Pre-metric Electromagnetism

By

D. H. Delphenich

For every good man chopping at the roots of evil there are a thousand more hacking at the leaves.

Henry David Thoreau

Leaf-hacking, being more labor-intensive than root-chopping, is more likely to get funding.

David Henry Delphenich

CONTENTS

LIST OF FIGURES

Introduction

The term "pre-metric electromagnetism" refers to the formulation of the mathematical theory of electromagnetism in a manner that not only does not assume the existence of a Lorentzian metric on the spacetime manifold to begin with, but also exhibits the appearance of such a geometrical structure as a natural consequence of investigating the manner in which electromagnetic waves propagate through that medium. Since the Lorentzian metric that appears – at least, under restricted conditions – is commonly taken to account for the presence of gravitation in the spacetime medium in the viewpoint of general relativity, one sees that in order to properly define the context of pre-metric electromagnetism one must really discuss not only electricity and magnetism, but also gravity, as well.

Hence, before we embark upon the detailed discussion of the mathematical and physical bases for the theory of pre-metric electromagnetism, we shall briefly recall the conceptual evolution of the three relevant physical phenomena of electricity, magnetism, and gravity.

1. The unification of electricity and magnetism [1]. In ancient times, it is unlikely that man ever suspected that the natural phenomena associated with electricity, such as lightning and static electricity on animal furs, could possibly be associated with magnetic phenomena, which were discovered somewhat later in history, and most likely in the Iron Age in the context of lodestones, which are magnetite deposits that have become magnetized from long-term exposure to the Earth's magnetic field.

However, when Europe emerged from the Dark Ages into the Renaissance the science of electricity gradually gave way to the development of batteries [2], wires, and currents, on the one hand, with the development of compasses for marine navigation on the other. It was only a matter of time before the early experimenters, primarily Hans Christian Oersted (1777-1851) and Michael Faraday (1791-1867), noticed that electrical currents could produce measurable magnetic fields around conductors, while time-varying magnetic fields could conversely induce electrical currents in current loops. Actually, this is not a precise converse, since a time-constant electrical current can induce a time-constant magnetic field, while a time-constant magnetic field will *not* induce an electrical current of any sort. This reciprocal set of phenomena was called *electromagnetic induction*.

One of Faraday's other innovations in the name of electromagnetism was the introduction of the concept of invisible force fields distributed in space that accounted to the forces of attraction or repulsion that "test" electric charges and magnetic dipoles experienced when placed at each point. Although nowadays the concept of vector fields

[1] For a comprehensive historical timeline of the theories of electricity and the ether, one should confer the tomely two-volume treatise of E. T. Whittaker [**1**].

[2] Apparently, the concept of a battery had been developed in a rudimentary form by ancient cultures, as pottery that seemed to involve weak acids and metal electrodes has been unearthed by archeologists.

seems rather commonplace and above dispute, nevertheless, in its day the ideas of invisible lines of force apparently seemed rather mystical and dubious. It is important to note that the key to defining such fields is the association of a purely dynamical notion – namely, force – with more esoteric ones, in the form of electrical and magnetic fields.

In addition to this key development, one must also observe the evolution of the theory of electrostatic forces from the early measurements by Charles Augustin de Coulomb (1736-1806) and the formulation of the empirical law that bears his name to its formulation in terms of potential theory by Laplace, Poisson, and the host of contributors to the theory of boundary-value problems in the Laplace or Poisson equation, such as Green, Cauchy, Dirichlet, Neumann, Robin, and many more. It was also found that although the static magnetic field was not apparently due to a magnetic charge monopole, but a magnetic dipole, nevertheless, the Laplace equation could still be used in the context of a magnetic vector potential.

Part of the shift in emphasis from Coulomb's law to the Poisson equation involves the introduction of electrical flux and the use of Gauss's law, which is due to the German mathematician Karl Friedrich Gauss (1777-1855). One can also introduce a corresponding notion of magnetic flux and obtain an analogous law that relates the electrical current in a loop that bounds a surface with the total magnetic flux through the surface; this law was due to the French physicist André Ampère (1775-1835).

The capstone of the unification of electricity and magnetism into a single unified field theory of electromagnetism was laid by James Clerk Maxwell (1831-1879) [**2**], when he postulated that the sort of electromagnetic induction that was discovered by Faraday might also work in reverse. That is, a time-varying electric flux through a surface with boundary might induce a *magnetomotive* force, or *displacement current*, in the boundary loop; effectively, this amounts to saying that it induces a magnetic field.

However, there was a significant difference between the two types of electromagnetic induction, namely, the one was 180^{o} out of phase with the other one. This took the form of a relative minus sign in the resulting field equations for the curl of the electric field strength vector field **E** and the curl of the magnetic field strength vector field **H**.

The resulting set of four first-order partial differential equations for the vector fields **E**, **B**, **D**, **H** [3]:

$$\nabla\times\mathbf{E} = -\frac{1}{c}\frac{\partial\mathbf{B}}{\partial t}, \qquad \nabla\cdot\mathbf{E} = 4\pi\rho, \qquad \nabla\times\mathbf{H} = +\frac{1}{c}\frac{\partial\mathbf{D}}{\partial t} + \frac{4\pi}{c}\mathbf{J}, \qquad \nabla\cdot\mathbf{B} = 0,$$

which constitute Maxwell's equations in one of their simplest forms, represent the culmination of Faraday's law, Coulomb's law, Ampère's law, combined with Maxwell's displacement current, and the non-existence of magnetic monopoles, respectively. The vector field **J** represents an electric current density, which can be associated with the electric charge density ρ by way of its motion with flow velocity vector field **v** in the form of $\rho\mathbf{v}$, or it can also be an electric source current that is independent of ρ.

Actually, the nature of the vector field **J** is not entirely arbitrary, since taking the divergence of the third equation gives the conservation of charge in the form:

[3] We present these equations in the form that they are given in the standard reference work by J. D. Jackson [**3**], but, as we shall discuss in due course, many other forms exist.

$$0 = \frac{\partial(\nabla \cdot \mathbf{D})}{\partial t} + 4\pi \nabla \cdot \mathbf{J}\,.$$

One can also regard this as an integrability condition for the over-determined system of linear first-order partial differential equations for **H** and **D** that the third set of Maxwell equations, in the form above, represents.

The constant *c* is the speed of propagation of electromagnetic waves *in vacuo*. However, its introduction at this point seems unrelated to general considerations, since the system of equations is valid in media other than the vacuum, and at this point the constant *c* serves as more of a units conversion constant than a fundamental property of an electromagnetic medium of a particular sort. We shall discuss these issues in more detail in the next section.

Moreover, in order to state Maxwell's equations in their traditional form, we have introduced two further vector fields **D** and **B**, which were classically referred to as the electric displacement and magnetic flux density vector fields. Hence, the system of equations as it was stated above is underdetermined; viz., we have defined nine component equations for the fifteen components of the vector fields **E**, **B**, **D**, **H**, **J**. The six remaining equations that we require take the form of the *electromagnetic constitutive law*:

$$\mathbf{D} = \mathbf{D}(\mathbf{E}, \mathbf{B}), \qquad \mathbf{H} = \mathbf{H}(\mathbf{E}, \mathbf{B}),$$

for the medium in question that relates the fields **E** and **B** to the fields **D** and **H** in a manner that is closely analogous to the way that mechanical constitutive laws couple the stress tensor for a material medium to its strain tensor. We shall discuss this aspect of Maxwell's equations in more detail in the next section of this introduction, as well, since it defines the key to understanding why pre-metric electromagnetism is based in established experimental physics, as well as purely mathematical refinements.

2. The evolution of geometrical optics into wave optics. Perhaps the most profound consequence of Maxwell's hypothesis was that it led to the existence of solutions of his equations that took the form of electromagnetic waves. This opened up a new realm of possibilities for explaining the optical phenomena that seemed beyond the reach of the earlier methods of geometrical optics.

The study of optics seems as old as man's awareness of the presence of light in the world. One can find writings on the subject as far back as Euclid, who also wrote about geometry more generally. Indeed, one sees that the concept of "light rays," which is at the root of geometrical optics, seems to be the first way of modeling the behavior of light.

One of the first enduring results of geometrical optics was due to Willebrod Snell (1581-1626), who experimentally derived the largely empirical result that the ratio of the sines of the angles of refraction and incidence at an optical interface could be expressed in terms of the ratio of two numbers that were characteristic of the materials. Although we now understand that these numbers are the indices of refraction of the materials and they are inversely proportional to the speeds of propagation of electromagnetic waves in them, at the time Snell had no such insight into the nature of the numbers. Interestingly,

after Snell's contribution the next significant advance in optics did not come about until fifty-two years after his death.

Similarly, although Sir Isaac Newton (1642-1727) published his celebrated treatise [**4**] in 1704, in which he suggested that light was due to the motion of infinitesimal "corpuscles," nonetheless, the Dutch physicist Christiaan Huygens (1629-95) had published a treatise [**5**] in 1678 in which he was introducing the notion that the behavior of light had more to do with envelopes of "elementary waves."

One also notes that after Newton's treatise the next major advance to optics did not seem to come until 104 years after its publication, which would be the work of Etienne-Louis Malus (1775 –1812), who did many experiments on the polarization of light and derived a law for the resulting intensity of a light beam that passes through a polarizer, as well as observing that reflected light is always polarized. His work [**6**] also made considerable use of what later become *line geometry*, which was first developed by the German geometer Julius Plücker (1801 – 1868).

As we will see, one of the foundations of pre-metric electromagnetism is due to the work of Augustin-Jean Fresnel (1788 – 1827), whose research in optics [**7**] further reinforced the wavelike conception that Huygens had previously introduced. The concept of the Fresnel wave surface or ray surface also reinforces the fundamental role of line geometry, as it admits such a formulation and interpretation quite naturally.

One the most far-reaching conceptual advances for geometrical optics was due to the contributions of Sir William Rowan Hamilton (1805–1865) to the Royal Irish Society [**8**] over a period of time from 1828 to1837. He essentially established that the same general mathematical techniques that allowed one to describe the trajectories of point particles in the motion of matter could suffice to determine the curves – viz., light rays – that were followed by light corpuscles. Perhaps some inkling of how long it took for physics to completely digest new ideas in those days was in the fact that when the German geometer Felix Klein (1849 – 1925), a former student of Plücker's who also advanced the cause of line geometry, commented upon the work of Hamilton some fifty-five years after his last supplement on the theory of rays had been published the title of his paper [**9**] (in translation) was "On *recent* English work in mechanics."

In roughly the same year as Klein's article, Pierre de Fermat (1601 or 1607/8 – 1665) began doing his own work on geometrical optics, which further reinforced the applicability of Hamilton-Jacobi theory by showing that one could derive the equations of geometrical optics by starting with a least-action principle, just as Hamilton's equations are derived from Hamilton's least action principle. The trick is to define an appropriate action functional for curves between points of space if one is to make the curves one has in mind – viz., light rays, in this case – take the form of the paths along which the action function has a minimum. The action that Fermat introduced for geometrical optics was the functional that associated each spatial curve segment with the time it took for light to go from one end to the other.

In 1895, H. Bruns published a lengthy treatise [**10**] in the papers of the mathematical physics class of the Saxon Academy of Sciences in which he introduced the notion of the *eikonal*, a function that embodied the essential information for the propagation of light rays. Klein also responded to this contribution in 1901 with a paper [**11**] in Zeitschrift für Mathematik und Physik in which he showed that the eikonal of Bruns was virtually the same thing as the *characteristic function* of Hamilton-Jacobi theory in mechanics.

Another tributary of research that was emerging in the later Nineteenth Century that eventually met up with both Hamiltonian mechanics and Hamiltonian optics was the method of contact transformations and contact geometry. They had been introduced by Marius Sophus Lie (1842-1899) in his monumental 1888 treatise [**12**] on transformation groups in the form of "Berührungstransformationen." Not only did one-parameter families of such transformations describe the motion of massive particles in Hamilton mechanics, but, as Ernest Vessiot (1865-1952) observed in 1906 [**13**] they also serve to describe the motion of light corpuscles in geometrical optics.

To return to the wave conception of light, around the time of Maxwell the general drift in thinking was towards a mechanical conception of light waves as behaving somewhat like elastic waves in a medium they referred to as the "ether." However, it eventually emerged that the optical wave constructions of Huygens, which predated the work of Maxwell by a considerable margin, were nonetheless of sufficient generality as to continue to apply to the electromagnetic waves that Maxwell's equations implied.

Consequently, the impact of Maxwell's equations on physics was not only in their complete unification of the theories of electricity and magnetism, but also the fact that they gave a precise physical basis for the notion of *wave optics*; i.e., the idea that the behavior of light in optical media was best explained by the propagation of electromagnetic waves, rather than light corpuscles.

However, since geometrical optics, namely, the optical constructions that were based in the Newtonian conception of corpuscular light, was not only adequate in its experimental accuracy in some contexts, such as reflection and refraction, but also more convenient, it was important to also establish the way that geometrical optics might emerge from wave optics, at least as an approximate class of solutions. One simply accepted that there were well-established optical phenomena, such as dispersion, diffraction, and polarization that seemed largely foreign to the methods of geometrical optics.

In order to account for diffraction, one must return to the wave optics of Huygens and regard each point of any spatial isophase surface (momentary wave front) as a potential source of elementary waves, such as expanding spherical waves. The form of these elementary waves is intimately related to the form of the Fresnel surfaces, and it is generally only isotropic optical media in which the elementary wave fronts are expanding sphere. One finds that if waves, in general, are characterized by a phase function, which defines the shape of the momentary wave fronts, and an amplitude function, which defines how the basic physical quantity is carried along by the wave fronts, then Huygens's principle by itself only allows one to propagate the momentary isophase surfaces.

In order to propagate the amplitude, as well, one usually must look more specifically at the nature of the wave equations that one is concerned with. When that equation is the linear wave equation, which again pertains to optically isotropic media, one often resorts to the methods of integral operators and fundamental solutions, such as Green functions, in order to propagate the amplitude. When one reduces from time-varying wave solutions to stationary solutions, the linear wave equation, after separating the time variable becomes the *Helmholtz equation*, after the German physicist and physician Hermann Ludwig Ferdinand von Helmholtz (1821 – 1894), who published his treatment of the problem of propagating amplitudes in 1859 [**14**].

As is often the case with integrals, general solutions are not possible, and one often must resort to approximations. The case of diffraction is no different and the approximate evaluation of the Helmholtz integral for the treatment of diffraction was first achieved by Sir George Gabriel Stokes (1819–1903) in his 1849 memoir [**15**] on the "Dynamical Theory of Diffraction" and refined by Gustav Robert Kirchhoff (1824–1887) in 1882 [**16**].

What eventually emerged was that even Kirchhoff's approximate integral was too complicated to admit general solutions, and the method of asymptotic approximations was introduced into the theory of diffraction. This method originated in previous work [**17**] of Henri Poincaré (1854–1912) on the three-body problem of celestial mechanics and involved the introduction of series expansions for solutions that did not have to converge, except in some asymptotic limit. One then obtains the solution for the propagation of amplitude in terms of successive contributions to the "classical" solution that one obtains from the geometrical optics approximation, which one then regards as the "diffracted fields" of each order.

It is mostly in the study of optical phenomena that one sees the full scope of the previous remark on the necessity of introducing electromagnetic constitutive laws for the medium in which one propagates electromagnetic waves. Due to the vast and increasing variety of optical media that have been studied up to this point in time, one sees that even though the constitutive laws are a system of algebraic equations, not a system of differential equations, nonetheless, one must leave open a considerable degree of generality concerning the properties that go into them depending upon the medium in question.

In particular, the first issue that seems to arise is that of linearity, namely, the possibility that the system in question is a system of linear equations. As usual, one finds that linearity is a simplifying approximation that one introduces only in order to make tangible progress in analyzing the equations, not a deep assumption about the laws of Nature. Indeed, if Nature has any fundamental law in that regard it would have to be: Linearity is always a simplifying approximation for something more nonlinear and complicated. In fact, nowadays, the field of nonlinear optics – i.e., the optics of nonlinear media – has grown to quite vast proportions in its own right in theory, experiment, and practical applications.

Other material properties that one must consider in the name of constitutive laws are homogeneity and isotropy. That is, the optical properties of a medium can vary from point to point, although often the change is a discontinuity at the interface between two different media, and the propagation of electromagnetic waves might even have a strong correlation with the direction of propagation. Furthermore, optical properties often depend upon the "state of polarization" of the wave, as well; by this term, we are referring to the fact that electromagnetic waves are generally polarized linearly, circularly, or elliptically.

Hence, one must realize that the model that one traditionally uses for the classical electromagnetic vacuum state is characterized by a constitutive law that is linear, isotropic, and homogeneous. Thus, one can concisely define it by two constants ε_0 and μ_0, which represent the electric permittivity (or dielectric constant) and magnetic permeability of the vacuum. It is important to note that speed of propagation c, which

was introduced above in a somewhat *ad hoc* way, then becomes a *derived* constant, by way of:

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}},$$

not merely a convenient basis for a unit conversion.

As we shall see later on, this situation is generic to optics, except that one does not always derive a single constant, such as *c*, but more generally a set of three functions of position.

3. Geometrization of gravity. Although the primary focus of this book is, of course, pre-metric *electromagnetism*, nonetheless, in order to fully account for the appearance of a spacetime metric, as well as to discuss its role in the equations of electromagnetism, one must unavoidable touch upon the fact that the spacetime metric plays the fundamental role in Einstein's theory of *gravitation*.

If the natural phenomena that pertained to magnetism seemed remote in their manifestation from those of electricity, it is just as much the case they are both remote from the phenomena that pertained to gravitation. Indeed, one encounters gravitational phenomena, such as falling, in primitive settings more frequently than one encounters electrical or magnetic ones.

However, as humanity learned more about nature, it eventually emerged that there was a close analogy between the law of gravitational attraction, as it was described by Newton, and the law of electrostatic attraction and repulsion, as it was described by Coulomb. Of course, there were some perplexing difficulties associated with making that analogy precise.

For one thing, the fact that electrostatic forces could be either attractive or repulsive, while gravitational forces were only observed to be attractive, suggested that there were two types of charges – viz., positive and negative ones – but only one type of gravitational mass. Indeed, if one postulates the existence of negative masses and examines their effect on Newton's equations of motion when the force is due to gravitation, one obtains contradictory conclusions. In particular, look at the one-dimensional picture, which is governed by:

$$G\frac{Mm}{r^2} = ma_m = -Ma_M,$$

in which M is one mass, m is the other, a_m and a_M are their respective accelerations, r is the distance between their center of mass, and G is Newton's gravitational constant.

Now, assume that the sign of M is positive, while that of m is negative. Although this implies the intriguing "anti-gravitational" possibility that the gravitational force between them is repulsive, nevertheless, if one assumes that Galileo's experiment is equally valid for negative masses as it is for positive ones – i.e., gravitational mass equals inertial mass, including their signs – then one must accept that the acceleration of a negative mass is in the *opposite* direction to the applied force, while the acceleration of the positive mass is in the *same* direction. Hence, the effect is that the negative mass seems

to accelerate in the same direction as the positive mass and they "chase each other" off to infinity, rather than flying apart as electric charges of the same sign would.

Of course, one way around this is to weaken Galileo's equality of gravitational and inertial mass to simply apply it to their *absolute values*. However, in the absence of experimental confirmation for the gravitational and inertial behavior of negative masses, such speculations are rather moot.

Another fascinating aspect of the analogy between electrostatics and gravitostatics is that historically no one pursued the possibility that it might extend to an analogy between electrodynamics and gravitodynamics, as well, until rather recent – i.e., post-Einsteinian – history. To be fair, this is because the gravitostatic force is quite small in its magnitude to begin with, except in astronomical contexts, and one usually expects the magnetic forces that are produced by moving charges to be considerably weaker than the electrostatic ones, as well. Hence, the only laboratory in which any "gravitomagnetic" forces produced by moving masses, in addition to the gravitostatic ones, might possibly be measurable would have to be astrophysical. It is only with the advances of high-precision measuring technology and orbiting space vehicles that science has managed to discover that there are expansions in the scope of Newtonian gravitation that logically precede the refinements of general relativity, which represents essentially a "strong-field" theory of gravitation, while Newton's theory represents its "weak-field" form.

Although nowadays general relativity is usually thought of as Einstein's theory of *gravitation*, one must understand the crucial and unavoidable fact that he did not originally set out to address gravity. Indeed, Einstein's earliest work on relativity grew out of his studies of *electromagnetism*, namely, "The Electrodynamics of Moving Bodies" (cf., [**18**]). The main issue at that point in time was the way that the equations of electromagnetism transformed from one measurer/observer to another, when one assumed that the medium itself was in its own state of "motion."

Now, when the electromagnetic medium is a material one, such as glass, it is easier to justify the assumption that it is in a state of motion relative to the measurer/observer. Indeed, there are experimentally established phenomena associated with such a relative motion; for instance, the Fresnel-Fizeau effect, which is due to the French physicists Fresnel and Armand Hippolyte Louis Fizeau (1819-1896). However, when one is considering an immaterial – i.e., massless – electromagnetic medium, such as the vacuum, it is harder to make the concept of its relative motion logically rigorous. Indeed, the negative result of Michelson and Morley concerning the possibility of measuring a difference between the speed of propagation of light in the direction of motion of the Earth, relative to the vacuum of space, and transverse to it showed that there were severe limitations to the mechanical analogy for electromagnetic wave propagation that the ether model had suggested.

As is well-known by now, the resolution of the paradox was effected by assuming that one had to add the time dimension to the spatial manifold to produce a four-dimensional spacetime manifold, and that the positive-definite, or "spherical," Euclidian metric had to be replaced with an indefinite hyperbolic one. This had the effect of replacing the origin as the sole "isotropic" point of Euclidian space with the light-cone of Minkowski space, which was named for Hermann Minkowski (1864-1909), who had observed that geometrical aspect of Einstein's theory of special relativity in his address to the 80th Assembly of German Natural Scientists and Physicians at Cologne in 1908 [**19**].

Largely at the suggestion of Marcel Grossman, Einstein then used the extension of Euclidian geometry to Minkowski geometry as the basis for a further extension to the differential geometry of curved spaces [**20**], which was primarily due to Georg Bernhard Friedrich Riemann (1826-1866) at that point in time, although the geometry of Riemann was actually still positive-definite in its signature. Nowadays, one refers to the spacetime manifold as a *Lorentzian manifold*, in honor of the Dutch physicist Hendrik Antoon Lorentz (1853-1928. cf., e.g., [**21**]), who actually did not define such manifolds, although his research in the relativity of electromagnetism led him to introduce transformations, namely, boosts, that augmented the Euclidian spatial rotations to give a six-dimensional group that is now called the *Lorentz group*. Hence, putting his name on the group or the manifolds is simply a gesture of respect for his fundamental role in relativity, like many of the names that get associated with units of measurement.

Eventually, Einstein showed that when one extended Minkowski geometry to Lorentzian geometry, the resulting curvature of spacetime, when coupled to the energy, momentum, and stress in the distribution of matter in spacetime, produced a set of field equations for gravitation. This curvature manifested itself primarily in the deviation of "geodesics" from straight lines into curves, and the solutions to these field equations then represented the Lorentzian metric tensor field, whose components took on the interpretation of gravitational potentials. Indeed, Einstein showed that one could recover the Newtonian law of gravitation in the weak-field limit in the form of the Poisson equation for the gravitational potential functions.

As the theoretical physics community became sufficiently comfortable with the new picture for gravitation as a manifestation of the geometry of the spacetime manifold, they returned to the problem of the relativistic formulation of electromagnetism, first, in Minkowski space and then in a Lorentzian manifold. Eventually, it was found that the most elegant way of representing Maxwell's equations was to first absorb the **E** and **B** field into a single second-rank antisymmetric tensor field F:

$$F = \tfrac{1}{2} F_{\mu\nu}\, dx^{\mu} \wedge dx^{\nu}, \qquad F_{0i} = E_i\,, \qquad F_{ij} = \varepsilon_{ijk} B^{k}.$$

In these expressions, the Greek induces range from 0 to 3, with 0 representing the time dimension of the spacetime manifold, at least for the particular choice of coordinate system, and the Latin indices range over the spatial values of 1, 2, 3. The Levi-Civita symbol ε_{ijk} is completely anti-symmetric and normalized such that $\varepsilon_{123} = 1$. The wedge symbol in the definition of F represents an anti-symmetrized tensor product of the 1-forms dx^{μ} and dx^{ν}; we shall discuss the definitions of all these terms in due course, but, for now, we refer to them casually.

Notice that although we are treating the components of **E** as if they are fundamentally those of a "spatial" covector field, nevertheless, we seem to treating the components of **B** as if they are fundamentally those of a spatial *vector* field that is associated with the components of a 2-form by way of the spatial "duality" that one obtains from the Levi-Civita symbol. As we shall see, this duality plays a key role in the geometry of electromagnetism.

One can then represent four of Maxwell's equations, namely, the three equations for the curl of **E** and the one for the divergence of **B**, quite concisely as:

$$dF = 0,$$

in which the d operator is the exterior derivative operator that acts on completely anti-symmetric covariant tensor fields, which are referred to as *exterior differential forms.*

In order to obtain the remaining four Maxwell equations, one similarly forms the 2-form:

$$G = \tfrac{1}{2} G_{\mu\nu}\, dx^{\mu} \wedge dx^{\nu}, \qquad G_{0i} = D_i\,, \qquad G_{ij} = \varepsilon_{ijk} H^k.$$

One then introduces the Hodge * isomorphism, which is an invertible linear map from 2-forms to 2-form that takes G to:

$$*G = \tfrac{1}{2} *G_{\mu\nu}\, dx^{\mu} \wedge dx^{\nu}, \qquad *G_{0i} = -H_i\,, \quad *G_{ij} = \varepsilon_{ijk} D^k.$$

Here, we have implicitly used the spatial Euclidian scalar product to raise and lower indices.

The last four of Maxwell's equations then take the form:

$$d*G = 4\pi *J \qquad (J = \rho dt + J_i\, dx^i),$$

or, if one defines the codifferential operator $\delta = *^{-1} d *$:

$$\delta G = 4\pi J.$$

One can then append the electromagnetic constitutive law in the form:

$$G = \kappa(F).$$

Before we return to criticize the necessity or desirability of introducing a Lorentzian metric in order to obtain the Maxwell equations in the present form, we return to our historical perspective in order to give further physical reasons for seeking a pre-metric form for them.

4. Attempts at unifying electromagnetism with gravitation. By the time that Einstein's theory of gravitation was becoming widely known to theoretical physicists, in combination with Maxwell's theory of electromagnetism, there was also an increasing suspicion – which mostly originated with Einstein himself – that perhaps the theories of electromagnetism and gravitation could be effectively unified, just as the theories of electricity and magnetism had been.

Many attempts at arriving at such a unification were subsequently made throughout the balance of the Twentieth Century [4], although after the 1930's the emphasis shifted away from the "classical" (more properly, neo-classical) problem of unifying gravitation

[4] An excellent historical reference on the various attempts that were made can be found in Vizgin [**22**]. The geometric aspects of Kaluza-Klein and Einstein-Schrödinger are covered in detail in Part II of Lichnerowicz [**23**].

and electromagnetism to the "quantum" problems that came about with the emergence of quantum field theory as the preferred approach to the theory of elementary particles and their interactions.

What all of the neo-classical models had in common was that they started with the assumption that one still had to consider the geometry of spacetime, and then enlarged the scope of the geometry in one manner or another. Generally, they were also of the Einstein-Maxwell type, in the sense that they attempted to account for the Einstein equations for gravitation, along with the Maxwell equations for electromagnetism. However, we shall treat the solutions to the Einstein field equations that involved finding the metric tensor for a four-dimensional Lorentzian spacetime when one includes the energy, momentum and stress of an electromagnetic field as a source for the curvature as being distinct from this class of unified field theories; such attempts were made by Rainich [**24**] and Wheeler [**25**].

One of the early set of attempts were made by the German mathematical physicist Hermann Weyl [**26**], the French mathematician Élie Cartan [**27**], who was also largely responsible for first emphasizing the utility of exterior differential forms in the expression of differential geometry, and the English astronomer and physicist Sir Arthur Stanley Eddington [**28**]. The basic expansion of scope in the geometry was in weakening the requirement that the spacetime connection, which allows one to define locally define the parallel translation of tangent vectors and frames, did not have to be a metric connection; that is, parallel translation did not have to preserve lengths and angles. This had the effect of introducing four additional components for the connection 1-form, which would then eventually be associated with the components of the electromagnetic potential 1-form $A = A_\mu \, dx^\mu$. Although the actual attempt at unification eventually failed – for instance, Einstein was concerned that the length of a tangent vector under parallel translation might depend upon the choice of curve used for the translation – nevertheless the mathematical apparatus that was introduced still drew considerable attention to the concerns of conformal geometry, as well as defining the tributary of research that eventually became gauge field theory.

Another attempt at unification that showed much promise, and is still being discussed in various forms, was made by Theodore Kaluza [**29**] and Oskar Klein [**30**]. The expansion of scope took the form of the addition of yet another dimension to the spacetime manifold and the extension of the Lorentzian metric $g^{\mu\nu}$, which locally takes on the form of the hyperbolic normal metric of Minkowski space η = diag[+1, –1, –1, –1] in an orthonormal frame field, to a Lorentzian metric G^{AB}, $A, B = 0, \ldots, 4$ with the signature type [+1, –1, –1, –1, –1]. This had the effect of introducing five more components to the metric tensor field, four of which, namely, $G^{\mu 0}$, indirectly accounted for the electromagnetic potentials.

An intriguing consequence of the use of this metric was that when one formed the d'Alembertian operator $\Box_5 = G^{AB} \, \partial^2 \, / \, \partial x^A \partial x^B$ that allows one to define the five-dimensional linear wave equation $\Box_5 \Psi = 0$, one found that separating out the fifth coordinate in the wave function by introducing a separation constant of the form k^2 produced a four-dimensional equation that took the form $\Box\psi + k^2 \psi = 0$ of the Klein-

Gordon equation [5]. In effect, the appearance of the mass in four-dimensions was a by-product of separating out the fifth coordinate when one started with a massless wave equation in five dimensions.

To say that the Kaluza-Klein program failed is somewhat unfair, since it did obtain both the Einstein equations for gravitation and the Maxwell equations of electromagnetism; hence, the theory did not fail in the sense of implying an unacceptable contradiction. The main issues that emerged as controversial were in the interpretation of the fifth coordinate x^4, as well as the remaining extra component of the metric, namely G^{00}, and the fact that no new couplings between electromagnetism and gravitation seemed to emerge. That is, there seemed to be no "induction" terms to examine experimentally that would say that electromagnetic fields might affect gravitational ones or vice versa. Hence, the unification of the field theories was really more of a concatenation.

Actually, there is good reason to assume that the fifth coordinate takes the form of the proper time parameter τ. This has the effect of making the fifth component of velocity vectors represent the speed of propagation of light, so instead of restricting velocity vectors to lie on the unit proper time hyperboloid in a four-dimensional Minkowski space, one requires them to lie on a light cone in a five-dimensional one, along with lightlike ones, which are associated with $\tau = 0$. This way of extending spacetime has an immediate interpretation in terms of manifolds of jets of differentiable curves, although we shall not return to that aspect of jet manifolds when we discuss their geometry later on.

In the process of accounting for fifth coordinate and the G^{00} component of the Kaluza-Klein metric, various researchers, notably Einstein and Mayer [**32**], Cartan [**33**], Oswald Veblen [**34**], and the Dutch mathematicians Jan Schouten and David van Dantzig [**35**] pursued the possibility that introducing the extra coordinate really represented the transition from inhomogeneous coordinates to homogeneous coordinates that one uses in projective geometry. Furthermore, this was consistent with the way that the Kaluza-Klein picture related to the Klein-Gordon wave equation. Of course, since it was still basically the same Kaluza-Klein model, the same criticisms applied, except that one could account for fifth coordinate and the extra metric component. However, to this day, projective relativity still manages to attract serious attention (cf., Schmutzer [**36**]).

Another attempt at Einstein-Maxwell unification that continues to attract modern attention was the one that Einstein, and later Mayer, made [**37**] under the name of *teleparallelism*. It was based the fact that a linear 4-coframe field $\{h_\mu, \mu = 0, \ldots, 3\}$ on a four-dimensional differentiable manifold has sixteen independent components $h_{\mu\nu}$, in general, which is also the total number of independent components that one expects from the metric tensor field $g_{\mu\nu}$ and the electromagnetic potential 1-form A_μ . Now, Roland Weitzenböck [**38**] had already discussed the geometry of parallelizable manifolds – which are sometimes called Weitzenböck spaces, for that reason – which are differentiable manifolds on which there exists a global frame field. For instance, given a choice of global frame field one can define a canonical volume element, Lorentzian

[5] Not only was this the same Klein that came up with the Kaluza-Klein theory, but one finds that his father Felix had also done some work on the subject of how arbitrary geodesic equations could be represented as null geodesic equation by introducing the extra coordinate as one would in projective geometry; this is discussed at length in Rumer [**31**].

metric, and linear connection. However, unlike the Levi-Civita connection that one obtains from a metric tensor, which has zero torsion and generally non-vanishing curvature, the Weitzenböck connection, which is defined to make the given frame field parallel, has just the opposite property. Hence, all of the geometry is in the torsion tensor field. One finds that not only are Lie groups important examples of parallelizable manifolds, but general parallelizable manifolds are, in a sense, "almost Lie groups." In particular, one replaces left-translation with parallel translation and the Maurer-Cartan connection with the Weitzenböck connection.

The failure of teleparallelism took the form of unacceptable contradictions, namely, it did not admit the Schwarzschild solution for the case of a spherically symmetric stationary purely gravitational field and it did admit a stable configuration of gravitating masses with no other forces acting to stop their mutual attraction. However, it is interesting that Einstein and Mayer made no mention of the topological obstructions to the existence of global frame fields, which are sufficiently severe that even such homogeneous spaces as most spheres (except in dimension 0, 1, 3, and 7) do not admit them. Indeed, the first definitive work on the subject of topological obstructions to parallelizability was made by Ethan Stiefel [**39**] in 1936, several years after Einstein and Mayer went on to other things. Stiefel constructed characteristic $\mathbb{Z}_2$-homology classes, whose dual $\mathbb{Z}_2$-cohomology classes are now referred to as *Stiefel-Whitney classes*, since Hassler Whitney preceded the work of Stiefel with his own derivation of the characteristic classes in 1935 [**40**]; however, the paper of Whitney does not explicitly address the problem of parallelizability.

The idea for such characteristic classes came to Stiefel as a generalization of the theorem of his advisor Heinz Hopf on obstructions to non-zero vector fields [**41**], which was also discussed by Henri Poincaré, in which the characteristic class one considers is the Euler class, whose integral, when it is represented by a differential form, over the manifold equals the Euler-Poincaré characteristic. Since the work of Stiefel and Whitney did not seem to find a path to Einstein and Mayer in that era, there was no attempt to extend the theory to the case of a non-parallelizable manifold; i.e., a singular frame field whose singularities might very well generate non-vanishing curvature of topological origin.

It is worth pointing out that in the eyes of pure mathematics the topological issue of whether the spacetime manifold is parallelizable or not is an unavoidable fundamental question to address. Hence, regardless of whether teleparallelism failed to unify gravitation and electromagnetism, the possible physical manifestation of these topological singularities will continue to be a fundamental problem.

Another way of weakening the hypotheses made on the spacetime connection that general relativity introduces to account for gravitation, besides dropping the requirement that it be a metric connection, is to weaken the assumption that it has vanishing torsion. Perhaps as a result of Einstein's prior exposure to non-zero torsion by way of teleparallelism, he [**42**], and later Erwin Schrödinger [**43**], pursued that extension of scope in the geometry of spacetime. Although the Einstein-Schrödinger unification program did not ultimately succeed, nevertheless, it did introduce the importance of considering spacetimes with non-vanishing torsion and the general role it plays, and interest in that topic has not vanished completely even to the present day. Perhaps to

some extent this is due to the seminal papers of Cartan [**44**] on the subject of the geometry of spaces with torsion and their application to the spacetime of general relativity, and partly to the fact that the geometry of connections with non-vanishing torsion and curvature had become quite established in the literature of the theory of plastic materials with continuous distributions of defects that are referred to as *dislocations* and *disclinations*, respectively [6]. Hence, the applicability of the mathematical concept is more general than spacetime structure. The reference by Lichnerowicz also contains a thorough discussion of the Einstein-Schrödinger models.

Yet another extension of Lorentzian spacetime geometry was made by John L. Synge [**46**] and Vranceanu [**47**]. To them, a promising way of interpreting the fifth coordinate in the Kaluza-Klein model was to imagine that actually the four-dimensional spacetime manifold was an *anholonomic hypersurface* in a five-dimensional manifold. This term is really a misnomer, in the same way that the phrase "non-inertial coordinate system" is not rigorously justified, and for the same reason. What one is really dealing with, as in the case of dynamical systems with anholonomic constraints, is a sub-bundle of the tangent bundle to a differentiable manifold that has corank one and is not assumed to be integrable as a differentiable system on the manifold. That is, one is associating a hyperplane in each tangent space in such a manner that there is no foliation of the manifold into leaves of codimension one that might represent spacetime manifolds. In effect, one sees them locally in the tangent spaces, even though they do not exist globally.

Although the method of anholonomic submanifolds attracted only passing attention, again the fact that it addresses fundamental issues, such as integrability and anholonomic constraints, suggests that it is still worthwhile to understand the possible implications of such considerations.

5. Rise of quantum electrodynamics. Eventually, Einstein himself developed some suspicions about the Einstein-Maxwell unification problem that were probably quite definitive. For one thing, the mainstream of physics was increasingly focusing on the problems of the emerging field of quantum physics, which was pursuing directions of approach to its problems and interpretations that were largely inconsistent with the very spirit of Einstein-Maxwell physics. Because the nature of electromagnetism at the atomic-to-subatomic level seemed to be seriously at odds with Maxwellian electromagnetism, Einstein increasingly suspected that the unification of electromagnetism with gravitation might be more meaningful when one introduced the quantum considerations.

One way of seeing that this is probably unavoidable is to note that the modern acceptance of gravito-electromagnetism as a valid extension of Newtonian gravitostatics to gravitodynamics suggests that really the Maxwell equations of electromagnetism – or gravito-electromagnetism, for that matter – are manifestly weak field equations, while the Einstein field equations of gravitation are mostly significant in the realm of large gravitational field strengths, such as in the neighborhood of compact astronomical bodies (neutron stars, black holes, and the like). Hence, one might expect that the Einstein

[6] This train of thought has roots in the early days of relativity theory and has generated a considerable body of literature. For a modern reference that contains citations to classical references one might confer Kleinert [**45**].

equations should be unified with some corresponding strong-field equations of electromagnetism that might pertain to the field strengths that one encounters in the close proximity to elementary charge distributions, such as electrons and nuclei.

Here, one encounters the most fundamental obstacle that separates Einstein-Maxwell field theories from quantum field theories, such as quantum electrodynamics, in particular. Because the nature of sub-atomic physics includes the idea that one will ever directly observe the inner structure of atoms, nuclei, and nucleons, the very methodology of quantum physics quickly took on an *ad hoc* sort of character that is now simply referred to as *phenomenology*. This philosophy is essentially a variation on the basic tenet of solipcism that says "to be is to be measured." Hence, one always treats the system that one is investigating as something of a "black box," such that all one can consider is the way that it responds to inputs. Eventually, one hopes that a sufficiently complete set of input-output relationships will allow one to construct a model for the states of the system inside the box and then speculate on the nature of the system itself. For instance, this is how geophysics constructs models for the Earth's interior out of seismic data and geomagnetic information.

One of the first leaps of faith that quantum electrodynamics made was to stop trying to model the fields of elementary particles as actual fields in the Einstein-Maxwell or continuum-mechanical sense and replace the emphasis on "fields of force" with a new emphasis on the "exchange particle" concept, which was largely due to Werner Heisenberg. Hence, it was only the *interactions* of particle fields that was important, not the fields themselves. A secondary effect of this was to increasingly focus on the scattering of particles as the primary source of information about their structure. At no point did anyone attempt to pose systems of partial differential equations for the fields and solve boundary-value problems for the static case or Cauchy problems for the time evolution of the fields.

