space is oriented, and some privileged straightline is oriented too, on its own. Then, any plane *transverse* to this line (i.e., thus placed that the intersection reduces to a single point) inherits an orientation, as follows: To know whether a frame in the plane is direct or skew, make a list of vectors composed of, in this order, (1) the line's director, (2) the vectors of the planar frame; hence an enlarged spatial frame, which is either direct or skew, which tells us about the status of the plane frame.

More generally, there is an interplay between the orientations of complementary subspaces and those of the encompassing space. Recall that two subspaces U and W of V are *complementary* if their *span* is all V (i.e., each v in V can be decomposed as $v = u + w$, with u in U and w in W) and if they are *transverse* ($U \cap W = \{0\}$, which makes the decomposition unique). We shall refer to V as the "ambient" space, and write $V = U + W$. If both U and W have orientation, this orients V, by the following convention: the frame obtained by listing the vectors of a direct frame in U first, then those of a direct frame in W, is direct. Conversely, if both U and V are oriented, one may orient W as follows: to know whether a given frame in W is direct or skew, list its vectors behind those of a direct frame of U, and check whether the enlarged frame thus obtained is direct or skew in V. This is a natural generalization of Ampère's rule.

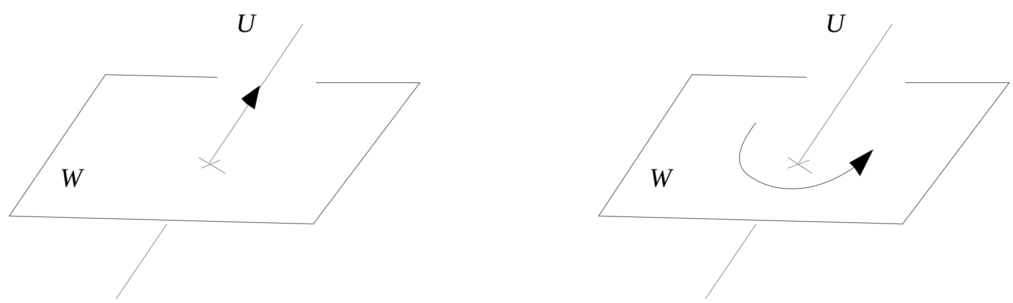

Figure 3. Left: Specifying a "crossing direction" through a plane W by inner-orienting a line U transverse to it. Right: Outer-orienting U, i.e., giving a sense of going around it, by inner-orienting W.

Now what if U is oriented, but ambient space is not? Is U's orientation of any relevance to the complement W? Yes, as Fig. 3 suggests (left): For instance, if W has dimension $n-1$, an orientation of the one-dimensional complement U can be interpreted as a crossing direction relative to W, an obviously useful notion. (Flow of something through a surface, for instance, presupposes a crossing direction.) Hence the concept of *external*, or *outer orientation* of subspaces of V: Outer orientation of a subspace is, by definition, an orientation of one[9] of its complements. Outer orientation of V itself is thus a sign, + or −. (For contrast and clarity, we shall call *inner* orientation what was simply "orientation" up to this point.) The notion (which one can trace back to Veblen [102], cf. [32] and [87]) passes to affine subspaces of an affine space the obvious way.

Note that *if* ambient space is oriented, outer orientation determines inner orientation

parallel to $\{0\}$, can therefore be oriented, which we shall mark by apposing a sign to it, + or −.

[9] Nothing ambiguous in that. There is a canonical linear map between two complements W_1 and W_2 of the same subspace U, namely, the "affine projection" π_U along U, thus defined: for v in W_1, set $\pi_U(v) = v + u$, where u is the unique vector in U such that $v + u \in W_2$. Use π_U to transfer orientation from W_1 to W_2.

(Fig. 4). But otherwise, the two kinds of orientation are independent. As we shall see, they cater for different needs in modelling.