Rather, one immediately turned to the problem of constructing the momentum-space kernel for the scattering operator, and eventually a vast set of largely algorithmic procedures for doing this emerged, including algorithms for dealing with the unphysical infinities that appeared as a consequence of the initial steps. Interestingly, what one is implicitly constructing in momentum space is not ultimately the kernel for a nonlinear differential operator in configuration space – indeed the method of Green functions and linear integral operators is not applicable to nonlinear differential operators – but the kernel for a *linear pseudo-differential operator* in configuration space. Hence, if one imagines that the most natural expansion of scope of Maxwell's equations is from linear to nonlinear differential equations as a consequence of going from linear to nonlinear constitutive laws then clearly the momentum-space constructions that follow from the method of Feynman diagrams are not giving one such an extension, at least directly.

Nevertheless, the phenomenological methods of quantum electrodynamics eventually reconnect with classical electromagnetic field theory by the use of *effective* models, such as the ones that were defined by Heisenberg, in conjunction with Hans Euler [**48**], as well Max Born and Leopold Infeld [**49**]. That is, although one usually starts by assuming that the action associated with the fields in question has a classical sort of form, after one has "quantized" it and renormalized it to something that is physical again, one generally finds that supplementary terms have appeared in the action that one identifies as "quantum corrections." For instance, if one quantizes the field theory with the method of functional

integrals, which evolved from the Feynman path integrals of quantum mechanics, then one often expands the Green function in a "loop expansion," which is a perturbation series expansion indexed by the number of loops in the Feynman diagram; the expansion parameter in a loop expansion is then Planck's constant $\hbar$. For instance, zero loops defines the "tree" level of expansion, and corresponds to the original classical theory, while one loop is analogous to the WKB approximation of quantum mechanics. Actually, the methodology of loop expansions is also based in the asymptotic expansions that Poincaré introduced, so geometrical optics represents essentially a "tree-level" approximation to wave optics, while diffraction introduces quantum corrections.

One sees that the effective actions, which usually suggest effective potentials, give one strongly-worded hints as to how one might expand the scope of one's "classical" model to include the quantum corrections. Of course, the ultimate challenge is not to simply refine the numerical accuracy, but to find a better model at the fundamental level. One is reminded of how long astronomy labored in its Ptolemaic phase of adding epicycles to cycles and then epi-epicycles in order to account for the motion of planets with increasing numerical accuracy when the elegant solution was to make the Copernican hypothesis at the outset. Perhaps one day the historians of physics will regard the Feynman diagrams as a sort of Ptolemaic algorithm for obtaining better numerical accuracy in one's agreement with experiments when the elegant solution was something of a Copernican revolution at the fundamental level.

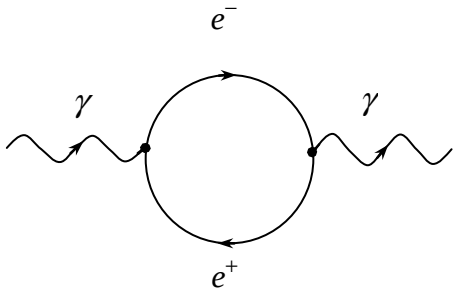
The ultimate challenge to quantum electrodynamics is then to respect the largely empirical nature of its greatest successes while using them as qualitative insights into how one might construct a more fundamental model of the physics at the atomic-to-subatomic level. One can see that the problem defined by the modeling of quantum – i.e., atomic-to-subatomic – phenomena is closely analogous to that of modeling the Earth's interior or that of the Sun in that one cannot expect direct information, so one must always propose effective models, and the measure of effectiveness for any effective model is in its conciseness, as well as its relationship to more general models of physics. For instance, concepts such as spontaneous symmetry breaking are just as important, as well as more directly observable, in the context of condensed matter physics as they are in theoretical particle physics.

In the name of qualitative lessons that one can learn from the empirical successes of quantum electrodynamics, one should start with the fact that every charged fermion, such as an electron or a proton, has an antiparticle that shares its properties except for having an opposite charge, while the photon, being a boson, has no anti-particle. Furthermore, the fact that electrons and positrons are anti-particles manifests itself in the phase transition that takes the form of pair annihilation, which makes anti-particle pairs combine to produce photons, which are generally in the gamma range of energies. The opposite process by which a photon of sufficient energy can split into an anti-particle pair is not as spontaneous, but still just as experimentally established. In particular, one usually needs an external electric or magnetic field in order to convert a photon into an anti-particle pair.

At photon energies slightly above 1 MeV (the rest energy of an electron and a positron), the particles will be electrons and positrons with the excess photon energy being converted into the kinetic energies of the resulting particles. When the photon energy exceeds the rest energy of a muon and its anti-muon (about 211 MeV), there is the

possibility that the particles created are a muon and an anti-muon. Once one reaches the energy level of the rest energy of a pion and an anti-pion (about 280 MeV), the fact that one is now involved with strongly-interacting particles implies that quantum electrodynamics is insufficient and one must recast the problem in the context of quantum chromodynamics. The converse process of pair annihilation is even more involved since an electron-positron collision at a sufficiently high total kinetic energy can create pion-anti-pion pairs, although their lifetimes will be brief, and ultimately one is left with various combinations of lower-energy electron-positron pairs and gamma rays as the stable particles.

These complementary processes of pair creation and annihilation then give rise to the possibility of *vacuum polarization*, in which the presence of photon as an intermediate stage in a particle interaction, such as an electromagnetic exchange particle, can be associated with the spontaneous creation and annihilation of "virtual" anti-particle pairs; i.e., intermediate steps in the process that are not observed directly. Since this sort of process can be represented graphically in the form:



one sees how the concept of loop expansions emerges naturally from elementary principles. In particular, there is nothing to say that the gamma photons in the diagram might not give rise to further virtual pairs with an increasing amount of loops.

The fact that vacuum polarization actually takes place in various quantum scattering and bound-state problems is well-established by numerous experiments by now. One of the most celebrated experimental verifications is the Lamb shift of atomic electron energy levels due to the presence of an anomalous magnetic moment to the electron, in addition to the one that is associated with its spin, since the origin of this anomalous magnetic moment is presumed to be vacuum polarization. Vacuum polarization can also affect the interaction of photons with charges, as well as other photons. The (Delbrück) scattering of photons by nuclear electric fields, which does not exist in Maxwellian electromagnetism, was established by the experiment, and although the scattering of photons by other photons, which is similarly non-existent in Maxwell's theory, was proposed theoretically in the early days of quantum electrodynamics. Interestingly, although the experimental physics community has been promising theoreticians that high-intensity laser technology has been "ten years" away from the critical field strengths for light pulses for quite some decades now, nevertheless, the actual experimental verification of photon-photon scattering has yet to emerge.

Besides the ubiquitous appearance of vacuum polarization in the quantum corrections to classical electromagnetic scattering and bound state problems, another fundamental concept that keeps asserting itself is that of a non-trivial electromagnetic "vacuum state."

Indeed, the very definition of that notion is a fundamental problem. One way of looking at the complexity of the problem is to recall that an electromagnetic wave can be regarded as a continuous distribution of coupled simple harmonic oscillators and then regard the quantum version of that wave – at least heuristically – as a continuous distribution of coupled *quantum harmonic oscillators.* Of course, one key difference between the simple and the quantum harmonic oscillator is the fact that the quantum harmonic oscillator has a small, but non-zero, ground state energy of $1/2\hbar\omega_n$, where ω_n is the natural frequency of the oscillation. Hence, one would expect this to imply that the lowest-energy quantum electromagnetic wave – i.e., photon – would have to also be of non-vanishing total energy, hence, of non-vanishing electric and magnetic field strengths. However, even though one can shift any scale of potential energy for a single oscillator to make the ground state energy precisely zero, nevertheless, when one extends this process to an infinitude of oscillators the necessary shift is not unique, and one finds that there is likely to be some – generally non-unique – *zero-point-field.* One finds that the existence of such a field, which presumably is not due to any specific charge or current source, has been experimentally observed in the form of the *Casimir effect* [**50, 51**]. This effect amounts to the statement that an uncharged perfectly conducting parallel plate capacitor will experience a measurable force of attraction between the plates. However, it is important to observe that the region between the plates is a compact manifold with boundary, so often the effect is attributed to that topological situation, as well as quantum electrodynamics.

The form that quantum electrodynamics eventually settled on (e.g., [**52**]) made it clear that one was dealing with a gauge field theory. That is, the fundamental field, namely, the photon, was best represented by a potential 1-form, which could also be regarded as a connection 1-form on a $U(1)$-principal bundle that one referred to as the *gauge structure* of the field theory; one also refers to the connection form as a *gauge field* or *gauge boson*, since the fact that one is also dealing with a covector field means that the spin of the representation of the Lorentz group that governs its frame transformations is one, and particles that are described by fields of integer spins must obey Bose-Einstein statistics in their energy levels. By contrast, the charged particles, such as electrons and positrons, are *fermions*, since they are represented by Dirac bi-spinor wave-functions, which are due to half-odd-integer spin representation of the Lorentz group and obey Fermi-Dirac statistics. This association of spin and statistics is widely regarded as one of the foundations upon which quantum field theories all rest (see, for instance, Streater and Wightman [**53**]).

The theoretical successes that followed, in the form of the further unification of the quantum theory of electromagnetism with that of the weak interaction, further established the methodology of gauge field theories as a powerful tool in the formulation of quantum field theories [**54, 55**]. In particular, the concept of spontaneous symmetry breaking in the vacuum state played a key role in this unification. Once again, the key qualitative notion is that of a non-unique vacuum ground state for the gauge field in question, although when the non-uniqueness follows from the existence of an unbroken vacuum symmetry group the set of vacuum states will have the geometrical and topological character of a "homogeneous space," such as circles, spheres, and torii.

6. Spacetime topology and electromagnetism. In addition to the increasing emphasis that was placed on the problems of quantum electrodynamics, another tributary of research in electromagnetism that grew more out of general relativistic consideration was the increasing interest in the role that spacetime topology played in electromagnetism.

Part of this increasing interest was due to the possible role of spacetime topology in the theory of gravitation, such as the spacetime singularities that one expected to find in the vicinity of black holes and the Big Bang singularity. Indeed, when one models spacetime as a differentiable manifold, it becomes unavoidable that one must deal with the global nature of its topology, and not just the local nature, which is still Euclidian.

Another contribution to the increasing interest in the role of topology in electromagnetism was due to the concerns of gauge field theory. In particular, when one is dealing with connections on principal fiber bundles from the outset one must similarly consider the global nature of the situation, such as whether the principal bundle is trivial or not; this is equivalent to the physical question of whether a global choice of gauge is possible. One finds that the issues associated with the triviality of G-principal bundles $P \to M$ over the spacetime manifold M, where G is the gauge group of the field theory, depend upon both the topology of M, as well as the topology of G, in a deep and complicated way.

Finally, the ambition to examine the role of topology in electromagnetism can arise at a much more elementary level that precedes the considerations of either general relativity or quantum electrodynamics, namely, the very nature of the constructions that one makes in classical electromagnetism itself.

For instance, the elementary concepts of charge, flux, and current can be most effectively defined in the language of topology, or, more to the point, *homology*. One finds that the sources of electromagnetic fields are usually represented by elementary finite chain complexes with real coefficients, such as finite sets of points or networks of current loops, while the concepts of flux and potential difference take the form elementary cochains with real coefficients. The introduction of field strength vector fields and covector fields with such cochains then follows naturally from the considerations of de Rham's theorem. Furthermore, electrostatics and potential theory can be elegantly formulated in the language of Hodge theory.

A key construction in all of this is that of an *orientation* on the spacetime manifold M (or really, its tangent bundle $T(M)$). This is because one must introduce a volume element on a manifold in order to define the integration of differential forms that allows one to speak of flux and the bilinear pairing of k-chains with k-cochains, and before one can introduce a volume element one must assume the orientability of the vector bundle $T(M)$. Since not every differentiable manifold is orientable, one must accept that there are topological obstructions to the existence of global orientations. However, every simply connected manifold is orientable and every manifold has a simply connected orientable covering manifold.

The existence of a volume element on an orientable manifold allows one to define a set of linear isomorphisms between k-vector fields and $n–k$-forms that one can call *Poincaré duality*. Although its role in homology theory is well-established by now, one should realize that it started off as a more projective-geometric sort of concept than a topological one. In particular, any k-dimensional subspace of an n-dimensional vector

space can either be spanned directly by k linearly independent vectors, whose exterior product is then a non-zero k-vector, or annihilated by $n–k$ linearly independent covectors, whose exterior product is then a non-zero n-k-form.

Of course, there is nothing unique about the choice of spanning vectors or annihilating covectors, but any other choice would affect the resulting k-vector or $n–k$-form only by a non-zero scalar multiplication. One then confronts the *Plücker-Klein* representation of k-dimensional vector subspaces in an n-dimensional vector space as lines through the origin in either the vector space of k-vectors over it or the vector space of $n–k$-forms over it; that is, one considers points in the projective spaces that these vector spaces define.

Hence, one sees that the same concept of orientation that is so fundamental to the basic concepts of electromagnetism has both a projective-geometric and a topological aspect to it. Mostly, we shall be concerned with the projective-geometric aspect in what follows. Indeed, a recurring theme is that just as metric differential geometry seems to be the natural language for the description of gravitation, projective differential geometry seems to be the natural language for the description of electromagnetism.

7. Pre-metric electromagnetism. As one sees, the mainstream of theoretical physics in the Twentieth Century largely split into the two warring camps of gravitational theorists, whose main concern was the geometry, and sometimes the topology of the spacetime manifold, and the quantum field theorists, whose main concern was to obtain the best possible agreement between the relativistic particle scattering amplitudes that were obtained by the method of Feynman diagrams and the actual scattering data that was being obtained by high-energy particle scattering experiments.

It is therefore not surprising that one of the lesser-discussed topics in electromagnetism that largely fell through the cracks between these two communities was based in the early observations of the German physicist Friedrich Kottler [**56**] that actually the metric tensor field that was so fundamental to Einstein's theory of gravitation seemed to play only an incidental role in formulating the Maxwell equations of electromagnetism. Indeed, one could formulate them without the necessity of introducing a metric.

Other physicists and mathematicians, such as Hargeaves [**57**], Cartan [**44**], and Bateman [**58**] made similar observations in the course of their own discussion of electromagnetism, and eventually van Dantzig [**59**] published a series of papers that expanded upon the formulation of Maxwell's equations in the absence of a metric. The topic of pre-metric electromagnetism did not die away completely, though, and occasionally resurfaced in the work of Murnaghan [**60**], Post [**61**], Truesdell and Toupin [**62**], Hehl, Obukhov, and Rubilar [**63**, **64**], and was masterfully presented in the modern language of exterior differential forms by Hehl and Obuhkov [**65**]. The present work comes about as the author's attempt to summarize the topics that he himself pursued in response to the work of Hehl and Obukhov, and is intended to only enlarge the scope of the subject of pre-metric electromagnetism with other specialized techniques and issues, and not to replace their definitive treatise.

The picture of electromagnetism that emerges is based in the idea that the Lorentzian structure g on the spacetime manifold enters into Maxwell equations only by way of the

Hodge * isomorphism, as it acts on 2-forms, in particular. Hence, one must take a closer look at this isomorphism from both a mathematical and a physical perspective.

One finds that it is best to factor the isomorphism *: $\Lambda^2 M \to \Lambda^2 M$ into a product of two isomorphisms, where $\Lambda^2 M \to M$ is the bundle of 2-forms over M. The first one $i_g \wedge i_g$: $\Lambda^2 M \to \Lambda_2 M$ is the one that "raises both indices" on 2-forms, where $\Lambda_2 M \to M$ is the bundle of bivectors on M. The second one #: $\Lambda_2 M \to \Lambda_2 M$ is the aforementioned Poincaré duality that one obtains from the assumption that M has a volume element on it.

Now, if one were to be more explicit about the introduction of an electromagnetic constitutive law into Maxwell's equations then one would find that the role of the isomorphism $i_g \wedge i_g$ could be absorbed into that constitutive law, as long as one represented it as a vector bundle isomorphism κ: $\Lambda^2 M \to \Lambda_2 M$, at least for linear media.

One then finds that in order to account for the introduction of a Lorentzian structure, one can first consider the dispersion law for electromagnetic waves that follows from the Maxwell equations, when formulated in this "pre-metric" fashion. What one obtains is generally a quartic hypersurface in the cotangent bundle $T^*M \to M$, rather than the usual quadratic one that derives from the light cones of a Lorentzian metric g. A further reduction of the homogeneous quartic polynomial that one obtains from the dispersion law into a square of a quadratic one of Lorentzian type is possible only if the constitutive law has the proper sort of symmetries, such as spatial isotropy.

When viewed in this light, one sees that in a real sense the gravitational structure of spacetime, as defined by the Lorentzian structure on its cotangent bundle, is a specialized *consequence* of assumptions that one makes about the electromagnetic constitutive laws of the spacetime manifold. Indeed, if one goes back to the historical progression that led Einstein from electromagnetism to gravitation then one sees that the connecting link between them related to the structure of the symbol of the wave operator that governed the propagation of electromagnetic waves in the spacetime manifold. Hence, one sees that gravitation relates to *light* cones and not purely gravitational ones that are unrelated to the propagation of electromagnetic waves.

What emerges from the foregoing picture is nothing short of a major paradigm shift in the consideration of spacetime geometry from the metric geometry of tangent vectors that pertains to gravitational geodesics to the projective geometry of 2-forms and bivectors (i.e., 2-planes) that pertains to the propagation of electromagnetic waves. This also suggests that one is redirecting one's focus from the geometry of Riemann to the geometry of Klein, who once asserted that "projective geometry is all geometry." Hence, there is something mathematically satisfying about the fact that one is simply moving to a higher plane of generality in one's discussion of geometry.

Another subtlety that appears is the fact that the Hodge * isomorphism has the property that if the metric g is Lorentzian then $*^2 = -I$ when * is applied to 2-forms. This means the linear operator * defines an "almost-complex structure" on the real vector bundle $\Lambda^2 M$. However, as observed by the author [**66**], if one starts with an electromagnetic constitutive law and defines $* = \# \cdot \kappa$ then one must note that not all of the physically meaningful laws give * isomorphisms with such a property. In particular, anisotropic optical media will not define almost-complex structure in this way. When one is dealing with such a class of media, though, one can also introduce concepts of

complex projective geometry accordingly. For instance, the complex projective space $\mathbb{C}P^2$ and its dual play an important role in the geometry of electromagnetic waves.

Since the geometry of metrics and geodesics seems to follow only after one passes from the pre-metric Maxwell equations to the characteristic equation that they define in the geometrical optics approximation, one sees that the transition from wave optics to geometrical optics implies a corresponding transition from wave geometry to geodesic geometry, just as one goes from wave mechanics to geometrical mechanics in quantum theory. The question then arises of how one might define wave geometry in general, and one finds that the solution of this problem is well-known: the geometry of waves is the contact geometry that was previously mentioned in the context of geometrical optics.

One begins to gain some inkling of how Einstein's suspicions about the unification of electromagnetism and gravitation were probably quite accurate: It would probably have to involve some simultaneous incorporation of quantum considerations and expansion of scope in spacetime geometry. Of course, the problem of resolving the growing gap between the very formalism, if not the natural philosophy, of relativity and gravitation with that of quantum field theory has always been regarded as every bit as perplexing as the Einstein-Maxwell unification problem. One might suspect that a better understanding of wave geometry might lead to a resolution of both the problem of reconciling general relativity with quantum physics and unifying the theory of electromagnetism with that of gravitation.

One also begins to see that gravity appears as an "emergent" phenomenon when one starts with the electromagnetic structure of spacetime; i.e., the dispersion law that follows from its electromagnetic constitutive law. Hence, one suspects that the best way to formulate the mathematics of gravity is in terms of 2-forms to begin with. Of course, this approach has been around since the 1960's in the form of "complex relativity," except, as we shall point out later, it becomes redundant to complexify the vector bundle $\Lambda^2 M$ when one has already introduced an almost-complex structure.

Another possible expansion of scope that is associated with pre-metric electromagnetism is the fact that the very significance of Lorentz invariance in physics is based on the assumption that one is concerned with a quadratic dispersion law for electromagnetism. One sees that the more general case of a quartic dispersion law implies that one might be required to alter one's conception of the invariance of the laws of Nature accordingly. This gives one a more tangible origin for the possible violation of Lorentz invariance.

So far, the emphasis on electromagnetic constitutive laws that result in quartic dispersion laws has mostly been based in macroscopic optical sorts of considerations. However, one finds that the effective models for quantum electrodynamics that take the form of the Heisenberg-Euler effective Lagrangian and the Born-Infeld both produce constitutive laws of the "bi-isotropic" type, in the language of Lindell [**67**], as well as dispersion laws that are of the "birefringent" type. Hence, one sees that pre-metric electromagnetism might give some new insight into the problem of the modeling of the electromagnetic vacuum state when one takes into account the considerations of quantum physics.

8. Summary of contents. Since the application of the calculus of exterior differential forms to the formulation of problems in electromagnetism as a substitute for vector calculus is better known to theoretical physicists who are usually only interested in Maxwellian electromagnetism as a springboard to more general gauge field theories, we begin our discussion of pre-metric electromagnetism with a chapter on that more recent [7] calculus. Furthermore, since most of the theoretical discussions are immediately concerned with topological matters – e.g., triviality of principal bundles – it is never emphasized that that exterior calculus can be applied at a much more elementary level in electromagnetism than the topological level, namely, at the level of a typical undergraduate course in electromagnetism. Hence, we shall attempt to exhibit the formulation of as many of the traditional non-theoretical topics in electromagnetism as possible in terms of differential forms to show that they are more than just a topological necessity.

Of course, the topological aspects of differential forms are still essential in any discussion of the fundamental concepts in electromagnetism, so Chapter II will summarize the usual issues in the topology of differential manifolds to the extent that is necessary for the discussion of those fundamental concepts. Although some attempt has been made at making the chapter self-contained and based only upon a knowledge of linear algebra and multi-variable calculus, nonetheless, it is suggested that readers with no prior exposure to differentiable manifolds will probably find the chapter a bit too concise.

In Chapter III, we discuss the topological nature of charge, flux, current, and field strengths in both the static electric and magnetic contexts. We also discuss the distinction between the field strength and the associated excitations of the medium that they produce (also called "inductions," by many researchers), as well as the introduction of potential functions for electric field strength 1-forms and potential 1-forms for magnetic field strength 2-forms, which then represent the classical "vector potentials".

Then, in Chapter IV, we show how one goes from electrostatics and magnetostatics to electrodynamics by the introduction of the two types of electromagnetic induction, namely, the ones that are described by Faraday's law and Maxwell's law. One can also assemble electric charge density and the electric current into a four-dimensional vector field, the vanishing of whose divergence is equivalent to the conservation of charge. We finally assemble all of the accumulated laws for electric and magnetic fields into the pre-metric Maxwell equations, which we express in both their three-dimensional time+space form, as well as the more general four-dimensional form. We also discuss the four-dimensional introduction of electromagnetic potential 1-forms for the field strength 2-form, which is at the basis of all gauge approaches to electromagnetism, whether classical or quantum.

In Chapter V, we discuss the largely phenomenological nature of electromagnetic constitutive laws, except that we continue to use the language of differential forms and multivector fields to the greatest extent possible, while most treatments of the subjects are oriented towards experimental physicists, whose preference is for the methods of vector

[7] Apparently, the transfer of wisdom from mainstream mathematics to mainstream physics is getting sufficiently sluggish that "new mathematics" generally encompasses most of the Twentieth Century. In particular, the concept of differential forms was discussed by pure mathematicians such as Darboux and Goursat in the late Nineteenth Century.

calculus. Among the examples of specific constitutive laws, we give not only ones familiar to experimental physicists, such as optical media and plasmas, but also more theoretical ones, such as Lorentzian and almost-complex media. The example of bi-isotropic media is shown to include both the constitutive laws of the Heisenberg-Euler effective model for the propagation of electromagnetic waves in the presence of background electromagnetic fields when one includes one-loop quantum corrections to the Maxwellian Lagrangian, and the Born-Infeld model, which is closely related to the Heisenberg-Euler model, but based in other considerations.

Because the pre-metric Maxwell equations give partial differential equations when one expresses them locally, Chapter VI is concerned with the formulation of the basic notions from the theory of partial of differential equations that relate to classical electromagnetism when one formulates those equations on differentiable manifolds. Since there is no universal agreement amongst mathematicians as to the "best" way to formulate partial differential equations on manifolds, we discuss three of the most popular approaches: differential operators on vector bundles, hypersurfaces in jet manifolds, and exterior differential systems, as well as how they relate to each other. We then attempt to formulate the traditional boundary-value problems of electrostatic or magnetostatic potential theory, along with the Cauchy problem of electrodynamics, in that language. Although the methods of Green functions and Fourier transforms are fundamentally limited to the treatment of linear differential equations, as well as being rather difficult to deal with specifically when the manifold in question does have a high degree of homogeneity (affine spaces, spheres, etc.), nevertheless, we attempt to at least define the nature of the mathematical problems involved.

In Chapter VII, we return to the physics of electromagnetism and discuss the issues that arise when one looks at all of the ways that electromagnetic fields and source distributions can interact with each other. However, since the interaction of sources and fields that takes the form of electromagnetic radiation is quite involved in its own right, we content ourselves with only a cursory discussion of some of the issues and defer a more detailed analysis of pre-metric radiation theory – if there even is such a thing – for a later monograph.

Chapter VIII represents the culmination of the pre-metric theory of electromagnetism in which one sees that the Lorentzian structure of spacetime can emerge from the dispersion laws for the medium that one obtains from the field equations when one specifies a constitutive law. However, the quadratic dispersion law that gives a Lorentzian metric is a degenerate case of the more general quartic polynomial law, and is usually associated with isotropic media.

In Chapter IX, one then finds that the geometrical optics approximation also represents an approximation to the geometry of waves by null geodesics in the same way that it represents an approximation to the physics of electromagnetic waves by light rays. Interestingly, this was also the driving force behind the early formulation of quantum wave mechanics, namely, to make its relationship to classical mechanics be analogous to the relationship of wave optics to geometrical optics. In optics, one usually attempts to go beyond the geometrical optics approximation by the introduction of diffraction effects and asymptotic expansions, so we discuss what that might mean in the context of pre-metric electromagnetism.

In Chapter X, we discuss the way that energy and action are associated with electromagnetic fields in both the static and dynamics contexts. We then show how one defines pre-metric Lagrangians for the static and dynamic field equations, as well as the equations of motion for an electric charge in the presence of the Lorentz force. We finally show that the usual statement of Fermat's principle for the variational formulation of the spatial light rays (spatial projections of null geodesics) is fraught with subtleties when one attempts to generalize from metric geometry to pre-metric geometry since even the nature of the projection of spacetime onto space needs to be considered in the language of projective geometry, along with a rethinking of the basic action functional that gives the elapsed-time along a curve.

The last three chapters essentially summarize some of the author's previous research into the extension of some traditional results in classical electromagnetism and general relativity. The expansion of scope that comes about when one attempts to examine the symmetries of the pre-metric field equations themselves is discussed in Chapter XI, when it is known that in the Lorentzian case, one must go from Lorentzian invariance to conformal Lorentzian invariance, among other extensions. Since projective geometry has been persistently suggesting itself as an unavoidable expansion of the scope of metric geometry, the topic is discussed at a more elementary level in Chapter XII and applied to elementary physical mechanics, as well as electromagnetism. A key construction is the Plücker-Klein association of linear planes in $\mathbb{R}^4$ with lines through the origin in either the vector space of bivectors or 2-forms over $\mathbb{R}^4$ by way of decomposable bivectors and 2-forms. Finally, Chapter XIII discusses the fact that the usual formulation of "complex relativity" in terms of 2-forms is naturally embedded in the present context, and indeed, is probably an essential part of incorporating "gravitational" considerations into pre-metric electromagnetism, in the form of more general spacetime connections than the Levi-Civita connection that is defined by its Lorentzian structure.

It would be inexcusably ungrateful of the author of this monograph to fail to mention the continued encouragement that he received from Friedrich Hehl at Cologne, without whose counsel this book would have ever materialized.

References

1. E. T. Whittaker, *A History of the Theories of Aether and Electricity*, 2 vols., Nelson, London. Vol. 1: *The Classical Theories* (1910, revised and enlarged 1951), Vol. 2: *The Modern Theories 1900-1926* (1953); reprinted by Harper Torchbooks, New York, 1960.
2. J. C. Maxwell, *Treatise on Electricity and Magnetism*, 2 vols., Dover, New York, 1954. (Reprint of 3rd ed., 1891).
3. Jackson, J.D., *Classical Electrodynamics*, 2nd ed., Wiley, New York, 1976.
4. I. Newton, *Optiks, or a treatise on reflexions, inflexions, and colours of light*, London, 1704. Reprinted by Dover, Mineola, N. Y., 1952.

5. C. Huygens, *Traité de la lumière; un discourse de la cause de pesanteur*,1690. English translation: *Treatise on Light*, by Silvanus P. Thompson, University of Chicago Press, 1950.
6. E. L. Malus, "Mémoire sur l'optique," J. l'école poly. **7** (1808), 1-44, 84-129.
7. A. J. Fresnel, "Supplement au mémoire sur la double refraction 1821," Oeuvres 2, pp. 343-367. Imprimerie Imperiale, Paris, 1858..
8. W. R. Hamilton, "Theory of systems of rays," Trans. Irish Acad. **15** (1828), pp. 69-178; supplements in ibid. **16** (1830), 1-61, 93-125; **17** (1837), 1-144.
9. F. Klein, "Über neuere englische Arbeiten zur Mechanik," Jber. Deutsch. Math.-Vereinig, Bd. 1 (1891/92) and in Ges. math. Abh. Bd II, pp. 601-602. Springer, Berlin, 1922.
10. H. Bruns, "Das Eikonal," Abh. math. phys. cl. sächs. Akad. Wiss. **21** (1895), 323-436.
11. F. Klein, "Über das BRUNS'SCHE Eikonal," Z. Math. u. physik. **46** (1901); Ges. math. Abh., Bd. II, pp. 603-606.
12. S. Lie, *Theorie der Transfromationsgruppen*, Leipzig, 1888.
13. E. Vessiot, "Sur l'interprétation mécanique des transformations de contact infinitesimales," Bull. Soc. Math. de France **34** (1906), 230-269.
14. H. Helmholtz, Journ. f. Math. **57** (1859), 7
15. G. G. Stokes, "Dynamical Theory of Diffraction," Trans. Camb. Phil. Soc. **9** (1849), 1.
16. G. Kirchhoff, Berl. Ber. (1882), 641; Ann. d. Phys. (2) **18** (1883), 163; Ges. Abh. Nachtr., pp. 22.
17. H. Poincaré, "Sur la problème des trios corps et les equations de la dynamique," Acta Math. **13** (1890), 5-270.
18. A. Einstein, "Zur Elektrodynamik bewegten Körper," Ann. Phys. (Leipzig) **17** (1905); English translation in *The Principle of Relativity. A collection of papers on the special and general theory of relativity.* Dover, New York, 1952. (Loc. cit.)
19. H. Minkowski, "Space and Time," Address to the 80th Assembly of German Natural Scientists and Physicians at Cologne, 1908; English translation in loc. cit.
20. A. Einstein, "Die Grundlage der allgemeinen Relativitätstheorie," Ann. Phys. (Leipzig) **49** (1916); English translation in loc. cit.
21. H. A. Lorentz, *Theory of Electrons*, Dover, NY, 1952.
22. V. P. Vizgin, *Unified Field Theories*, Birkhäuser, Boston, 1994.
23. A. Lichnerowicz, *Théorie relativiste de la gravitational et de l'electromagnetisme*, Masson and Co., Paris, 1955.
24. G. Y. Rainich, "Electrodynamics in the General Relativity Theory," Proc. N. A. S. **10** (1924), 124-127; "Electrodynamics in the general relativity theory," Trans. Am. Math. Soc. **27** (1925), 106-136.
25. C. W. Misner and J. A. Wheeler, "Physics as Geometry," Ann. Phys. (New York) **2** (1957), 525-660; reprinted in *Geometrodynamics*, Academic Press, New York, 1962.
26. H. Weyl, *Space, Time, and Matter*, Dover, New York, 1952. (English translation of *Raum-Zeit-Materie*, 4th ed., Springer, Berlin, 1921.)
27. E. Cartan, "Une classe d'espaces de Weyl," Ann. Ec. Norm. **60** (1943), 1-16.

28. A. S. Eddington, *The Mathematical Theory of Relativity*, 3rd ed., Chelsea, New York, 1975. (First published by Cambridge University Press in 1923.)
29. T. Kaluza, "Zum Unitätsproblem der Physik," Sitz. d. preuss. Akad. Wiss. Berlin (1921), 966; English translation in *Modern Kaluza-Klein Theories*, ed., T. Applequist, A. Chodos, and P. G. O. Freund, Addison-Wesley, Menlo Park, 1987.
30. O. Klein, "Quantentheorie und fünf-dimenzionale Relativitätstheorie," Zeit. f. Phys. **37**, 895 (1926); English translation in *Modern Kaluza-Klein Theories*, ed., T. Applequist, A. Chodos, and P. G. O. Freund, Addison-Wesley, Menlo Park, 1987.
31. Y. B. Rumer, *Investigations in 5-optics*, Gosudarstvennoe Izdatel'stvo Tekhniko-teoreticheskoi Literatur', Moscow, 1956 (Russian).
32. A. Einstein and W. Mayer, "Einheitliche Theorie von Gravitation and Elektrizität," Sitz. preuss. Akad. Wiss. 25 (1931), 541-557. also, A. Einstein, "Zur Kaluza's Theorie des Zusammenhanges von Gravitation und Elektrizität," Sitz. preuss. Akad. Wiss (1927), 32-40.
33. E. Cartan, "Le theorie unitaire de Einstein-Mayer," Oeuvres Complètes, Ed. du C. N. R. S., Paris, 1984.
34. O. Veblen, *Projektive Relativitätstheorie*, Springer, Berlin, 1933.
35. J. A. Schouten and D. van Dantzig, "On projective connections and their application to the general field theory," Ann. Math. **34** (1933), 271-312.
36. E. Schmutzer, *Projektive Einheitliche Feldtheorie mit Anwendungen in Kosmologie und Astrophysik*, Harri Deutsch, Frankfurt, 2004.
37. A. Einstein, Sitz. Preuss. Akad. Wiss. (1928), 217-221; (1929), 2-7; 156-159; (1930), 18-23; and W. Mayer, 110-120, 410-412.
38. R. Weitzenböck, "Differentialinvarianten in den Einstein'sche Theorie der Fernparallelismus," Sitz. preuss. Akad. Wiss. (1928), 466; *Invariantentheorie*, Noordhoff, Gronigen, 1923.
39. E. Stiefel, "Richtungsfelder und Fernparallelismus in *n*-dimensionalen Mannigfaltigkeiten," Comm. Math. Helv., **8** (1936), 3-51.
40. H. Whitney, "Sphere spaces," Proc. Nat. Acad. Sci. **21** (1935), 462-468; "On the theory of sphere bundles," ibid. **26** (1940), 148-153.
41. H. Hopf, "Vektofelder in *n*-dimensionalen Mannigfaltigkeiten," Math. Ann. **96** (1926/27), 225-250.
42. A. Einstein and B. Kaufmann, "A new form of the general relativistic field equations," Ann. Math. **62** (1955), 128-138.
43. E. Schrödinger, *Space-time Structure*, Cambridge University Press, 1954.
44. E. Cartan, *On manifolds with an affine connection and the theory of relativity*, Bibliopolis, Napoli, 1986 (English translation by A. Ashtekar of a series of French articles from 1923 to 1926).
45. H. Kleinert, *Gauge Fields in Condensed Matter*, World Scientific, Singapore, 1989.
46. J. L. Synge, "Geodesics in non-holonomic geometry," Math. Ann. **99** (1928) 738-751.
47. G. Vranceneau, *Les espaces non holonomes*, Mem. Math. Sci., fasc. **76**, Gauthier Villars, Paris, 1936.
48. W. Heisenberg and H. Euler, "Folgerungen aus der Dirac'schen Theorie des Positrons," Zeit. f. Phys., **98** (1936), 714-732.

49. M. Born and L. Infeld, "Foundations of a New Field Theory," Proc. Roy. Soc. A, **144** (1934), 425-451; M. Born, "Théorie non-linéaire du champs électromagnétique, Ann. Inst. H. P., **7** (1937), 155-265.
50. H. B. G. Casimir, Proc. Kon. Nederl. Akad. Wet. **51** (1948), 793-796.
51. V. M. Mostepanenko and N. N. Trunov, *The Casimir Effect and its Applications*, Clarendon Press, Oxford, 1997.
52. V.B. Berestetskii, E.M. Lifschitz, L.P. Pitaevskii, Quantum Electrodynamics, 2nd ed., Elsevier, Amsterdam, 1984.
53. R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and all that*, Benjamin/Cummings, Reading, MA, 1964.
54. E. S. Abers and B. W. Lee, "Gauge Theories," Phys. Rep. **9** (1973), 1-141.
55. S. Pokorski, *Gauge Field Theory*, Cambridge Univ. Press, Cambridge, 1990.
56. F. Kottler, "Maxwell'sche Gleichungen und Metrik," Sitz. Akad. Wien IIa, **131** (1922), 119-146.
57. R. Hargreaves, "On integral forms and their connection with physical equations," Camb. Phil. Soc. Trans. **21** (1908), 116.
58. H. Bateman, "The transformation of the electrodynamical equations," Proc. London Math. Soc. [2] **8** (1910) 223-264.
59. D. van Dantzig, "The fundamental equations of electromagnetism, independent of metrical geometry," Proc. Camb. Phil. Soc. **30** (1934), 421-427.
60. F. D. Murnaghan, "The absolute significance of Maxwell's equations," Phys. Rev. (2) **17** (1921), 73-88; *Vector Analysis and the Theory of Relativity*, Johns Hopkins Press, Baltimore, 1922.
61. E. J. Post, *Formal Structure of Electromagnetics*, Dover, New York, 1997.
62. Truesdell and Toupin, "The Classical Field Theories," in *Handbuch der Physik* III/1, ed. S. Flügge, Springer, Berlin, 1960, pp. 226-793.
63. F. W. Hehl, Y. N. Obukhov, and G. F. Rubilar, "Spacetime metric from linear electrodynamics II," Ann. Phys. (Leipzig) **9** (2000), Special issue, SI-71-SI-78.
64. Y. N. Obukhov and G. F. Rubillar, "Fresnel Analysis of the wave propagation in nonlinear electrodynamics," arXiv.org, gr-qc/0204028.
65. F. W. Hehl and Y. N. Obukhov, *Foundations of Classical Electrodynamics*, Birkhäuser, Boston, 2003.
66. D. H. Delphenich, "On linear electromagnetic constitutive laws that define almost-complex structures," Ann. Phys. (Leipzig) **16** (2007), 207-217.
67. I. Lindell, *Differential Forms in Electromagnetics*, IEEE Press, Wiley-Interscience, N. J., 2004.

Chapter I

Calculus of exterior differential forms

Although at the present time it is traditional for mathematicians – and even some physicists – to first introduce the calculus of exterior differential forms in the context of differentiable manifolds, nevertheless, since the purpose of this study is to begin the discussion of electromagnetism in the same place that conventional physics does, it seems more appropriate to illustrate that there are really three advantages to the use of differential forms over vector calculus: computational conciseness, dimensional generality, and their sensitivity to the topology of the space in question. Hence, the sole purpose of this first chapter will be to show how the calculus of exterior differential forms on vector spaces neatly subsumes the main results of vector calculus. Chapter II will then address the topological aspects of differential forms as a separate issue.

Most of the material in this chapter can be found, in one form or another, in the introductory chapters of most books on differential manifolds [**1-3**], the later chapters of some books on advanced calculus [**4-6**], and various books on differential forms themselves [**7**, **8**]. The list of references at the end of the chapter is only a sample of these possibilities.

1. Multilinear algebra. The topic of *exterior algebra* is actually a special case of the more general topic of *tensor algebra*, or, as it is sometimes called, *multilinear algebra* [**9**]. Hence, we shall begin by summarizing a few generalities from the latter branch of mathematics.