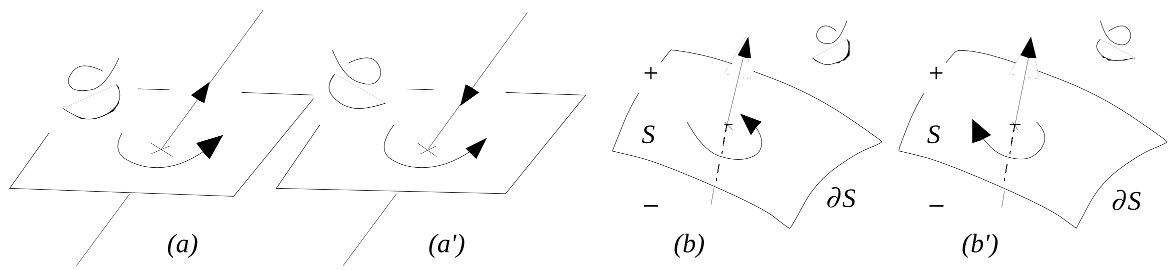

Figure 4. Left: How an externally oriented line acquires inner orientation, depending on the orientation of ambient space. (Alternative interpretation: if one knows both orientations, inner and outer, for a line, one knows the ambient orientation.) Right: Assigning to a surface a crossing direction (here from region "−" below to region "+" above) will not by itself imply an inner orientation. But it does if ambient space is oriented, as seen in *(b)* and *(b′)*. Figures 4a and 4b can be understood as an explanation of Ampère's rule, in which the ambient orientation is, by convention, the one shown here by the "right corkscrew" icon.

1.3.2 Oriented manifolds

Orientation can be defined for other figures than linear subspaces. Connected parts of affine subspaces, such as polygonal faces, or line segments, can be oriented by orienting the supporting subspace (i.e., the smallest one containing them). Smooth lines and surfaces as a whole are oriented by attributing orientations to all their tangents or tangent planes in a consistent way.

"Consistent"? Let's explain what that means, in the case of a surface. First, subspaces parallel to the tangent planes at all points in the neighborhood $N(x)$ of a given surface point x have, if $N(x)$ is taken small enough, a common complement, characterized by a director $n(x)$ (not the "normal" vector, since we have no notion of orthogonality at this stage, but the idea is the same). Then $N(x)$ is consistently oriented if all these orientations correspond via the affine projection along $n(x)$ (cf. Note 9). But this is only *local* consistency, which can always be achieved, and one wants more: *global* consistency, which holds if the surface can be covered by such neighborhoods, with consistent orientation in each non-empty intersection $N(x) \cap N(y)$. This may not be feasible, as in the case of a Möbius band, hence the distinction between (internally) orientable and non-orientable manifolds.

Cells, as defined above, are inner orientable, thanks to the fact that Dc does not vanish. For instance (cf. Fig. 5), for a path c, i.e., a smooth embedding $t \to c(t)$ from [0, 1] to A_n, the tangent vectors $\partial_t c(t)$ determine consistent orientations of their supporting lines, hence an orientation of the path. (The other orientation would be obtained by starting from the "reverse" path, $t \to c(1-t)$.) Same with a patch $\{s,t\} \to S(s,t)$ on the triangle $T = \{\{s,t\} : 0 \leq s,\ 0 \leq t,\ s+t \leq 1\}$: The vectors $\partial_s S(s,t)$ and $\partial_t S(s,t)$, in this order, form a basis at $S(s,t)$ which orients the tangent plane, and these orientations are consistent.

As for piecewise smooth manifolds, finally, the problem is at points x where cells join, for a tangent subspace may not exist there. But according to our conventions, there must be a neighborhood homeomorphic to a ball or half-ball, which *is* orientable, hence a way to check whether tangent subspaces at regular points in the vicinity of x have consistent orientations, and therefore, to check whether the manifold as a whole is or is not orientable.

Similar considerations hold for external orientation. Outer-orienting a surface consists in giving a (globally consistent) crossing direction through it. For a line, it's a way of "turning around" it, or "gyratory sense" (Fig. 3, right). For a point, it's an orientation of the space in its neighborhood. For a connected region of space, it's just a sign, + or −.

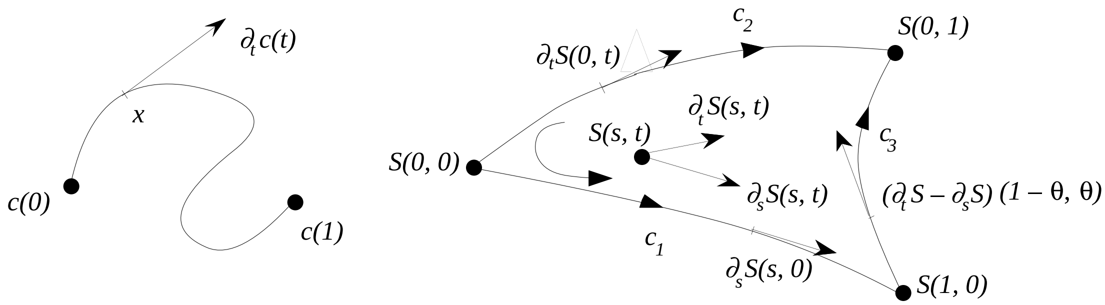

Figure 5. A path and a patch, with natural inner orientations. Observe how their boundaries are themselves assemblies of cells: $\partial c = c(0) - c(1)$ and $\partial S = c_1 - c_2 + c_3$, with a notation soon to be introduced more formally. Paths c_i are $c_1 = s \to S(s, 0)$, $c_2 = t \to S(0, t)$, and $c_3 = \theta \to S(1 - \theta, \theta)$, each with its natural inner orientation.

1.3.3 Induced orientation

Surfaces which enclose a volume V (which one may suppose connected, though the boundary ∂V itself need not be) can always be outer oriented, for the "inside out" crossing direction is always globally consistent. Let us, by convention, take this direction as defining the canonical outer orientation of ∂V. No similarly canonical *inner* orientation of the surface results, as could already be seen on Fig. 4, since there are, in the neighborhood of each boundary point, two eligible orientations of ambient space. But if V is inner oriented, this orientation can act in conjunction with the outer one of ∂V to yield a natural inner orientation of V's boundary about this point. For example, on the left of Fig. 6, the 2-frame $\{v_1, v_2\}$ in the tangent plane of a boundary point is taken as direct because, by listing its vectors behind an outward directed vector ν, one gets the direct 3-frame $\{\nu, v_1, v_2\}$. Consistency of these orientations stems from the consistency of the crossing direction. Hence V's inner orientation *induces* one on each part of its boundary.

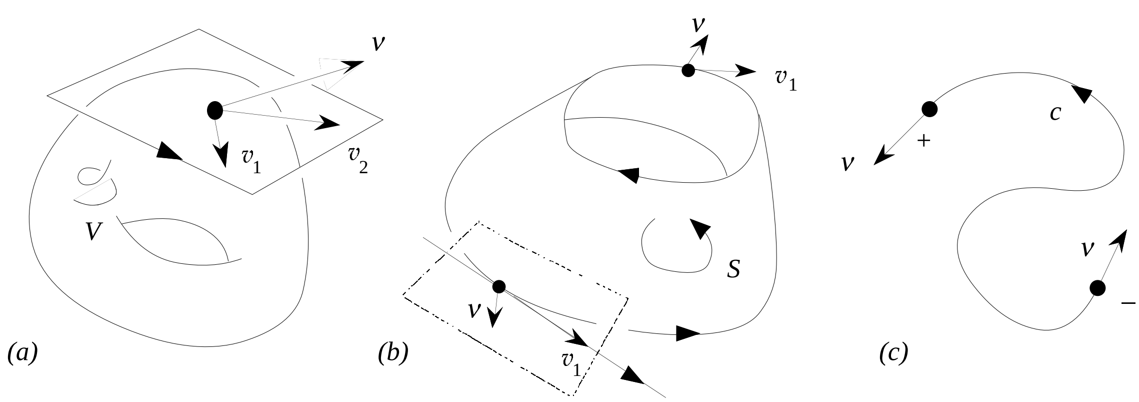

Figure 6. Left: Induced orientation of the boundary of a volume (v_1 and v_2 are tangent to ∂V, ν points outwards). Middle: The same idea, one dimension below. The tangent to the boundary, being a complement of (the affine subspace that supports) ν, with respect to the plane tangent to the surface (in broken lines), inherits from the latter an inner orientation. Right: Induced orientation of the endpoints of an oriented curve.

The same method applies to manifolds of lower dimension p, by working inside the affine p-subspace tangent to each boundary point. See Fig. 6*b* for the case $p = 2$. The p-manifold, thus, serves as ambient space with respect to its own boundary, for the purpose of inducing orientation.

In quite a similar way (Fig. 7), *outer* orientation of a manifold induces an *outer* orientation of each part of its boundary. (For a volume V, the induced outer orientation of ∂V is the inside-out or outside-in direction, depending on the outer orientation, + or −, of V.)