Let V represent a vector space of dimension n whose scalars come from the field $\mathbb{R}$ of real numbers; later, we shall return to some of the same notions when it becomes necessary to discuss complex vector spaces. We also let V^* represent the dual vector space to V – viz., the vector space of all linear functionals on V. Hence, if $\phi \in V^*$, $\mathbf{v}, \mathbf{w} \in V$, and $\alpha, \beta \in \mathbb{R}$ then we will always have:

$$\phi(\alpha\mathbf{v} + \beta\mathbf{w}) = \alpha\,\phi(\mathbf{v}) + \beta\,\phi(\mathbf{w}). \tag{I.1}$$

a. Bilinear functionals. A function T: $V \times V \to \mathbb{R}$ is said to be *bilinear* if it is linear in each variable separately:

$$T(\alpha\mathbf{u} + \beta\mathbf{v}, \mathbf{w}) = \alpha\,T(\mathbf{u}, \mathbf{w}) + \beta\,T(\mathbf{v}, \mathbf{w}), \tag{I.2a}$$
$$T(\mathbf{u}, \alpha\mathbf{v} + \beta\mathbf{w}) = \alpha\,T(\mathbf{u}, \mathbf{v}) + \beta\,T(\mathbf{u}, \mathbf{w}), \tag{I.2b}$$

If $\{\mathbf{e}_i, i = 1, \ldots, n\}$ is a basis for V, so $\mathbf{u} = u^i\,\mathbf{e}_i$, $\mathbf{v} = v^j\,\mathbf{e}_j$ (the summation convention is always in effect unless specified to the contrary) then from bilinearity one will have:

$$T(\mathbf{u}, \mathbf{v}) = u^i v^j \, T(\mathbf{e}_i, \mathbf{e}_j). \tag{I.3}$$

If we introduce the notation $T_{ij} = T(\mathbf{e}_i, \mathbf{e}_j)$, which one refers to as the *components of T with respect to the basis* $\mathbf{e}_i$, then we can write (I.3) in the component form:

$$T(\mathbf{u}, \mathbf{v}) = T_{ij} \, u^i v^j \, . \tag{I.4}$$

Suppose we perform a change of basis $\mathbf{e}_i \rightarrow \overline{\mathbf{e}}_i$, which defines an invertible linear transformation of V with a matrix A_i^j relative to the basis $\mathbf{e}_i$ that is defined by the basic equations:

$$\overline{\mathbf{e}}_i = A_i^j \mathbf{e}_j \, . \tag{I.5}$$

The components $\overline{u}^i$ of $\mathbf{u}$ with respect to the transformed basis $\overline{\mathbf{e}}_i$ are then obtained from the assumption that both of the linear combinations $u^i \mathbf{e}_i$ and $\overline{u}^i \overline{\mathbf{e}}_i = \overline{u}^i A_i^j \mathbf{e}_j$ must define the same vector, namely, $\mathbf{u}$. Hence, one must have:

$$\overline{u}^i = \tilde{A}_j^i u^j \, , \tag{I.6}$$

in which we are using a tilde to denote the matrix inverse to A_i^j. That is, the transformation of components for vectors is inverse to the transformation of bases, a situation that one sometimes refers to as *contragredience*, while the vector $\mathbf{u}$ is said to be *contravariant*. Since the purely component-oriented formulation of tensor algebra usually performs the transformation of components first, modern mathematics adapted by usually assuming that bases – or *frames* – are the objects that transform by the inverse.

The effect of this change of basis on the components T_{ij} is obtained by inserting $\mathbf{u} = \overline{u}^i \overline{\mathbf{e}}_i$ and $\mathbf{v} = \overline{v}^j \overline{\mathbf{e}}_j$ into $T(\mathbf{u}, \mathbf{v})$ in (I.4), which gives:

$$\overline{T}_{ij} \overline{u}^i \overline{v}^j = T_{ij} \, u^i v^j \, . \tag{I.7}$$

By means of (I.6), and a similar equation for the v^j, this gives:

$$\overline{T}_{ij} = A_i^k A_j^l T_{kl} \, . \tag{I.8}$$

Since the effect of the frame change in V shows up in the components of T directly, not inversely, one says that the bilinear functional T on V is *covariant.* Often, one hears T referred to as a *(doubly-) covariant second rank tensor* on V.

When one considers bilinear functionals on V^* one obtains formulas that are analogous to those of the covariant case, and which then represent contravariant second rank tensors on V. In particular, a *coframe* for V^* is a basis $\{\theta^i, \, i = 1, \ldots, n\}$, so any covector $\alpha \in V^*$ can be expressed in the form $\alpha = \alpha_i \, \theta^i$. Relative to this coframe, the value of a contravariant bilinear functional $\mathbf{T}(\alpha, \beta)$ when evaluated on two covectors α and β is:

$$\mathbf{T}(\alpha, \beta) = T^{ij}\, \alpha_i\, \beta_j\,. \tag{I.9}$$

When one chooses a frame $\mathbf{e}_i$ on V there is a unique coframe θ^i that it defines on V^* by the requirement that one must have:

$$\theta^i(\mathbf{e}_j) = \delta^i_j\,; \tag{I.10}$$

one calls this coframe the *reciprocal (or inverse) coframe to* $\mathbf{e}_i$. Since the reciprocal coframe $\overline{\theta}^i$ to any other frame $\overline{\mathbf{e}}_i$ must satisfy the analogue of (I.10), as well, one sees that under the transformation from $\mathbf{e}_i$ to $\overline{\mathbf{e}}_i = A^j_i \mathbf{e}_j$, the reciprocal coframe θ^i must go to:

$$\overline{\theta}^i = \tilde{A}^i_j \theta^j\,. \tag{I.11}$$

Hence, coframes transform contragrediently to frames. The effect of the transformation of $\mathbf{e}_i$ on the components of covectors must then be:

$$\overline{\alpha}_i = A^j_i \alpha_j\,. \tag{I.12}$$

This then has the effect of making the transformation of the components of $\mathbf{T}$ a contravariant one:

$$\overline{T}^{ij} = \tilde{A}^i_k \tilde{A}^j_l T^{kl}\,. \tag{I.13}$$

b. Multilinear functionals. In order to go from bilinearity to multilinearity, all that one has to do is to require that a multilinear functional $V \times \ldots \times V \to \mathbb{R}$, with k factors of V in each the Cartesian product, be linear in of its factors individually.

If we choose a basis $\mathbf{e}_i$ for V then the components:

$$T_{ij\ldots k} = T(\mathbf{e}_i\,, \mathbf{e}_j\,, \ldots, \mathbf{e}_k) \tag{I.14}$$

of T for this basis will have k indices, so:

$$T(\mathbf{u}, \mathbf{v}, \ldots, \mathbf{w}) = T_{ij\ldots k}\, u^i\, v^j \ldots w^k\,. \tag{I.15}$$

Under a change of basis on V, the components will transform covariantly:

$$\overline{T}_{ij\cdots k} = A^l_i A^m_j \cdots A^n_k\, T_{lm\cdots n}\,. \tag{I.16}$$

Analogously, a multilinear functional $\mathbf{T}$: $V^* \times \ldots \times V^* \to \mathbb{R}$ will have components with respect to the reciprocal coframe θ^i to the frame $\mathbf{e}_i$ that are defined by:

$$T^{ij\ldots k} = \mathbf{T}(\theta^i\,, \theta^j\,, \ldots, \theta^k) \tag{I.17}$$

and transform contravariantly:

$$\overline{T}^{ij\cdots k} = \tilde{A}^i_l \tilde{A}^j_m \cdots \tilde{A}^k_n T^{lm\cdots n}. \tag{I.18}$$

The set of all multilinear functionals on V of rank k can be made into a vector space in its own right, since $\mathbb{R}$ is a vector space. One defines a scalar combination $\alpha T_1 + \ldots + \beta T_m$ of m multilinear functionals by letting each T_a, $a = 1, \ldots, m$ act on the m-tuple of vectors $(\mathbf{u}, \ldots, \mathbf{w})$ to produce m numbers $T_a(\mathbf{u}, \ldots, \mathbf{w})$ and then forming the corresponding scalar combination of real numbers:

$$(\alpha T_1 + \ldots + \beta T_m)(\mathbf{u}, \ldots, \mathbf{w}) = \alpha T_1(\mathbf{u}, \ldots, \mathbf{w}) + \ldots + \beta T_m(\mathbf{u}, \ldots, \mathbf{w}). \tag{I.19}$$

Since the vector space of linear functionals on V is denoted by V^*, we denote the vector space of k-linear functionals by $V^*\otimes\ldots\otimes V^*$, for consistency, and refer to it as the *tensor product* of k copies of V^*. Hence, the tensor product of a finite number of vector spaces is another vector space[8]. Its dimension is n^k, which can be seen by defining a basis $\{\theta^i \otimes\ldots\otimes \theta^j\}$ for $V^*\otimes\ldots\otimes V^*$ by forming all *tensor products* of the basis vectors θ^i for V*. Although it is possible to give a mathematically rigorous definition of the tensor product of vectors or covectors (see [**9**]), for our purposes it is only necessary to know that it is bilinear, so the tensor product of k vectors is k-linear, and the tensor product of vectors belongs to a higher-dimensional vector space.

One can then express a k-linear functional T in component form relative to this tensor product basis as:

$$T = T_{i\ldots j}\, \theta^i \otimes\ldots\otimes \theta^j. \tag{I.20}$$

Similarly, one can form the tensor product $V\otimes\ldots\otimes V$ to represent the vector space of all k-linear functionals on V^*. It is also n^k-dimensional and has a basis $\{\mathbf{e}_i \otimes\ldots\otimes \mathbf{e}_j\}$ that is defined by all tensor products of the basis elements $\mathbf{e}_i$. In fact, the basis $\{\theta^i \otimes\ldots\otimes \theta^j\}$ on $V^*\otimes\ldots\otimes V^*$ is easily seen to be reciprocal to the basis $\{\mathbf{e}_i \otimes\ldots\otimes \mathbf{e}_j\}$ on $V\otimes\ldots\otimes V$ when θ^i is reciprocal to $\mathbf{e}_i$.

A general k-linear functional $\mathbf{T}$ on V* can be represented in component form relative to this basis as:

$$\mathbf{T} = T^{i\ldots j}\, \mathbf{e}_i \otimes\ldots\otimes \mathbf{e}_j\,. \tag{I.21}$$

2. Exterior algebra. Whenever a multilinear functional acts on a Cartesian product of copies of the same vector space – such as V or V^* – with itself, one can consider the issue of whether the functional is symmetric under permutations of the vectors that it acts upon.

a. Antisymmetric multilinear functionals. For instance, a bilinear functional T on V is *symmetric* iff $T(\mathbf{u}, \mathbf{v}) = T(\mathbf{u}, \mathbf{v})$ and *antisymmetric (or skew-symmetric)* iff $T(\mathbf{u}, \mathbf{v}) = -$

[8] Some members of the physics community find this ambivalence between the use of the words "vector" and "tensor" confusing, since tensors have more indices than vectors, to them. Hopefully, if one can understand that vector spaces are the general concept and tensor products of vector spaces are a specialization of the concept of a vector space then this confusion will pass in time.

$T(\mathbf{u}, \mathbf{v})$. This has the effect of making its components with respect to any frame $\mathbf{e}_i$ on V be symmetric or antisymmetric under permutation, as well:

$$T_{ij} = \pm\, T_{ji}\ . \tag{I.22}$$

One finds that the symmetric bilinear functionals on V form a vector subspace in $V^*\otimes V^*$, as do the antisymmetric ones. This follows from the fact that scalar combinations of (anti-)symmetric bilinear functionals are also (anti-)symmetric. In fact, one can polarize any bilinear functional into a symmetric part T_+ and an antisymmetric part T_- :

$$T = T_+ + T_-, \quad T_\pm(\mathbf{u}, \mathbf{v}) \equiv \tfrac{1}{2}[T(\mathbf{u}, \mathbf{v}) \pm T(\mathbf{v}, \mathbf{u})]. \tag{I.23}$$

Correspondingly, the components of T polarize in the form:

$$T_{ij} = T_{(ij)} + T_{[ij]}, \qquad T_{(ij)} \equiv \tfrac{1}{2}[T_{ij} + T_{ji}], \qquad T_{[ij]} \equiv \tfrac{1}{2}[T_{ij} - T_{ji}]\ . \tag{I.24}$$

One can then think of the association of T with T_+ and T_- as defining linear projections of $V^*\otimes V^*$ onto subspaces that we denote by $V^* \odot V^*$ and $V^* \wedge V^*$, respectively. This means that we can express $V^* \otimes V^*$ as the direct sum:

$$V^*\otimes V^* = (V^* \odot V^*) \oplus (V^* \wedge V^*). \tag{I.25}$$

We think of the vector space $V^* \odot V^*$ as the *symmetric* tensor product of V^* with itself and the vector space $V^* \wedge V^*$as the *exterior product* of V^* with itself.

We shall be primarily concerned with the antisymmetric case in what follows.

If $\alpha, \beta \in V^*$ then one defines their *exterior product* by antisymmetrizing their tensor product:

$$\alpha \wedge \beta = \tfrac{1}{2}(\alpha\otimes\beta - \beta\otimes\alpha). \tag{I.26}$$

Relative to the coframe θ^i, this takes the component form:

$$\alpha \wedge \beta = \tfrac{1}{2}(\alpha_i \beta_j - \alpha_j \beta_i)\ \theta^i \wedge \theta^j. \tag{I.27}$$

One notices that the antisymmetry of the "wedge" product makes:

$$\alpha \wedge \alpha = 0 \tag{I.28}$$

in any case. Another of way of looking at this is to observe that the tensor product $\alpha\otimes\alpha$ is always symmetric, so its projection into the space of antisymmetric tensors must be zero.

The vector space $V^* \wedge V^*$is seen to have dimension $n(n - 1)/2$ if one considers the number of linearly independent antisymmetric combinations of basis covectors $\theta^i \wedge \theta^j$. For instance, if V^* is of dimension 1, 2, 3, 4, resp. then $V^* \wedge V^*$ is of dimension 0, 1, 3, 6,

resp. By contrast, the (complementary) dimension of $V^* \odot V^*$ is $n(n + 1)/2$, which is zero only when $n = 0$.

In terms of bilinear functionals, the action of $\alpha \wedge \beta$ on any two vectors $\mathbf{u}, \mathbf{v} \in V$ gives:

$$(\alpha \wedge \beta)(\mathbf{u}, \mathbf{v}) = \tfrac{1}{2}(\alpha \otimes \beta - \beta \otimes \alpha)(\mathbf{u}, \mathbf{v}) = \tfrac{1}{2}[\alpha(\mathbf{u})\, \beta(\mathbf{v}) - \beta(\mathbf{u})\alpha(\mathbf{v})], \tag{I.29}$$

which has the component form:

$$(\alpha \wedge \beta)(\mathbf{u}, \mathbf{v}) = \tfrac{1}{2}(\alpha_i\, \beta_j - \beta_i\, \alpha_j)\, u^i\, v^j; \tag{I.30}$$

one could also obtain this from (I.27).

When one goes beyond the symmetry of second-rank covariant tensors, one sees that polarization no longer gives a simple dichotomy of the higher-rank tensor product spaces; i.e., there are more than two types of symmetry since there is more than one way to permute pairs of vectors. (Indeed, this is at the root of the decomposition of tensor product representations of groups into irreducible representations.) In order to define the exterior algebra over a vector space, one need only focus on the completely antisymmetric subspaces of the higher-rank tensor products. That is, any transposition of a pair of vectors that the multilinear functional acts on will change the sign of its value.

More generally, if π: {1, 2, …, k} → {1, 2, …, k} is a permutation (i.e., a bijection) then if T is a k-linear functional on V one says that T is *completely antisymmetric* iff:

$$T(\mathbf{v}_{\pi(1)}, \ldots, \mathbf{v}_{\pi(k)}) = \mathrm{sgn}(\pi)\, T(\mathbf{u}_1, \ldots, \mathbf{u}_k) \tag{I.31}$$

where sgn(π) is + when the number of transpositions in π is even and – when it is odd.

One finds that the set of all completely antisymmetric k-linear functionals forms a vector under the scalar combinations that are defined by the obvious extension of (I.19). We shall denote this vector space by $\mathrm{A}^k(V)$ and refer to its elements as *algebraic k-forms* on V. It has a dimension that vanishes when $k > n$ and is given by the binomial coefficient $\left\{ \begin{matrix} n \\ k \end{matrix} \right\}$ for $k \leq n$. This can be seen from the fact that $\mathrm{A}^k(V)$ has a basis that is given by the linearly independent k-fold exterior products $\theta^{i_1} \wedge \ldots \wedge \theta^{i_k}$ of the θ^i, and from the fact that any permutation of the indices $[i_1 \ldots i_k]$ will change the functional thus defined by at most a sign, while there are $k!$ such permutations. One can express the arbitrary k-form α in terms of this redundant basis as:

$$\alpha = \frac{1}{k!} \alpha_{i \cdots j} \theta^i \wedge \cdots \wedge \theta^j \tag{I.32}$$

or in terms of a non-redundant basis by using only those k-fold products for which the indices are in ascending order:

$$\alpha = \sum_{i_1 < \cdots < i_k} \alpha_{i_1 \cdots i_k} \theta^{i_1} \wedge \cdots \wedge \theta^{i_k} \ . \tag{I.33}$$

Although the term "redundant" sounds pejorative, there are actually times when calculations are more conveniently carried out in the redundant basis.

Note that in order for the right-hand side of (I.32) to make sense the indices of $\alpha_{i\ldots j}$ need to have the same permutation symmetry as the k-fold wedge product of the θ^i:

$$\alpha_{i_{\pi(1)}\cdots i_{\pi(k)}} = \mathrm{sgn}(\pi)\,\alpha_{i_1\cdots i_k} \,. \tag{I.34}$$

Indeed, even if one had formed the linear combination in question when starting with components $\alpha_{i\ldots j}$ of unspecified symmetry, the complete antisymmetry of the exterior product of basis elements would select out only the completely antisymmetric part of $\alpha_{i\ldots j}$; for instance, $\alpha_{ij}\,\theta^i \wedge \theta^j = 0$ when $\alpha_{ij} = \alpha_{ji}$.

Due to the symmetry $\left\{ {n \atop k} \right\} = \left\{ {n \atop n-k} \right\}$, one sees that the dimension of $\mathrm{A}^k(V)$ is the same as the dimension of $\mathrm{A}^{n-k}(V)$. Hence, they are linearly isomorphic, but not canonically so; a typical way of defining the isomorphism would be to choose bases for both spaces. For instance, when $n = 3$, one has $\mathrm{A}^0(V) \cong \mathrm{A}^3(V) \cong \mathbb{R}$, $\mathrm{A}^1(V) \cong \mathrm{A}^2(V) \cong \mathbb{R}^3$.

b. Volume elements. One notes that in the extreme case $k = n$, one always has $\mathrm{A}^n(V) \cong \mathbb{R}$. A basis for this one-dimensional vector space is simply a non-zero n-form $\mathcal{V}$. In terms of a basis θ^i for V^*, it can be written as either:

$$\mathcal{V} = \theta^1 \wedge \ldots \wedge \theta^n \tag{I.35}$$

or:

$$\mathcal{V} = \frac{1}{n!}\varepsilon_{i_1\cdots i_n}\,\theta^{i_1} \wedge \cdots \wedge \theta^{i_n} \,. \tag{I.36}$$

The ε symbol represents the usual Levi-Civita symbol with n indices, which equals +1 when $i_1 \ldots i_n$ is an even permutation of $12\ldots n$, equals –1 when it is an odd permutation, and is zero otherwise.

One generally refers to a choice of $\mathcal{V}$ as a *volume element* on V since the effect of applying the n-linear function $\mathcal{V}$ to an n-tuple of vectors in V is:

$$\mathcal{V}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \frac{1}{k!}\varepsilon_{i_1\cdots i_n} v_1^{i_1} \cdots v_n^{i_n} = \det(v_j^i) \,, \tag{I.37}$$

where v_j^i is the $n\times n$ matrix whose columns are the components of $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Hence, the value of $\mathcal{V}$ when applied to an n-tuple of vectors in V is to give the volume of the parallelepiped that they span. This volume will vanish unless they are all linearly independent.

Under a change of coframe from θ^i to $\bar{\theta}^i = A^i_j \theta^j$, the redundant form (I.36) shows that $\mathcal{V}$ goes to:

$$\bar{\mathcal{V}} = \det(A)\, \mathcal{V}. \tag{I.38}$$

This situation is often described by physicists as implying that *n*-forms are "pseudo-scalars."

c. Exterior products of general k-forms. So far, we have really only discussed the exterior products of covectors as a means of defining completely antisymmetric multilinear functionals on vector spaces. Now, we shall extend the exterior product to include the product of a *k*-form α and an *l*-form β. If we demand that the result be a *k*+*l*-form then we see that although a simple tensor product of α and β will produce a tensor of the required rank, it will only have the desired anti-symmetry if one completely anti-symmetrizes the resulting tensor product accordingly. Naively, this involves (*k*+*l*)! permutations of the factors, but since α and β are antisymmetric to begin with *k*! of those permutations will affect only the sign of α and *l*! will affect only the sign of β; that is, one only needs to anti-symmetrize the transpositions that are not already antisymmetric. The result is:

$$(\alpha \wedge \beta)(\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_{k+l}) = \sum_\pi \frac{(k+l)!}{k!\,l!} \operatorname{sgn}(\pi)\, \alpha(\mathbf{v}_{\pi(1)}, \ldots, \mathbf{v}_{\pi(k)}) \beta(\mathbf{v}_{\pi(k+1)}, \ldots, \mathbf{v}_{\pi(k+l)}). \tag{I.39}$$

The effect on components is similar:

$$(\alpha \wedge \beta)_{1\ldots k,\, k+1\ldots k+l} = \sum_\pi \frac{(k+l)!}{k!\,l!}\, \operatorname{sgn}(\pi)\, \alpha_{\pi(1)} \ldots_{\pi(k)} \beta_{\pi(k+1)} \ldots_{\pi(k+l)}. \tag{I.40}$$

For instance if α is a 1-form and β is a 2-form then the components of $\alpha \wedge \beta$ will take the form:

$$(\alpha \wedge \beta)_{ijk} = (l+1) \sum_\pi \operatorname{sgn}(\pi) a_{\pi(i)} b_{\pi(j)\pi(k)}\,. \tag{I.41}$$

Hence, if we define the direct sum $\mathrm{A}^*(V) = \mathrm{A}^0 \oplus \mathrm{A}^1 \oplus \ldots \oplus A^n$ of all the vector spaces $\mathrm{A}^k(V)$, which we abbreviate by A^k when *V* is unambiguous, then we obtain a vector space of dimension 2^n whose elements then represent finite sums of multilinear functionals on *V*. When all of the functionals in a sum have the same rank *k*, one calls the linear combination *homogeneous*, and it will define an element of A^k; the other elements are called *mixed* forms.

The exterior product, as we have extended it, defines a bilinear map $\mathrm{A}^* \times \mathrm{A}^* \to \mathrm{A}^*$, $(\alpha, \beta) \mapsto \alpha \wedge \beta$, which then makes A^*, with this bilinear product, into an *algebra* over the vector space A^*. It is associative:

$$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma), \qquad (\text{all } \alpha, \beta, \gamma) \tag{I.42}$$

and has a unity element – namely, the real number $1 \in A^0$ – but it is not commutative, since if α is a k-form and β is an l-form, one has:

$$\alpha \wedge \beta = (-1)^{kl} \beta \wedge \alpha. \tag{I.43}$$

Hence, although we have obtained the exterior product by a process of complete anti-symmetrization, the exterior product itself can be either antisymmetric or symmetric. In particular, as long as either k or l is even, the product will be symmetric.

Since the product of exterior algebraic forms always has a rank that is greater than or equal to the individual ranks, and the unity element 1 has rank 0, one sees that the only possible *units* in the algebra – i.e., elements that have multiplicative inverses under the wedge product – will be non-zero real numbers. Similarly, from (I.43), one sees that the only way that a k-form α can commute with all l-forms β, regardless of l, is if $kl = 0$ for all l; hence, $k = 0$. This says that the *center* of the algebra A^* – viz. the set of all elements in A^* that commute with all other elements – is defined by the elements of A^0.

The fact that the exterior product takes $A^k \times A^l$ to A^{k+l} means that, by definition, the algebra over A^* that is defined by the exterior product, which one calls the *exterior algebra over* V^*, is a *graded* algebra.

d. The algebra of multivectors. One can also define an exterior algebra over the vector space V itself by an analogous process of the complete anti-symmetrization of tensor products. For instance, the exterior product of vectors $\mathbf{v}$ and $\mathbf{w}$ in V is the *bivector:*

$$\mathbf{v} \wedge \mathbf{w} = \tfrac{1}{2}(\mathbf{v} \otimes \mathbf{w} - \mathbf{w} \otimes \mathbf{v}). \tag{I.44}$$

If $\mathbf{e}_i$ is a basis for V then all $\mathbf{e}_i \wedge \mathbf{e}_j$ will define a redundant basis for $A_2(V)$, and in order to eliminate the redundancy, one must use only exterior products with $i < j$. The components of $\mathbf{v} \wedge \mathbf{w}$ are then:

$$(\mathbf{v} \wedge \mathbf{w})^{ij} = v^i w^j - v^j w^i , \tag{I.45}$$

and:

$$\mathbf{v} \wedge \mathbf{w} = \tfrac{1}{2}(v^i w^j - v^j w^i)\, \mathbf{e}_i \wedge \mathbf{e}_j . \tag{I.46}$$

One could regard a bivector as an antisymmetric linear functional on A^2, since the expression:

$$(\mathbf{v} \wedge \mathbf{w})(\alpha \wedge \beta) = (\alpha \wedge \beta)(\mathbf{v} \wedge \mathbf{w}) = (\alpha \wedge \beta)(\mathbf{v}, \mathbf{w}) = \tfrac{1}{2}[\alpha(\mathbf{v})\beta(\mathbf{w}) - \alpha(\mathbf{w})\beta(\mathbf{v})], \tag{I.47}$$

can be extended by linearity to all finite linear combinations of 2-forms, but it is generally preferable to regard a non-zero bivector as more like a pair of linearly independent vectors in V, although this is true only in the *simple* case, where the bivector is of the form $\mathbf{v} \wedge \mathbf{w}$ for some pair of vectors $\mathbf{v}$ and $\mathbf{w}$, which is not, however, unique.

One defines the vector spaces $A_k(V)$ in an analogous manner to the way that one defined the $A^k(V)$, only in terms of vectors in V, not covectors in V^*. The elements of $A_k(V)$ – or just A_k , for short – are called *k-vectors* or *multivectors*, in general. One similarly defines the exterior product of k-vectors and l-vectors, and ultimately the

exterior algebra A_* becomes a graded algebra that is isomorphic to A^* as an algebra by means of any choice of basis for V. More specifically, each vector space A_k is linearly isomorphic to the corresponding A^k. However, none of these isomorphisms are canonical – i.e., defined uniquely in the absence of further assumptions.

e. Interior products. Just as (I.47) can be generalized to give a bilinear pairing $A^k \times A_k \to \mathbb{R}$, $(\alpha, \mathbf{A}) \mapsto \alpha(\mathbf{A})$, which represents the evaluation of a k-form on a k-vector, one can generalize this construction, as well, to give a bilinear pairing $A_l \times A^k \to A^{k-l}$ when $k > l$ and $A_l \times A^k \to A_{k-l}$ when $k < l$.

The construction begins by looking at how it works for the bilinear pairing of a vector $\mathbf{v}$ and a simple 2-form $\alpha \wedge \beta$. We define:

$$i_{\mathbf{v}}(\alpha \wedge \beta) = (i_{\mathbf{v}}\alpha)\,\beta - \alpha\,(i_{\mathbf{v}}\beta) = \alpha(\mathbf{v})\beta - \beta(\mathbf{v})\alpha. \tag{I.48}$$

and extend to a simple k-form similarly:

$$i_{\mathbf{v}}(\alpha_1 \wedge \ldots \wedge \alpha_k) = \sum_{m=1}^{k} (-1)^{m+1} \alpha_m(\mathbf{v})(\alpha_1 \wedge \cdots \wedge \hat{\alpha}_m \wedge \cdots \wedge \alpha_k), \tag{I.49}$$

in which the caret over the α_m means that the indicated 1-form is omitted from the exterior product.

Since any k-form can be expressed as a finite linear combination of simple k-forms – e.g., the basis elements – the action of $\mathbf{v}$ on simple k-forms can be extended by linearity to all of A^k, which then gives a bilinear pairing $A_1 \times A^k \to A^{k-1}$ that one refers to as the *interior product* of a k-form with a vector.

One can further extend this interior product from vectors to simple l-vectors by iteration:

$$i_{\mathbf{v} \wedge \ldots \wedge \mathbf{w}}\alpha = (-1)^{\tau}\, i_{\mathbf{v}}(\ldots i_{\mathbf{w}}\alpha), \tag{I.50}$$

and to all l-vectors by linearity. The sign comes from the number τ of transpositions that it takes to permute the sequence $\mathbf{v} \ldots \mathbf{w}$ into its reverse sequence $\mathbf{w} \ldots \mathbf{v}$. For instance, **vw** goes to **wv** by one transposition, as does **uvw** to **wvu**, while it takes two transpositions to take **tuvw** to **wvut**. Therefore, in the cases that will be of recurring interest to us:

$$i_{\mathbf{v} \wedge \mathbf{w}}\alpha = -\, i_{\mathbf{v}}(i_{\mathbf{w}}\alpha), \qquad i_{\mathbf{u} \wedge \mathbf{v} \wedge \mathbf{w}}\alpha = -\, i_{\mathbf{u}}(i_{\mathbf{v}}\,(i_{\mathbf{w}}\alpha), \qquad i_{\mathbf{t} \wedge \mathbf{u} \wedge \mathbf{v} \wedge \mathbf{w}}\alpha = i_{\mathbf{t}}(i_{\mathbf{u}}(i_{\mathbf{v}}(i_{\mathbf{w}}\alpha), \tag{I.51}$$

Hence, we now have our bilinear pairing of $A_l \times A^k \to A^{k-l}$, when $k > l$.

In order to define the bilinear pairing when $k < l$, we start by defining the interior product of a simple k-vector by a 1-form, analogously to (I.49):

$$i_{\alpha}(\mathbf{v}_1 \wedge \ldots \wedge \mathbf{v}_k) = \sum_{m=1}^{k} (-1)^{m+1} \alpha(\mathbf{v}_m)(\mathbf{v}_1 \wedge \cdots \wedge \hat{\mathbf{v}}_m \wedge \cdots \wedge \mathbf{v}_k), \tag{I.52}$$

and then extend this by linearity to all k-vectors. This gives a bilinear pairing $\mathrm{A}_k \times \mathrm{A}^1 \to \mathrm{A}_{k-1}$. We then extend to a bilinear pairing $\mathrm{A}_k \times \mathrm{A}^l \to \mathrm{A}_{l-k}$ for $k < l$ analogously to (I.50):

$$i_{\alpha \wedge \ldots \wedge \beta}\mathbf{A} = (-1)^{\tau}\, i_{\alpha}(\ldots\, i_{\beta}\mathbf{A}). \tag{I.53}$$

f. Poincaré duality. An important consequence of the bilinear pairings that we discussed in the last section is that whenever one chooses a volume element $\mathcal{V} \in \mathrm{A}^n$ for V one then defines a linear map #: $\mathrm{A}_k \to \mathrm{A}^{n-k}$, $\mathbf{A} \mapsto i_{\mathbf{A}}\mathcal{V}$ for each $0 \le k \le n$. These linear maps are seen to be, in fact, linear isomorphisms.

Conversely, if one chooses a volume element $\boldsymbol{\mathcal{V}} \in \mathrm{A}_n$ for V^* then one can define linear maps in the opposite directions #′: $\mathrm{A}^k \to \mathrm{A}_{n-k}$, $\alpha \mapsto i_{\alpha}\boldsymbol{\mathcal{V}}$, which are also linear isomorphisms. In fact, as long as one chooses $\boldsymbol{\mathcal{V}}$ to be the n-vector that makes $\mathcal{V}(\boldsymbol{\mathcal{V}}) = 1$, for consistency, the isomorphism #′ will be the inverse to #.

The complete set of linear isomorphisms that are defined by a choice of volume element on V is then referred to as *Poincaré duality*. Its geometric origin is in projective geometry, as we shall discuss later, and amounts to the idea that a k-dimensional subspace of V can either be spanned by k linearly independent vectors in V or annihilated by $n - k$ linearly independent covectors in V^*. It plays an essential role in the foundations of electromagnetism, as we shall see.

We illustrate the nature of Poincaré duality by showing how it works in the low dimensional vector spaces in terms of frame and coframes.

In two dimensions, if $\{\mathbf{e}_1, \mathbf{e}_2\}$ is a basis and $\{\theta^1, \theta^2\}$ is its reciprocal basis then we can define $\boldsymbol{\mathcal{V}} = \mathbf{e}_1 \wedge \mathbf{e}_2$ and $\mathcal{V} = \theta^1 \wedge \theta^2$. The dual of a 0-vector $\lambda \in \mathbb{R}$ is the form $\lambda\mathcal{V}$, and conversely. As for the duals of the basis elements, by direct calculation, one verifies that:

$$\#\mathbf{e}_1 = i_{\mathbf{e}_1}(\theta^1 \wedge \theta^2) = \theta^2, \qquad \#\mathbf{e}_2 = -\,\theta^1. \tag{I.54}$$

More generally, one can say:

$$\#\mathbf{e}_i = \varepsilon_{ij}\,\theta^j, \qquad \varepsilon_{ij} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{I.55}$$

Although this all looks somewhat trivial at this point, actually when one goes from real to complex scalars, as we shall do later on, one finds that these isomorphisms are of great help in understanding the representation of Lorentz transformations in terms of $SL(2, \mathbb{C})$, which acts naturally on $\mathbb{C}^2$.

The three-dimensional case pertains directly to conventional vector algebra, as we shall discuss in more detail shortly. The non-trivial isomorphism is now #: $\mathrm{A}^1 \to \mathrm{A}^2$, which one can think of as taking "polar" vectors to "axial" vectors in the traditional physics terminology. If $\boldsymbol{\mathcal{V}} = \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$ and $\mathcal{V} = \theta^1 \wedge \theta^2 \wedge \theta^3$ now, then we have:

$$\#\mathbf{e}_1 = \theta^2 \wedge \theta^3, \qquad \#\mathbf{e}_2 = \theta^3 \wedge \theta^1, \qquad \#\mathbf{e}_3 = \theta^1 \wedge \theta^2 , \tag{I.56}$$

or:

$$\#\mathbf{e}_i = \tfrac{1}{2}\, \varepsilon_{ijk}\, \theta^j \wedge \theta^k. \tag{I.57}$$

In four dimensions, both the isomorphisms $A_1 \to A^3$ and $A_2 \to A^2$ are non-trivial:

$$\begin{aligned} \#\mathbf{e}_1 &= \theta^2 \wedge \theta^3 \wedge \theta^4, \qquad \#\mathbf{e}_2 = \theta^3 \wedge \theta^4 \wedge \theta^1, \\ \#\mathbf{e}_3 &= \theta^4 \wedge \theta^1 \wedge \theta^3, \qquad \#\mathbf{e}_4 = \theta^1 \wedge \theta^2 \wedge \theta^3 \end{aligned} \tag{I.58a}$$

$$\begin{aligned} &\#(\mathbf{e}_1 \wedge \mathbf{e}_2) = \theta^3 \wedge \theta^4, \quad \#(\mathbf{e}_1 \wedge \mathbf{e}_3) = \theta^4 \wedge \theta^2, \quad \#(\mathbf{e}_1 \wedge \mathbf{e}_4) = \theta^3 \wedge \theta^2, \\ &\#(\mathbf{e}_2 \wedge \mathbf{e}_3) = \theta^1 \wedge \theta^4, \quad \#(\mathbf{e}_2 \wedge \mathbf{e}_4) = \theta^3 \wedge \theta^1, \quad \#(\mathbf{e}_3 \wedge \mathbf{e}_4) = \theta^1 \wedge \theta^2, \end{aligned} \tag{I.58b}$$

or, more concisely:

$$\#\mathbf{e}_i = \varepsilon_{ijkl}\, \theta^j \wedge \theta^k \wedge \theta^l, \qquad \#(\mathbf{e}_i \wedge \mathbf{e}_j) = \varepsilon_{ijkl}\, \theta^k \wedge \theta^l. \tag{I.59}$$

These will be the isomorphisms that are most useful for electrodynamics.

g. Algebraic operators defined on exterior algebras. The concepts of exterior product and interior product allow one to define various linear operators on A_* and A^* that prove useful in treating more elaborate topics.

Since the exterior product takes $A^k \times A^l$ to A^{k+l}, one sees that by fixing a k-form α, one can define a linear map e_α: $A^l \to A^{k+l}$, $\beta \mapsto \alpha \wedge \beta$, which one might think of as the *adjoint map* defined by α in the algebra, or simply, *left multiplication* by α.

Naively, the dimension of the image of A^l – i.e., the rank of e_α – under this map is less than or equal to $\min\{\dim(A^l), \dim(A^{k+l})\}$. In order for the dimension to be $\dim(A^l)$ the map would have to be an injection, which would imply that its kernel would have to vanish. This, in turn, would have to imply that $\alpha \wedge \beta \neq 0$ as long as $\beta \neq 0$. However, this is not always the case. For instance, in the elementary case of a 1-form α acting on other 1-forms, one sees that all β of the form $\lambda\alpha$ for a real scalar λ will give $\alpha \wedge \beta = 0$. Hence, the kernel of e_α: $A^1 \to A^2$ is 1-dimensional, so its image is n – 1-dimensional; its restriction to any hyperplane in $A^1 = V^*$ that is transverse to α will then be injective.

As we have seen, the interior products define a bilinear pairing $A_k \times A^l \to A^{l-k}$, when $k < l$ and $A_k \times A^l \to A_{k-l}$ when $k > l$. Hence, when one fixes a k-vector $\mathbf{A}$ one defines a linear map $i_{\mathbf{A}}$: $A^l \to A^{l-k}$, for $k < l$, and when one fixes an l-form α one defines a linear map i_α: $A_k \to A_{k-l}$ when $k > l$.

A crucial property of the interior product operator is that when α is a k-form and β is an l-form one has for any vector $\mathbf{v} \in V$:

$$i_{\mathbf{v}}(\alpha \wedge \beta) = i_{\mathbf{v}}\alpha \wedge \beta + (-1)^k\, \alpha \wedge i_{\mathbf{v}}\beta\,. \tag{I.60}$$

Suppose $\mathbf{e}_i$ is a basis for V and θ^i is the reciprocal basis for V^*. Any vector $\mathbf{v} \in V$ can be expressed as:

$$\mathbf{v} = v^i \mathbf{e}_i = \theta^i(\mathbf{v})\mathbf{e}_i = (e_{\mathbf{e}_i} \circ i_{\theta^i})(\mathbf{v})\,. \tag{I.61}$$

Hence, $e_{\mathbf{e}_i} \circ i_{\theta^i} = I$ when applied to vectors in V. In fact, each individual $e_{\mathbf{e}_i} \circ i_{\theta^i}$ (i not summed) acts as a projection of $\mathbf{v}$ onto the line in the direction $\mathbf{e}_i$. Therefore, the sum of terms over i represents a sort of spectral decomposition of the identity operator into projections onto the basis elements.

Similarly, $e_{\theta^i} \circ i_{\mathbf{e}_i} = I$ when applied to covectors in V^*.

Now, let us examine the opposite operator $i_{\theta^i} \circ e_{\mathbf{e}_i}$. When applied to $\mathbf{v}$ it gives:

$$(i_{\theta^i} \circ e_{\mathbf{e}_i})(\mathbf{v}) = i_{\theta^i}(\mathbf{e}_i \wedge \mathbf{v}) = \theta^i(\mathbf{e}_i)\mathbf{v} - \mathbf{e}_i\, \theta^i(\mathbf{v}) = n\mathbf{v} - \mathbf{v} = (n-1)\mathbf{v}. \tag{I.62}$$

This means that when we sum the operators thus defined, we get:

$$e_{\mathbf{e}_i} \circ i_{\theta^i} + i_{\theta^i} \circ e_{\mathbf{e}_i} = nI. \tag{I.63}$$

More generally, we would also find that for any non-zero vector $\mathbf{v}$ and covector α such that $\alpha(\mathbf{v}) = 1$, one would have:

$$I = e_{\mathbf{v}} \cdot i_\alpha + i_\alpha \cdot e_{\mathbf{v}}\,. \tag{I.64}$$

The first term represents a projection of a vector in V onto the direction [$\mathbf{v}$] spanned by $\mathbf{v}$, so the second term represents a projection of that vector onto the hyperplane Ann(α) in V that is annihilated by α, which is transverse to the line through $\mathbf{v}$. Hence, the pair ($\mathbf{v}$, α) defines a direct sum decomposition $V = [\mathbf{v}] \oplus \mathrm{Ann}(\alpha)$, and (I.64) corresponds to the fact that one will have a unique decomposition $\mathbf{w} = \mathbf{w}_\parallel + \mathbf{w}_\perp$ of $\mathbf{w} \in V$ into projections parallel to $\mathbf{v}$ and transverse to it.

One finds that the operator $e_{\mathbf{v}} \cdot i_\alpha + i_\alpha \cdot e_{\mathbf{v}}$ also acts as the identity operator on all of the other spaces A_k. Analogous statements to the preceding ones apply to the action of the operator $I = e_\alpha \cdot i_{\mathbf{v}} + i_{\mathbf{v}} \cdot e_\alpha$ on A^k. These operators will prove essential later when treating the effect of defining a timelike observer on the spacetime manifold; i.e., a space + time decomposition of the tangent bundle to the spacetime manifold.

3. Exterior derivative. From multi-variable calculus [**4**, **5**], one presumably knows how to define the differential df of a differentiable function f on a vector space V with real values and the fact that it represents a linear functional on V.

A choice of basis $\mathbf{e}_i$ for V defines a linear isomorphism $V \to \mathbb{R}^n$, $\mathbf{x} \mapsto (x^1(\mathbf{x}), \ldots, x^n(\mathbf{x}))$ which can also be treated as a global coordinate system on V. Each of the coordinate functions x^i on V is presumed to be differentiable and defines a differential dx^i. This makes:

$$df = \frac{\partial f}{\partial x^i}\, dx^i = f_{,i}\, dx^i, \tag{I.66}$$

in which we have introduced a common notation for partial derivatives as being indicated by a comma.

If one regards smooth functions on V as *(differential) 0-forms* and their differentials as *(differential) 1-forms* then one sees that the effect of differentiation is to take 0-forms to 1-forms. However, since functions take their values in $\mathbb{R}$ without actually being elements of $\mathbb{R}$, and the components of differential 1-forms are functions, not constants, in general, we see that we are not really dealing with A^0 and A^1 directly, but rather with smooth functions on V that take their values in these vector spaces.

In general, we will then denote the vector space of all smooth functions on V with values in A^k by $\Lambda^k(V)$, or Λ^k, for short. For instance, elements of Λ^0 will be smooth functions on V, elements of Λ^1 will be smooth maps from V to V^*, and elements of Λ^2 will be of the form:

$$F = \tfrac{1}{2} F_{ij}(x)\, dx^i \wedge dx^j \,. \tag{I.67}$$

The exterior derivative operator is an extension of the differential operator $d\colon \Lambda^0 \to \Lambda^1$ to a more general linear operator $d\colon \Lambda^k \to \Lambda^{k+1}$. As it turns out (see [**1**]), besides requiring linearity and agreement with the differential on 0-forms, this operator is uniquely defined by further requiring only that d be an anti-derivation with respect to the exterior product and that its square be zero, in any case. That is, if α is a k-form and β is an l-form then:

$$d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^k\, \alpha \wedge d\beta \,. \tag{I.68a}$$

$$d^2 = 0. \tag{I.68b}$$

For instance, let us see what this produces for the exterior derivatives of 1-forms. If $\alpha = \alpha_i(x)\, dx^i$ then a direct application of the rules gives:

$$d\alpha = d\alpha_i \wedge dx^i + \alpha_i\, d^2x^i = (\alpha_{i,j}\, dx^j) \wedge dx^i = -\tfrac{1}{2}(a_{i,j} - a_{j,i})\, dx^i \wedge dx^j \,. \tag{I.69}$$

If $F = \tfrac{1}{2} F_{ij}\, dx^i \wedge dx^j$ is a 2-form then:

$$\begin{aligned} dF &= \tfrac{1}{2}\, dF_{ij} \wedge dx^i \wedge dx^j = \tfrac{1}{2}(F_{ij,k}\, dx^k) \wedge dx^i \wedge dx^j \\ &= \tfrac{1}{3}(F_{ij,k} + F_{jk,i} + F_{k,ij})\, dx^i \wedge dx^j \wedge dx^k. \end{aligned} \tag{I.70}$$

The fact that we are making d^2 vanish in any case is really a generalization of the fact that the second derivative of f defines a symmetric bilinear functional on V at each point of V; i.e., the mixed partial derivatives $\partial^2 f / \partial x^i\, \partial x^j$ are symmetric in i and j, since we are assuming that f has continuous second derivatives. Hence, the antisymmetric part of the bilinear functional must vanish. To illustrate how the symmetry of mixed partial derivatives implies the vanishing of d^2, apply it to a smooth function f:

$$d(df) = d(f_{,i}\, dx^i) = -\tfrac{1}{2}(f_{,i,j} - f_{j,i})\, dx^i \wedge dx^j = 0. \tag{I.71}$$

4. Divergence operator. When one has chosen a volume element $\mathcal{V}$ for a vector space V, one can use Poincaré duality to define a differential operator on multivector fields that is "adjoint" to d. For our immediate purposes, a k-vector field on V will be a smooth function from V to A_k. Hence, if $\{\partial_i, i = 1, \ldots, n\}$is a basis for V then a k-vector field $\mathbf{A}$ on V has the local form [9]:

$$\mathbf{A} = \frac{1}{k!} A^{ij\cdots m}(x)\, \partial_i \wedge \partial_j \wedge \cdots \wedge \partial_m . \tag{I.72}$$

We use Poincaré duality and the exterior derivative operator to define a divergence operator δ: $\Lambda^k \to \Lambda^{k-1}$ by way of:

$$\delta = \#^{-1} \cdot d \cdot \#. \tag{I.73}$$

This can be represented schematically by a "commutative diagram:"

$$\begin{array}{ccc} \Lambda^{n-k} & \xrightarrow{\ d\ } & \Lambda^{n-k+1} \\ \#\uparrow & & \#\uparrow \\ \Lambda_k & \xrightarrow{\ \delta\ } & \Lambda_{k-1} \end{array}$$

The properties of δ are not as convenient as those of d, although some of them are derived from properties of d, such as linearity, and the fact that:

$$\delta^2 = \#^{-1} \cdot d^2 \cdot \# = 0. \tag{I.74}$$

An immediate consequence of (I.73) is the useful relation:

$$d\# = \delta\#. \tag{I.75}$$

We also have:

$$\delta\partial_i = \#^{-1} \cdot d \cdot \#\partial_i = \varepsilon_{ij\ldots m}\, \#^{-1} \cdot d\,(dx^j \wedge \ldots \wedge dx^m) = 0. \tag{I.76}$$

Hence, as long as the coordinate basis ∂_i for V is reciprocal to the differential basis dx^i for Λ^1, the vanishing of the divergences of these basis vectors follows from $d^2 = 0$.

The relation (I.75) suggests that, in some sense, the linear operator δ is "adjoint" to the operator d. In order to clarify the sense in which this is meaningful, we introduce the bilinear pairing $\langle .,. \rangle$: $\mathrm{A}^k \times \mathrm{A}_k \to \mathrm{A}^n$, that takes $(\alpha, \mathbf{A})$ to:

$$\langle \alpha, \mathbf{A} \rangle = \alpha(\mathbf{A})\, \mathcal{V} = \alpha \wedge \#\mathbf{A} . \tag{I.77}$$

Now, let $\alpha \in \mathrm{A}^{k-1}$, $\mathbf{A} \in \mathrm{A}_k$ and compare $\langle d\alpha, \mathbf{A} \rangle$ to $\langle \alpha, \delta\mathbf{A} \rangle$:

[9] Although we shall eventually wish to regard the symbols ∂_i as relating to directional derivative operators, for the present, we shall only regard them as vectors in V.

$$<d\alpha, \mathbf{A}> = d\alpha \wedge \#\mathbf{A} = d(\alpha \wedge \#\mathbf{A}) - (-1)^k\, \alpha \wedge d\#\mathbf{A}$$
$$= d(\alpha \wedge \#\mathbf{A}) - (-1)^k\, \alpha \wedge \#\delta\mathbf{A}, \tag{I.78}$$

which leads to:

$$<d\alpha, \mathbf{A}> + (-1)^k <\alpha, \delta\mathbf{A}> = d(\alpha \wedge \#\mathbf{A}). \tag{I.79}$$

Once we have discussed the integration of differential forms and Stokes's theorem, we can show this leads to a "graded" form of adjointness between d and δ, at least for compact, orientable manifolds without boundary.

What complicates the use of δ in computations is that since # is a linear isomorphism, but not an algebra isomorphism – viz., it does not take $\alpha \wedge \beta$ to $\#\alpha \wedge \#\beta$ – the divergence operator is not an anti-derivation. One does, however, have the following useful result when f is a smooth function and $\mathbf{A}$ is a k-vector field:

$$\delta(f\mathbf{A}) = \#^{-1}(df \wedge \#\mathbf{A}) + f\, \delta\mathbf{A} \tag{I.80}$$

For $k = 1$, δ produces the usual divergence of a vector field:

$$\delta\mathbf{v} = \delta(v^i\, \partial_i) = \#^{-1}(dv^i \wedge \#\partial_i) + v^i\, \delta\partial_i = v^i{}_{,j}\, (dx^j \wedge \#\partial_i)(\mathbf{V}) = v^i{}_{,i}\,. \tag{I.81}$$

In the last step, we made use of what will later be described as an "orthogonality" condition:

$$dx^j \wedge \#\partial_i = \delta^j_i\, \mathcal{V}. \tag{I.82}$$

As we will also discuss later, when one is dealing with a Riemannian – or even Lorentzian – manifold, which has a metric tensor field $g_{ij}\, dx^i\, dx^j$ defined on it, the local form of the volume element $\mathcal{V}$ will also include a factor of $\sqrt{g}$, where $g \equiv |\det g_{ij}|$; hence, the volume element $\boldsymbol{\mathcal{V}}$ will have a factor of $1/\sqrt{g}$. One then sees that (I.81) takes on the customary form:

$$\delta\mathbf{v} = \frac{1}{\sqrt{g}} \frac{\partial(\sqrt{g}\, v^i)}{\partial x^i}. \tag{I.83}$$

Although the discussion of such matters at this point seems somewhat premature, the reason for mentioning it is to point out that in its most fundamental form the divergence operator is related solely to volume elements, not metrics, and this fact is essential to understanding pre-metric electromagnetism. In the next section, we shall see that vector fields with vanishing divergence generate flows of volume-preserving diffeomorphisms.

5. Lie derivative. Now that we have defined the exterior derivative and the interior product, we have enough tools to construct the Lie derivative operator. However, before we do so, in order to interpret that operator we first mention a few elementary notions from the theory of systems of ordinary differential equations (see [**10**] for all references to matters concerned with that theory).

a. Systems of first order ordinary differential equations. A vector field $\mathbf{v}$: $V \to V$, $x \mapsto \mathbf{v}(x) = v^i(x)\, \partial_i$ on an n-dimensional vector space V defines a system of n first order ordinary differential equations by starting with the assumption that $\mathbf{v}(x)$ represents the velocity vector field of a differentiable curve γ: $\mathbb{R} \to V$, $\tau \mapsto \gamma(\tau) = x^i(\tau)\, \partial_i$ at each point. The system of equations then takes the form:

$$\frac{dx^i}{d\tau} = v^i(x^j(\tau)) \tag{I.84}$$

relative to this choice of coordinate system on V.

Actually, this system is more general than it looks on first glance. For one thing, any ordinary differential equation of order k can be converted into a system of k first order equations, so the order of the system is no restriction. Furthermore, although the fact that $\mathbf{v}$ is only indirectly a function of τ implies that the system is autonomous – i.e., time-invariant – one can extend $\mathbf{v}$ to a vector field on $\mathbb{R} \times V$ that gives a system that is autonomous on that extended space.

From the theorem of existence and uniqueness of such systems of ordinary differential equations, as long as one assumes that $\mathbf{v}$ is continuously differentiable one always has the existence of *local flows* about each point $x \in V$; that is, there is always a sufficiently small neighborhood $(-\varepsilon, +\varepsilon)$ of $0 \in \mathbb{R}$ and a neighborhood U of x such that:

a. There is a one-parameter family of differentiable maps Φ: $(-\varepsilon, +\varepsilon) \times U \to V$, $(\tau, x) \mapsto \Phi_\tau(x)$, such that for each τ the map Φ_τ: $U \to V$ is a diffeomorphism onto its image. That is, it is invertible onto its image, and both it and its inverse are continuously differentiable.

b. The differentiable curves that are defined through each $x_0 \in U$ by way of $\gamma_0(\tau) = \Phi_\tau(x)$ are solutions to the equation $d\gamma / d\tau = \mathbf{v}$.

Because of the uniqueness of solutions in U these curves cannot intersect within U, and one finds that U is "foliated" by a *congruence* of integral curves; i.e., U is partitioned into distinct curves that are each a local solution of the system of ordinary differential equations that is defined by $\mathbf{v}$.

Two obvious questions to ask about the local flows that we have defined are those of whether $(-\varepsilon, +\varepsilon)$ can be extended to all of $\mathbb{R}$ for a given U and whether U can be extended to all of V, for a given ε. In the former case, the local flow would be called *complete*, and in the latter case, it would be a *global* flow. Neither possibility is obtained in all cases, but the analytical details are beyond the scope of the present discussion, so we refer the reader to the reference cited above.

b. Differentiation along a flow. To return to the calculus of exterior differential forms, the concept of a Lie derivative is based in the notion of looking at the time rate of change of something as it flows along a given integral curve for a system of first order ordinary differential equations. When our "something" takes the form of a smooth

function F: $\mathbb{R} \times V \to W$, $(\tau, x) \mapsto F(\tau, x)$, where W is a vector space of dimension N, this derivative with respect to τ along a curve $x(\tau)$ becomes:

$$\left.\frac{dF}{d\tau}\right|_{(\tau_0, x(\tau))} = \lim_{\tau \to 0} \frac{1}{\tau}[F(\tau_0 + \tau, x(\tau_0 + \tau)) - F(\tau_0, x(\tau_0))] = \left(\frac{\partial F^A}{\partial \tau} + \frac{dx^i}{d\tau} \frac{\partial F^A}{\partial x^i} \right) \partial_A , \qquad \text{(I.85)}$$

in which $\{\partial_A , A = 1, \ldots, N\}$ is the basis for W defined by a choice of coordinates.

Not surprisingly, this sort of derivative plays a key role in continuum mechanics, where it called variously the "material derivative" or the "substantial derivative." If we set $v^i = dx^i/d\tau$ then we shall also regard this derivative as the *Lie derivative of F with respect to* **v**, and denote it by $L_{\mathbf{v}}F$. We shall state without proof some useful facts about Lie derivatives, and refer the curious to other references for the proofs (e.g., [**1-3**])

The Lie derivative of a vector field $\mathbf{X} = X^j \partial_j$ on V with respect to another vector field $\mathbf{v} = v^i \partial_i$ takes the form:

$$L_{\mathbf{v}}\mathbf{X} = [\mathbf{v}, \mathbf{X}] = \left(v^j \frac{\partial X^i}{\partial x^j} - X^j \frac{\partial v^i}{\partial x^j} \right) \partial_i . \qquad \text{(I.86)}$$

The square brackets that we introduced are referred to as the *Lie bracket* of the vector fields in question. Because they are bilinear, anti-symmetric:

$$[\mathbf{X}, \mathbf{Y}] = -[\mathbf{Y}, \mathbf{X}], \qquad \text{(I.87)}$$

and satisfy the *Jacobi identity:*

$$[\mathbf{X}, [\mathbf{Y}, \mathbf{Z}]] + [\mathbf{Y}, [\mathbf{Z}, \mathbf{X}]] + [\mathbf{Z}, [\mathbf{X}, \mathbf{Y}]] = 0, \qquad \text{(I.88)}$$

the product [.,.]: $\mathfrak{X}(V) \times \mathfrak{X}(V) \to \mathfrak{X}(V)$ defines, by definition, a *Lie algebra* on the (infinite-dimensional) vector space $\mathfrak{X}(V)$ of all vector fields on V.

In fact, the usual vector (or cross) product on $\mathbb{R}^3$ satisfies these requirements and thus defines a Lie algebra that is isomorphic to the Lie algebra of all infinitesimal Euclidian rotations in space. As we shall see, all that one needs to do to make the cross product useful in relativistic physics is to define it on $\mathbb{C}^3$, instead. The Lie algebra thus defined is then isomorphic to that of the infinitesimal Lorentz transformations.

The Lie derivative of more general multivector field on V is obtained from (I.86) by adding the assumption that the Lie derivative with respect to **v** acts on $\Lambda^*(V)$ as a *derivation* with respect to the exterior product:

$$L_{\mathbf{v}}(\mathbf{A} \wedge \mathbf{B}) = L_{\mathbf{v}}\mathbf{A} \wedge \mathbf{B} + \mathbf{A} \wedge L_{\mathbf{v}}\mathbf{B} . \qquad \text{(I.89)}$$

For instance, when one takes the Lie derivative of a simple bivector field $\mathbf{X} \wedge \mathbf{Y}$, one gets:

$$L_{\mathbf{v}}(\mathbf{X} \wedge \mathbf{Y}) = [\mathbf{v}, \mathbf{X}] \wedge \mathbf{Y} + \mathbf{X} \wedge [\mathbf{v}, \mathbf{Y}]. \qquad \text{(I.90)}$$

The Lie derivative of a covector field $\alpha = \alpha_i\, dx^i$ with respect to **v** is given by:

$$\mathrm{L}_{\mathbf{v}}\alpha = (i_{\mathbf{v}}d + di_{\mathbf{v}})\alpha = i_{\mathbf{v}}d\alpha + d\alpha(\mathbf{v}) = \left(v^j \frac{\partial \alpha_i}{\partial x^j} + \frac{\partial (v^j \alpha_j)}{\partial x^i} \right) dx^i , \qquad \text{(I.91)}$$

and can be similarly extended to all *k*-forms by assuming that $\mathrm{L}_{\mathbf{v}}$ also acts as a derivation with respect to the exterior product on differential forms:

$$\mathrm{L}_{\mathbf{v}}(\alpha \wedge \beta) = \mathrm{L}_{\mathbf{v}}\alpha \wedge \beta + \alpha \wedge \mathrm{L}_{\mathbf{v}}\beta . \qquad \text{(I.92)}$$

However, one finds that the expression in parentheses in the first equality of (I.90) is general to all differential forms. One obtains what is sometimes referred to as *Cartan's Magic Formula:*

$$\mathrm{L}_{\mathbf{v}} = i_{\mathbf{v}}d + di_{\mathbf{v}} . \qquad \text{(I.93)}$$

Of particular interest is the Lie derivative of a volume element $\mathcal{V}$ – or any other *n*-form – along the integral curves of a vector field **v**:

$$\mathrm{L}_{\mathbf{v}}\mathcal{V} = i_{\mathbf{v}}d\mathcal{V} + di_{\mathbf{v}}\mathcal{V} = d\#\mathbf{v} = \#\delta\mathbf{v} = (\delta\mathbf{v})\mathcal{V}. \qquad \text{(I.94)}$$

(In the second step, we implicitly used the fact that $d\mathcal{V}$ is an n+1-form on an n-dimensional vector space, hence, zero.) This gives, as a consequence the fact that the flow of **v** preserves the volume element iff **v** has vanishing divergence.

6. Integration of differential forms. Since a differential *n*-form $\phi = f(x^i)\, dx^1 \wedge \ldots \wedge dx^n$ looks suspiciously reminiscent of the integrand $f(x^i)\, dx^1 \ldots dx^n$ for the integration of a function f: $V \to \mathbb{R}$ over some *n*-dimensional region *B* in a vector space *V* of dimension at least *n*, it should come as no surprise that the integral of a differential *n*-form over such an *n*-dimensional region can be defined in much the conventional way that is defined in multivariable calculus by replacing the expression $f(x^i)\, dx^1 \wedge \ldots \wedge dx^n$ with the expression $f(x^i)\, dx^1 \ldots dx^n$. The possible change in sign that comes about by changing the order of the factors in $dx^1 \wedge \ldots \wedge dx^n$ then simply represents the possibility that ultimately there will be two orientations for *B*, and opposite orientations give opposite signs for the integral.

There is another possibility, namely, that *B* is not orientable, in the first place. In that case, *B* does not admit an *n*-form that is non-zero everywhere. We shall discuss the subject of orientation in more detail in the next chapter when we can give it its proper topological context.

We denote the integral of an *n*-form *f* over an *n*-dimensional region *B* by:

$$\int_B \phi = \int_B f(x^1, \cdots, x^n)\, dx^1 \cdots dx^n . \qquad \text{(I.95)}$$

It is important to recognize the integral of an n-form is defined only over an n-dimensional region. This is because one expects the integral of anything over B to be independent of the parameterization of B, and the only n for which n-forms transform properly is $n = \dim(B)$; in effect, this is the only dimension in which the differential map of a change in parameterization contributes only a determinant to the transformation of the n-form components. That is, if $y^i = y^i(x^j)$ represents a change in parameterization – i.e., a diffeomorphism – for the region B then the n-form $f(y^i)\, dy^1 \wedge \ldots \wedge dy^n$ goes to $f(x^i)\, J(x)\, dx^1 \wedge \ldots \wedge dx^n$, in which:

$$J(x) = \det\left(\frac{\partial y^i}{\partial x^j}\right) \tag{I.96}$$

is the *Jacobian* of the diffeomorphism.

Note that since the differential map to a diffeomorphism must be invertible (this follows from the chain rule for differential maps), one must have that $J(x) \neq 0$ everywhere. If $J(x) > 0$ everywhere then the change in parameterization is *orientation-preseving*; if it is < 0 everywhere then it is *orientation-reversing*.

Ultimately, one can only take line integrals of 1-forms, surface integrals of 2-forms, and volume integrals of 3-forms, etc. Although this may sound trivial, actually, it is common practice in physics and engineering to integrate components of vector fields and tensor fields whose rank is largely unrelated to the dimension of the region, such as when one defines the total linear momentum or total angular momentum of a three-dimensional extended object by integrating the components of the corresponding densities. One is cautioned that the resulting integrals are not invariant under changes of frames that are not constant at all points, such as the transition from a holonomic to an anholonomic frame. Hence, just as differentiation can produce fictitious – i.e., frame-dependent – velocities and accelerations, integration can produce "fictitious moments."

If α is a k-form then $d\alpha$ is a k+1-form. Hence, if B is an orientable k+1-dimensional region with a k-dimensional boundary ∂B then it is possible to define the integral of α over ∂B and $d\alpha$ over B invariantly. If $\mathcal{V}$ is the k+1-dimensional volume element on B and $\mathbf{n}$ is the unit normal to the k-dimensional boundary (assuming that V is equipped with a Euclidian metric) then the k-form $i_{\mathbf{n}}\mathcal{V}$, when restricted to the points of ∂B gives it a k-dimensional volume element. In an "adapted" coordinate system for B, if $\mathcal{V} = dx^1 \wedge dx^2 \wedge \ldots \wedge dx^{k+1}$ and x^1 is adapted to the normal direction then $i_{\mathbf{n}}\mathcal{V} = dx^2 \wedge \ldots \wedge dx^{k+1}$.

The two integrals are related in a simple, but far-reaching, way by the fact that:

$$\int_{\partial B} \alpha = \int_B d\alpha\,. \tag{I.97}$$

Depending upon the dimension of B this equality can be interpreted as the fundamental theorem of calculus (dim = 1), Green's theorem (dim = 2), or Stokes's theorem (dim = 3). We shall discuss this more in the next section.

In fact, Gauss's theorem (i.e., the divergence theorem) also relates to the three-dimensional form, by way of Poincaré duality. We shall have more to say about this in Chapter III after we have defined charge and flux, but, for now, suppose that $\alpha = \#\mathbf{a}$. Then (I.97) becomes:

$$\int_{\partial B} \#\mathbf{a} = \int_B (\delta\mathbf{a})\mathcal{V} . \tag{I.98}$$

Proof:

$$\int_{\partial B} \#\mathbf{a} = \int_B d\,\#\mathbf{a} = \int_B \#\,\delta\mathbf{a} = \int_B (\delta\mathbf{a})\mathcal{V}$$

In the standard treatments on differential forms, the generic term for this theorem is *Stokes's theorem* [10]. Since the topological significance of this relationship is quite subtle, but important, we shall have more to say about this theorem in Chapter II.

One can further note that, from (I.97), whenever a k-form α is *closed* – viz., $d\alpha = 0$ – the integral of α over the boundary of a k+1-dimensional region must always vanish.

We can now return to the discussion of adjointness between d and δ that we suspended in a previous section. We simply redefine our bilinear pairing to be $<.,.> : A^k \times A_k \to \mathbb{R}$, which takes $(\alpha, \mathbf{A})$ to:

$$<\alpha, \mathbf{A}> = \int_B \alpha \wedge \#\mathbf{A} , \tag{I.99}$$

when B is a compact, orientable n-dimensional manifold with boundary.

By integration and an application of Stokes's theorem, (I.79) takes the form:

$$<d\alpha, \mathbf{A}> + (-1)^k <\alpha, \delta\mathbf{A}> = \int_{\partial B} \alpha \wedge \#\mathbf{A} . \tag{I.100}$$

Hence, when B has no boundary, one can say that:

$$<d\alpha, \mathbf{A}> = -(-1)^k <\alpha, \delta\mathbf{A}>. \tag{I.101}$$

In particular, in even dimensions d and δ are skew-adjoint, while in odd dimensions they are self-adjoint.

7. Relationship to vector calculus. The first thing that one notices about the relationship between exterior differential forms and vector analysis is that the exterior product subsumes the vector cross product, as it defined in $V = \mathbb{R}^3$.

To see this, give $A_1 = \mathbb{R}^3$ the canonical basis $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0)$, $\mathbf{e}_3 = (0, 0, 1)$, which one depicts as column vectors, so the reciprocal basis θ^i has the same triples of numbers as row vectors. One can give A_2 the basis $\mathbf{e}_1 \wedge \mathbf{e}_2$, $\mathbf{e}_2 \wedge \mathbf{e}_3$, $\mathbf{e}_3 \wedge \mathbf{e}_1$, which means:

$$\mathbf{e}_i \wedge \mathbf{e}_j = \varepsilon_{ijk}\, \mathbf{e}_k . \tag{I.102}$$

[10] Vladimir Arnol'd may have made a valid point in [**11**] when he suggested that if one wished to be consistent with the modern tendency to hyphenate all of the names that contributed to the general form of a modern theorem then one might wish to call it the *Newton-Leibniz-Gauss-Green-Ostragradskii-Stokes-Poincaré theorem!* (…and perhaps abbreviate this to the acronym NLGGOSP!)

Hence, one can think of the symbol ε_{ijk} as essentially the matrix of the isomorphism of A_1 with A_2.

Now, strictly speaking, this isomorphism only comes about because we chose a basis for $\mathbb{R}^3$. In point of fact, the natural frame-invariant isomorphism is the Poincaré duality between A_1 and A^2:

$$\theta^i \wedge \theta^i = \varepsilon^{ijk}\, \mathbf{e}_k \,. \tag{I.103}$$

The volume element on V that defines this is given by the 3-form $\theta^1 \wedge \theta^2 \wedge \theta^3$.

In order to make this a frame-invariant isomorphism of A_1 with A_2, one must define either an isomorphism of A_1 with A^1 or A_2 with A^2, such as one usually gets from the Euclidian metric. However, the spirit of pre-metric electromagnetism is to treat the spacetime metric as something that shows up at a later stage than the most fundamental assumptions about the spacetime manifold. Indeed, we shall find that both the electric permittivity and the magnetic permeability by themselves can also define such isomorphisms.

For now, we notice that the components of $\mathbf{a} \wedge \mathbf{b}$ with respect to the stated basis are:

$$(\mathbf{a} \wedge \mathbf{b})_{ij} = a_i\, b_j - a_j\, b_i \,, \tag{I.104}$$

and those of $\mathbf{a} \times \mathbf{b}$ are:

$$(\mathbf{a} \times \mathbf{b})_i = \tfrac{1}{2}\, \varepsilon_{ijk}\, (\mathbf{a} \wedge \mathbf{b})_{jk} \,. \tag{I.105}$$

Hence, one is basically associating the two sets of components by way of the isomorphism of vectors and bivectors that we have chosen.

Of course, the advantage of exterior forms then becomes the generality of their application in the eyes of dimension, while the cross product can only be defined in dimension three.

The triple product $\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}$, which also involves the Euclidian scalar product, is related to an even more elementary exterior product:

$$\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c} = (\mathbf{a} \cdot \mathbf{b} \times \mathbf{c})\boldsymbol{\mathcal{V}} = \det[\mathbf{a} \mid \mathbf{b} \mid \mathbf{c}]\boldsymbol{\mathcal{V}}. \tag{I.106}$$

One finds that the exterior derivative operator immediately subsumes both the gradient operator, as applied to smooth functions, and the curl operator, as applied to vector fields, and we have already showed that the divergence operator, as we have defined it, agrees with the classical one on vector fields:

$$(df)_i = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right) = (\nabla f)_i \,, \tag{I.107a}$$

$$(dv)_{ij} = v_{i,j} - v_{j,i} = \varepsilon_{ijk} (\nabla \times \mathbf{v})_k \,, \tag{I.107b}$$

$$\delta \mathbf{v} = \nabla \cdot \mathbf{v} \,. \tag{I.107c}$$

However, as is usually the case in three-dimensional computations, one is always implicitly defining the Euclidian scalar product on $\mathbb{R}^3$ and using it to associate vectors

with covectors. For instance, the gradient of a function is usually thought of as a vector field, not a covector field. Consequently, in the component formulation of vector and tensor analysis, there is no attention paid to whether one is implicitly raising or lowering indices as it suits the calculations; indeed, all indices are usually written as subscripts, for simplicity. This is really a bad habit to get into if one is also going to be concerned with spaces of other dimensions and metrics of other signature types, since it often involves remembering when such tacit constructions were being carried out, if one is to replace them with the more general constructions.

The fact that $d^2 = 0$, in any case, subsumes the vanishing of $\nabla\times\nabla f$ for any f and $\nabla\cdot(\nabla\times\mathbf{v})$ for any $\mathbf{v}$; that is, the curl of a gradient and the divergence of a curl vanish identically. Ordinarily, one assumes that the vanishing of $\nabla \times \mathbf{v}$ implies that $\mathbf{v} = \nabla f$ for some (non-unique) function f, but, as we shall see in the next chapter, this really goes back to the fact that we are looking at vector fields and k-forms on a vector space, which is, among other things, simply connected. That is, whether $d\alpha = 0$ implies that there is a ϕ such that $\alpha = d\phi$ depends upon deeper topological considerations, which we will discuss in Chap. III.

In order to relate the general notation for the integration of a k-form α over a k-dimensional region B in an n-dimensional vector space V to the various notations for the integral of a function or vector field over regions of $\mathbb{R}^3$ that are one, two, and three-dimensional, one mostly has to show how the free index of a vector field $\mathbf{A}(x) = (A^1(x), A^2(x), A^3(x))$ gets absorbed to produce a k-form that looks like $f\mathcal{V}_k$.

For line integrals, it is sufficient to turn the vector field $\mathbf{A}$ into a 1-form $A = A_i\, dx^i$ by means of the Euclidian scalar product ($A_i = \delta_{ij}A^i$) and show that this is equal to:

$$A = \mathbf{A}\cdot d\mathbf{s}\,, \tag{I.108}$$

in which:

$$d\mathbf{s} = \mathbf{v}\, d\tau = (dx^1, dx^2, dx^3) \tag{I.109}$$

is the differential element of arc length; however, the result is immediate by this definition. Hence, we can unambiguously write:

$$\int_C A = \oint_C \mathbf{A}\cdot d\mathbf{s}\,. \tag{I.110}$$

In two dimensions, the elimination of the free vector index is accomplished by means of taking the scalar product of $\mathbf{A}$ with the normal vector field $\mathbf{n} = \#^{-1}dS$ to the surface S over which the integration is performed; here dS refers to the surface element on S (i.e., a non-vanishing 2-form in $\Lambda^2(S)$). We can also associate the vector field $\mathbf{A}$ with the 2-form:

$$\#\mathbf{A} = i_{\mathbf{A}}\mathcal{V} = A_n\, i_{\mathbf{n}}\mathcal{V} = A_n\, dS. \tag{I.111}$$

Indeed, the only 2-forms that can be integrated over S have to be tangent to S, and therefore of the form $A_n\, dS$.

If one uses an adapted coordinate system in the neighborhood of S that makes:

$$\mathcal{V} = n \wedge dS \qquad (n_i = \delta_{ij}\, n^j) \tag{I.112}$$

then one has:

$$\#\mathbf{A} = i_{\mathbf{A}}\mathcal{V} = n(\mathbf{A})\, dS - n \wedge i_{\mathbf{A}}\, dS\,. \tag{I.113}$$

and the part of this that is tangential to S is:

$$n(\mathbf{A})\, dS = (\mathbf{A} \cdot \mathbf{n})\, dS\,. \tag{I.114}$$

Thus, in order to make sense of the equality:

$$\int_S A = \oiint_S (\mathbf{A} \cdot \mathbf{n})\, dS \tag{I.115}$$

the 2-form A and the vector field $\mathbf{A}$ must be related by:

$$A = P_S(\#\mathbf{A}), \tag{I.116}$$

in which P_S refers to the projection onto the tangential 2-forms. This implies that, although the map # is a bijection, nonetheless, a sizable subspace of vector fields will produce the same tangential 2-form, namely the ones for which $i_{\mathbf{A}}dS$ is non-vanishing; i.e., ones that differ by a vector that is tangent to S.

As for a three-dimensional B, the only things that can be integrated invariantly are pseudo-scalars of the form $f\mathcal{V}_3$. A non-algebraic way of turning $\mathbf{A}$ into a scalar is to take its divergence, which gives Gauss's theorem (I.99) the form:

$$\oiint_{\partial B} (\mathbf{A} \cdot \mathbf{n})\, dS = \oiiint_B (\nabla \cdot \mathbf{A})\, \mathcal{V}\,. \tag{I.117}$$

Something that one begins to notice after working with the calculus of exterior differential forms long enough is that even though it is possible to find differential form equivalents for many of the tabulated formulas of vector calculus beyond the elementary ones that were discussed above, nevertheless, part of the beauty of differential forms is in the way that many of these vector calculus identities become largely unnecessary, since one generally only needs them as lemmas for the proofs of more fundamental results that can be proved more directly with differential forms.

References

1. F. Warner, *Differentiable Manifolds and Lie Groups,* Scott Foresman, Glenview, IL, 1971.
2. S. Sternberg, *Lectures on Differential Geometry 2nd ed.,* Chelsea, New York, 1983.
3. R. L. Bishop and S. Goldberg, *Tensor analysis on Manifolds,* Dover, New York, 1980.

4. H. K. Nickerson, D. C. Spencer, and N. E., Steenrod, *Advanced Calculus*, Van Nostrand, Princeton, 1959.
5. L. H. Loomis and S. Sternberg, *Advanced Calculus*, Addison Wesley Longman, NY, 1968.
6. M. Spivak, *Calculus on Manifolds*, Westview Press, CO, 1971.
7. H. Cartan, *Differential Forms*, Dover, NY, 2006.
8. H. Flanders, *Differential Forms*, Academic Press, NY, 1963.
9. W. Greub, *Multilinear Algebra*, Springer, Berlin, 1967.
10. V.I. Arnol'd, *Ordinary Differential Equations*, The MIT Press, MA, 1975.
11. V.I. Arnol'd, *Mathematical Methods in Classical Mechanics*, Springer, Berlin, 1978.

Chapter II

Topology of differentiable manifolds

From the standpoint of physics, there are two main reasons for extending the calculus of exterior differential forms from the formulation that was presented in the previous chapter for vector spaces to a formulation that pertains to more general differentiable manifolds:

1. The introduction of coordinate systems is inevitable in physics and differentiable manifolds are the mathematical structures that have evolved to address that situation.

2. The physical foundations of electromagnetism, such as charge and flux, are manifestly topological in character.

Consequently, the purpose of this chapter is to carry out that extension so that the aforementioned foundations of electromagnetism can be presented in their topological form in the next chapter. We shall present only the elements of point-set topology to begin with, and then we shall introduce some of the more advanced topological notions in a form that is adapted to the nature of the problem at hand.

1. Differentiable manifolds. In the previous chapter, we made use of coordinate systems on vector spaces, which often came about as a consequence of defining a basis for the vector space. Actually, such a construction will produce only "rectilinear" or "Cartesian" coordinate systems. If one wishes to discuss "curvilinear" coordinate systems then one will have to deal with the fact that they usually come about as an adaptation to the demands of dealing with nonlinear spaces such as cylinders, spheres, and ellipsoids.

First, we need to recall some useful generalities at the level of point-set topology.

a. Topological spaces [**1, 2**]. In mathematics, one of the most useful nonlinear generalizations of a vector space is that of *topological space*. Such a space is a set S that is given a *topology* – viz., a collection of *open subsets* that satisfy certain axioms regarding unions and intersections. In particular, the union of any family of open subsets is another open subset and the intersection of any *finite* family is another open subset [11]. One can then define a *closed* subset as the set complement of an open set and obtain a collection of closed subsets that satisfy the complementary axioms (finite unions, any intersections). Conversely, one can start with closed subsets and define open subsets by complementation. Note that a given subset of S can be open, closed, both, or neither.

Some elementary elementary examples of topological spaces that will be of interest in what follows are the real line $\mathbb{R}$ and the Cartesian product $\mathbb{R}^n$ of ordered n-tuples of real numbers. The topology that is commonly used for $\mathbb{R}$ is defined by means of its total

[11] Often, one encounters the axioms that S and the empty set are open subsets, although to some authors, these axioms follow logically from the two that were stated.

ordering <, namely, the interval topology, whose open subsets are unions of open intervals of the form $(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$. The closed subsets are then intersections of finite unions of closed intervals $[a, b] = (a, b) \cup \{a, b\}$. One can also define a topology for $\mathbb{R}$ by means of the metric $d(x, y) = | x - y |$, which allows one to define open "ε-balls" $B_x(\varepsilon)$ about each point by way of $B_x(\varepsilon) = \{y \in \mathbb{R} \mid | x - y | < \varepsilon\}$. Since these will also define open intervals of the form $(x - e, x + e)$, one sees that the two topologies just defined amount to the same open subsets; hence, they are topologically equivalent. As for $\mathbb{R}^n$, one can extend the interval topology on $\mathbb{R}$ by means of the product topology on $\mathbb{R}^n$, which makes every open subset the union of finite intersections of products $(a_1, b_1) \times \ldots \times (a_n, b_n)$ of open intervals in $\mathbb{R}$. An equivalent topology is given by extending the concept of ε-balls to n dimensions by extending the norm in the Euclidian manner $\| \mathbf{v} \| = (v_1^2 + \cdots + v_n^2)^{1/2}$. Of course, one needs to show that every ε-ball is the union of n-cubes or vice versa, but, as it turns out, in order to show the topological equivalence of the two topologies, it is only necessary to show that there is a continuous, invertible map from one to the other that also has a continuous inverse, as we shall discuss shortly. Such a map is given by radially projecting the points of one onto the other.

The effect of defining a topology is to define an abstract way of characterizing "closeness" without introducing an actual numerical way of defining the concept, such as a metric. A *neighborhood* of a point $x \in S$ is any subset $U \subset S$ that contains an open subset that includes x. Hence, any point of S defines a localization of the topology on S to a system of neighborhoods of x. The partial ordering of subset inclusion allows one to the think of the relative "closeness" of a point y to x, as compared to another point z as being related to whether there is a neighborhood of x that contains one point, but not the other.

One can use the partial ordering of subset inclusion to define a partial ordering on the set of topologies over a given set. A topology τ on S is *finer* than another topology τ' iff every open subset in τ' is also an open subset in τ. Hence, τ potentially has "more open subsets" than τ'. One also says that the topology τ' is *coarser* than the topology τ.

Another issue that is related to the "fineness" of the topology is the question of whether there is always an open neighborhood of x that excludes any other given point of S. This gets one into the *separation* axioms, the most common of which is the Haussdorff axiom: A topological space is *Haussdorff* iff given any two distinct points there are disjoint open neighborhoods of each point. Another axiom that shows up in the study of manifolds is normality: A topological space is *normal* iff any two disjoint closed subsets have disjoint open neighborhoods. One sees that as long as subsets that consist of only individual points are always closed normal spaces must be Haussdorff.