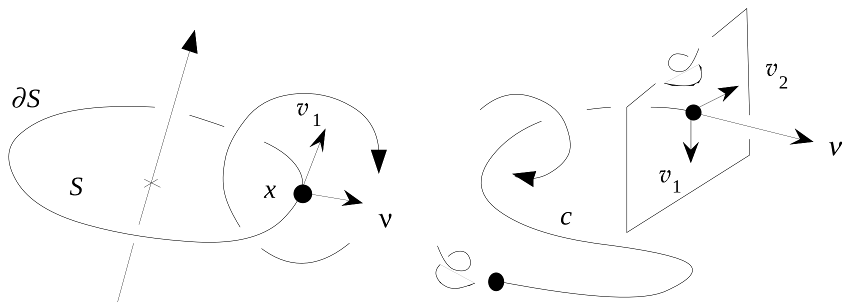

Figure 7. Left: To outer-orient ∂S is to (consistently) inner-orient complements of the tangent, one at each boundary point x. For this, take as direct the frame $\{v_1, \nu\}$, where $\{v_1\}$ is a direct frame in the complement of the plane tangent to S at x, and ν an outward directed vector tangent to S. That $\{v_1\}$ is direct is known from the outer orientation of S. Right: Same idea about the boundary points of line c. Notice that ν is now appended *behind* the list of frame vectors. Consistency stems from the consistency of ν, the inside-out direction with respect to S. The icons near the endpoints are appropriate, since outer orientation of a point is inner orientation of the space in its vicinity.

1.3.4 Inner vs outer orientation of submanifolds

We might (but won't, as the present baggage is enough) extend these concepts to submanifolds of ambient manifolds other that A_3, including non-orientable ones. A two-dimensional example will give the idea (Fig. 8): Take as ambient manifold a Möbius band M, and forget about the 3-dimensional space it is embedded in for the sake of the drawing. Then it's easy to find in M a line which (being a line) is inner orientable, but cannot consistently be outer oriented. Note that the band by itself, i.e., considered as its own ambient space, can be outer oriented, by giving it a sign: Indeed, outer orientation of the tangent plane at each point of M, being inner orientation of this point, is such a sign, so consistent orientation means attributing the same sign to all points. (By the same token, any manifold is outer orientable, with respect to itself as ambient space.)

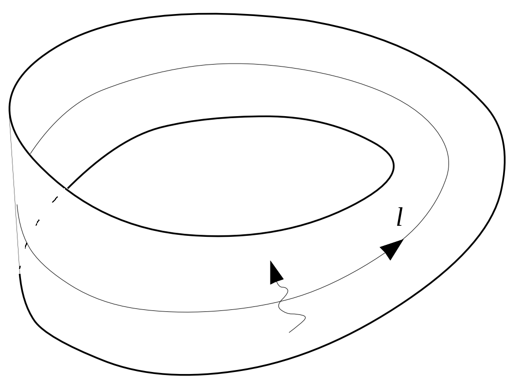

Figure 8. Möbius band, not orientable. As the middle line l does not separate two regions, it cannot be assigned any consistent crossing direction, so it has no outer orientation with respect to the "ambient" band.

For completeness, let us give another example (Fig. 9), this time of an outer-orientable surface without inner orientation, owing to non-orientability of the ambient manifold. The latter (whose boundary is a Klein bottle) is made by sticking together the top and bottom sides of a vertical cube, according to the rule of Fig. 9a. The ribbon shown in *(b)* is topologically a Möbius band, a non-(inner)orientable surface. Yet, it plainly has a consistent set of transverse vectors. (Follow the upper arrow as its anchor point goes up and reenters at the bottom, and notice that the arrow keeps pointing in the direction of AB in the process. So it coincides with the lower arrow when this passage has been done.) Contrast with the ordinary ribbon in *(c)*, orientable, but not outer orientable with respect to this ambient space.

The two concepts of orientation are therefore essentially different.