So far, the nature of topological spaces seems rather set-theoretic and prosaic. The real power of the topological structure on a set emerges when one uses it to define the continuity of maps. A map $f: A \to B$ between topological spaces A and B is said to be *continuous* iff the inverse image $f^{-1}(V) = \{x \in S \mid f(x) \in V\}$ of any open subset of B is an

open subset of A. All that one needs to do to reconcile this with the "ε–δ" definition that one encounters in calculus is to consider the topologies on normed vector spaces V, W that are given by considering open balls $B_\delta(x)$ in V and $B_\delta(y)$ in W. Hence, one of the subtle sources of the power of topology is that it simultaneously generalizes elements of both geometry and analysis.

An even stronger condition on the map f than continuity is that it should be invertible and have a continuous inverse. Such a map between topological spaces is called a *homeomorphism*. They are fundamental because they represent topological equivalences, since one has not only a one-to-one correspondence between the points of the spaces one also has a one-to-one correspondence between the open subsets of the topologies.

For instance, the map f might be the identity map on a set S that has been given two different topologies. Although the map is clearly invertible, whether it is continuous in one direction or another is another way of characterizing the fineness of one topology relative to the other one. For instance, the previous example of $\mathbb{R}^n$, when given both the product topology and the norm topology, has the property that the identity map is continuous, along with its inverse, since every open subset of one topology is an open subset of the other one; hence, we can say that the two topologies are homeomorphic.

Having defined topological structures and maps that relate to them, one also wishes to know what sort of properties of topological spaces are preserved by such maps. Since homeomorphic spaces have equivalent topologies, it is essentially a matter of definition to say that a topological property of a space is one that is equally true for homeomorphic spaces. The properties that are preserved by merely continuous maps, at least on their images, are somewhat more definitive.

One such property of continuous maps is *compactness:* a topological space S is called *compact* iff it is Haussdorff and every open covering of it can be reduced to a finite sub-covering; note that this does not say that any space S that can be covered by a finite number of open subsets is compact, since S itself covers any topological space, in any case. One finds that any closed subset of a compact space is compact, and any compact subset of a Haussdorff space is closed. By the *Heine-Borel theorem*, the compact subsets of $\mathbb{R}^n$, when given the Euclidian norm topology, are the closed and bounded subsets. The fact that the image of a compact topological space by a continuous map is compact also explains why any continuous real-valued function on a compact space, such as a sphere, must have a maximum value, along with a minimum value, at some points of the space.

Another property that is preserved by continuous maps, which also gets generalized to more sophisticated methods in topology, such as algebraic topology, is connectedness. A topological space is *connected* iff it is not the union of two disjoint open subsets; equivalently, the only subsets that are both open and closed are the empty set and the entire set. The fact that continuous real-valued function from a connected space S onto the interval (a, b) in the real line must attain each intermediate value between a and b is due to the fact that the connected subsets of $\mathbb{R}$ are the intervals.

Conversely, a continuous map $\gamma\colon [a, b] \to S$ must have a connected image. Such a map is described as a *path* in S from $\gamma(a)$ to $\gamma(b)$. If any two points of S admit at least one path that connects them then S is called *path-connected.* The image of a path-connected

space under a continuous map is clearly path-connected. One also finds that any path-connected space is connected, although the converse is not necessarily true.

A particular class of paths in a space that is of fundamental interest is the class of all loops: A *loop* in S is a path $\gamma:[a, b] \to S$ such that $\gamma(a) = \gamma(b)$. Note that this includes the possibility that the image of γ is just one point.

b. Homotopy theory [**1-3**]. A fundamental question about a topological space is whether every loop can be continuously deformed to a point. That is, two continuous maps $f, g: A \to B$ between topological spaces are called *homotopic* iff there is a continuous map $F: A \times [0, 1] \to B$ such that $F(x, 0) = f(x)$ and $F(x, 1) = g(x)$. Hence, if f is homotopic to g then one can continuously deform f into g by means of a one-parameter family of maps in between. In effect, this is also like defining a path in the "topological space" of continuous maps from A to B, although defining the topology on the space that would make this statement well-defined proves to be the less practical way of looking at homotopies. One finds that homotopy is an equivalence relation between continuous maps, since every map is homotopic to itself, if f is homotopic to g then g is homotopic to f, and if f is homotopic to g while g is homotopic to h the f is homotopic to h. Hence, one can partition the set of all continuous maps from A to B into equivalences classes, namely, *homotopy classes*, which we denote by $[f]$.

In the case of homotopies of loops in a topological space S, one finds that the set $\pi_1(S)$ of homotopy classes $[\gamma]$ of loops can often have a very elementary character to it. When there is only one such class – i.e., all loops are homotopic to constant loops – one calls S *simply-connected*, and denotes this by $\pi_1(S) = 0$. For instance, any vector space is simply-connected. The notion of simple-connectedness is neither stronger nor weaker than path-connectedness, since a plane minus a point – i.e., a *punctured plane* – is path-connected, but not simply-connected, while a pair of disjoint discs in the plane is simply-connected, but not path-connected. The homotopy classes of loops in a circle – i.e., the homotopy classes $[f]$ of continuous maps $f: S^1 \to S^1$ – are in one-to-one correspondence with the integers $\mathbb{Z}$, by way of the "winding number" of the map f.

The set $\pi_1(S)$ can also be given a group structure by composing subsequent loops, and with that group structure one calls $\pi_1(S)$ the *fundamental group* of S. The subscript "1" on $\pi_1(S)$ is suggestive of a sequence of other such groups. Indeed, one can define *higher homotopy groups* for a topological space S by defining the k^{th} homotopy group $\pi_k(S)$ to consist of homotopy classes of continuous maps of the k-dimensional sphere S^k into S. The group structure is somewhat more difficult to describe (cf. [**3**]), although, as it turns out, homotopy groups are always Abelian in dimensions higher than one. Interestingly, although clearly $\pi_n(S^n) = \mathbb{Z}$, by way of a higher-dimensional winding number argument, and $\pi_1(S^n) = 0$ whenever $n > 1$, nonetheless, finding the higher homotopy groups of spheres in general is more difficult than it sounds.

For the sake of completeness, one usually denotes the set of all path connected components of S by $\pi_0(S)$, although it is not generally given a group structure.

Since the composition of continuous maps is continuous, one finds that the image of a loop γ in A by a continuous map $f: A \to B$ is a loop $f \cdot \gamma$ in B. Since homotopic loops in A will have homotopic images in B under f – in particular, homotopic loops – a continuous

map will take $\pi_1(A)$ to a subgroup of $\pi_1(B)$. Indeed, this situation extends to the higher dimensions, and when two topological spaces are homeomorphic they will have isomorphic homotopy groups in all dimensions. The converse, however, is not always true, although it has finally been proven for spheres, which was the generalized Poincaré conjecture. (The original conjecture was for three-dimensional spheres, which was also the last dimension in which the conjecture was proved.)

The most extreme form of a homotopy is a *contraction:* A topological space is called *contractible* iff the identity map is homotopic to a constant map; i.e., the whole space is homotopic to one of its points. For example, any vector space is contractible, as is a space that consists of only one point, to begin with. The homotopy groups of a contractible space vanish in every dimension, since this is the case for a point. It is a deep result of homotopy theory that the converse statement is also true: Any topological space whose homotopy groups vanish in every dimension is contractible.

c. Differential structures [**4-9**]. The next step beyond considering topological spaces that are homeomorphic to vector spaces, which are only so topologically interesting, is considering topological spaces that are locally homeomorphic to vector spaces. That is, a *topological manifold* is a topological space M such that every point $x \in M$ has a neighborhood U that is homeomorphic to $\mathbb{R}^n$. One can think of the homeomorphism ϕ: $U \to \mathbb{R}^n$, $p \mapsto (x^1(p), \ldots, x^n(p))$ as defining a *coordinate system* on U, and we call the pair (U, ϕ) a *coordinate chart* on M about x.

Examples of such spaces are spheres, tori, polygons, polyhedra, and many algebraic sets. Counter-examples would be such things as intersecting curves, a pair of tangent spheres, and a disk with a line segment attached to its perimeter at one endpoint. In effect, since $\mathbb{R}^n$ is homeomorphic to any open n-ball (regardless of radius), to say that a point of M has a neighborhood that looks like $\mathbb{R}^n$ topologically is to say that it has a neighborhood that looks like a (sufficiently small) open n-ball.

If one has another neighborhood V of x and another homeomorphism ψ: V: $\to \mathbb{R}^n$, $p \mapsto (y^1(p), \ldots, y^n(p))$ then on the overlap $U \cap V$ one can invert ϕ and compose it with ψ to obtain a homeomorphism $\psi \cdot \phi^{-1}$: $\mathbb{R}^n \to \mathbb{R}^n$, $(x^1, \ldots, x^n) \mapsto (y^1, \ldots, y^n)$ which then represents a *coordinate transformation*.

From the examples given above, it is clear that topological equivalence is really quite vague in character, since it ignores many of the geometric properties of spaces. For instance, every point of a circle has a unique tangent, but the vertices of a triangle, which is homeomorphic to a circle, have double tangents. Hence, if one wishes to refine the equivalence classes of topological equivalence, a next step might be to include some consideration for how tangent spaces are associated with the points of the space. Since the main point of differentiation is to locally linearize nonlinear functions by associating them with linear ones, to some degree of approximation, this brings one into the realm of differential topology.

In order to refine the structure of a topological manifold, since the differentiation of maps from $\mathbb{R}^n$ to $\mathbb{R}^n$ can be defined (when they are differentiable) we restrict the class of allowable coordinate transformations, as defined above, to the ones that are differentiable and have a differentiable inverse; i.e., to *diffeomorphisms* of $\mathbb{R}^n$. Hence, since such a coordinate transformation $\psi \cdot \phi^{-1}: \mathbb{R}^n \to \mathbb{R}^n$ will have a differential map $D(\psi \cdot \phi^{-1})|_y : \mathbb{R}^n \to \mathbb{R}^n$ that is defined at each $y \in \mathbb{R}^n$, and since this is an invertible linear map it defines an element of the group $GL(n)$ of all invertible $n \times n$ real matrices (by means of the standard basis on $\mathbb{R}^n$) we can define a *transition function* g_{UV}: $U \cap V \to GL(n)$, $x \mapsto D(\psi \cdot \phi^{-1})|_{\phi(x)}$. If the coordinates that are defined by ϕ are x^i and those defined by ψ are y^i then the coordinate transformation takes the form $y^i = y^i(x^j)$ and the transition function takes the form:

$$\left[g_{U\cap V}(x)\right]^i_j = \left.\frac{\partial y^i}{\partial x^j}\right|_x . \tag{II.1}$$

Notice that is it the coordinate *transformations* that we are restricting, not the coordinate *charts*. However, having fewer allowable transformations will lead to having fewer charts.

A *differential structure* on a topological manifold M is a collection $\{(U_\alpha, \phi_\alpha), \alpha \in \Gamma\}$ of charts – which one calls an *atlas* – such that:

1. The atlas covers M, i.e., every point of M is contained in at least one chart.
2. All of the coordinate transformations are diffeomorphisms of $\mathbb{R}^n$.
3. The atlas is maximal, in the sense that any chart whose intersections with the existing charts gives coordinate transformations that are diffeomorphisms must already be a chart of the atlas.

A topological manifold that has been given a differential structure is called a *differentiable manifold.* If $f: M \to N$ is a map from a differentiable manifold M to another one N then for every chart (U, ϕ) about a point $x \in M$ and every chart (V, ψ) about the point $f(x) \in N$ one can define the map $\psi \cdot f \cdot \phi^{-1}: \mathbb{R}^m \to \mathbb{R}^n$ by inversion and composition.

If this map is differentiable for every possible x, (U, ϕ), and (V, ψ) then one calls the map f itself *differentiable.* Furthermore, if f is invertible, differentiable, and has a differentiable inverse then one calls f a *diffeomorphism* of M and N. This is a stronger statement than homeomorphism, since every differentiable map is continuous, but not always the converse.

When M is given two differential structures, one can consider the differentiability of the identity map. If it is a diffeomorphism then the two differential structures are equivalent. Although the definition of differential structure sounds sufficiently general as to encompass all of the possibilities for a given M, that is not the case. Even in the simplest example of differential structures on $\mathbb{R}^n$ itself, it is now known (thanks to a theorem of Simon Donaldson in 1983) that all of the spaces $\mathbb{R}^n$ have a unique differential

structure except for $\mathbb{R}^4$, which has an infinitude of inequivalent differential structures that are often called "fake $\mathbb{R}^4$'s." It had been known for a long time, from the work of Milnor, that the seven-dimensional sphere admits two inequivalent differential structures: the conventional one that it inherits from $\mathbb{R}^8$ and an "exotic" one.

More specifically, when an n-sphere is defined as a point set in Euclidian $\mathbb{R}^{n+1}$ in the usual way, one cannot cover it with one chart since S^n is compact, but $\mathbb{R}^n$ is not. At the very least, one can cover it with two charts, one that is homeomorphic to an open n-disc centered at the North pole and covers everything except the South pole and another one with the poles reversed. Their intersection is everything except for the two poles, which is an open subset that is homeomorphic to the "cylinder" $\mathbb{R} \times S^{n-1}$, with S^{n-1} playing the role of "equator."

A closely related compact differentiable manifold to S^n is the n-dimensional real projective space $\mathbb{R}\mathrm{P}^n$, which we shall return to later in this work. It consists of all lines through the origin in $\mathbb{R}^{n+1}$, and since every such line intersects S^n (when centered at the origin) in two antipodal points, one can also think of $\mathbb{R}\mathrm{P}^n$ as consisting of all pairs of antipodal points on S^n. At the very minimum, one can cover $\mathbb{R}\mathrm{P}^n$ with n+1 charts, which are defined by the fact that the projection of *homogeneous coordinates* $(x^0, \ldots, x^n) \in \mathbb{R}^{n+1}$ for a point of $\mathbb{R}\mathrm{P}^n$ onto its *inhomogeneous coordinates* $(X^1, \ldots, X^n)$ can be accomplished in n+1 ways by choosing a non-zero coordinate x^k and defining $X^i = x^i/x^k$, all $i \neq k$.

d. Tangency, tangent spaces [**4-9**]. In multivariable calculus, one learns that differentiation allows one to associate a straight line with each point of a differentiable curve and, more generally, a k-dimensional vector space with each point of a differentiable k-dimensional hypersurface by way of the concept of tangency. The notion of a differential structure on a topological manifold also allows one to associate an n-dimensional vector space T_xM with each point x of an n-dimensional differentiable manifold M by a more general notion of tangency.

The key to this construction is to note that when the images of two curves in M in $\mathbb{R}^n$ for some coordinate chart (U, ϕ) have a common tangent line in at a particular point $x \in U \subset M$ the same will be true in any other coordinate chart; i.e., tangency is a coordinate-invariant concept. One can then define an equivalence class $[\gamma]_x$ of all curves through x M that have a common tangent in $\mathbb{R}^n$ at x for some, and therefore all, coordinate charts about x. One calls this equivalence class a *tangent vector* to M at x, since it is associated with a vector in $\mathbb{R}^n$. The set of all tangent vectors to M at x is called the *tangent space* to M at x and is usually denoted by T_xM.

One can give T_xM the structure of an n-dimensional vector space by way of the fact that $\mathbb{R}^n$ has that structure, but here is where differential topology takes a convenient, but non-obvious, turn and associates a tangent vector $\mathbf{v}_x \in T_xM$, not merely with an n-tuple of real numbers, such as $(v^1, \ldots, v^n)$, that is defined by a choice of coordinate chart (U, x^i) about $x \in M$, but with the *directional derivative operator* that these components define:

$$\mathbf{v}_x = v^i \frac{\partial}{\partial x^i} . \tag{II.2}$$

The key to making this association work is to see that this directional derivative also has a coordinate invariant character, as well, from the chain rule for differentiation. In another coordinate system (V, y^i) about x, one has:

$$\mathbf{v}_x = \overline{v}^i \frac{\partial}{\partial y^i} = \overline{v}^i \frac{\partial x^j}{\partial y^i} \frac{\partial}{\partial x^j} , \tag{II.3}$$

which is consistent with (II.2) as long as:

$$v^j = \left. \frac{\partial x^j}{\partial y^j} \right|_x \overline{v}^j = [g^{-1}_{U \cap V}]^i_j \, \overline{v}^j . \tag{II.4}$$

Note that the components of $\mathbf{v}$ transform by means of the inverse of the transition function that is defined by the coordinate transformation. Hence, one can think of transition functions as transformations that act on tangent objects to $\mathbb{R}^n$, while coordinate transformations act on points of $\mathbb{R}^n$.

Now that we have defined a notion of differential maps and functions on M and tangent vectors as directional derivatives, we can give an invariant sense to the action of the tangent vector $\mathbf{v}_x \in T_xM$ on a differentiable function f: $M \to \mathbb{R}$, which actually only needs to be defined in some neighborhood of x:

$$\mathbf{v}_x f = v^i \left. \frac{\partial f}{\partial x^i} \right|_x . \tag{II.5}$$

There is a slight generalization of the n operators $\partial_i = \partial/\partial x^i$ that one uses as a generalized basis for T_xM, at least relative to a choice of chart, that has far-reaching applications and fundamental significance to physics, in general. That is the concept of a tangent *frame* in T_xM , namely, a set $\{\mathbf{e}_i, i = 1, \ldots, n\}$ of n linearly independent tangent vectors in T_xM. Hence, one can represent any tangent vector $\mathbf{v}_x \in T_xM$ uniquely in the form:

$$\mathbf{v}_x = v^i \, \mathbf{e}_i \tag{II.6}$$

relative to this frame.

Since the members of the frame $\mathbf{e}_i$ are themselves tangent vectors, when one chooses a coordinate chart (U, x^i) about x one can express them as:

$$\mathbf{e}_i = A_i^j \partial_j \tag{II.7}$$

for a unique invertible matrix of components A_i^j. Hence, the n operators ∂_i can just as well be regarded as defining a tangent frame in T_xM. Since they are only defined by a choice of coordinate chart, one refers to the frame $\{\partial_i\}$ as the *natural frame* at x that is defined by (U, x^i).

e. Vector fields [**4-9**]. So far, the components of a tangent vector, such as $\mathbf{v}_x$, relative to some chosen tangent frame are just constant scalars. Since a choice of coordinate chart (U, x^i) allows us to define a tangent vector at every $x \in U$, one can extend the concept of a tangent vector at x to a *tangent vector field* on U – or *vector field,* for short – whose components relative to the natural frame ∂_i on $\mathbb{R}^n$ will be functions on U*:*

$$\mathbf{v}(x) = v^i(x)\, \partial_i \,. \tag{II.8}$$

In order to extend the concept of a vector field on U to a vector field on all of M, we need to introduce one more general construction, in the form of the *tangent bundle* to M. It is a differentiable manifold $T(M)$ of dimension $2n$ that one obtains by taking the disjoint union of all T_xM as x ranges over all of the points of M. Hence, as a disjoint union, when x and y are distinct points of M the vector spaces T_xM and T_xM will be distinct from each other, as well, no matter how "close" the two points are. There is a natural projection $T(M) \to M$, $\mathbf{v}_x \mapsto x$, and the set of all $\mathbf{v}_x$ that project to a given x, which is, of course T_xM, is called the *fiber* of the projection over x. One can also restrict the projection to $T(U) \to U$ for any $U \subset M$.

The coordinate systems about each tangent vector $\mathbf{v}_x \in T(M)$ take the form of homeomorphisms $T(U) \to U \times \mathbb{R}^n$ for some, but not all, open neighborhoods U about each $x \in M$, such that the projection of $T(U)$ onto U corresponds to the projection of $U \times \mathbb{R}^n$ onto its first factor. Such a coordinate chart is called a *local trivialization* of $T(M)$ over U, since the most elementary sort of *fibration* $A \to B$ of one topological space A over another B (i.e., a projection that is local trivial) is the projection $M \times N \to M$ of a product space onto one of its factors; one calls such a fibration *trivial*. In that sense, fiber bundles do not become topologically interesting until one considers the ones for which the local trivializations cannot be extended beyond some limit to global trivializations. For instance, even on a manifold as elementary as the two-dimensional sphere S^2 the tangent bundle $T(S^2)$ is not globally trivializable, although one can define a local trivialization over an open subset $U \subset S^2$ that omits only a single point. Indeed, the only dimensions for which spheres are trivializable are dimensions 0, 1, 3, and 7. (This is actually closely related to the fact that the only real "division algebras," i.e., algebras with inverses, are defined over $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{R}^4$, and $\mathbb{R}^8$.)

Once one has the concept of the fibration $T(M) \to M$ to work with, one can define the opposite concept of a (global) *section* of this fibration, namely, a differentiable map $\mathbf{v}$: $M \to T(M)$, $x \mapsto \mathbf{v}(x)$ such that when one projects $T(M)$ back to M each $\mathbf{v}(x)$ goes back to x; this is equivalent to the statement that $\mathbf{v}(x) \in T_xM$ for each $x \in M$. One can also define local sections over any $U \subset M$ by the same means. Since the local sections correspond to what we called local vector fields above, we see that a global section of the tangent bundle fibration is a reasonable candidate for a global vector field on M. Now, more general fiber bundles, whose fibers are not always vector spaces, do not have admit global sections, while $T(M) \to M$ always has at least one global section, namely the *zero section* Z: $M \to T(M)$ that takes each $x \in M$ to the origin of T_xM. However, as the example of S^2 shows, one cannot always find a global *non-zero* section of the tangent bundle fibration, since every vector field on S^2 must have at least one zero, a result that is usually described rather colorfully as the "Hairy Ball Theorem," when one thinks of the tangent vectors as hairs.

One can now extend the concept of a tangent frame in T_xM to a *local frame field* on $U \subset M$, namely, a set of n vector fields $\mathbf{e}_i : U \to T(U)$, $x \mapsto \mathbf{e}_i(x)$ that are linearly independent at each x. A local vector field $\mathbf{v} : U \to T(U)$ can then be expressed in terms of this local frame field as:

$$\mathbf{v}(x) = v^i(x)\, \mathbf{e}_i(x), \tag{II.9}$$

in which it is important to see that we are allowing the frame members themselves to vary over the points of U. In fact, if U admits a coordinate system x^i the local frame field can be expressed in the form:

$$\mathbf{e}_i(x) = A_i^j(x)\, \partial_j\,, \tag{II.10}$$

in which the component matrices $A_j^i(x)$ are now differentiable functions on U, and are invertible at each $x \in U$. Among other things, this means that every coordinate chart defines a local frame field by way of ∂_i that one calls the *natural frame field* defined by this chart. However, not every local frame field is the natural frame field for some coordinate chart. This gets one into the difference between holonomic and anholonomic local frame fields, which we shall discuss later in this book.

Although global vector fields always exist on a general manifold, global frame fields do not, in general. In fact, the existence of a global frame field on M is equivalent to the possibility that $T(M)$ admits a global trivialization. When this is possible, one calls M *parallelizable*, because one then has unique linear isomorphisms between each pair of tangent spaces in $T(M)$ that allow one to clarify the meaning of parallelism between tangent vectors at finitely-separated points.

In the case of the 2-sphere, it is clear that any manifold that does not admit a non-zero vector field cannot admit a global frame field, but it is not true that the existence of a non-zero vector field implies the existence of a global frame field. Hence, the existence

of global frame fields is a much stronger condition than the existence of non-zero vector fields [12].

By defining vector fields on M in terms of directional derivative operators, we introduce the possibility of composing the action of two vector fields **X** and **Y** on a smooth function f on M, such as **XY**f, which has the local form:

$$\mathbf{XY}f = X^i \frac{\partial}{\partial x^i}\left(Y^j \frac{\partial f}{\partial x^j} \right) = X^i \frac{\partial Y^j}{\partial x^i}\frac{\partial f}{\partial x^j} + X^i Y^j \frac{\partial^2 f}{\partial x^i \partial x^j}. \tag{II.11}$$

However, this construction does not have the coordinate-invariant (or really, *frame*-invariant) character that we desire for objects that are defined on manifolds. Since it is the second term in the final expression that is undesirable, we find that in order to obtain a frame-invariant object, we need only to subtract **YX**f and obtain:

$$[\mathbf{X}, \mathbf{Y}]\, f = \mathbf{XY}f - \mathbf{YX}f = \left(X^i \frac{\partial Y^j}{\partial x^i} - Y^i \frac{\partial X^j}{\partial x^i} \right)\frac{\partial f}{\partial x^j}, \tag{II.12}$$

so we can define the vector field:

$$[\mathbf{X}, \mathbf{Y}] = \mathbf{XY} - \mathbf{YX} = \left(X^i \frac{\partial Y^j}{\partial x^i} - Y^i \frac{\partial X^j}{\partial x^i} \right)\frac{\partial}{\partial x^j}, \tag{II.13}$$

which can also be defined globally.

This means that not only is the set $\mathfrak{X}(M)$ of all vector fields on M an infinite-dimensional vector space under pointwise finite linear combinations, but with this bracket, it also becomes an infinite-dimensional Lie algebra. Note that one also has a local Lie algebra $\mathfrak{X}(U)$ for each $U \subset M$. A deep problem in differential topology is that of determining how much detail concerning the differential structure on M can be expressed in terms of the algebraic structure of $\mathfrak{X}(M)$. For instance, not all n-dimensional Lie algebras on $\mathbb{R}^n$ can be represented in $\mathfrak{X}(M)$, or even $\mathfrak{X}(U)$. In particular, the Abelian Lie algebra on $\mathbb{R}^n$ is often hard to find globally.

A basic property of natural frame fields is that they are *holonomic*:

$$[\partial_i, \partial_j] = 0, \tag{II.14}$$

which follows from the equality of the mixed partial second derivatives. (Simply set $X^i = Y^i = \text{const.}$ in (II.13).)

[12] One can even introduce an intermediate notion of *k-parallelizability*, which means that there are k globally linearly independent vector fields on M, but not k+1. One calls k the *degree of parallelizability* of M. Not surprisingly, the topological nature of this gets quite involved.

e. Covector fields [**4-9**]. There are dual constructions to those of the preceding subsection that have just as much, if not more, application to the problems of physics. One starts by defining a *covector* α_x at $x \in M$ to be a linear functional on T_xM. Since T_xM is presumed to be a vector space, in its own right, it has a dual vector space consisting of all linear functionals on T_xM, and we denote it by T_x^*M ; one calls this vector space the *cotangent space* to M at x. Hence, we can say $\alpha_x \in T_x^*M$, and if $\mathbf{v}_x \in T_xM$ then we write the evaluation of α_x on $\mathbf{v}_x$ as $\alpha_x(\mathbf{v}_x)$.

We can basically jump ahead to define the *cotangent bundle on M* as the disjoint union of all the cotangent spaces, which then has a projection $T^*M \to M$, and is locally trivial by way of local trivializations $T^*U \to U \times \mathbb{R}^{n*}$ over all of the same open subsets that locally trivialize $T(M)$, if one chooses a linear isomorphism of $\mathbb{R}^n$ with its dual; for instance, one could map the standard frame on $\mathbb{R}^n$ to its reciprocal coframe. Similarly, this means that the bundle $T(M)$ is isomorphic to the bundle T^*M, but not canonically so[13]. One also has restrictions $T^*U \to U$ over any open subset $U \subset M$.

Going in the opposite direction, a *local covector field* on U is a section $\alpha: U \to T^*U$, $x \mapsto \alpha_x$ and a *global covector field* on M is a section $\alpha: M \to T^*M$. If U carries a coordinate system x^i then the coordinate differentials dx^i define a *coframe* on $\mathbb{R}^{n*}$ that is reciprocal to the frame defined by the ∂_i :

$$dx^i(\partial_j) = \delta^i_j \, . \tag{II.15}$$

Any local covector field $\alpha: U \to T^*U$ can then be expressed in local form as:

$$\alpha(x) = \alpha_i(x)\, dx^i \, . \tag{II.16}$$

in which the components $\alpha_i(x)$ are presumed to be smooth functions on U.

Global covector fields are defined predictably, and the same caveat that applies to the existence of global non-zero vector fields applies to existence of global non-zero covector fields. A local coframe field $\theta^i : U \to T^*U$ will be reciprocal to a unique local frame field $\mathbf{e}_i$ on U, and if U carries a local coordinate system x^i then one can express θ^i in the form:

$$\theta^i(x) = \tilde{A}^i_j(x)\, dx^j \, , \tag{II.17}$$

where $\tilde{A}^i_j(x)$ is the matrix inverse to $A^i_j(x)$ at each x, and one also has:

$$\mathbf{e}_i(x) = A^i_j(x)\, \partial_i \, . \tag{II.18}$$

---

[13] By "vector bundle isomorphism," we simply mean that the manifolds $T(M)$ and T^*M are diffeomorphic by a diffeomorphism that takes each fiber $T_x(M)$ to the fiber T^*_x linearly.

One might think that the isomorphism of $T(M)$ and T^*M would imply that anything that is true of one bundle is true of the other. However, this shows one the danger of extending equivalence relations beyond their natural limits. For instance, the infinite-dimensional vector space of sections $\alpha: M \to T^*M$, i.e., global covector fields, does not admit a natural Lie algebra structure, as does $\mathfrak{X}(M)$. Conversely, as we shall see, one can define the exterior derivative of a covector field, but not a vector field. Interestingly, these last two statements are weakly related.

f. Vector bundles in general [**4-9**]. Since the fibers of $T(M)$ and T^*M over any point $x \in M$ are both vector spaces, it is useful to generalize to the concept of (differentiable) *vector bundle*, which is a differentiable manifold E that projects onto M in such a manner that the fiber E_x over each $x \in M$ is a vector space whose dimension n – which is assumed constant over M – is called the *rank* of the vector bundle [14]. Furthermore, one assumes that E is locally trivial, which means that each $x \in M$ has some open neighborhood U such that $E(U)$ is diffeomorphic to $U \times \mathbb{R}^n$ in such a way that the projection of $E(U)$ onto U corresponds to the projection of $U \times \mathbb{R}^n$ onto its first factor. These local trivializations also define the atlas of coordinate charts for the manifold E.

Although the mathematics of vector bundles can go off into abstruse, esoteric realms that often seem hopelessly divorced from the problems of physics, the same can be said of group theory; so the challenge to the theoretical physicist is to isolate from the potentially vast set of irrelevant sidetracks that one "critical path" that is most immediately relevant to the problem at hand, which is the spirit of Occam's Razor. For our purposes, the main objective of generalizing to vector bundles is merely to define the bundle of exterior differential forms on M and the bundle of multivector fields on M. Hence, we shall only point out some of the elementary constructions on vector bundles that will give us that much.

Some of the same constructions that one carries out with vector spaces can be applied to vector bundles, as well. In particular, just as one can form direct sums $V \oplus W$ and tensor products $V \otimes W$ of vector spaces, one can form direct – or *Whitney* – sums $E \oplus F$ and tensor products $E \otimes F$ of vector bundles that are both fibered over the same manifold M. All that is basically necessary is to do these operations on each fiber individually; that is, the fiber of $E \oplus F$ over $x \in M$ is $E_x \oplus F_x$, and the fiber of $E \otimes F$ is $E_x \otimes F_x$. Indeed, one can extend these constructions to direct sums of families of vector bundles and tensor products of finite families.

Both of these constructions can be applied to the vector spaces $\Gamma(E)$ and $\Gamma(F)$ of sections of the vector bundles E and F. The result is that $\Gamma(E) \oplus \Gamma(F)$ represents the vector space of sections of $E \oplus F \to M$, and $\Gamma(E) \otimes \Gamma(F)$ represents the vector space of sections of the vector bundle $E \otimes F \to M$.

One can also define the dual E^* to a vector bundle E by looking at all linear functionals on the vectors of E.

[14] A non-constant rank tends to suggest a non-connected base manifold M, which we shall mostly exclude from our considerations.

Some of things that one expects of tangent and cotangent bundles that no longer have any meaning for more general vector bundles are: the interpretation of elements in the fibers as directional derivatives, the existence of a natural Lie algebra on the vector space of sections, and the existence of a naturally-defined exterior derivative operator.

2. Differential forms on manifolds [4-10]. When the basic constructions that we discussed above are applied to the vector bundles $T(M)$ and T^*M, one finds that it is straightforward to define the bundles $\Lambda_k M \to M$ and $\Lambda^k M \to M$, which represent the k^{th} exterior products of the vector bundles $T(M)$ and T^*M, respectively. Hence, the fibers of these vector bundles are the vector spaces that one obtains by taking the k^{th} exterior power of the vector spaces T_xM and T_x^*M at each $x \in M$.

A section of the bundle $\Lambda_k M \to M$ is called a *k-vector field* on M and a section of its dual $\Lambda^k M \to M$ is called a *differential k-form*. At each point $x \in M$, the value of such a field will be a k-vector in $\Lambda_{k,x}$ or an algebraic k-form in Λ_x^k, respectively. Hence, one can think of a non-zero k-vector field on M as associating k linearly independent vectors in the tangent space to each point (at least in the simple case), and a differential k-form associates a completely anti-symmetric k-linear functional on the tangent vectors at x.

The local forms of k-vector fields $\mathbf{A}$ and k-forms α over $U \subset M$ are:

$$\mathbf{A}(x) = \frac{1}{k!} A^{ij\cdots m}(x)\, \partial_i \wedge \partial_j \wedge \cdots \partial_m, \qquad \alpha(x) = \frac{1}{k!} \alpha_{ij\cdots m}(x)\, dx^i \wedge dx^j \wedge \cdots \wedge dx^m \tag{II.19}$$

when U carries a coordinate system x^i and, more generally:

$$\mathbf{A}(x) = \frac{1}{k!} A^{ij\ldots m}(x)\, \mathbf{e}_i \wedge \mathbf{e}_j \wedge \ldots \wedge \mathbf{e}_m, \quad \alpha(x) = \frac{1}{k!} \alpha_{ij\ldots m}(x)\, \theta^i \wedge \theta^j \wedge \ldots \wedge \theta^m \tag{II.20}$$

when U carries a local frame field $\mathbf{e}_i$ and its reciprocal coframe field θ^i. (Of course, the component functions will not generally be the same in these two cases.)

One can define bilinear pairings of k-forms and l-vector fields for the cases of $k < l$, $k = l$ and $k > l$. However, one must be careful in the case of $k = l$ to notice that although the evaluation of an algebraic k-form on a k-vector is a *number*, nonetheless, when this is going on at each point of a manifold, the evaluation of a differential k-form on a k-vector field gives a *smooth function* on M.

The basic rules for interior products that were discussed in Chapter I still apply with only a new interpretation. For instance:

$$i_{\mathbf{v}}(\alpha_1 \wedge \ldots \wedge \alpha_k) = \sum_{i=1}^{k} (-1)^{i+1} \alpha_i(\mathbf{v})\, \alpha_1 \wedge \cdots \wedge \hat{\alpha}_i \wedge \cdots \wedge \alpha_k, \tag{II.21a}$$

$$i_{\alpha}(\mathbf{v}_1 \wedge \ldots \wedge \mathbf{v}_k) = \sum_{i=1}^{k} (-1)^{i+1} \alpha(\mathbf{v}_i)\, \mathbf{v}_1 \wedge \cdots \wedge \hat{\mathbf{v}}_i \wedge \cdots \wedge \mathbf{v}_k. \tag{II.21b}$$

The exterior derivative operator that we previously defined in Chapter 1 can be extended to a linear anti-derivation d: $\Lambda^k \to \Lambda^{k+1}$ that not only agrees with df on smooth functions, but also agrees with the exterior derivative operator that we defined in Chapter I for local k-forms over open subsets $U \subset M$ that carry coordinate systems or local coframe fields. Hence, if $\alpha \in \Lambda^k$ and $\beta \in \Lambda^l$ then:

$$d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^k \alpha \wedge d\beta . \tag{II.22}$$

Sometimes it useful to have a way of evaluating exterior derivatives without having to introduce local coframe fields and components. For that purpose, one has the intrinsic formula for the exterior derivative. Although it can be generalized (see [**1**]), we shall present it only for 1-forms. Basically, one must evaluate the 2-form $d\alpha$ on an arbitrary pair of vector fields **X**, **Y** and obtain:

$$d\alpha(\mathbf{X}, \mathbf{Y}) = \mathbf{X}(\alpha(\mathbf{Y})) - \mathbf{Y}(\alpha(\mathbf{X}) - \alpha([\mathbf{X}, \mathbf{Y}]). \tag{II.23}$$

Here, we see our first inkling of the relationship between Lie brackets of vector fields and exterior derivatives of 1-forms.

The Lie derivative of a k-vector field **A** or a k-form α with respect to a vector field **v** still works as it did in Chapter I, as well. One has, for instance:

$$\mathrm{L}_{\mathbf{v}}(\mathbf{X} \wedge \mathbf{Y}) = \mathrm{L}_{\mathbf{v}}\mathbf{X} \wedge \mathbf{Y} + \mathbf{X} \wedge \mathrm{L}_{\mathbf{v}}\mathbf{Y} = [\mathbf{v}, \mathbf{X}] \wedge \mathbf{Y} + \mathbf{X} \wedge [\mathbf{v}, \mathbf{Y}], \tag{II.24a}$$

$$\mathrm{L}_{\mathbf{v}}\alpha = (i_{\mathbf{v}}d + di_{\mathbf{v}})\, \alpha . \tag{II.24b}$$

3. Differentiable singular cubic chains [11-13]. In order to define the integration of differential forms on differentiable manifolds, one basically has to show how the integrals on manifolds can be converted – viz., "pulled back" – to conventional multiple integrals over regions in $\mathbb{R}^n$ in a coordinate-invariant manner. Hence, it helps to be mapping regions of $\mathbb{R}^n$ into the manifold in question in a differentiable manner that permits such a pull-back.

A particularly useful class of differentiable mappings of regions in $\mathbb{R}^n$ into a manifold M that is adapted to this purpose is that of *differentiable singular cubic chains*. These objects also have bonus that they define elementary building blocks for the topology of the manifold, as well.

To begin with, a *k-cube* $I^k \subset \mathbb{R}^k$ is simply any region of $\mathbb{R}^k$ that is homeomorphic to the k-fold Cartesian product $[0, 1] \times \ldots \times [0, 1]$. Note that this means that a k-cube could just as well be a closed k-ball or a k-simplex, which is the k-dimensional generalization of a triangle or a tetrahedron. Of course, since we are going to be dealing with objects in the differentiable category, it is important to recall that the corners and edges of a cube prevent it from being *diffeomorphic* to a ball.

A *singular cubic k-simplex* in a topological space M is a continuous map σ_k: $I^k \to M$. Note that since we did not require that it be a homeomorphism onto – i.e., a topological

embedding of a k-cube – there is nothing to say that the image of I^k is still k-dimensional; indeed, it could very well be merely a point. This is why one calls such simplexes [15] "singular." As it turns out the effect of such degenerate cases eventually disappears when one goes on to homology. A *differentiable singular cubic k-simplex* is then a singular cubic k-simplex for which the defining map is differentiable. Since the k-cube is not really a differentiable manifold, in the first place, the way that one addresses this detail is to extend σ_k to a differentiable map on *any* open neighborhood of I^k in $\mathbb{R}^k$. Because differentiation is a purely local process, the differentials of all of the extensions must agree everywhere on I^k, as well as the functions themselves.

Once again, differentiable k-cubes, which is how we will abbreviate the more cumbersome expression, do not have to be diffeomorphisms onto. Also, one finds that in the eyes of homotopy theory, differentiability is no severe restriction, since a "smoothing" argument shows that any homotopy class $[\sigma_k]$ of continuous maps of I^k into M contains a differentiable representative.

A *differentiable singular cubic k-chain*, or *differentiable k-chain*, for short, is a formal linear combination $\sum_{m=1}^{N} \alpha_i \sigma_i$ of a finite number of k-cubes σ_i with real coefficients α_i . We should point out that singular homology (see, [**11, 12**]) usually starts with coefficients in more general rings than $\mathbb{R}$, such $\mathbb{Z}$, and the resulting homology has more detail than what we will be considering, but we are using real coefficients to be consistent with the de Rham cohomology that we shall discuss in a later section. If one is justifiably suspicious of all "formal" constructions as having no rigorous basis, then be assured that these formal linear combinations can be easily made rigorous. All one needs to do is define the "free $\mathbb{R}$-vector space" $C_k(M; \mathbb{R})$ that has the set of all k-cubes in M for its basis. Suffice it to say the main consequence of this construction is that one can regard the set of all k-cubes in M as the basis for an uncountably-infinite dimensional vector space. The miracle of homology is that one can usually reduce this uncountable infinitude to a finite set of equivalence classes that pertains to the topology of M.

Since non-zero real numbers have signs, one can think of a k-cube as being oriented by the sign of its coefficient. The actual numerical value of the coefficient should be regarded as simply a scalar that one associates with the k-cube, like a "charge."