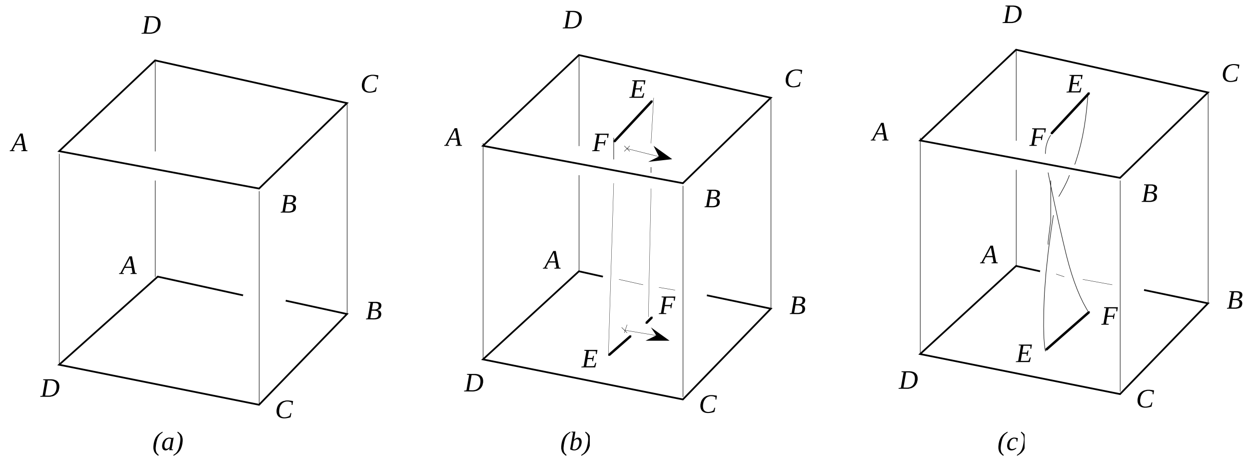

Figure 9. Left: Non-orientable 3-manifold with boundary: Identify top and bottom by matching upper A with lower A, etc. Middle: Embedded Möbius band, with a globally consistent crossing direction. Right: Embedded ribbon.

In what follows, we shall use the word "twisted" (as opposed to "straight") to connote anything that is to do with outer (as opposed to inner) orientation.

1.4 Chains, boundary operator

It may be convenient at times to describe a manifold M as an assembly of several manifolds, even if M is connected. Think for example of the boundary of a triangle, as an assembly of three edges, and more generally of a piecewise smooth assembly of cells. But it may happen—so will be the case here, later—that these various manifolds have been *independently* oriented, with orientations which may or may not coincide with the desired one for M. This routinely occurs with boundaries, in particular. The concept of chain will be useful to deal with such situations.

A *p-chain* is a finite family $\mathcal{M} = \{M_i : i = 1, \ldots, k\}$ of oriented connected p-manifolds,[10] to which we shall loosely refer below as the "components" of the chain, each loaded with a weight μ^i belonging to some ring of coefficients, such as $\mathbb{R}$ or $\mathbb{Z}$ (say $\mathbb{R}$ for definiteness, although weights will be signed integers in most of our examples). Such a chain is conveniently denoted by the "formal" sum $\sum_i \mu^i M_i \equiv \mu^1 M_1 + \cdots + \mu^k M_k$, thus called because the + signs do not mean "add" in any standard way. On the other hand, chains themselves, as whole objects, can be added, and there the notation helps: To get the sum $\sum_i \mu^i M_i + \sum_j \nu^j N_j$, first merge the two families $\mathcal{M}$ and $\mathcal{N}$, then attribute weights by adding the weights each component has in each chain, making use of the convention that $\mu M'$ is the same chain as $-\mu M$ when M' is the same manifold as M with opposite orientation. If all weights are zero, we have the *null chain*, denoted 0. All this amounts, as one sees, to handling chains according to the rules of algebra, when they are represented via formal sums, which is the point of such a notation. *Twisted* chains are defined the same way, except that all orientations are external. (Twisted and straight chains are not to be added, or otherwise mixed.)

If M is an oriented piecewise smooth manifold, all its cells c_i inherit this orientation, but one may have had reasons to orient them on their own, independently of M. (The same cell may well be part of several piecewise smooth manifolds, for instance.) Then, it is natural to associate with M the chain $\sum_i \pm c_i$, also denoted by M, with i-th weight -1 when the orientations of M and c_i differ. (Refer back to Fig. 5 for simple examples.)

Now, the boundary of an oriented piecewise smooth $(p+1)$-manifold M is an assembly of p-manifolds, each of which we assume has an orientation of its own. Let us assign each

[10] For instance, cells. But we don't request that. Each M_i may be a piecewise smooth manifold already.

of them the weight ± 1, according to whether its orientation coincides with the one inherited from M. (We say the two orientations *match* when this coincidence occurs.) Hence a chain, also denoted ∂M. By linearity, the operator ∂ extends to chains: $\partial(\sum_i \mu^i M_i) = \sum_i \mu^i \, \partial M_i$. A chain with null boundary is called a *cycle*. A chain which is the boundary of another chain is called, appropriately, a *boundary*. Boundaries are cycles, because of the fundamental property

$$\partial \circ \partial = 0, \tag{2}$$

i.e., the boundary of a boundary is the null chain. A concrete example, as in Fig. 10, will be more instructive here than a formal proof.

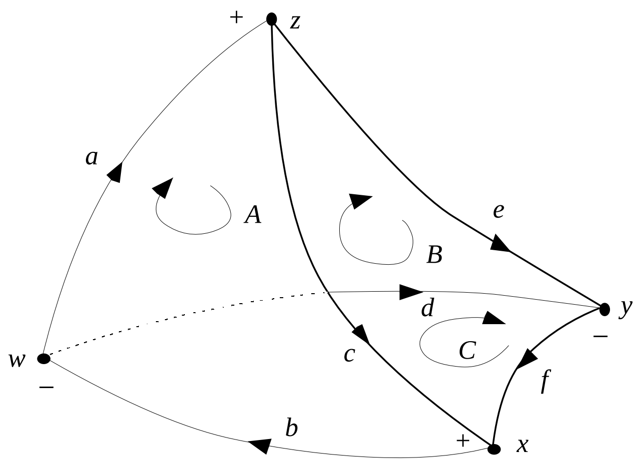

Figure 10. Piecewise smooth surface S, inner oriented (its orientation is taken to be that of the curved triangle in the fore, marked A), represented as the chain $A - B - C$ based on the oriented curved triangles A, B, C. (Note the minus signs: B's and C's orientations don't match that of S.) One has $\partial A = a + b + c$, $\partial B = e + a - d$, $\partial C = b + d + f$, where a, b, c, d, e, f are the boundary curves, arbitrarily oriented as indicated. Now, $\partial S = \partial(A - B - C) = c - e - f$: Observe how the "seams" a, b, c automatically receive null weights in this 1-chain, whatever their orientation, because they appear twice with opposite signs. Next, since $\partial c = x - z$, $\partial e = -y - z$, and $\partial f = x + y$, owing to the (arbitrary) orientations assigned to points w, x, y, z, one has $\partial\partial S = \partial(c - e - f) = 0$, by the same process of cancellation by pairs. The reader is invited to work out a similar example involving twisted chains instead of straight ones.

Remark. Beyond its connection with assemblies of oriented cells, no too definite intuitive interpretation of the concept of chain should be looked for. Perhaps, when $p = 1$, one can think of the chain $\sum_i \gamma_i c_i$, with integer weights, as "running along each c_i, in turn, $|\gamma_i|$ times, in the direction indicated by c_i's orientation, or in the reverse direction, depending on the sign of γ_i". But this is a bit contrived. Chains are better conceived as algebraic objects, based on geometric ones in a useful way—as the example in Fig. 10 should suggest, and as we shall see later. However, we shall indulge in language abuse, and say that a closed curve "is" a 1-cycle, or that a closed surface "is" a 2-cycle, with implicit reference to the associated chain. ◊

So boundaries are cycles, after (2). Whether the converse is true is an essential question. In affine space, the answer is positive: A closed surface encloses a volume, a closed curve (even if knotted) is the boundary of some surface (free of self-intersections, amazing as this may appear), called a Seifert surface ([89], [2], p. 224). But in some less simple ambient manifolds, a cycle need not bound. In the case of a solid torus, for instance, a meridian circle is a boundary, but a parallel circle is not, because none of the disks it bounds in A_3 is entirely contained in the torus. Whether cycles are or aren't boundaries is therefore an issue when investigating the global topological properties of a manifold. Chains being algebraic objects then becomes an asset, for it makes possible to harness the power of algebra to the study of topology. This is the gist of *homology* [50, 51], and of algebraic topology in general.

1.5 Metric notions

Now, let us equip V_n with a dot product: $u \cdot v$ is a real number, linearly depending on vectors

u and v, with symmetry ($u \cdot v = v \cdot u$) and strict positive-definiteness ($u \cdot u > 0$ if $u \neq 0$). Come from this, first the notions of orthogonality and angle, next a norm $|u| = (u \cdot u)^{1/2}$ on V_n, then a distance $d(x, y) = |y - x|$, translation-invariant by construction, between points of the affine associate A_n.