The nature of "free" constructions in mathematics, such as free groups, rings, modules, etc., is that the only role that the set of generators plays is due to solely to its cardinality. That is, the free $\mathbb{R}$-vector space over a set of three apples is no different from the one over a set of three oranges; they are both linearly isomorphic to $\mathbb{R}^3$. In particular,

[15] Although in the early days of homology everyone was content to form the plural of "simplex" just as they formed the plural of "complex" – with an "es" – sometime in the 1960's mathematicians apparently had second thoughts because "vertex" forms its plural with "ices," and took to saying "simplices;" note that the spelling checker on most word processors does not approve of that construction. One could say it is really just a matter of personal taste, but that is also the sort of attitude towards the English language that gave us "Canterbury Tales."

the topology of M does not affect the free $\mathbb{R}$-vector space of all k-chains in M, at this point. The way that topology enters the picture is by way of defining a *boundary map* for each k-chain in $C_k(M; \mathbb{R})$. This means that one is regarding the k-cubes in a given k-chain as being "glued" to each other on their mutual boundaries in some way.

In general, the boundary map is a linear map ∂: $C_k(M; \mathbb{R}) \to C_{k-1}(M; \mathbb{R})$ with the property that $\partial^2 = 0$. Since one assumes linearity, it is sufficient to define the effect of the boundary operator on k-cubes. By definition, one regards any 0-cube – i.e., any point in M – as having boundary zero. The boundary of a 1-cube – i.e., an oriented path AB in M – is $\partial(AB) = B - A$. If one represents a 2-cube in M by the images of its vertices as $ABCD$ then one sees that its boundary is the 1-chain:

$$\partial(ABCD) = AB + BC + CD + DA. \tag{II.25}$$

A second application of ∂ to this 1-chain gives:

$$\partial^2(ABCD) = (B - A) + (C - B) + (D - C) + (A - D) = 0, \tag{II.26}$$

as expected. We illustrate this situation in Fig. 1.

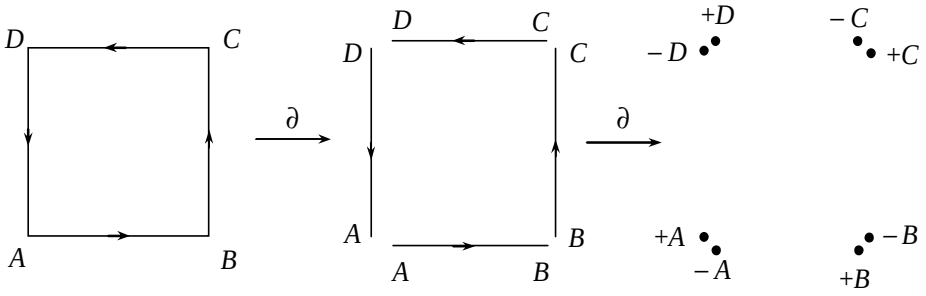


Figure 1. The boundary of a boundary of a 2-cube.

An elementary example that shows how the boundary operator attaches k-cubes along their boundaries is that of a circle, which we think of as the 1-chain $AB + BA$, whose boundary is then $B - A + A - B = 0$.

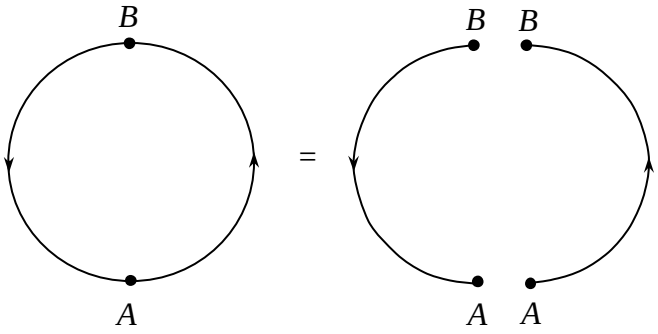


Figure 2. The representation of a circle as a 1-chain with boundary zero.

However, this means that we must regard the 1-cube BA as basically distinct from $-AB$, since otherwise the 1-chain itself would vanish, as well. We illustrate this situation in Fig. 2.

Many two-dimensional examples can be represented as squares with sides and corners identified by means of the boundary operator. We depict some of them in Fig. 3.

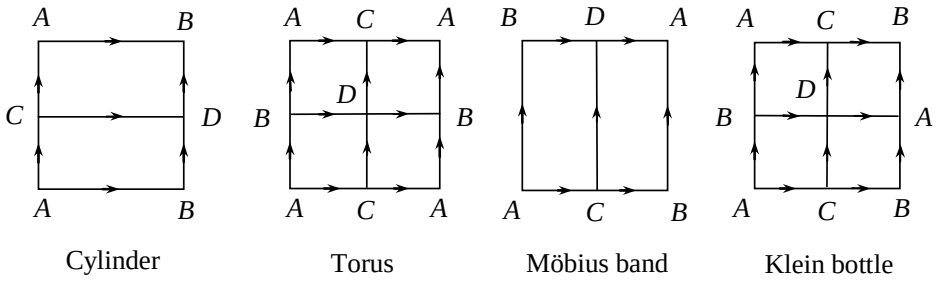


Figure 3. Two-dimensional spaces that are described by 2-chains.

We can represent these four examples by 2-chains c_2 and boundary operators as follows:

Cylinder: $$\begin{aligned}\partial c_2 = {} & CA + AB - BD - CD \\ & + AC - AB - DB + CD \\ = {} & AC + CA - (BD + DB),\end{aligned}$$

which then represents the formal difference of two circles.

Torus: $$\begin{aligned}\partial c_2 = {} & BA + AC - CD - BD \\ & - AB + CA + DC - DB \\ & + AB - AC - DC + BD \\ & - BA - AC + CD + DB \\ = {} & 0,\end{aligned}$$

which is what one would expect for a torus. This time, one sees that the 1-cycles $AC + CA$ and $AB + BA$ do not bound any 2-chains, nor does their difference.

Möbius band: $$\begin{aligned}\partial c_2 = {} & AB - BD - CD - CA \\ & - AB + CD + DA - BC \\ = {} & BD + DA - (CA + BC).\end{aligned}$$

If one replaces $BD + DA$ with the homologous arc BA, while replacing $CA + BC$ with $-AB$ then one sees that the boundary of a Möbius band is a circle, which is completed twice.

Klein bottle: $$\begin{aligned}\partial c_2 = {} & BA + AC - CD - BD \\ & - AB + DC + CB - DA \\ & + AB - AC - DC + BD \\ & - BA + CD - CB + DA \\ = {} & 0\end{aligned}$$

One can think of a Klein bottle as a twisted cylinder with its ends identified in the same way that a Möbius band is a twisted rectangle with its ends identified.

It is important to understand that, in some sense, the boundary operator must be defined "by hand" to reflect the topological nature of the manifold M and that it is not, by any means, canonically defined by the bases for the vector spaces $C_k(M; \mathbb{R})$.

We have now introduced the fact that there are three different types of k-chains c_k as far as the boundary operator is concerned: the ones for which ∂c_k is non-vanishing, the ones for which ∂c_k is vanishing, and the ones for which $c_k = \partial c_{k+1}$ for some k+1-chain c_{k+1}. Since $\partial^2 = 0$, the last possibility implies the second-to-last one. The first type of k-chain has no special name, but those of the second type are called *k-cycles*, and those of the third type are called *k-boundaries*. Since ∂: $C_x(M; \mathbb{R}) \to C_{k+1}(M; \mathbb{R})$ is linear, its image is a vector subspace of $C_{k+1}(M; \mathbb{R})$ that represents all k+1-boundaries, and we denote it by $B_{k+1}(M; \mathbb{R})$. Similarly, its kernel is a vector subspace of $C_x(M; \mathbb{R})$ that consists of all k-cycles, and we denote it by $Z_k(M; \mathbb{R})$.

The power of homology theory to define elementary building blocks for the topology of M is that even though both $Z_k(M; \mathbb{R})$ and $B_k(M; \mathbb{R})$ are generally infinite-dimensional vector spaces, nonetheless, their quotient vector space [16] $H_k(M; \mathbb{R}) = Z_k(M; \mathbb{R}) / B_k(M; \mathbb{R})$ is often finite-dimensional; indeed, it is often 0-dimensional. One calls the vector space $H_k(M; \mathbb{R})$ the *k^{th} homology space of M*. Ordinarily, one uses coefficients in a more general ring and obtains homology modules, but, once again, we shall be mostly concerned with the field of real coefficients, for the sake of de Rham cohomology. In order for a homology space to vanish in dimension k all of the k-cycles must bound k+1-cycles. A convenient way of regarding the basis elements for the homology space $H_k(M; \mathbb{R})$ is to think of them as essentially "k-dimensional holes" in M, such as circles that do not bound disks and spheres that do not bound balls.

The equivalence relation of homology is actually more tangible than it might seem. One sees that two k-cycles z_k and z_k' are *homologous* iff their difference is the boundary of a k-chain c_{k+1}: $z_k - z_k' = \partial c_{k+1}$. As we saw above, the boundary of a cylinder is a pair of circles. Hence, since circles are 1-cycles, one can say that the two boundary components of a cylinder are homologous, and the 2-chain that effects this homology is the cylinder itself. When the boundary of a k-chain has only one component, such as the boundary of a disk, one says that this component is *homologous to zero*. Often, physics uses a more general homology to describe this sort of situation that one calls *cobordism*. The main difference at the elementary level is that one does not assume a triangulation of the differentiable manifold whose boundary components are being connected together, but

[16] Recall that the quotient A/B of two vector spaces $B \subset A$ is the set of all translates of the subspace B by the vectors in A. One can also say that its elements are equivalence classes $[a]$ of vectors $a_1, a_2 \in A$ whose difference $a_1 - a_2$ is some vector $b \in B$.

the effect in the long run is that generally much less is known about the structure of the cobordism ring of compact submanifolds of a given manifold than is known about its singular homology. Generally, the power of cobordism is more necessary when one is dealing with purely mathematical aspects of differentiable manifolds, such as defining equivalence classes of them and classifying them.

One of the advantages of using cubes for ones basic topological objects is that a homotopy of a singular k-cube is itself a singular k+1-cube. This follows from the fact that, by definition, if σ_k: $I^k \to M$ is a singular k-cube in a topological space M then a homotopy of σ_k is a continuous map F: $I^k \times I \to M$, that is, a continuous map F: $I^{k+1} \to M$, which is then simply a singular k+1-cube in M. The maps $F(x, 0)$ and $F(x, 1)$ then represent two opposite k-faces of the k+1-cube F.

As we saw above, a 1-cycle can take the form of a loop, up to homeomorphism. If a loop bounds a 2-chain, which is then homeomorphic to a two-dimensional disk, then the boundary is homotopic to any point of the disk. Conversely, if no such disk exists then the boundary loop cannot be contracted to any point in M. Hence, one sees that there is clearly a close relationship between homology and homotopy, at least in dimension one. In fact, they are related in all dimensions, but not as conveniently as one might wish. The closest that one can come, at present, is Hurwitz's theorem, which says, for our purposes, that the first non-zero homotopy group – say $\pi_k(M)$ – has as many generators as the dimension of $H_k(M; \mathbb{R})$; for $k = 1$, one must specify that one is looking at generators of the "Abelianization" of $\pi_1(M)$, if it is not already Abelian. For all dimensions above that dimension, one only has a possibly many-to-one map of generators for $\pi_m(M)$ to basis vectors for $H_m(M; \mathbb{R})$.

For instance, as Fig. 2 shows, the homology of the circle S^1 is $H_0(M; \mathbb{R}) = \mathbb{R}$, $H_1(M; \mathbb{R}) = \mathbb{R}$, $H_k(M; \mathbb{R}) = 0$, for $k > 1$. For any contractible topological space – whose homotopy groups also vanish – all homology vector spaces vanish, as well. Such topological spaces are sometimes called *acyclic*.

It is always true that if the dimension of a manifold M is n then $H_m(M; \mathbb{R}) = 0$ for all $m > n$.

As an example of one of the limitations to doing homology with coefficients in a field such as $\mathbb{R}$, consider the difference between the 2-sphere S^2 and the two-dimensional projective space $\mathbb{R}\mathrm{P}^2$. Whereas S^2 is simply-connected, $\mathbb{R}\mathrm{P}^2$ has a fundamental group that is isomorphic to $\mathbb{Z}_2$, which comes from the two-to-one projection of S^2 onto $\mathbb{R}\mathrm{P}^2$. Had we considered homology with integer coefficients, as one usually does in singular homology, we would have found that $H_1(S^2) = 0$, but $H_1(\mathbb{R}\mathrm{P}^2) = \mathbb{Z}_2$. However, the use of coefficients in a field does not permit the appearance of "finite cyclic summands" in the homology modules, and we then find that $H_1(S^2; \mathbb{R}) = H_1(\mathbb{R}\mathrm{P}^2; \mathbb{R}) = 0$. Therefore, one sees that there is something topologically "incomplete" about looking at homology with real coefficients.

4. Integration of differential forms [4-10]. Before going on to a discussion of the duals of the homology spaces, we first return to the more elementary business of integrating differential forms over manifolds.

In effect, we shall be restricting ourselves to manifolds that can be "triangulated," i.e., ones that are homologically equivalent to finite chain complexes with suitably-defined boundary operators. In fact, this is not much of a restriction, since it can be shown (cf., [**14**]) that any compact differentiable manifold can be triangulated.

One first notes that the only things that can be integrated over an n-dimensional compact manifold M are n-forms, which must, moreover, take the form $f\mathcal{V}_n$, where $\mathcal{V}_n$ is a globally non-zero volume element on M. Of course, this presupposes that such a volume element actually exists, which means that first we must address the issue of orientability for differentiable manifolds.

The fibers of $\Lambda^n(M) \to M$ are each one-dimensional, so if one takes away the zero section from $\Lambda^n(M)$ then what will be left looks like the disjoint union $(-\infty, 0) \cup (0, +\infty)$ at each point of M. Up to homotopy, one can contract each such union to a pair of disjoint points, which we denote by $\{-, +\}$. This defines a fibration $\mathcal{O}(M) \to M$ whose fibers all look like $\{-, +\}$, but it does not have to be trivial. In fact, it is trivial iff it admits a global section, which we will then call an *orientation* for M. (Really, it is an orientation for $T(M)$.)

Not all manifolds are orientable. For instance, the aforementioned Klein bottle and Möbius band, as well as all of the projective spaces, are all non-orientable. Since all of them have $\mathbb{Z}_2$ for a fundamental group, one suspects that this is relevant; it is, but we shall not go into the matter of "Stiefel-Whitney classes" at the moment. All simply-connected manifolds are orientable, but not all orientable manifolds are simply-connected (counterexample: torus). In any event, $\mathcal{O}(M)$ itself is always an orientable manifold, and one refers to it as the *orientable covering manifold* of M.

Actually, although orientability of a manifold is necessary and sufficient for the existence of a volume element, nonetheless, the choice of volume element is still open to one's discretion. Hence, one has to specify that a manifold M be not only orientable, but also oriented, and a choice of volume element has made before one can define integrals on it.

Having made such restrictions on a compact manifold M, as well as a choice of triangulation, $M = c_n = \sum \alpha_i \sigma_i$ one then defines the integral of an n-form $f\mathcal{V}_n$ over M by assuming that integration is linear with respect to linear combinations of n-chains:

$$\int_M f\mathcal{V}_n = \sum \alpha_i \int_{\sigma_i} f\mathcal{V}_n \,. \tag{II.27}$$

One then defines the integral over any basis simplex σ_i: $I^n \to M$ by pulling back the integral over its image in M to an integral over I^n:

$$\int_{\sigma_i} f\mathcal{V}_n = \int_0^1 \cdots \int_0^1 f(\sigma_i) J(\sigma_i)\, dx^1 \cdots dx^n \,, \tag{II.28}$$

in which $J(\sigma_i)$ is the Jacobian of the map σ_i. Hence, in order for this to be non-vanishing, one must consider only the *non-singular* n-simplexes; i.e., the ones that are diffeomorphisms onto their images, which makes them *embeddings*.

Stokes's theorem can be generalized to the integration of n-forms over n-chains in a manner that has deep topological significance, as we shall see in the next section. If $\alpha \in \Lambda^n(M)$ and $M = \partial c_{n+1}$ is an n-chain then Stokes's theorem says that:

$$\int_{\partial c_{n+1}} \alpha = \int_{c_{n+1}} d\alpha \ . \tag{II.29}$$

We introduce the following bracket notation for the integration of an n-form α on an n-chain c_n :

$$<\alpha, c_n> = \int_{c_n} \alpha . \tag{II.30}$$

With this notation, Stokes's theorem takes the form:

$$<\alpha, \partial c_{n+1}> = <d\alpha, c_{n+1}>. \tag{II.31}$$

In this form, Stokes's theorem appears to be asserting that in some sense the exterior derivative operator on differential forms is "adjoint" to the boundary operator on chains. In the next section, we shall clarify the extent to which this is indeed the case.

An important consequence of Stokes's theorem is that if z_k and z_k' are homologous k-cycles then the integral of a closed k-form α is the same over both z_k and z_k'. Abstractly, this only says that the value of the integral $<\alpha, z_k>$ can be expressed as $<\alpha, [z_k]>$.

At the more practical level of physics applications, this invariant pairing of closed k-forms and k-dimensional homology classes by way of integrals serves as the basis for certain *integral invariants* when the homology of two k-cycles comes about in less abstract ways. For instance, the interpolating k+1-chain c_{k+1} – that is, one that makes $z_k - z_k' = \partial c_{k+1}$ – might take the form of a differentiable homotopy of z_k to z_k', or even a one-parameter family of diffeomorphisms that represent the flow of a vector field on c_{k+1}.

5. De Rham's theorem [5, 10]. From (II.30), we see that for a fixed $\alpha \in \Lambda^k$ the bracket $<\alpha, c_k>$ defines a linear functional on the vector space $C_k(M; \mathbb{R})$. If we denote the dual vector space to $C_k(M; \mathbb{R})$ by $C^k(M; \mathbb{R})$ then we can say that $<\alpha, \ . \ > \in C^k(M; \mathbb{R})$. Since the elements of $C_k(M; \mathbb{R})$ are called k-chains, we call the elements of $C^k(M; \mathbb{R})$ *cochains.*

We can define a *coboundary* operator on cochains that is adjoint to the boundary operator under the bilinear pairing $(c^k, c_k) = c^k(c_k)$ of k-cochains with k-chains that amounts to the evaluation of a linear functional on a vector:

$$(\delta c^k, c_{k+1}) \equiv (c^k, \partial c_{k+1}). \tag{II.32}$$

Note the resemblance to (II.31).

Hence, from the way that we defined things the coboundary operator δ: $C^k(M; \mathbb{R}) \rightarrow C^{k+1}(M; \mathbb{R})$ is a linear operator with the property that $\delta^2 = 0$. This means that most of what we said in the context of chains applies just as well to cochains. For instance, there are three types of cochains in the eyes of δ: cochains for which $\delta c^k \neq 0$, those for which $\delta c^k = 0$, and those for which there is a k–1-cochain c^{k-1} such that $c^k = \delta c^{k-1}$. The first type has no special name, but the second type is called a *cocycle*, while the last type is called a *coboundary*. We denote the vector space of all k-cocycles by $Z^k(M; \mathbb{R})$ and the vector space of all k-coboundaries by $B^k(M; \mathbb{R})$; the former space is the kernel of δ: $C^k(M; \mathbb{R}) \rightarrow C^{k+1}(M; \mathbb{R})$, while the latter is the image of δ: $C^{k-1}(M; \mathbb{R}) \rightarrow C^k(M; \mathbb{R})$.

Just as we defined the real singular homology vector spaces by quotients of spaces of cycles by spaces of boundaries, we now do the same thing with cocycles and coboundaries. We call the vector space $H^k(M; \mathbb{R}) = Z^k(M; \mathbb{R}) / B^k(M; \mathbb{R})$ the *real singular cohomology vector space* in dimension k. One of the advantages of using real coefficients that partially compensates for the fact that we cannot consider the "torsion" summands in our homology or cohomology modules – i.e., finite cyclic Abelian groups – is the fact that one does indeed have that the vector space $H^k(M; \mathbb{R})$ is the dual to the vector space $H_k(M; \mathbb{R})$; in particular, they have the same dimension, so if one of them vanishes then the other one does.

Now, since the exterior derivative operator d: $\Lambda^k \rightarrow \Lambda^{k+1}$ works in a manner that is clearly analogous to the coboundary operator, we see that we can define vector spaces $Z^k_{dR}(M)$, $B^k_{dR}(M)$, and $H^k_{dR}(M) = Z^k_{dR}(M) / B^k_{dR}(M)$ by means of the kernel, image, and quotient constructions. One calls the vector space $Z^k_{dR}(M)$, the space of closed k-forms, $B^k_{dR}(M)$, the space of exact k-forms, and $H^k_{dR}(M)$, the *de Rham cohomology vector space* in dimension k. A de Rham cohomology class can then be represented by a closed k-form, and two closed k-forms α, α' are *cohomologous* if they differ by an exact k-form: $\alpha - \alpha' = d\beta$, for some k–1-form β. One must note that β is not unique, since one can add any closed k–1-form γ to it and still produce the same exterior derivative: $d(\beta + \gamma) = d\beta$.

Since a de Rham cohomology class $[\alpha] \in H^k_{dR}(M)$ can be represented by a closed k-form α and a k-form defines a singular cochain by way of $<\alpha, .>$, we see that we have a linear map from $Z^k_{dR}(M)$ to $Z^k(M; \mathbb{R})$. One finds that this map also "descends to cohomology" since it commutes with the action of the exterior derivative and coboundary operators. Hence, there is a corresponding linear map of $H^k_{dR}(M)$ to $H^k(M; \mathbb{R})$. It was the profound and far-reaching contribution of Georges de Rham that this map could be shown to be an isomorphism. Basically, it meant that the otherwise analytic issue of whether a closed k-form was also an exact k-form, which amounts to a question of

integrability for the operator d, was rooted in purely topological matters, such as whether the manifold M has "holes" in dimension k.

In dimension one this usually takes the form of multiple connectedness. Hence, one finds many classical references to the difference between "single-valued" potentials and "many-valued" potentials in the context of vector calculus, such as in electromagnetism or hydrodynamics. One should be advised that since $d^2 = 0$ follows from the equality of mixed partial derivatives of functions that are twice continuously differentiable, the introduction of many-valued potentials is one way of getting around that identity by dropping the assumption that the second derivative exists everywhere. Rather than deal with topology by way of closed forms that are not exact, one introduces topology in the form of the singular subset of the many-valued potential ϕ for which $d^2\phi \neq 0$, which would be the points at which $d^2\phi$ has jump discontinuities.

If M is a contractible space, such as any vector space, then any closed k-form will be exact, and this will be true in each dimension k. Since every point of any manifold has a neighborhood U that is homeomorphic to $\mathbb{R}^n$ – hence, contractible – one then sees that $H^k(U; \mathbb{R}) \cong H^k_{dR}(U)$ vanishes in every dimension, and one can say that every closed k-form on M is *locally exact;* this is referred to as the *Poincaré lemma.*

One way in which cohomology differs from homology, despite the linear isomorphism of the spaces, is in the fact that there is a natural ring structure on cohomology that is does not appear in homology, in general. It is easiest to account for the ring structure in terms of de Rham cohomology because the "cup" product $[\alpha] \cup [\beta]$ of a cohomology class $[\alpha]$ in dimension k with another one $[\beta]$ in dimension l is simply the $k+l$-dimensional cohomology class $[\alpha \wedge \beta]$. In order to show that this definition does not depend upon the choice of closed k-form α and closed l-form β, one actually only needs to show that $\alpha \wedge \beta$ is also closed; of course, this is an elementary calculation.

Now, suppose that M is orientable and given a choice of volume element $\mathcal{V} \in \Lambda^n$. We have already seen that this implies isomorphisms of the vector spaces A_k and A^{n-k}, which define the fibers of the vector bundles in question. Hence, there is an isomorphism of the vector bundles themselves #: $\Lambda_k \to \Lambda^{n-k}$, $\mathbf{A} \mapsto i_{\mathbf{A}}\mathcal{V}$ that one calls *Poincaré duality*, as well.

In the singular homology of orientable manifolds (see [**11, 12**]), the phrase "Poincaré duality" generally refers to a set of isomorphisms of the homology modules $H_k(M; R)$ with the cohomology modules $H^{n-k}(M; R)$, where R is a more general coefficient ring. We could use de Rham's theorem to replace $H^{n-k}(M; R)$ with $H^{n-k}_{dR}(M)$ for the case where $R = \mathbb{R}$, but we find that on orientable manifolds it is also possible to define homology vector spaces that are directly dual to the de Rham cohomology spaces, and which make Poincaré duality even more straightforward.

We simply define our *de Rham k-chains* [17] to be k-vector fields and our boundary operator to be the divergence operator δ: $\Lambda_k \to \Lambda_{k-1}$, which still takes the same form $\delta =$

[17] Strictly speaking, de Rham did not define the homology that was dual to his cohomology in terms of multivector fields and the divergence operator, but in terms of *currents*, which amount to continuous linear functionals on k-forms – i.e., k-form distributions. However, since the structure of the homology that we

$\#^{-1} \cdot d \cdot \#$, as it did in the previous chapter. (From now on, we drop the use of δ to mean the coboundary operator in singular cohomology.) The kernel of δ, which we denote by $Z_k^{dR}(M)$, consists of divergenceless k-vector fields, which play the role of de *Rham k-cycles*. The image of δ: $\Lambda_{k-1} \to \Lambda_k$, which we denote by $B_k^{dR}(M)$, then plays the role of *de Rham k-boundaries.* The quotient vector space $H_k^{dR}(M)$ is then the *de Rham homology* vector space in dimension k, and is isomorphic to $H_{dR}^k(M)$, but not canonically, and is therefore also isomorphic to the real singular homology vector space $H_k(M; \mathbb{R})$.

One sees that Poincaré duality $\#: H_k^{dR}(M) \to H_{dR}^{n-k}(M)$ follows naturally from the way that we defined δ, since it implies that:

$$d\# = \#\delta. \tag{II.33}$$

Hence, # takes de Rham k-cycles to n–k-cocycles and k-boundaries to n–k-boundaries, which allows one conclude that the isomorphism $\#: C_k^{dR}(M) \to C_{dR}^{n-k}(M)$ "descends to homology."

One finds that there is another intriguing difference between homology and cohomology in the fact that there is a Lie algebra structure defined over vector fields that is not defined over covector fields. In fact, the Lie bracket of divergenceless vector fields is also divergenceless, so one can define the Lie bracket of de Rham homology classes in dimension one. It is even possible to extend the Lie bracket to include k-vector fields for $k > 1$, but since the divergence operator is not an anti-derivation on k-vector fields, it is more difficult to establish whether the Lie bracket of divergenceless k-vector fields is still divergenceless. Furthermore, the possible role that topology might play in a Lie algebra structure on the homology spaces would have to be explored.

6. Hodge theory [4, 15]. In the study of Riemannian manifolds, one has another isomorphism of k-vector fields and k-forms at one's disposal. It is rooted in the fact that a Riemannian metric on a differentiable manifold M, which is a positive-definite symmetric non-degenerate second-rank covariant tensor field g on M, defines an isomorphism $i_g : T(M) \to T^*M$, $\mathbf{v} \mapsto i_g\mathbf{v}$, of the tangent bundle with the cotangent bundle. In this definition, we intend that the covector field $i_g\mathbf{v}$ is defined by:

$$(i_g\mathbf{v})(\mathbf{w}) = g(\mathbf{v}, \mathbf{w}) \tag{II.34}$$

for any tangent vector $\mathbf{w}$ on M.

This isomorphism of bundles defines a corresponding isomorphism of the vector space $\mathfrak{X}(M) = \Lambda_1(M)$ of vector fields with the space $\Lambda^1(M)$ of covector fields (i.e., 1-forms). By tensor product, and consequently, exterior product, one can extend this to an isomorphism $i_g \wedge \ldots \wedge i_g : \Lambda_k \to \Lambda^k$ of k-vector fields and k-forms.

---

are defining is clearly derived from corresponding concepts in de Rham cohomology, we shall make a minor abuse of terminology.

If $g = g_{\mu\nu}\, dx^\mu\, dx^\nu$ in some local coordinate system (U, x^μ) then:

$$i_g \mathbf{v} = (g_{\mu\nu} v^\nu)\, dx^\mu = v_\mu\, dx^\mu. \tag{II.35}$$

Hence, the isomorphism i_g amounts to what is usually called "lowering the index" in this case, as well as in its extension to k-vector fields, which then amounts to lowering all indices. In particular, for bivector fields, we have:

$$(i_g \wedge i_g)\mathbf{A} = \tfrac{1}{2} A_{\mu\nu}\, dx^\mu \wedge dx^\nu = \tfrac{1}{2}(g_{\mu\alpha} g_{\nu\beta} - g_{\mu\beta} g_{\nu\alpha})\, A^{\alpha\beta}\, dx^\mu \wedge dx^\nu . \tag{II.36}$$

Therefore, we can think of the linear map $i_g \wedge i_g$ as having the matrix:

$$[i_g \wedge i_g]_{\mu\nu\alpha\beta} = g_{\mu\alpha} g_{\nu\beta} - g_{\mu\beta} g_{\nu\alpha}, \tag{II.37}$$

relative to this choice of local coframe field.

Understanding this isomorphism is crucial to all of what follows, since the whole point of pre-metric electromagnetism is to replace this isomorphism with one that is defined not by the spacetime metric, but by the electromagnetic constitutive laws of the medium.

When one composes the inverse $[i_g \wedge \ldots \wedge i_g]^{-1}$: $\Lambda^k \to \Lambda_k$ of the metric isomorphism with the Poincaré isomorphism #: $\Lambda_k \to \Lambda^{n-k}$ one gets a linear isomorphism *: $\Lambda^k \to \Lambda^{n-k}$:

$$* = \# \cdot [i_g \wedge \ldots \wedge i_g]^{-1} \tag{II.38}$$

that one refers to as *Hodge duality*.

Since * is defined for every k from 0 to n one can take its square and obtain:

$$*^2 = (-1)^{k(n-k)}\, I. \tag{II.39}$$

Of particular interest is the middle dimension k in an even-dimensional space, for which $*^2 = I$, so the eigenvalues of * in this case are +1 and −1. This allows one to decompose Λ^k into a direct sum of eigen-bundles that one calls the bundles of *self-dual* and *anti-self-dual* k-forms, respectively. However, we shall be generalizing from the case of a *Lorentzian g*, which is not positive-definite, and whose eigenvalues in the case of 2-forms on a four-dimensional manifold are imaginary, not real. A decomposition of Λ^2 into essentially "real" and "imaginary" sub-bundles is not canonical, and must be specified, much as one chooses a frame for a vector space.

The metric isomorphism allows one to map the divergence operator on k-vector fields over to a *codifferential* operator δ: $\Lambda^k \to \Lambda^{k-1}$ on k-forms:

$$\delta = i_g \cdot \delta \cdot i_g^{-1} = *^{-1} d * = (-1)^{n(k+1)} * d * \tag{II.40}$$

Clearly, one has $\delta^2 = 0$, as a consequence of the fact that $d^2 = 0$. A k-form α for which $\delta\alpha$ vanishes is called *co-closed* and one for which there is a k+1-form β such that $\alpha = \delta\beta$ is called *co-exact*.

Using the operators d and δ, one can form a second-order differential operator Δ: $\Lambda^k \to \Lambda^k$: on k-forms:

$$\Delta = d\delta + \delta d \tag{II.41}$$

that one calls the *Laplacian operator* for the Riemannian manifold (M, g).

A k-form α for which $\Delta\alpha = 0$ is called *harmonic*. Clearly, it is sufficient that a harmonic k-form be both closed and co-closed:

$$d\alpha = 0, \qquad \delta\alpha = 0. \tag{II.42}$$

In fact, if M does not have a boundary then this is also necessary.

This follows from the fact that if $\alpha, \beta \in \Lambda^k$ then:

$$\alpha \wedge *\beta = \beta \wedge *\alpha \tag{II.43}$$

is always an n-form, and if we assume that M is compact then we can define the following symmetric, bilinear functional on k-forms:

$$(\alpha, \beta) = \int_M \alpha \wedge *\beta\,, \tag{II.44}$$

which is also positive-definite.

From (II.39) and (II.43), * becomes a (graded) *self-adjoint* operator under this inner product.

$$(*\alpha, \beta) = (-1)^{k(n-k)}(\alpha, *\beta). \tag{II.45}$$

Suppose $\alpha \in \Lambda^k$ and $\beta \in \Lambda^{k+1}$, so that $(d\alpha, \beta)$ and $(\alpha, \delta\beta)$ make sense. Now:

$$d(\alpha \wedge *\beta) = d\alpha \wedge *\beta + (-1)^k\, \alpha \wedge d*\beta = d\alpha \wedge *\beta - \alpha \wedge *\delta\beta\,, \tag{II.46}$$

so, by Stokes's theorem, one has:

$$(d\alpha, \beta) - (\alpha, \delta\beta) = \int_{\partial M} \alpha \wedge *\beta\,. \tag{II.47}$$

Hence, when M is closed (= compact, without boundary) d and δ are adjoint to each other.

This makes:

$$\begin{aligned}(\Delta\alpha, \beta) &= (d\delta\alpha, \beta) + (\delta d\alpha, \beta) \\ &= (-1)^k[(\delta\alpha, \delta\beta) + (d\alpha, d\beta)] + \int_{\partial M} (\alpha \wedge *d\beta + \delta\alpha \wedge *\beta)\,.\end{aligned} \tag{II.48}$$

Therefore, if $\Delta\alpha = 0$, which is a generalization of the Laplace equation, and one specializes this last formula by setting $\beta = \alpha$ then since the inner product is positive-definite, we see that when M has no boundary one must have the vanishing of $d\alpha$ and $\delta\alpha$ as a consequence of the vanishing of $\Delta\alpha$.

Another formula that is useful in the solution of boundary-value problems for the Laplace equation is *Green's formula* (one of many, actually; cf., [**16, 17**]):

$$(\Delta\alpha, \beta) - (\alpha, \Delta\beta) = \int_{\partial M} [\alpha \wedge *d\beta - \beta \wedge *d\alpha + \delta\alpha \wedge *\beta - \delta\beta \wedge *\alpha]. \quad \text{(II.49)}$$

Once again, we see that whether the Laplacian operator itself is self-adjoint depends upon the vanishing of the boundary of M.

The *Hodge decomposition theorem* says that on a compact, orientable manifold M *without boundary* the vector spaces Λ^k can all be expressed as direct sums $Z^k \oplus Y^k \oplus \mathcal{H}^k$, in which Y^k represents the space of all co-closed k-forms and $\mathcal{H}^k$ represents the space of all harmonic k-forms. In other words, any k-form can be expressed as the sum of a closed form, a co-closed form, and a harmonic form.

A well-known consequence of this decomposition is the fact that if M is a compact, orientable manifold without boundary then each de Rham cohomology class $[\alpha] \in H^k_{dR}(M)$ contains a unique harmonic representative. In particular, if M is a vector space then there should be no harmonic forms in any dimension. This has, as a consequence, the well-known result of vector calculus that is called *Helmholtz's theorem*, which says that any vector field (i.e., 1-form α) can be uniquely expressed as the sum of an irrotational (i.e., closed) vector field and a solenoidal (i.e., co-closed) one. Actually, the way that this changes in the case of non-vanishing cohomology is that one simply has to specify the generators γ_a, $a = 1, \ldots, b_k$ of $H^k_{dR}(M)$, or equivalently, the values of the $<\alpha, \gamma_a>$, in addition to the irrotational and solenoidal parts; b_k refers to the *Betti number* in dimension k, which is the dimension of $H^k_{dR}(M)$. This decomposition was discovered by Lord Kelvin in the case of multiply-connected spaces ($k = 1$, $b_1 > 0$).

Of course, if one has been exposed to the rich variety of harmonic functions that arise by way of solving boundary-value problems in the Laplace equation – also known as potential theory – then one will find it disappointing that Hodge theory gives such a result. The resolution of the discrepancy is in the fact that the Hodge decomposition theorem is purely related to manifolds *without boundary*, which obviously eliminates the possibility of posing boundary-value problems.

There are, however, some results for the case of manifolds with boundary. They mostly involve going to what one calls *relative homology*, in which a chain in M is a *relative cycle* modulo ∂M iff its boundary is a chain in ∂M. Since this is going beyond the scope of the present study, we simply refer the curious to the work of Duff and Spencer [**16, 17**].

Another physical context in which Hodge theory does not apply is when one generalizes the construction of the Laplacian operator on k-forms to metrics of more general signature type, such as Lorentzian manifolds. One finds that the proof of the Hodge decomposition theorem breaks down due to the fact that the kernel of the d'Alembertian operator, which is hyperbolic, is infinite-dimensional, while the kernel of the Laplacian, which is elliptic for Riemannian manifolds, is finite-dimensional.

Basically, Hodge theory will apply to the physics of static or stationary fields, in which the four-dimensional hyperbolic picture reduces to a three-dimensional Euclidian one. However, this reduction must be treated with care, as it bears upon fundamental

issues in relativistic physics, such as simultaneity, and fundamental issues in mathematics, such as integrability.

7. Space-time splittings of the spacetime manifold. Although the subtle geometrical, topological, and physical issues that are involved with space-time splittings (see [**18**] for more discussion and references), since our treatment of the foundations of electromagnetism begins in the traditional realm of static fields, it is unavoidable that we make at least some perfunctory remarks about the subject.

If the spacetime manifold M takes the form of a four-dimensional vector space V, as it does in special relativity, then the main geometrical issue associated with decomposing spacetime into a three-dimensional spatial manifold Σ and a one-dimensional time line L is simply one of decomposing V into a direct sum $L \oplus \Sigma$.

This implies that any vector $\mathbf{v} \in V$ can be uniquely expressed as a sum $\mathbf{v}_t + \mathbf{v}_s$ of a *temporal* vector $\mathbf{v}_t \in L$ and a *spatial* vector $\mathbf{v}_t \in \Sigma$. One also has canonically-defined projections P_t: $V \to L$, $\mathbf{v} \mapsto \mathbf{v}_t$ and P_s: $V \to \Sigma$, $\mathbf{v} \mapsto \mathbf{v}_s$.

Usually, it is the line L that is defined first, in the form of a line that is tangent to the motion of the measurer/observer that defines a rest space. In other words, one is "modding out" the motion of the measure/observer by defining a comoving frame field along its world line. The complementary spatial vector space Σ then becomes essentially the normal space V/L to L in V. However, this normal space is really a subspace of V^*, not a subspace of V, so if one is to define a complement to L in V, one must either choose it arbitrarily or introduce – say – a scalar product, such as the Minkowski scalar product, and define Σ to be the orthogonal complement to L. Of course, that is not in the spirit of pre-metric physics, but fortunately there is much that can still be said about space-time splittings at a pre-metric level, as we shall see.

Now, when one goes from a vector space V to a more general manifold M there are actually two ways that one can speak of space-time separability. The most stringent form is to demand that the manifold M take the form of a product manifold $L \times \Sigma$, where L is a one-dimensional temporal manifold and Σ is a three-dimensional spatial manifold. Hence, the time manifold might be either $\mathbb{R}$, as in the vector space, or possibly the circle S^1, which would be more appropriate to situations in which periodicity featured prominently.

An immediate consequence of assuming that M has a product structure is the fact that the tangent bundle $T(M)$ can be given a direct sum – i.e., *Whitney* – decomposition $L(M) \oplus \Sigma(M)$, which then means that the tangent space T_x at any point x M admits a direct sum decomposition $L_x \oplus \Sigma_x$ as a vector space, with the previous consequences applying locally. Similarly, one has a Whitney sum decomposition of the cotangent bundle T^*M into $L^*M \oplus \Sigma^*M$.

One then has direct sum decompositions of the various tensor products of tangent and cotangent bundles. For instance:

$$T(M) \otimes T(M) = [T(L) \otimes T(L)] \oplus [T(L) \otimes T(\Sigma)] \oplus [T(\Sigma) \otimes T(L)] \oplus [T(\Sigma) \otimes T(\Sigma)],$$

and similarly for $T^*M \otimes T^*M$.

One often hears the projections T_{tt}, T_{ts}, T_{st}, T_{ss} of a second-rank tensor field T when it is decomposed in this way as its *time-time*, *time-space*, *space-time*, and *space-space* projections. (See, for instance, Cattaneo [**19**]).

Under symmetrization or anti-symmetrization, one finds that the sub-bundle $[T(L) \otimes T(\Sigma)] \oplus [T(\Sigma) \otimes T(L)]$ becomes either $T(L) \odot T(\Sigma)$ or $T(L) \wedge T(\Sigma)$, respectively. In particular, the exterior algebra $\Lambda^*(M)$ of T^*M becomes:

$$\begin{aligned}
&\Lambda^0 M = \Lambda^0 L \otimes \Lambda^0 \Sigma, \\
&\Lambda^1 M = \Lambda^1 L \oplus \Lambda^1 \Sigma, \\
&\Lambda^2 M = [\Lambda^1 L \wedge \Lambda^1 \Sigma] \oplus \Lambda^2 \Sigma, \\
&\Lambda^3 M = [\Lambda^1 L \wedge \Lambda^2 \Sigma] \oplus \Lambda^3 \Sigma, \\
&\Lambda^4 M = \Lambda^1 L \wedge \Lambda^3 \Sigma.
\end{aligned}$$

Note that since L is one-dimensional, $\Lambda^k L = 0$ for all $k > 1$, and similarly, $\Lambda^k \Sigma = 0$ for all $k > 3$. It is particularly convenient that each of the bundles $\Lambda^k L$ for $k = 1, 2, 3$ admits a decomposition into only a temporal sub-bundle and a spatial one, where "temporal" in this case means that it has $\Lambda^1 L$ as an exterior factor.

One can then extend the projection operators P_t and P_s to $\Lambda^k M$, $k = 1, 2, 3$ such that P_t: $\Lambda^k M = \Lambda^1 L \wedge \Lambda^{k-1} \Sigma$, P_s: $\Lambda^k M = \Lambda^k \Sigma$.

The specific forms that k-forms in each dimension then take are:

$$f(t, x) = T(t)S(x), \tag{II.49a}$$
$$\phi = \phi_0\, dt + \phi_s, \tag{II.49b}$$
$$F = dt \wedge E_s + F_s, \tag{II.49c}$$
$$G = dt \wedge F_s + G_s, \tag{II.49d}$$
$$\rho = \rho_0\, dt \wedge \rho_s, \tag{II.49e}$$

in which the subscript s denotes k-forms in $\Lambda^*\Sigma$ in each case.

A subtlety that is easy to overlook is the fact that since the component functions for k-forms on M are functions on M the component functions for the temporal and spatial parts of any k-form will still be functions on M, in general, not purely temporal or spatial functions. For instance, one will have:

$$\phi = \phi_0(t, x)\, dt + \phi_i(t, x)\, dx^i . \tag{II.50}$$

This is especially important to keep in mind when differentiating.

Of course, there is an analogous decomposition for the exterior algebra $\Lambda_* M$ of multivector fields on M that amounts to lowering all of the superscripts in the decompositions above.

One finds that the exterior derivative operator d: $\Lambda^k(M) \to \Lambda^{k+1}(M)$, $k = 0, 1, 2, 3$ admits a decomposition into a temporal part d_t: $\Lambda^k(M) \to \Lambda^1 L \wedge \Lambda^k \Sigma$, and a spatial one d_s: $\Lambda^k(M) \to \Lambda^{k+1} \Sigma$. One simply composes the operator d with the projections P_t and P_s, respectively. For instance:

$$d\phi = d_t\phi + d_s\phi = -\phi_{0,i}\, dt \wedge dx^i - \tfrac{1}{2}[\phi_{i,j} - \phi_{j,i}]\, dx^i \wedge dx^j. \tag{II.51}$$

By means of such compositions of operators with projections, one can similarly decompose the divergence operator $\delta\colon \Lambda_k(M) \to \Lambda_{k-1}(M)$, $k = 1, 2, 3, 4$, as well as the various exterior multiplication and interior multiplication operators on k-forms or k-vector fields.

As we mentioned above, there are two ways of imposing a space-time decomposition, the most demanding of which is a product structure $L \times \Sigma$ on the manifold M. One finds that quite often since most of the physical constructions are local and define differential equations, the reason that one needs a space-time decomposition first asserts itself in the local context. That is, one does not always require that M to decompose as a manifold, but only that $T(M)$ decompose as a vector bundle.

When $T(M)$ is given a Whitney sum decomposition $T(L) \oplus T(\Sigma)$, with no other restrictions on M, one says that M has been given an *almost-product* structure. Whether such an almost-product structure implies a product structure on M is a deep and subtle matter of integrability. For further discussion, one might confer the author's work in [**20**], which includes references to other approaches to the problem.

References

12. J. Munkres, *Topology,* Prentice Hall, NJ, 2000.
13. D.B. Fuks, V.A. Rokhlin, *Beginner's Course in Topology,* Springer, Berlin, 1984.
14. P.J. Hilton, *An Introduction to Homotopy Theory,* Cambridge University Press, Cambridge, 1953.
15. F. Warner, *Differentiable Manifolds and Lie Groups,* Scott Foresman, Glenview, IL, 1971.
16. G. de Rham, *Differentiable Manifolds,* Springer, Berlin, 1984.
17. R. L. Bishop and S. Goldberg, *Tensor analysis on Manifolds,* Dover, New York, 1980.
18. R. L. Bishop and R. J. Crittenden, *Geometry of Manifolds,* Academic Press, New York, 1964.
19. S. Sternberg, *Lectures on Differential Geometry 2nd ed.,* Chelsea, New York, 1983.
20. S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry,* Interscience, London, 1964.
21. R. Bott and L. Tu, *Differential Forms in Algebraic Topology,* Springer, Berlin, 1982.
22. J. Vick, *Homology Theory,* Academic Press, New York, 1973.
23. M. Greenberg, *Lectures on Algebraic Topology,* Benjamin-Cummings, Boston, 1967.
24. J. Rotman, *An Introduction to Algebraic Topology,* Springer, Berlin, 1988.
25. J. Munkres, *Elementary Differential Topology,* Princeton University Press, NJ, 1966.
26. W.V.D. Hodge, *The Theory and Applications of Harmonic Integrals,* Cambridge Mathematical Library, Cambridge, 1989.

27. G.F.D. Duff, “Differential Forms on Manifolds With Boundary,” Ann. Math., 56 (1952), 115-127.
28. G.F.D. Duff and D.C. Spencer, “Harmonic Tensors on Riemannian Manifolds With Boundary,” Ann. Math., 56(1952), 128-156.
29. D. H. Delphenich, “Proper Time Foliations of Lorentz Manifolds,” arXiv.org preprint, gr-qc/0211066.
30. E. Cattaneo, “Proiezione naturali e derivazione transversali in una varietá riemanniane a metrico iperbolica normale,” *Ann. di Mat. Pura e Appl.* **48** (1959), 361-386.
31. D. H. Delphenich, “Nonlinear connections and 1+3 splittings of spacetime,” arXiv.org preprint, gr-qc/0702115.

Chapter III

Static electric and magnetic fields

In addressing the laws of nature in their most fundamental form, however enigmatic, one must first address the issue of what constitute the most directly observable phenomena of nature. In physics, these phenomena tend to be the ones that appeal to one's vision, hearing, and sense of touch. (Taste and smell seem to be more chemical in character.) Not surprisingly, much of physics is concerned with light, sound, and motion, and, more generally, wave phenomena.

One can immediately distinguish classes of physical phenomena, such as passive and reactive phenomena, or static and dynamic ones. One sees that, in effect, these two ways of distinguishing phenomena are really the same. Passive or static phenomena are the ones that are observable in the absence of input from other parts of the system, including the observer, such as light from the distant stars or the equilibrium state of a mechanical structure. The reactive or dynamic phenomena are then the ones that pertain to the response of the state of a system to an input from elsewhere in the system, which then includes the measurements performed by an observer.

At the root of the physics of electricity and magnetism, one must realize that such a fundamental concept as charge does not represent a directly observable phenomenon, but only an indirectly observable one. The directly observable phenomenon that pertains to charge is the acceleration of some – but not all – matter as a result of the presence of some – but not all – other matter, in a manner that cannot be accounted for by the other known forces, such as gravitation. One only *postulates* the existence of an invisible, intangible quality, such as charge, to account for the dynamics of that interaction, just as thermodynamics once postulated the concept of "phlogiston" to account for the fact that some materials were flammable and others were not.

Of course, nowadays one must add the atomic hypothesis to the list of basic axioms, namely, that not only does charge come in three basic types +, –, and 0, but furthermore it does not really exist on an entire continuum of possible values. Rather, one always considers integer multiples of a nearly infinitesimal basic charge unit that is due to the presumed "indivisibility" of elementary matter. Nonetheless, just as the study of ideal gases is not best approached by looking at an ensemble of individual molecules whose cardinality is Avogadro big ($> 10^{23}$), but by replacing it with suitable totals and density functions defined on the volume in question, similarly, macroscopic electromagnetism is usually defined in terms of total and average quantities that are traceable to microscopic origins.

An unexpected consequence of the history of Twentieth Century physics is that it seems that the success of the atomic hypothesis reached its zenith during the early years of quantum physics, when it successfully accounted for the structure of periodic table. After that, it became clear that atoms were not indivisible, and neither were nuclei, or even nucleons. Hence, in the present era, one must quibble over whether the most elementary unit of electric charge is due to the electron, which is observable in a free state, or the quarks whose charge has an absolute value of one-third that of the electron, but are, apparently, observable only in bound states. Thus, one can only think of quarks

as having a somewhat more tentative character than electrons, positrons, protons, and the like. Furthermore, if one regards "elementary" particles as synonymous with their fields, and fields belong to an infinite-dimensional space, then one must realize that matters are not getting simpler as one goes deeper into the realm of elementary matter, but more complicated! Hence, we shall try to be objective about the existence of a fundamental duality between discreteness and continuity in the manifestation of matter, and use one or the other as it seems convenient.

Another issue that affects the way that one perceives physics at its most fundamental level – viz., the observation and measurement of physical phenomena in nature – is that there seems to be an unavoidable "duality" between statics and dynamics. That is, should one think of dynamics as the response of statics to "external" (i.e., external to the subsystem in question) perturbations, or should one think of statics as the limiting state of dynamics in the absence of external perturbations? This question actually bears upon the difference in spirit between quantum physics and relativistic (but non-quantum) physics. In the former approach to physics, one is essentially looking at the response of static – or, at least, *stationary* – systems to the perturbations – viz., measurements – of an external "measurer." Indeed, most of one's early exposure to quantum mechanics is exclusively concerned with the structure of the stationary states. In the latter approach, one must regard the four-dimensional world of dynamics as more fundamental in character, and the static world only comes about as the rest space that is defined by a choice of "observer."

Perhaps, the best way to reconcile these two viewpoints is to keep both of them in mind and consider the observation of Max Born that all measurements are made in the rest space of the measuring devices. Hence, there is something unavoidable, indeed *fundamental*, about the role of measurer/observer in the description of physical phenomena.

1. Electric charge [1-4]. We shall start with the concept of electric charge – with all of the aforementioned caveats – as the new logical primitive that we add to the laws of kinematics and dynamics as a starting point for our theory of electrostatics. The axioms are then that *total electric charge* is a real number $Q[V]$ that one associates with a volume V of space Σ (which is a differentiable manifold of dimension one, two, or three) like an *extensive* variable, in the language of thermodynamics. That is, when one combines disjoint volumes, one adds the total electric charges that they contain:

$$Q\left[\sum_{i=1}^{N} V_i\right] = \sum_{i=1}^{N} Q[V_i]\,. \tag{III.1}$$

This amounts to saying that total electric charge defines a *measure* on space, in the language of measure theory, as long as the volume V is a measurable subset.

The value $Q[V]$ of charge is an integer multiple of an elementary charge, which we call e. Interestingly, although e is very small in units of Coulombs, that fact is really a criticism of the applicability of the latter unit, since one Coulomb of static charge is harder to configure than it sounds. (It is, nevertheless, quite dynamically practicable. Consider one Ampere of current flowing through the cross-section of a wire, which amounts to one Coulomb every second.)

Although a region of space that contains no elementary charge sources will be associated with zero charge, the converse is not true. In fact, most macroscopic matter, such as stars, seems to have close to zero total charge, despite the fact that it contains an enormous number of elementary charges. That is why gravitational forces seem to be dominant at the astronomical level, even though the force of gravitation is almost infinitesimal by comparison to that of electrostatic attraction/repulsion.

The intensive variable that is associated with total electric charge is *electric charge density*, which one first obtains as an average charge density in a given volume by dividing the total charge in a volume by its volume:

$$\bar{\rho}[V] = \frac{1}{V} Q[V] \tag{III.2}$$

and then passing to the limit of zero volume while regarding each V as a neighborhood of each of its points:

$$\rho(x) = \lim_{V_x \to 0} \bar{\rho}(V_x) \,. \tag{III.3}$$

Of course, we are then assuming that such a limit actually exists. This is where the concept of a point charge introduces a predictable infinity in the form of the electric charge density at the point where the charge is localized. If one takes the position that infinity never manifests itself in physical reality (i.e., in measurements) then one must regard the point charge as merely a convenient approximation to something more difficult to fathom, namely, a charge distribution that is concentrated in a volume that is too small to probe directly. Furthermore, the fact that quantum mechanics routinely regards electrons as wavelike in character suggests that they have finite spatial extent.

Here, we introduce the methods of the last chapter by restricting ourselves to volumes that take the form of differentiable singular cubic n-chains, where $n = \dim(\Sigma)$. Hence, the additivity that we postulated in (III.1) suggests that we might wish to regard total electric charge as an n-cochain; i.e., a linear functional on n-chains. All that we need to add is the requirement that it also be homogeneous of degree one with respect to scalar multiplication by a real number λ:

$$Q[\lambda V] = \lambda Q[V]. \tag{III.4}$$

Admittedly, the concept of multiplying a volume by a real scalar is less physically intuitive than that of adding disjoint volumes, but we shall see that this is a small price to pay for the utility of homology in describing charge and flux.

Since we are clearly assuming that Σ is orientable, oriented, and endowed with a volume element $\mathcal{V}$, we already have one linear functional on n-chains that is defined by the volume functional. On any n-cube σ_n it takes the form:

$$V[\sigma_n] = \int_{\sigma_n} \mathcal{V} \,. \tag{III.5}$$

One then extends this definition to more general n-chains by the assumption of linearity.

From (III.2), one can then define the average electric charge density in the region described by σ_n as the ratio:

$$\overline{\rho}[\sigma_n] = \frac{Q[\sigma_n]}{V[\sigma_n]}. \tag{III.6}$$

It is in passing to the limit that one must deal with the existence of elementary charges, since that fact suggests that ultimately a volume must contain either one elementary charge (with its sign) or none at all. If an elementary charge is distributed over a finite volume then, in a sense, it is not really elementary, since it is presumed to be further reducible. If it is pointlike then one must accept the non-existence of a finite limiting charge density for the neighborhoods that it contains.

One can consider some sort of "confinement" hypothesis for charges less than e, similar to the hypotheses that make free quarks unobservable. However, in the next section, we shall resolve the issue by a simple topological device, namely, by removing the point to which the volumes are converging from space itself. This then has the effect of saying that the concept of electric charge density is useful only as a macroscopic approximation.

Hence, we accept that the concept of electric charge density is only so fundamental at the elementary level and assume the electric charge density $\rho(x)$ at a point $x \in \Sigma$ actually exists as a function. One can then think of it as the "distribution kernel" of the electric charge functional:

$$Q[\sigma_n] = \int_{\sigma_n} \rho \mathcal{V}. \tag{III.7}$$

The non-existence of a limit to the electric charge density for a point charge is closely related to the concept of the Dirac delta function, which is not really a function, at all, but a fictitious kernel for a much-better-defined distribution called the "evaluation" functional, which we briefly describe.

If S is a set and $F(S; \mathbb{R})$ is the vector space of real-valued functions on S then the *evaluation functional* that is associated with each $x \in S$ is the linear functional E_x: $F(S; \mathbb{R}) \to \mathbb{R}$, $f \mapsto E_x[f] = f(x)$, which simply evaluates the function f: $S \to \mathbb{R}$ at the point x. We can extend this to a linear functional on n-forms E_x: $\Lambda^n \to \mathbb{R}$ by assuming that all elements of Λ^n take the form $f\mathcal{V}$, so $E_x[f\mathcal{V}] = f(x)\mathcal{V}_x$. Despite the fact that this distribution does not have a kernel, we follow tradition and formally write:

$$E_x[f\mathcal{V}] = \int_\Sigma \delta(x, y) f(y) \mathcal{V}_y = f(x)\mathcal{V}_x, \tag{III.8}$$

in which $\delta(x, y)$ is the *Dirac delta function.* Here, we are using the notation $\mathcal{V}_y$ to suggest that the integration is over the y variable, while the x is held fixed.

If we integrate over a subset $V \subset \Sigma$, instead of Σ itself, then we can define the *selection functional* for $x \in V$ using this delta function:

$$\psi_x[V] = \int_V \delta(x,y)\mathcal{V}_y = \begin{cases} 1 & x \in V \\ 0 & \text{otherwise.} \end{cases} \tag{III.9}$$

Hence, the selection functional for x identifies all subsets that have x as an element; i.e., all neighborhoods of x. Of course, this makes perfect sense as an n-cochain, as long as we restrict ourselves to subsets V that can be parameterized by k-chains.

Usually, when Σ is a vector space, one defines $\delta(x)$ to give evaluation at the origin and then replaces x with $x - y$ to give evaluation at $x = y$, but one can see that this definition breaks down when Σ does not have an affine structure that would allow us to make sense of the expression $x - y$ for two points $x, y \in \Sigma$. This is a recurring theme for the problems that are associated with generalizing linear constructions to nonlinear manifolds.

By means of this fictitious kernel, one can define electric charge densities for finite sets of point charges of charge Q_i at points $x_i \in \Sigma$ as linear combinations of delta functions:

$$\rho(y) = \sum_{i=1}^{N} Q_i \delta(x_i, y)\,. \tag{III.10}$$

Although our sum over points suggests that we are defining a 0-cochain, from (III.9), we see that this sum makes better sense as an n-cochain, which only means multiplying both sides by $\mathcal{V}_y$ and integrating over an n-chain V:

$$Q[V] = \int_V \rho\mathcal{V} = \sum_{i=1}^{N} Q_i \int_V \delta(x_i, y)\mathcal{V}_y = \text{total charge in } V. \tag{III.11}$$

2. Electric field strength and electric flux [1-4]. If a point charge Q at a point $x \in \Sigma$ experiences a force $F(x; Q) \in T_x^*M$, which we presume to be of electrostatic origin, then we define the *electric field strength 1-form for this charge* to be the covector field E: $\Sigma \rightarrow T^*(\Sigma)$:

$$E(x; Q) = \frac{1}{Q} F(x; Q). \tag{III.12}$$

However, this definition has two unsavory aspects to it:

1. We are defining it for point charges, and we have already expressed doubts about the limits of that approximation.

2. We are defining E in terms of a particular charge Q, which suggests that a different charge Q' might define a different E.

The traditional way of getting around the second objection is to introduce the *test particle hypothesis:* For every real number Q, no matter how small, there exists a charged particle in nature whose charge is Q. Clearly, this is totally inconsistent with the assumption that there is a non-zero minimum charge.

If one accepts this hypothesis then one can factor out the effect of Q by passing to the limit – if there is one – and defining:

$$E(x) = \lim_{Q\to 0} \frac{1}{Q} E(x; Q). \tag{III.13}$$

If one does not wish to introduce test particles then one can define $E(x)$ to be the force on a *unit charge*.

$$E(x) \equiv F(x; 1). \tag{III.14}$$

Now, we can deal with an even more subtle issue concerning the nature of the electrostatic force: Is $E(x; Q)$ going to be this same 1-form for $Q \neq 1$, and if not, how does it change? At first, this suggests the question of the linearity of the field equations for E in disguise; i.e., the principle of superposition. However, in order for the nature of the electrostatic force to be consistent with the independence of $E(x)$ from Q, it is necessary and sufficient to assume that $F(x, Q)$ is homogeneous of degree one in Q:

$$F(x; \lambda Q) = \lambda F(x; Q), \tag{III.15}$$

which then makes:

$$F(x; Q) = QF(x; 1) = QE(x). \tag{III.16}$$

We shall simply work with the definition of E that we get from (III.14) and keep in mind what that might entail.

If we represent a path in Σ by a 1-chain c_1 then since the work done by a force F along that path is the 1-cochain:

$$\Delta U[c_1] = \int_{c_1} F\,, \tag{III.17}$$

we define the *change in potential energy* along c_1 due to the influence of $F(x; Q)$ on a charge Q to be:

$$\Delta U[c_1; Q] = \int_{c_1} F(x;Q)\,, \tag{III.18}$$

and the *change in electric potential* along c_1 due to the influence of $F(x; Q)$ to be the change in potential energy experienced by a unit charge:

$$\mathcal{E}[c_1] = \Delta U[c_1; 1] = \int_{c_1} F(x;1) = \int_{c_1} E\,. \tag{III.19}$$

We are using the notation $\mathcal{E}$ to be consistent with the classical term "electromotive force," but that term, despite it continued popularity, is a misnomer, since the quantity being described has units of *work* per unit charge, not force per unit charge.

Usually, in elementary physics, one regards the issue of whether this work done is independent of the path between two points as being equivalent to the question of whether it vanishes around any loop (i.e., any 1-cycle). Of course, that is because one is

usually concerned with life in $\Sigma = \mathbb{R}^3$, which is contractible, and therefore simply connected, *a fortiori*.

Since we are focusing more on topological issues in this study, we shall clearly distinguish between the two conditions. Hence, one must treat the cases of loops that do or do not bound 2-chains separately. If z_1 is a bounding 1-cycle $z_1 = \partial c_2$ then, by Stokes's theorem:

$$\mathcal{E}[\partial c_2] = \int_{\partial c_2} E = \int_{c_2} dE \,. \tag{III.20}$$

Hence, if the work done on a unit charge by an electrostatic force F vanishes around any loop in Σ that bounds a 2-chain then one must have:

$$dE = 0. \tag{III.21}$$

This amounts to the statement that the electric field strength 1-form E is closed; one can also say that it is *irrotational*. Hence, E defines a de Rham cohomology class in dimension 1: $[E] \in H^1_{dR}(M)$.

Of course, if Σ is simply connected then every 1-cycle bounds a 2-chain and (III.21) is equivalent to the condition that E be exact; we shall return to this concept in a later section.

Another way of looking at (III.21) is that it represents an integrability condition for the exterior differential system that is defined by $E = 0$. That is, since E is a 1-form, at each point of Σ there is a hyperplane $\mathrm{Ann}(E)_x$ in $T_x\Sigma$ that consists of the tangent vectors $\mathbf{v}$ at x that are annihilated by E: $E(\mathbf{v}) = 0$. This set of hyperplanes defines a sub-bundle of $T(M)$ of codimension one that calls a *differential system* on M. It is completely integrable iff every point $x \in \Sigma$ has a hypersurface that passes through it, and its tangent hyperplane agrees with $\mathrm{Ann}(E)_x$. Such a hypersurface is called an *integral hypersurface* or *integral submanifold* and Σ is said to be *foliated* by these integral submanifolds.

The most general necessary and sufficient condition for the complete integrability of $\mathrm{Ann}(E)$ is given by *Frobenius's theorem*, which says that $\mathrm{Ann}(E)$ is completely integrable iff it is *involutory*. By definition, this means that the vector space of vector fields on Σ that take their values in the fibers of $\mathrm{Ann}(E)$ is closed under the Lie bracket; i.e., it is a Lie subalgebra of $\mathfrak{X}(\Sigma)$. An equivalent condition is that:

$$E \wedge dE = 0, \tag{III.22}$$

which is equivalent to the condition that there be a 1-form η such that:

$$dE = \eta \wedge E \,. \tag{III.23}$$

Hence, (III.21) is a stronger condition than Frobenius requires.

An even stronger condition is that E be exact, which gets us into the realm of electric potential functions and shows us that the nature of the integral hypersurfaces is simply that of equipotentials. We shall return to this topic shortly.

3. Electric excitation (displacement) [1-4]. When an electric field E is present in a macroscopic medium that is composed of atomic ions and electrons that are bound into a crystal lattice or molecules that are bound into a more amorphous solid or fluid, the charges in the medium will react to E to a greater or lesser degree. In conductors, for which the electron mobility is high, the response of the medium will generally take the form of an electric current – i.e., the translational motion of the electrons. In an insulator, for which the electron mobility is low, the effect will often be the alignment of its electric dipoles, which are spatially separated charge pairs of the form $\{+Q, -Q\}$. If the spatial separation is described by the displacement vector **d** that points from $+Q$ to $-Q$ then the *electric dipole moment* of the pair is $Q\mathbf{d}$.

When this is going on macroscopically, so we can consider an electric charge density ρ in place of Q, we can replace $Q\mathbf{d}$ with a vector field **D** that represents an electric dipole density. We shall call this vector field the *electric excitation* of the medium; it is also called the *electric displacement*, following Maxwell's terminology.

Actually, one usually associates such a vector field with E even in the otherwise uninteresting case of the electromagnetic vacuum. However, the association, which we write in the form:

$$\mathbf{D} = \varepsilon_0\, i_g E \qquad (D^i = \varepsilon_0\, g^{ij} E_j) \tag{III.24}$$

has some suspicious features that get passed over when one is only doing topology, geometry, and physics in $\mathbb{R}^3$:

1. To characterize the response of the vacuum state to E by the constant ε_0, which one calls the *electric permittivity* or *dielectric constant* of the vacuum, is to assume that the vacuum is a linear, homogeneous, isotropic dielectric medium, which ignores a lot of lessons from quantum electrodynamics, such as vacuum polarization. (We shall discuss this in more detail in a later chapter.)

2. We are using the spatial Euclidian metric g on Σ as if it were given independently of electrostatic considerations. One of the main assumptions of pre-metric electromagnetism is that the four-dimensional Lorentzian structure on spacetime is a consequence of the electromagnetic constitutive laws as they relate to the propagation of electromagnetic waves, so we wonder if this also extends to the electrostatic level, as well.

For now, we simply assume that there is a diffeomorphism ε_x: $\Lambda^1_x \to \Lambda_{1,x}$ that takes each fiber of Λ^1 at $x \in \Sigma$ to the corresponding fiber of Λ_1 at x. When this diffeomorphism is a linear isomorphism, the map ε: $\Lambda^1 \to \Lambda_1$ is a vector bundle isomorphism. Of course, this would leave out the possibility of nonlinear electrostatics, such as one encounters in the effective field theories of quantum electrodynamics [**2**].

Since we are assuming that Σ is orientable and given a volume element $\mathcal{V}$, the vector field **D** can be associated with an n–1-form:

$$\#\mathbf{D} = i_{\mathbf{D}}\mathcal{V} = \frac{1}{3!}\,\varepsilon_{ijk}\, D^k\, dx^i \wedge dx^j \tag{III.25}$$

by Poincaré duality.

We call this n–1-form the *electric flux density*. The corresponding integral of #**D** over an n–1-chain:

$$\Phi_{\mathbf{D}}[c_{n-1}] = \int_{c_{n-1}} \#\mathbf{D} \tag{III.26}$$

is then the *total electric flux though* c_{n-1}. This makes it clear that the total electric flux defines an n–1-chain on Σ.

If $c_{n-1} = \partial c_n$ then from Stokes's (viz., Gauss's, in this case) theorem

$$\Phi_{\mathbf{D}}[c_{n-1}] = \int_{c_n} d\,\#\mathbf{D} = \int_{c_n} (\delta\mathbf{D})\mathcal{V} = Q[c_n], \tag{III.27}$$

in which we need only set our charge density ρ equal to:

$$\delta\mathbf{D} = \rho \tag{III.28}$$

in order to make (III.26) valid for every bounding n-chain.

Hence, one can say that the total electric flux through the boundary of an n-cycle is equal to the total electric charge that is contained in the n-cycle itself. (We are omitting a multiplicative constant, such as 4π, which can be absorbed into the definition of $\mathcal{V}$.)

One can also include the constitutive law that connects **D** with E and express (III.28) as first order partial differential equation in E:

$$\delta \cdot \varepsilon(E) = \rho. \tag{III.29}$$

If ε is expressed by local component functions of the form $\varepsilon^{ij}(x, E)$ then (III.29) can be expressed in local form as:

$$\tilde{\varepsilon}^{ij}\frac{\partial E_j}{\partial x^i} + \frac{\partial \varepsilon^{ij}}{\partial x^i}E_j = \rho, \tag{III.30}$$

in which we have introduced the notation:

$$\tilde{\varepsilon}^{ij} = \varepsilon^{ij} + \frac{\partial \varepsilon^{ik}}{\partial E_j}E_k\,. \tag{III.31}$$

Therefore, we conclude that the inhomogeneity in ε affects E directly, while the nonlinearity affects the spatial derivatives of E.

We see from (III.28) that outside the support of ρ (supp(ρ) = closure of the set of points at which ρ is non-vanishing) the vector field **D** is divergenceless. Hence, if one thinks of ρ as the generator of a one-parameter family of local diffeomorphisms of Σ then outside of supp(ρ) these diffeomorphisms preserve the volume element $\mathcal{V}$. The integral curves of the vector field **D** are what one commonly refers to as *electric field lines*, or, more precisely, *electric flux lines.* In the case of $\mathbf{D} = \varepsilon\, i_g E$, where ε is a smooth function, **D** will also represent the direction of the force and acceleration that acts on a charge Q of

non-zero mass at each given point. It is not, in general, the direction of the velocity vector for the subsequent motion.

Since any $n-1$-chain that bounds will be an $n-1$-cycle, the question arises of what happens to the total electric flux through a non-bounding $n-1$-cycle. Of course, in order for this to be possible $H_{n-1}(\Sigma; \mathbb{R})$ must not vanish. From Poincaré duality, this is equivalent to the non-vanishing of $H^1(\Sigma; \mathbb{R}) \cong H_1(\Sigma; \mathbb{R})$, so it is sufficient that Σ be non-simply connected.

Another possibility for making $H_{n-1}(\Sigma; \mathbb{R})$ non-vanishing is that Σ is simply connected, but the first non-vanishing homotopy group is in dimension $n-1$, which is the case with $\mathbb{R}^n - \{0\}$. A generator for $H_{n-1}(\Sigma; \mathbb{R}) \cong \mathbb{R}$ in that case is then defined by any $n-1$-sphere (i.e., any $n-1$-cycle) that includes the origin in the ball that it bounds.

When z_{n-1} is not a bounding cycle, Stokes's theorem no longer applies, so $\Phi_D[z_{n-1}]$ can be non-vanishing even when the total charge contained in the region "bounded" by z_{n-1} can no longer be defined, since there is no such region. This allows one to give a purely topological origin to charge, which is probably more appropriate at the elementary level than the macroscopic one that involves a more statistically-defined charge density: Elementary charge is due to the presence of non-bounding $n-1$-cycles in Σ; i.e., the non-vanishing of its homology in dimension $n-1$.

Furthermore, if $H_{n-1}(\Sigma; \mathbb{R})$ is non-vanishing then so is $H_{dR}^{n-1}(\Sigma)$, and there are closed $n-1$-forms that are not exact. Hence, it is possible to find $\#\mathbf{D}$ such that $d\#\mathbf{D} = 0$, but $\Phi_{\mathbf{D}}[z_{n-1}] \neq 0$; for such a $\mathbf{D}$, one clearly has $\delta\mathbf{D} = 0$. This is essentially the spirit of what Wheeler and Misner [**5**] called "charge without charge;" i.e., non-vanishing flux with vanishing divergence.

For example, the Coulomb $\mathbf{D}$-field:

$$\mathbf{D}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\mathbf{r}}, \tag{III.32}$$

has this property.

4. Electrostatic field equations [2]. If we combine the equations that we proposed for E and $\mathbf{D}$ so far then we have:

$$dE = 0, \qquad \delta\mathbf{D} = \rho, \qquad \mathbf{D} = \varepsilon(E). \tag{III.33}$$

As we have seen, the presence of ρ tends to be more of interest in macroscopic problems, in which the concept of a charge density makes sense as an approximation to a large ensemble of elementary charges. When dealing with elementary charges, it is probably preferable to consider the homogeneous form of these equations, which is also valid outside $\text{supp}(\rho)$:

$$dE = 0, \qquad \delta\mathbf{D} = 0, \qquad \mathbf{D} = \varepsilon(E). \tag{III.34}$$

In this form, one can see the manifestly topological nature of the first two equations, since the first one says that E defines a de Rham cohomology class for Σ in dimension one and the second one says that $\mathbf{D}$ defines a de Rham homology class in dimension one; equivalently, $\#\mathbf{D}$ defines a de Rham cohomology class in dimension $n - 1$. This also introduces a topological origin for elementary charges by the possibility that either of – hence, both of – $H^1_{dR}(\Sigma)$ and $H^{n-1}_{dR}(\Sigma)$ are non-trivial vector spaces. For instance, Σ might be multiply connected or it could be simply connected and have its first non-vanishing homotopy group in dimension $n - 1$, like $\mathbb{R}^n - \{0\}$, among other possibilities.

Although one often encounters the equations of electrostatics as differential equations, one can also express them in integral form:

$$\int_{\partial c_2} E = 0, \qquad \int_{\partial c_3} \#\mathbf{D} = \int_{c_3} \rho\,\mathcal{V}, \qquad \mathbf{D} = \varepsilon(E). \tag{III.35}$$

One can think of the first two equations as expressing conservation of energy and conservation of charge, respectively.

The integral form of differential equations is equivalent to regarding the differential equations as having distributions for solutions, which is necessary when one is dealing with solutions that have singularities, such as jump discontinuities or poles.

5. Electrostatic potential functions [2]. Previously, we only required that the work done by an electrostatic force around a loop that bounded a 2-cycle vanished, and showed that if this were true for all 1-boundaries then one could conclude that E was closed. A closely-related issue associated with work and potential energy, independently of the nature of F as we discussed above, is the issue of whether the work done along a path connecting two points A and B is path-dependent or not. However, one must keep in mind that if c_1 is a path from A to B, we must have $\partial c_1 = B - A$. This suggests that we might define ΔV as a 0-cochain; i.e.:

$$\Delta V[A, B] = \int_{\partial c_1} V = \int_{c_1} dV = \int_{c_1} E\,. \tag{III.36}$$

If this is true for all 1-chains then we can conclude that E is exact:

$$E = dV. \tag{III.37}$$

Of course, the function V is not defined uniquely, but only up to an additive constant. This is because it is only the difference:

$$\Delta V[A, B] = V(B) - V(A) \tag{III.38}$$

that is uniquely defined.

If one chooses a point, such as A, and a value $V(A)$ to associate with A then one can convert the two-point function $\Delta V[A, B]$ on $\Sigma \times \Sigma$ to a one-point function on Σ, at least, under the assumption that it is path-connected:

$$V(x) = V(A) + \Delta V[A, x]. \tag{III.39}$$

One calls such a function V an (electrostatic) *potential function* for E.

Note that (III.38) implies that the work done around any 1-cycle – bounding or not – vanishes. One simply puts two points A and B on the loop and treats the two arcs AB and BA as two paths from A to B, one which has a negative orientation with respect to the other.

As we pointed out above, exactness is the strongest integrability requirement that one can put on the exterior differential system $E = 0$. Basically, it makes the integral submanifolds take the form of the level hypersurfaces $V(x) = \text{const.}$ of the function V; a different choice of $V(A)$ will not change the nature of the hypersurfaces, only the value of V on each of them.

If E admits a potential function then we can combine the three field equations (III.33) into a single one:

$$\Delta_\varepsilon V = \rho, \tag{III.40}$$

in which:

$$\Delta_\varepsilon = \delta \cdot \varepsilon \cdot d \tag{III.41}$$

is the generalized Laplacian operator on 0-forms that is defined by ε. If ε is a nonlinear map then Δ_ε will be a nonlinear second-order elliptic differential operator.

From (III.30), all that we have to do to get the local form of (III.40) is substitute $E_i = V_{,i}$:

$$\tilde{\varepsilon}^{ij} \frac{\partial^2 V}{\partial x^i \partial x^j} + \frac{\partial \varepsilon^{ij}}{\partial x^i} \frac{\partial V}{\partial x^j} = \rho. \tag{III.42}$$

6. Magnetic charge and flux [1, 4]. If we continue our argument that the most fundamental level of any force of nature involves the observation/measurement of changes in the state of a natural systems that are consistently correlated with some other natural phenomenon then we admit that magnetism first manifests itself as forces on some, but not all, materials (viz., magnetic materials) as a result of being exposed to other influences that we simply identify as magnetic forces. Of course, this all sounds very tautological, unless we get more specific about the origin of magnetic forces.

Here, the atomic hypothesis is only so helpful. If we break a bar magnet into a geometric progression of smaller pieces then once again the series will terminate with electrons, but not in the same way as for electrostatic forces. If a bar magnet is a macroscopic ensemble of elementary magnetic dipoles then one sees that the most elementary manifestation of a magnetic field is due to a *dipole*, not a pair of monopoles. Namely, the ultimate source of magnetic fields seems to be the *relative motion* of elementary electric charges. In the case of elementary dipoles, this usually takes the form of electron or nucleon spin.

Of course, at this point in the history of physics, the possible existence of magnetic monopoles is a hotly-contested subject. They were first postulated by Dirac in the 1930's as a possible consequence of the formulation of electromagnetism as a $U(1)$ gauge theory. In particular, if the $U(1)$-principal bundle on spacetime that represented the gauge structure for electromagnetism happened to be non-trivial – so no global choice of $U(1)$ gauge was possible – then this would suggest topological obstructions (viz., the first Chern class of the bundle) that might bring about magnetic monopoles. Because these topological considerations seem to be so fundamental and unavoidable, the fact that, to date, no experimental evidence for the existence of magnetic monopoles has emerged has not deterred theoretical physics from continuing to publish thousands of books and papers on the subject. Nowadays, one can always use spontaneous symmetry breaking as a convenient excuse for the non-existence of theoretically-predicted phenomena in experiments: One simply postulates that the energy level of spontaneous symmetry breaking for the phenomenon in question is many orders of magnitude beyond any reasonable experimental bounds (e.g., 10^{15} or 10^{19} GeV vs. 10^{3} GeV for terrestrial particle colliders, 10^{6} GeV for ultra-high energy cosmic rays). This usually means that everything unobservable on Earth probably existed in the first nanosecond of the Big Bang when the universe was that energetic.

We shall err on the side of caution in this presentation and take the experimentally established position that the source of all magnetic fields is the relative motion of electric charges, whether that takes the form of the rectilinear motion of charges in currents or the rotational motion of spinning charge distributions. Even then, we must caution the reader that, as Jackson [**6**] pointed out, there does not always exist a rest frame for an observer that makes a given magnetic field disappear. There is certainly such a frame for a single elementary charge moving along a curve in spacetime; one simply chooses a co-moving frame, in which the only field is the electrostatic field. However, as the example of a bar magnet shows, when one is concerned with macroscopic magnetic fields, trying to find a co-moving frame gets hopeless. In fact, there is always the question of whether a measurer/observer that is orbiting around a rotating charge distribution with the same orbital period as the angular period of the rotation will similarly see no magnetic field, since there are fictitious forces that come about due to the orbital motion.

One sees that if one truly believes that the origin of all magnetic forces is the relative motion of electric charges then the existence of magnetic monopoles seems to be a patent absurdity. We shall express this situation topologically by saying that elementary charges represent non-trivial generators of $\pi_2(\Sigma)$ – i.e., monopoles, in the language of topological defects [**7, 8**] – and elementary magnetic sources represent non-trivial generators of $\pi_1(\Sigma)$ – i.e., vortex defects [18]. That is, the spatial sources of magnetic fields will, for us, exist as curves in Σ, not individual points.

One immediately sees that the apparent non-existence of magnetic monopoles makes it harder to define a magnetic field in Σ than it was to define the electric field E. We cannot speak of the magnetic force on a unit magnetic charge if there is no such thing as a magnetic charge. Since magnetic fields exert torques on magnetic dipoles, one might

[18] Although these are also called *string defects*, since the use of the term "string" in theoretical physics is so well-established in the literature of Big-Bang cosmology and life beyond the Planck scale of space, time, and energy, we shall avoid its usage in the context of tangible physics that is easily observed in terrestrial experiments.

define the magnetic field strength $B(x)$ at a point $x \in \Sigma$ to be the torque on a unit dipole. This has the advantage of making B a 2-form on Σ, by its nature. However, a magnetic dipole moment $\boldsymbol{\mu}$ has to have a direction in space, as well as a magnitude, since it is a vector, not a scalar.