Definition. Euclidean space, E_n, *is the structure composed of* A_n*, plus a dot product on its associate* V_n*, plus an orientation.*

Saying "the" structure implies that two realizations of it (with two different dot products and/or orientations) are isomorphic in some substantial way. This is so: For any other dot product, " $\bullet$ " say, there is an invertible linear transform L such that $u \bullet v = Lu \cdot Lv$. Moreover,[11] one may have L "direct", in the sense that it maps a frame to another frame of the same orientation class, or "skew". Therefore, two distinct Euclidean structures on A_n are linked by some L. In the language of group actions, the linear group GL_n, composed of the above L's, acts transitively on Euclidean structures, i.e., with a unique orbit, which is our justification for using the singular. (These structures are said to be *affine equivalent*,[12] a concept that will recur.) The point can vividly be made by using the language of group actions: the isotropy group of $\{\cdot, Or\}$ "cannot be any larger". (More precisely, it is maximal, as a subgroup, in the group of direct linear transforms.)

In dimension 3,[13] dot product and orientation conspire in spawning the *cross product*: $u \times v$ is characterized by the equality

$$|u \times v|^2 + (u \cdot v)^2 = |u|^2 |v|^2 \tag{3}$$

and the fact that vectors u, v and $u \times v$ form, in this order, a direct frame. The *3-volume* of the parallelotope built on vectors u, v, w, defined by $\text{vol}(u, v, w) = (u \times v) \cdot w$, is equal, up to sign, to the above volumic measure, with equality if the frame is direct.[14] Be well aware that $\times$ doesn't make any sense in *non*-oriented three-space.

We shall have use for the related notion of *vectorial area* of an outer oriented triangle T, defined as the vector $\vec{\text{T}} = \text{area}(\text{T})\, n$, where n is the normal unit vector that provides the crossing direction. (If an ambient orientation exists, two vectors u and v can be laid along two of the three sides, in such a way that $\{u, v, n\}$ is a direct frame. Then, $\vec{\text{T}} = {}^1/_2\, u \times v$. Figure 11 gives an example.) More generally, an outer oriented surface of E_3 has a vectorial area: Chop the surface into small adjacent triangular patches, add the vectorial areas of these, and pass to the limit. (This yields 0 for a closed surface.)

For later use, we state the relations between the structures induced by $\{\cdot, Or\}$ and $\{\bullet, \mathbf{Or}\}$, where $\mathbf{Or} = \pm Or$, the sign being that of $\det(L)$. (There is no ambiguity about "$\det(L)$", understood as the determinant of the matrix representation of L: its value is the same in any basis.) The norm $(u \bullet u)^{1/2}$ will be denoted by $\|u\|$. The corresponding cross product is defined by $\|u \boldsymbol{\times} v\|^2 + (u \bullet v)^2 = \|u\|^2 \|v\|^2$ as in (3) (plus the request that $\{u, v, u \boldsymbol{\times} v\}$ be $\mathbf{Or}$-direct), and the new volume is $\mathbf{vol}(u, v, w) = (u \boldsymbol{\times} v) \bullet w$. It's a simple exercise to show that

$$\|u\| = |Lu|, \; L(u \boldsymbol{\times} v) = Lu \times Lv, \; \mathbf{vol}(u, v, w) = \det(L)\, \text{vol}(u, v, w). \tag{4}$$

(It all comes from the equality $\det(Lu, Lv, Lw) = \det(L)\, \det(u, v, w)$, when u, v, w, and L are represented in some basis, a purely affine formula.) Notice that, for any w, one has

[11] L is not unique, since UL, for any *unitary* U (i.e., such that $|Uv| = |v|\, \forall v$), will work as well. In particular, one might force L to be self-adjoint, but we won't take advantage of that.

[12] Such equivalence is what sets Euclidean norms apart among all conceivable norms on V_n, like for instance $|v| = \sum_i |v^i|$. As argued at more length in [12], choosing to work in a Euclidean framework is an acknowledgment of another observed symmetry of the world we live in: its *isotropy,* in addition to its homogeneity.

[13] A binary operation with the properties of the cross product can exist only in dimensions 3 and 7 [90, 37].

[14] An n-volume could directly be defined on V_n, as a map $\{v_1, \ldots, v_n\} \to \text{vol}(v_1, \ldots, v_n)$, multilinear and null when two vectors of the list are equal. Giving an n-volume implies an orientation (direct frames are those with positive n-volumes), but no metric (unless $n = 1$).

$L^a L(u \boldsymbol{\times} v) \cdot w = L(u \boldsymbol{\times} v) \cdot Lw = \det(L)\,(u \times v) \cdot w$, where L^a denotes the *adjoint* of L (defined by $Lu \cdot v = u \cdot L^a v$ for all u, v), hence an alternative formula:

$$u \boldsymbol{\times} v = \det(L)\,(L^a L)^{-1}(u \times v). \tag{5}$$

As for the vectorial area, denoted $\vec{\mathbf{T}}$ in the "bold" metric, one will see that

$$\vec{\mathbf{T}} = |\det(L)|\,(L^a L)^{-1}\vec{\mathrm{T}}, \tag{6}$$

with a factor $|\det(L)|$, not $\det(L)$, because $\vec{\mathrm{T}}$ and $\vec{\mathbf{T}}$, both going along the crossing direction, point towards the same side of T.

We shall also need a topology on the space of p-chains, in order to define differential forms as *continuous* linear functionals on this space. As we shall argue later, physical observables such as electromotive force, flux, and so forth, can be conceived as the values of functionals of this kind, the chain operand being the idealization of some measuring device. Such values don't change suddenly when the measurement apparatus is slightly displaced, which is the rationale for continuity. But to make precise what "slightly displaced" means, we need a notion of "nearness" between chains—a topology.

First thing, nearness between manifolds. Let us define the distance $d(M, N)$ between two of them as the greatest lower bound (the infimum) of $d_\phi(M, N) = \sup\{x \in M : |x - \phi(x)|\}$ with respect to all orientation-preserving piecewise smooth diffeomorphisms (OPD) ϕ that exist between M and N. There may be no such OPD, in which case we take the distance as infinite, but otherwise there is symmetry between M and N (consider ϕ^{-1} from N to M), positivity, d can't be zero if $M \neq N$, and the triangle inequality holds. [*Proof:* Take M, N, P, select OPD's ϕ and ψ from P to M and N, and consider x in P. Then $|\phi(x) - \psi(x)| \leq |\phi(x) - x| + |x - \psi(x)|$, hence $d_{\psi \circ \phi^{-1}}(M, N) \leq d_\phi(M, P) + d_\psi(N, P)$, then minimize with respect to ϕ and ψ.] Nearness of two manifolds, in this sense, does account for the intuitive notion[15] of "slight displacement" of a line, a surface, etc. The topology thus obtained does not depend on the original dot product, although d does.

Next, on to chains. The notion of convergence we want to capture is clear enough: a sequence of chains $\{c_n = \sum_{i=1,\dots,k} \mu^i_n M_{i,n} : n \in \mathbb{N}\}$ should certainly converge towards the chain $c = \sum_{i=1,\dots,k} \mu^i M_i$ when the sequences of components $\{M_{i,n} : n \in \mathbb{N}\}$ all converge, in the sense of the previous distance, to M_i, while the weights $\{\mu^i_n : n \in \mathbb{N}\}$ converge too, towards μ^i. But knowing some convergent sequences is not enough to know the topology. (For that matter, even the knowledge of *all* convergent sequences would not suffice, see [43], p. 161.) On the other hand, the finer the topology, i.e., the more open sets it has, the more difficult it is for a sequence to converge, which tells us what to do: Define the desired topology as the finest one which (1) is compatible with the vector space structure of p-chains (in particular, each neighborhood of 0 should contain a convex neighborhood) (2) makes all sequences of the above kind converge.

The space of straight [resp. twisted] p-chains, as equipped with this topology, will be denoted by $\mathcal{C}_p$ [resp. $\tilde{\mathcal{C}}_p$]. Both spaces are purely affine constructs, independent of the Euclidean structure, which only played a transient role in their definition.

It now makes sense to ask whether the linear map ∂ is continuous from $\mathcal{C}_p$ to $\mathcal{C}_{p-1}$. The answer is by the affirmative, thanks to the linearity of ∂ and the inequality $d(\partial M, \partial N) \leq d(M, N)$. [*Proof:* The restriction to ∂M of an OPD ϕ is an OPD which sends it to ∂N, so $d(\partial M, \partial N) \leq \inf_\phi \sup\{x \in \partial M : |\phi(x) - x|\} \leq \inf_\phi \sup\{x \in M : |\phi(x) - x|\} = d(M, N)$.]

[15] More refined topologies, involving the derivatives of the charts, could be defined in the case of smooth manifolds, but the very roughness of the present one will be an asset for what we have in view, which is, keeping the topological dual of $\mathcal{C}_p$ reasonably small.