If one looks at the way that the torque $\boldsymbol{\tau}$ is coupled to $\mathbf{B}$ and $\boldsymbol{\mu}$, namely:

$$\boldsymbol{\tau} = \boldsymbol{\mu} \times \mathbf{B} = *(\boldsymbol{\mu} \wedge \mathbf{B}) \tag{III.43}$$

then one sees that although one cannot simply divide $\boldsymbol{\tau}$ by $\boldsymbol{\mu}$ to get $\mathbf{B}$, nevertheless, in principle, a complete knowledge of the $\boldsymbol{\tau}$ as a function of $\boldsymbol{\mu}$ will give a unique vector $\mathbf{B}$.

This is three-step process for a given "test dipole" $\boldsymbol{\mu}$:

1. Obtain the magnitude B of $\mathbf{B}$ from the maximum value of τ/μ as the direction of $\boldsymbol{\mu}$ varies over all possibilities.
2. Obtain the direction of $\mathbf{B}$ – up to orientation – from the direction of $\boldsymbol{\mu}$ that gives a null to $\boldsymbol{\tau}$, since $\boldsymbol{\mu} \times \mathbf{B} = 0$ iff $\boldsymbol{\mu}$ is parallel to $\mathbf{B}$.
3. Since $\boldsymbol{\tau}$ cannot be parallel to a non-zero $\boldsymbol{\mu}$, the plane of $\boldsymbol{\tau}$–$\boldsymbol{\mu}$ then divides $\mathbb{R}^3$ into two disjoint halves. By choosing one of them to contain the positive normal to that plane, one can define that to also contain the direction of $\mathbf{B}$.

In order for this construction to produce a vector field $\mathbf{B}(x)$ on Σ, one simply understands that the aforementioned process is concerned with tangent vectors at each point of Σ.

We point out that it is generally more traditional to define $\mathbf{B}$ in terms of the Lorentz linear force density $\mathbf{F} = \mathbf{I} \times \mathbf{B}$ that acts on a "test" current vector $\mathbf{I}$, but currents do not seem to suggest the same kind of minimal unit as the Bohr magneton defines for the magnetic dipole moment of the electron.

The total magnetic flux through a 2-chain is then:

$$\Phi_{\mathbf{B}}[c_2] = \int_{c_2} \#\mathbf{B}\,. \tag{III.44}$$

When c_2 is a bounding 2-cycle ∂c_3 Stokes's theorem gives:

$$\Phi_{\mathbf{B}}[\partial c_3] = \int_{c_3} d\,\#\mathbf{B} = \int_{c_3} (\delta\mathbf{B})\,\mathcal{V} = Q_M[c_3]. \tag{III.45}$$

Since this is supposed to equal the total "magnetic charge" that is contained in the region of Σ defined by c_3, to say that there are no magnetic monopoles is to say that $\Phi_{\mathbf{B}}[\partial c_3]$ must vanish for all bounding 2-cycles, which leads to the differential equation:

$$\delta\mathbf{B} = 0. \tag{III.46}$$

Of course, if there are non-bounding 2-cycles z_2, which implies that $H_2(\Sigma; \mathbb{R})$ is non-vanishing, it is entirely possible that $\Phi_{\mathbf{B}}[z_2]$ is non-vanishing even though one still has (III.46). Hence, this is another way of accounting for "magnetic charge without magnetic

charge" by a purely topological device. Furthermore, in the gauge formulation of electromagnetism, a necessary, but not sufficient, condition for the non-triviality of the $U(1)$-principal bundle that defines the gauge structure is that $H_2(\Sigma; \mathbb{R}) \cong H^2_{dR}(M)$ be non-vanishing, since one is concerned with the first Chern class of the bundle, and this will be a closed 2-form; i.e., a de Rham cohomology class in dimension two.

7. Magnetic excitation (induction). When a magnetic material is exposed to a magnetic field, there is a tendency for the elementary magnetic dipoles to align themselves to a greater degree. As opposed to the situation described in the context of electric fields there is no corresponding issue of the translational motion of the elementary magnetic monopoles, only the rotational motion of the dipoles. There is a resulting magnetic dipole moment density associated with the medium, which we refer to as the *magnetic excitation* of the medium, and describe by a 1-form H on Σ. Although the term "magnetic induction" is also used, since that tends to lead to confusion in the context of electromagnetic induction, we shall use the more precise term of "excitation."

The response H of a given medium to an applied **B** field will then be described by a *magnetic constitutive law:*

$$H = \mu^{-1}(\mathbf{B}), \tag{III.47}$$

in which μ: $\Lambda^1 \to \Lambda_1$ is a diffeomorphism of each cotangent space with its corresponding tangent space. Here, we see that classical electromagnetism seemed to be treating the vector field **B** and the 1-form H as having the opposite roles to what later seemed to work better in the eyes of relativistic electromagnetism. Hence, we shall have to use the inverse of the map μ, instead of the map itself.

In the case of the classical electromagnetic vacuum, which is linear, isotropic, and homogeneous with respect to its magnetic response, this relation takes the form:

$$H = \frac{1}{\mu_0}\, i_g^{-1}\mathbf{B}, \qquad (H_i = \frac{1}{\mu_0}\, g_{ij} B^j). \tag{III.48}$$

The constant μ_0 is referred to as the *magnetic permeability* of the vacuum. This, too, is subject to vacuum polarization in the eyes of quantum electrodynamics.

When one integrates the 1-form H along a 1-chain c_1 the resulting number:

$$\mathcal{M}[c_1] = \int_{c_1} H \tag{III.49}$$

is called the *magnetomotive force*, which is also an unfortunate term, since the units of the quantity are work done per unit magnetic charge, not force.

If the 1-chain is a bounding 1-cycle ∂c_2 then Stokes's (or rather, Green's) theorem gives:

$$\mathcal{M}[\partial c_2] = \int_{c_2} dH = \int_{c_2} \#\mathbf{i} = I[c_2], \tag{III.50}$$

in which **i** is the *electric current density* – or electric charge flux density – in that region of Σ so $I[c_2]$ then represents the *total electric current* through c_2. In magnetostatics, this law is referred to as *Ampère's Law.*

If this relationship (III.50) is valid for all bounding 1-cycles then one can deduce the differential equation:

$$dH = \#\mathbf{i}\ . \tag{III.51}$$

If one expresses $H = \#\mathbf{H}$, which would make **H** a bivector field, then this latter equation can be put into the form:

$$\delta\mathbf{H} = \mathbf{i}\ . \tag{III.52}$$

In the macroscopic case, the vector field **i**, which represents the source of the **H** field, commonly takes the form:

$$\mathbf{i} = \rho\,\mathbf{v}, \tag{III.53}$$

in which ρ is the electric charge density and **v** is the velocity of the charge cloud. Both ρ and **v** then have the same support.

The vector field **i**, or equivalently **v**, defines a congruence of integral curves. Outside the support of **i**, the field equations (III.52) become:

$$\delta\mathbf{H} = 0. \tag{III.54}$$

In local form, with $\mathbf{B} = B^i\,\partial_i$, $H = H_i\,dx^i$, (III.51) becomes:

$$H_i = \tilde{\mu}_{ij}\,B^j, \qquad \tfrac{1}{2}(H_{i,j} - H_{j,i}) = -\,\varepsilon_{ijk}\,i^k, \tag{III.55}$$

and (III.52) becomes:

$$H^{ij} = \varepsilon^{ijk}\,H_k, \qquad H^{ki}{}_{,i} = i^k. \tag{III.56}$$

If one thinks of Σ as being the disjoint union of supp(**i**) and its complement $\Sigma' = \Sigma -$ supp(**i**) then the restriction of **H** to Σ′ is a de Rham homology class in dimension two. However, since the integral of #**H** around a 1-cycle in not in Σ′, which bounds in Σ, but Σ′, is non-zero, we must conclude that **H** is divergenceless, but not the divergence of a 3-vector field. This is analogous to the way that Coulomb **D** field (III.32), which is not defined at the origin, has zero divergence, even though its integral over any sphere centered at the origin is Q/ε_0 . In either case, when one removes the source points from Σ, the same 2-cycles that bound in Σ might no longer bound in Σ′, and Stokes's theorem no longer applies.

In the extreme case of a point charge moving along a single differentiable curve, Σ′ will be missing that curve, which might have the effect of introducing a non-trivial generator for the fundamental group of Σ′, as well as $H_1(\Sigma'; \mathbb{R})$. Hence, the source singularity can be regarded as a topological defect in Σ′ of vortex type.

Perhaps the fact that the elementary sources of electric fields are of monopole type and the elementary sources of magnetic fields are of vortex type might shed some light

on the reason for the continued absence of magnetic monopoles from the world of experimental physics.

8. Magnetostatic field equations. We can summarize the field equations of magnetostatics that we have obtained up to this point as differential equations:

$$\delta\mathbf{B} = 0, \qquad dH = \#\mathbf{i}, \qquad H = \mu^{-1}(\mathbf{B}). \tag{III.57}$$

However, for the sake of introducing a vector potential, which we will do in the next section, it is also convenient to give them the form:

$$dB = 0, \qquad \delta\mathbf{H} = \mathbf{i}, \qquad \mathbf{H} = \mu^{-1}(B), \tag{III.58}$$

in which $B = \#\mathbf{B}$ is now a 2-form and $\mathbf{H} = \#^{-1}H$ is a bivector field. The constitutive map $\mu^{-1}\colon \Lambda^2 \to \Lambda_2$, $B \mapsto \mathbf{H}$, is usually represented by the corresponding map of vector fields to 1-forms $\# \cdot \mu^{-1} \cdot \# : \Lambda_1 \to \Lambda^1$, $\mathbf{B} \mapsto H$, which one locally denotes by:

$$H_i = \mu_{ij}B^j. \tag{III.59}$$

Hence, we can represent the local components of the map on 2-forms in the form:

$$\mu^{ijkl} = \varepsilon^{ijm}\varepsilon^{kln}\mu_{mn}\,. \tag{III.60}$$

One can also express the differential equations in integral form:

$$\int_{\partial c_2} H = \int_{c_2} \#\mathbf{i}\,, \qquad \int_{\partial c_3} \#\mathbf{B} = 0\,. \tag{III.61}$$

The second of these admits the topological interpretation that the (singular) 2-cochain $\mathbf{\Phi_B}[.]$, which associates the total magnetic flux through a 2-chain, is a 2-cocycle. In terms of de Rham cohomology, this is equivalent to the statement that the 2-form $B = \#\mathbf{B}$ is closed, which is the first equation in (III.58). When $H_2(\Sigma; \mathbb{R})$ does not vanish, neither does $H^2(\Sigma; \mathbb{R})$ or $H^2_{dR}(\Sigma)$, and it is possible for there to be closed – but non-bounding – surfaces through which the total magnetic flux is non-zero. This is equivalent to the possibility that there are closed 2-forms B that are not exact. We shall return to this in the next section.

The first of equations (III.61) takes on a similar topological interpretation outside the support of $\mathbf{i}$, where the 1-cochain $\mathcal{M}[.]$ vanishes on every bounding 1-cycle. Hence, it too becomes a 1-cocycle, and the 1-form H becomes a closed 1-form; i.e., a de Rham cohomology class in dimension one. However, since we have changed the topology of Σ by deleting supp($\mathbf{i}$), we see that $H \in H^1_{dR}(\Sigma')$, and if the deletion of supp($\mathbf{i}$) renders this cohomology vector space non-trivial then the fact that H is closed but not exact implies

that $\mathcal{M}[z_1]$ is non-vanishing. Although Stokes's theorem does not apply when z_1 is not a boundary, one can treat the non-vanishing of $\mathcal{M}[z_1]$ as a topological source for the field H that surrogates for the missing current **i**.

9. Magnetostatic potential 1-forms. Now that we have drawn attention to the possible non-exactness of the closed 2-form B, let us consider the case in which B is exact. Recall that from the Poincaré lemma this is always possible locally about any point of Σ. Hence, we are assuming that there is a 1-form A such that:

$$B = \#\mathbf{B} = dA. \tag{III.62}$$

Such a 1-form is called a *potential 1-form* for B, or more imprecisely, a *vector potential*. It is clearly not uniquely defined since the addition of any closed 1-form α to A will produce the same B. This defines an equivalence relation on 1-forms, which one calls *gauge equivalence*, by the requirement that gauge-equivalent 1-forms A and A' must differ by a closed 1-form:

$$A' - A = \alpha \in Z^1(\Sigma). \tag{III.63}$$

If one is dealing with α locally, or if Σ is simply connected, then α is also exact – say, $\alpha = d\lambda$ – and one obtains the conventional form of a gauge transformation (of the second kind):

$$A \mapsto A + d\lambda\,. \tag{III.64}$$

From Stokes's theorem, the integral of A around a bounding 1-cycle ∂c_2 is:

$$\int_{\partial c_2} A = \int_{c_2} dA = \int_{c_2} B = \Phi_{\mathbf{B}}[c_2]; \tag{III.65}$$

i.e., it gives the total magnetic flux through the 2-chain.

When B is expressed in terms of a potential 1-form A, one can consolidate the field equations (III.58) into a single second-order equation:

$$(\delta \cdot \mu^{-1} \cdot d)A = \mathbf{i}\,. \tag{III.66}$$

If the action of μ^{-1} on 2-forms is locally expressed by the action of a matrix of functions of the form $\mu^{ijkl}(x, B)$ then (III.66) can be given the local form:

$$\tilde{\mu}^{ijkl}\frac{\partial^2 A_k}{\partial x^i \partial x^j} + \frac{\partial \mu^{ijkl}}{\partial x^i}\frac{\partial A_k}{\partial x^j} = i^l, \tag{III.67}$$

in which we have introduced:

$$\tilde{\mu}^{ijkl} = \mu^{ijkl} + \frac{\partial \mu^{ijkl}}{\partial B_{rs}} A_{r,s}\,. \tag{III.68}$$

When the medium is linear, homogeneous, and isotropic, such as the classical magnetic vacuum, one can express μ_{mn} as $1/\mu_0\ \delta_{mn}$, and (III.60) becomes:

$$\mu^{ijkl} = \frac{1}{\mu_0}\,\varepsilon^{ijm}\varepsilon^{klm} = \frac{1}{2\mu_0}\,(\delta^{ik}\delta^{jl} - \delta^{il}\delta^{jk}), \tag{III.69}$$

while $\tilde{\mu}^{ijkl} = \mu^{ijkl}$, $\mu^{ijkl}{}_{,i} = 0$, which makes (III.66) take the form:

$$\delta dA = \mu_0\,\mathbf{i}\,. \tag{III.70}$$

and (III.67) takes the form:

$$\delta^{ij}\,\frac{\partial^2 A^k}{\partial x^i \partial x^j} = \mu_0\, i^k. \tag{III.71}$$

We have implicitly raised the index of A_k by means of the Euclidian spatial metric that is naturally associated with any linear, isotropic (but not necessarily homogeneous) magnetic constitutive law.

When magnetostatics has the luxury of a spatial metric at its disposal – or, at least, a Hodge * operator that is defined in all dimensions of Λ^k – one can choose the gauge potential A to be one that has vanishing codifferential:

$$\delta A = 0, \tag{III.72}$$

which allows one to write (III.70) in the form of Poisson's equation for A:

$$\Delta A = \mu_0\, i\ ; \tag{III.73}$$

now, we have lowered the index on the source current vector field **i** in order to make it a 1-form.

Of course, in pre-metric magnetostatics, one must simply deal with (III.70) directly.

10. Field-source duality [3]. There has been a recurring theme in the foregoing presentation: As a general rule, the most elementary fields are not defined at their sources, and, as a result, one must deal with the topology of the space that is complementary to the region in which the source is defined. Frequently, this deletion will render the applicability of Gauss's theorem invalid, and one will be dealing with non-zero fluxes over k-cycles that are associated with vector fields that nonetheless have zero divergence.

For instance, Coulomb's law of electrostatics breaks down at the origin – if that is where the point charge is located. However, one notices that there is a further equivalence between the field that is exterior to a spherically-symmetric extended charge distribution and that of a point charge. Since a ball is contractible to a point, one sees that in the eyes of homotopy it is only the absence of a single point from the space in which the field exists that dictates the appearance of the field.

One can interpret the deletion of a point from a closed n-ball $\bar{B}(x;r)$ in a topological space Σ from either the standpoint of homotopy or homology. In the context of homotopy, the absence of a point from $\bar{B}(x;r)$ means that its boundary n–1-sphere is (probably) no longer homotopic to a point. Hence, there is at least one non-trivial generator to $\pi_{n-1}(\Sigma)$, depending upon what is happening with the topology of Σ, more globally. In the context of homology, it means that the boundary n–1-cycle is (probably) not the boundary of an n-chain, which implies that there is a non-trivial generator for $H_{n-1}(\Sigma; \mathbb{R})$.

Actually, when one goes to the case of one-dimensional field sources, such as line charges and currents in (non-closed) infinite wires, one finds that, up to homotopy, deleting a straight line from $\mathbb{R}^n$ is the same thing as deleting a point from $\mathbb{R}^{n-1}$. That is because the line being deleted is contractible to a point by a contraction that takes the rest of $\mathbb{R}^n$ to $\mathbb{R}^{n-1}$ minus the point that the line turns into. For instance, $\mathbb{R}^3$ minus the z-axis contracts to $\mathbb{R}^2$ minus the origin, and $\mathbb{R}^2$ minus the y-axis contracts to two non-zero points on the x axis.

Of course, $\mathbb{R}^n$ minus a circle is an entirely different matter. When $n = 2$, one sees that a plane minus a circle is homeomorphic to the disjoint union of an open disc and a plane minus a closed disc, which is then homotopically equivalent, by contraction, to the disjoint union of a point and a circle. When $n = 3$, one already sees that the resulting space is rather homotopically complicated. In addition to contractible loops, there are also non-contractible ones that encircle the missing circle, but one cannot retract the space to the circle without deleting – say – the z-axis, which must then be added as a "line at infinity."

When one deletes $\mathbb{R}^2$ from $\mathbb{R}^n$ the resulting space is homotopically equivalent to $\mathbb{R}^{n-2}$ minus a point. For instance, $\mathbb{R}^3$ minus a plane is homotopically equivalent to two points. This is useful in approaching the electric field of a charge distribution on an infinite plane.

Similarly to the situation in the second-to-last paragraph, the deletion of a 2-sphere – i.e., a basic 2-cycle – from $\mathbb{R}^n$ is not generally equivalent to the deletion of a plane. For instance, in the case of $\mathbb{R}^3$, the resulting space is contractible to the disjoint union of a point and a 2-sphere, which is reminiscent of the deletion of a circle from a plane, only with the next higher dimension of sphere. More generally, deleting an n-sphere from $\mathbb{R}^{n+1}$ will produce a space that is homotopically equivalent to a point and an n-sphere.

Since a point is a 0-cycle, we begin to see that the sources of electric and magnetic fields seem to take the form of a finite set $\{z_i, i = 1, \ldots, N\}$ of singular cycles of varying dimensions. In the most elementary case, this is a finite set of points. Furthermore, we can associate a number Q_i with each cycle that represents an electric charge in the case of

an electrostatic field and an electric current in the case of a magnetostatic field. Hence, one can form the linear combination:

$$Q = \sum_{i=1}^{N} Q_i z_i \,, \tag{III.74}$$

which is then a cycle of mixed dimension, in general, and we call this cycle the *source complex*. For instance, when one has a set of N charged points it will be a 0-cycle.

Hence, we shall always think of the field – say, **D** – that is produced by any source complex as being defined in the space $\Sigma - Q$ that is complementary to (the carrier of) Q in Σ. When we take the total flux $\Phi_{\mathbf{D}}[z_{n-1}]$ of **D** over an n–1-cycle z_{n-1} that bounds in Σ, but not in $\Sigma - Q$, we shall *define* the value of $\Phi_{\mathbf{D}}[z_{n-1}]$ to be Q, even though Gauss's theorem is in applicable in $\Sigma - Q$.

In fact, that is exactly what happens in the case of Coulomb's law: the field **D** is not defined at the point charge Q, but the total flux over any sphere that includes it in its interior equals Q, even though **D** has zero divergence. Since the total flux functional is a singular cocycle, in the event that **D** has vanishing divergence we shall call this coupling of the homological information in the source complex to the cohomological information in the field space *field-source duality.*

References

32. D.H. Delphenich, "On the Axioms of Topological Electromagnetism," Ann. d. Phys. (Leipzig) **14** 347 (2005), and hep-th/0311256.
33. D. H. Delphenich, "Nonlinear electrostatics: steps towards a neoclassical electron model," arXiv:0708.4194.
34. D. H. Delphenich, "Field/source duality in topological field theories," arXiv:hep-th/0702105.
35. F. Hehl and Y. Obukhov, *Foundations of Classical Electrodynamics*, Birkhäuser, Boston, 2003.
36. C. W. Misner and J. A. Wheeler, "Classical Geometry as Geometry," Ann. Phys. (New York) **2** (1957), 525-660.
37. N.D. Mermin, "The Topological Theory of Defects in Ordered Media," Rev. Mod. Phys., **51** (1979), 591-648.
38. L. Michel, "Symmetry Defects and Broken Symmetries. Configurations. Hidden Symmetries." Rev. Mod. Phys. **52** (1980), 617-651.
39. J.D. Jackson, *Classical Electrodynamics*, 2nd ed., Wiley, New York, 1976.
40. F. Rohrlich, *Classical Charged Particles*, Addison-Wesley, Reading, MA, 1965.

Chapter IV

Dynamic electromagnetic fields

Much of the modern fascination, if not obsession, with the unification of the theories of the fundamental forces of nature is probably traceable to the stellar success of the unification of the theories of electricity and magnetism into a single theory of electromagnetism, which was not only capable of explaining both theories as they existed independently up to that point in history, but was also capable of explaining natural phenomena that had not been recognized as being explainable in electromagnetic terms, namely, optical phenomena.

Here, we see a perfect example of the difference between a unification of theories and a mere "concatenation" of them. Generally, the unification of theories that pertain to phenomena that do not seem to be otherwise related will produce unpredicted consequences due to unexplored interactions between the individual theories. In general, one calls these interactions *inductions,* because the state of one field is inducing a response in the state of the other field.

1. Electromagnetic induction. One can distinguish static inductions from dynamic ones by the constraint that with a static induction the field itself induces a response, whereas with a dynamic response, it is one of the derivatives of the field that is provoking the response. For instance, a time-varying magnetic field induces a time-varying electric field. However, although it is true that a time-constant electric current will induce a time-constant magnetic field, nevertheless the converse statement is not true.

a. Faraday's law of induction [**1**]. Historically, the first dynamical coupling of electric and magnetic fields to be discovered was the coupling of time-varying magnetic fields to time-varying electric fields. More precisely, the first manifestation of that phenomenon was the coupling of time-varying magnetic fields to electric *currents* that was observed by Michael Faraday. However, if one regards the current in a conductor as being a response to the imposition of an electric field on the medium then one sees that it is more intrinsic to look at the coupling of time-varying magnetic fields and electric fields.

One first encounters Faraday's law of induction in the form of the coupling of time-varying magnetic *flux* to an induced *electromotive force:*

$$\mathcal{E}[\partial c_2] = -\frac{d\Phi_{\mathbf{B}}[c_2]}{d\tau}. \tag{IV.1}$$

In words, this says: When a loop ∂c_2 bounds a 2-chain c_2, the time-variation of the magnetic flux that links the 2-chain will induce an electromotive force in the loop that is 180 degrees out of phase with the time-variation in the magnetic flux.

When expressed in integral form, (IV.1) becomes:

$$\int_{\partial c_2} E = -\frac{d}{d\tau}\int_{c_2} \#\mathbf{B}\,, \tag{IV.2}$$

and an application of Stokes's theorem makes this take the form:

$$\int_{c_2} dE = -\frac{d}{d\tau}\int_{c_2} \#\mathbf{B}\,. \tag{IV.3}$$

Of course, when phrased in the manifestly topological form (IV.1), one notices certain features that are never discussed in a first exposure to Faraday's law:

1. The loop has to be a bounding 1-cycle. This leaves open the possibility that when the spatial manifold Σ is not simply connected there might non-zero emf's in loops that do not bound.

2. In some sense, we are defining a "differentiable curve" in the infinite-dimensional vector space $Z^2(\Sigma; \mathbb{R})$ that the 2-cocyle $\Phi_{\mathbf{B}}[.]$ lives in, although we have not said anything up till now about putting a topology or differential structure on it.

3. We now have a good example of the non-conservation of energy in an electromagnetic system that has nothing to do with dissipation in the thermodynamic sense. Basically, the reason that E cannot be exact anymore is because it is not closed and the non-vanishing of dE is due to the time-variation of $\Phi_{\mathbf{B}}[.]$. Hence, energy is being added to the loop in such a manner that the work done on a unit charge as it goes around the loop does not vanish.

4. If E were the covelocity 1-form for a fluid flow then one would say that the time-variation of $\Phi_{\mathbf{B}}[.]$ is acting like a torque that induces vorticity in the flow around the loop ∂c_2.

We see that what (IV.1) accomplishes is to effectively generalize our previous law of electrostatic fields, which becomes the limiting form of Faraday's law when the magnetic flux is constant in time.

Of course, one always learns that the minus sign in Faraday's law is a law unto itself, namely, *Lenz's law.* One can think of the induced emf as a sort of "damping" term that shows up to oppose the changes that create it, just as viscous damping in a fluid medium acts to oppose the change in position of an object immersed in the medium.

Even though Faraday's law is usually applied to sinusoidally changing magnetic fluxes, it works just fine for linearly and exponentially changing fluxes, as well. In particular, a linearly growing/decaying magnetic flux induces a constant emf, while an exponentially growing/decaying flux induces an exponentially growing/decaying emf.

Due to the derivative coupling, the actual magnitude of the change in flux does not have to be large to produce potentially enormous emf's. Indeed, a serious issue in most practical situations is what happens as the change in flux approaches a jump discontinuity, such as a switching transient, for which the magnitude of the derivative can grow quite large, even though the current itself is small.

Now suppose that the induced emf in ∂c_2 induces a current $I[\partial c_2]$, which is coupled to an induced magnetic flux $\Phi_{\mathbf{B}}[c_2]$ in the 2-chain that is bounded by the current loop. If the coupling of current in the loop to the resulting magnetic flux in the 2-chain it bounds is simply:

$$\Phi_{\mathbf{B}}[c_2] = L\, I[\partial c_2] \tag{IV.4}$$

in which the constant L is the *self-inductance* of the loop, then we can write Faraday's law as:

$$\mathcal{E}[\partial c_2] = -\, L\, \frac{d}{d\tau} I[\partial c_2]\,. \tag{IV.5}$$

Notice that now the coupling takes place solely on the bounding 2-cycle.

Part of the subtle profundity in Faraday's law is based in the fact the emf that is induced in a loop does not depend upon it being composed of a conducting medium, and, in fact, emf's can be induced in a vacuum, just the same. Of course, in that event one usually thinks in terms of induced E fields directly. Hence, we need to examine the process of going from the integral form of Faraday's law to its differential form.

Since the magnetic flux integral can be regarded as the evaluation $<\Phi_{\mathbf{B}}, c_2>$ of a linear functional on a vector – or rather, the evaluation $<\#\mathbf{B}, c_2>$ of a 2-form on a 2-chain – if one wishes to make the resulting number time-varying then one can make both $\mathbf{B}$ and c_2 time-varying, in their own right. However, since allowing c_2 to vary in time is more interesting to engineering applications, such as motors and generators, we shall simply consider the possibility that only $\mathbf{B} = \mathbf{B}(\tau, x^i)$ is a (differentiable) function of time.

This makes the magnetic flux derivative take the form:

$$\frac{d\Phi_{\mathbf{B}}[c_2]}{d\tau} = \int_{c_2} \frac{\partial(\#\mathbf{B})}{\partial\tau}\,. \tag{IV.6}$$

We can now write Faraday's law (IV.3) in the form:

$$\int_{c_2} [dE + \partial_\tau(\#\mathbf{B})] = 0, \tag{IV.7}$$

and if this is true for any possible 2-chain c_2 then one obtains the differential equation:

$$dE + \partial_\tau(\#\mathbf{B}) = 0. \tag{IV.8}$$

Of course, this means that we are regarding both E and $\mathbf{B}$ as time-varying fields, now.

b. Maxwell's law of induction [**1**]. It was Maxwell who put the capstone on the classical theory of electromagnetism by the intuitive leap of assuming that there is a partial converse to Faraday's law of induction that takes the form of saying that a time-varying electric flux through a 2-chain c_2 will induce a magnetomotive force in its boundary loop. However, this time the induction is in phase:

$$\mathcal{M}[\partial c_2] = +\frac{d\Phi_{\mathbf{D}}[c_2]}{d\tau}\,. \tag{IV.9}$$

In integral form, this is:

$$\int_{\partial c_2} H = +\frac{d}{d\tau}\int_{c_2} \#\mathbf{D}\,, \tag{IV.10}$$

and by an application of Stokes's theorem, this becomes:

$$\int_{c_2} dH = +\frac{d}{d\tau}\int_{c_2} \#\mathbf{D}\,. \tag{IV.11}$$

If (IV.10) is to serve as a generalization of Ampère's law (III.49) then we must add a contribution to the right-hand side that accounts for the current that is also producing the field:

$$\mathcal{M}[\partial c_2] = +\frac{d\Phi_{\mathbf{D}}[c_2]}{d\tau} + I[c_2], \tag{IV.12}$$

in which:

$$I[c_2] = \int_{c_2} \#\mathbf{i} \tag{IV.13}$$

is the total electric current through c_2 . Keep in mind that the actual support of the electric current density **i** will generally be different from c_2 , so the integral will involve only the intersection of those two sets.

Once again, if we consider only the possibility that **D** is time varying, but not c_2 itself, this leads to the differential equation:

$$dH - \partial_\tau \#\mathbf{D} = \#\mathbf{i}\,. \tag{IV.14}$$

It was this reciprocity in the coupling between time-varying electric fields and time-varying magnetic fields that eventually led to the possibility of wavelike solutions to the electromagnetic field equations, and the wave theory of optics.

2. Conservation of charge [**1**]**.** If we wish to go beyond the static perspective on electric and magnetic fields, we need to also go beyond the static perspective on their sources. Of course, we have actually made a first step in this direction by pointing out that electric currents already represent electric charges in a state of relative motion, but in order to produce static magnetic fields it was necessary to consider only time-invariant currents.

Now we shall consider the situation in its full generality, and assume that we have a time-varying electric charge density $\rho = \rho(\tau, x)$ that is defined on the spatial manifold Σ, or really, on $\mathbb{R} \times \Sigma$. On the support $\mathrm{supp}(\rho)$ of this density, we also have a time-varying vector field $\mathbf{v} = \mathbf{v}(\tau, x)$ that is defined, and which represents the flow velocity vector field of the charge distribution.

Together, these two data define another time-varying vector field:

$$\mathbf{i} = \rho\mathbf{v}, \tag{IV.15}$$

whose support is clearly the same as that of ρ and **i**. We call this vector field the *electric current density* for the distribution ρ.

The sense in which we think of the charge as being carried along by the motion defined by the vector field **v** – i.e., its congruence of integral curves – is defined by looking at the time rate of change for the total charge $Q[c_3]$ that is contained in a spatial 3-chain c_3. (By "spatial," we mean it is a 3-chain in $\tau_0 \times \Sigma$ for some fixed value τ_0 of τ.):

$$\frac{dQ[c_3]}{d\tau} = \frac{d}{d\tau}\int_{c_3} \#\rho = \int_{c_3}\left(\frac{\partial\rho}{\partial\tau}\right)\mathcal{V}\,. \tag{IV.16}$$

We couple this to **i** by the assumption that the physical meaning of the total flux of **i** through the boundary of c_3:

$$\Phi_{\mathbf{i}}[c_3] = \int_{\partial c_3} \#\mathbf{i} \tag{IV.17}$$

is the time rate at which charge is flowing out of the region described by c_3:

$$\frac{dQ[c_3]}{d\tau} = -\,\Phi_{\mathbf{i}}[\partial c_3]. \tag{IV.18}$$

This coupling of total charge with total charge current through the boundary is the most fundamental form of the *law of charge conservation*.

In integral form, this is:

$$\int_{c_3}\left(\frac{\partial\rho}{\partial\tau}\right)\mathcal{V} = -\int_{\partial c_3}\#\mathbf{i}\,, \tag{IV.19}$$

and an application of Stokes's theorem makes this become:

$$\int_{c_3}\left(\frac{\partial\rho}{\partial\tau} + \delta\mathbf{i}\right)\mathcal{V} = 0, \tag{IV.20}$$

after a few self-evident manipulations.

If (IV.20) is to be true for all possible spatial 3-chains then one must have the validity of the following differential equation:

$$\frac{\partial\rho}{\partial\tau} + \delta\mathbf{i} = 0. \tag{IV.21}$$

Hence, the differential form of (IV.18) amounts to the statement that the divergence of the vector field **i**, when restricted to the boundary of a 3-chain, is minus the time rate of change of the electric charge density at each point.

In local form, with $\mathbf{i} = \rho v^i \partial_i$, one has:

$$\frac{\partial\rho}{\partial\tau} + \frac{\partial(\rho v^i)}{\partial x^i} = 0. \tag{IV.22}$$

3. Pre-metric Maxwell equations. Let us summarize the equations that we have accumulated for the fields E, $\mathbf{D}$, H, and $\mathbf{B}$. In integral form, they are:

Gauss's law for $\mathbf{D}$:	$\Phi_{\mathbf{D}}[\partial c_3] = Q[c_3]$,	(IV.23a)
Gauss's law for $\mathbf{B}$:	$\Phi_{\mathbf{B}}[\partial c_3] = 0$,	(IV.23b)
Faraday's law of induction:	$\mathcal{E}[\partial c_2] = -\partial_\tau \Phi_{\mathbf{B}}[c_2]$,	(IV.23c)
Maxwell's law of induction:	$\mathcal{M}[\partial c_2] = +\partial_\tau \Phi_{\mathbf{D}}[c_2] + I[c_2]$,	(IV.23d)

and in differential form, they are:

Gauss's law for $\mathbf{D}$:	$\delta\mathbf{D} = \rho$,	(IV.24a)
Gauss's law for $\mathbf{B}$:	$\delta\mathbf{B} = 0$,	(IV.24b)
Faraday's law of induction:	$dE + \partial_\tau \# \mathbf{B} = 0$,	(IV.24c)
Maxwell's law of induction:	$dH - \partial_\tau \# \mathbf{D} = \# \mathbf{i}$.	(IV.24d)

Furthermore, we must add the constitutive laws that we have been using, so far, namely:

$$\mathbf{D} = \varepsilon(E), \qquad H = \mu(\mathbf{B}). \tag{IV.25}$$

Collectively, (IV.24a-d) represent the formulation of Maxwell's equations in terms of differential forms, at least in the eyes of non-relativistic physics. The first two equations basically express the divergences of $\mathbf{D}$ and $\mathbf{B}$, while the last two equations express their curls, if one takes into account the constitutive laws (IV.25) that couple them.

One of the most important advances to these classical equations came from the work of Lorentz, Poincaré, Minkowski, and others to give these non-relativistic equations a relativistic form (as one might find in [**2-8**]). Basically, one must replace the four-dimensional time + space manifold $\mathbb{R} \times \Sigma$, in which the time dimension really represents the proper time parameter τ that would be measured by a particular measurer/observer, with a more general four-dimensional spacetime manifold M. Now, the time dimension only shows up as one of the coordinates x^μ in a choice of coordinate chart (U, x^μ) about each point of M.

It was eventually established that the most intuitive and computationally useful formulation of Maxwell's equations on M came about by consolidating E and $\mathbf{B}$ together into a single 2-form on M:

$$F = dt \wedge E - \# \mathbf{B}, \tag{IV.26}$$

while consolidating D and $\mathbf{H}$ into the bivector field:

$$\mathfrak{H} = \partial_t \wedge \mathbf{H} + \#^{-1} D \, . \tag{IV.27}$$

We relate $\mathfrak{H}$ to F by a more general constitutive law than (IV.25):

$$\mathfrak{H} = \kappa(F) \, . \tag{IV.28}$$

We shall discuss the nature of four-dimensional constitutive laws in the next chapter, but, for now, we just treat it as a diffeomorphism of each fiber of $\Lambda^2(M)$ at each $x \in M$ with the corresponding fiber of $\Lambda_2(M)$ at the same point.

Furthermore, we consolidate the sources ρ and **i** into a single four-dimensional current vector:

$$\mathbf{J} = \rho\, \partial_t + \mathbf{i}\,. \tag{IV.29}$$

With these consolidations, the Maxwell equations can be expressed in the form:

$$dF = 0, \qquad \delta\mathfrak{h} = \mathbf{J}, \qquad \mathfrak{h} = \kappa(F). \tag{IV.30}$$

One sees that conservation of charge follows as an unavoidable consequence:

$$\delta\mathbf{J} = 0. \tag{IV.31}$$

One can also treat this equation as an integrability – or compatibility – condition for a given electric current four-vector field **J** to be the source of a bivector field **H**.

If one expresses the 2-form F, the bivector field $\mathfrak{h}$, and the electric current **J** in local form as:

$$F = \tfrac{1}{2} F_{\mu\nu}\, dx^\mu \wedge dx^\nu, \qquad \mathfrak{h} = \tfrac{1}{2} H^{\mu\nu}\, \partial_\mu \wedge \partial_\nu, \qquad \mathbf{J} = J^\mu\, \partial_\mu \tag{IV.32}$$

then Maxwell's equations take on their familiar local form:

$$F_{\mu\nu,\kappa} + F_{\nu\kappa,\mu} + F_{\kappa\mu,\nu} = 0, \qquad H^{\mu\nu}{}_{,\nu} = J^\mu, \qquad H^{\mu\nu} = \kappa(F_{\mu\nu}). \tag{IV.33}$$

When the constitutive map k is linear, the last equation can be expressed in local form as:

$$H^{\mu\nu} = \kappa^{\mu\nu\alpha\beta} F_{\alpha\beta}\,. \tag{IV.34}$$

We point out that at no point has it been necessary to introduce a metric of any signature in order to define the system of equations (IV.30), only a volume element. The usual role of the Lorentzian metric g in the formulation of Maxwell's equations in terms of differential forms is purely confined to the definition of the Hodge duality isomorphism *, and only for 2-forms, moreover.

In order to see that we have essentially replaced the necessity for g by means of the constitutive map κ; we first rephrase the middle equation in (IV.30) as:

$$d\#\kappa(F) = \#\mathbf{J}\,. \tag{IV.35}$$

If we consider the case of a linear map κ and define *: $\Lambda^2 M \to \Lambda^2 M$, $F \mapsto \#\kappa(F)$ then we see that * is an isomorphism of 2-forms with 2-forms. We can then put (IV.35) into the form:

$$d{*}F = \#\mathbf{J}. \tag{IV.36}$$

This is consistent with the metric form of Maxwell's equations that gets presented in most texts on relativistic electrodynamics, as long as our isomorphism * has one of the key properties of the Hodge isomorphism, at least, as it acts on 2-forms on a four-dimensional Lorentzian manifold, namely:

$$*^2 = -I, \tag{IV.37}$$

which actually means that the * isomorphism defines an "almost-complex structure" on the vector bundle $\Lambda^2 M$. We shall have more to say about this later, but, for now, we point out that generally all that one can assume is that:

$$*^2 = \#\kappa\#\kappa. \tag{IV.38}$$

The integral form of Maxwell's equations is then:

$$\Phi_F[\partial c_3] = 0, \quad \Phi_{\mathfrak{h}}[\partial c_3] = \Phi_{\mathbf{J}}[c_3], \qquad \mathfrak{h} = \kappa(F). \tag{IV.39}$$

Once again, we see that we can give topological interpretations to the field equations of electromagnetism (IV.30), just as we did for the static field equations of electric and magnetic fields individually. Now, we see that F must be a de Rham cocycle in dimension 2, while $\mathfrak{h}$, outside the support of the source current $\mathbf{J}$, is a de Rham cycle in dimension two.

4. Electromagnetic potential 1-forms. The first of Maxwell's equations – viz., $dF = 0$ – takes the form of saying that the 2-form F is closed. Depending upon whether $H^2_{dR}(M)$ is trivial or not, F might be globally exact or locally exact, resp. That is, there will be a 1-form A such that:

$$F = dA. \tag{IV.40}$$

One calls this 1-form an *electromagnetic potential 1-form*, or a choice of *gauge* for the electromagnetic field F.

a. Gauge equivalence. As before, A is not unique because any other 1-form A' that differs from A by a closed 1-form z^1 will give the same 2-form F under exterior differentiation. The relation $A \sim A'$ iff $dA = dA'$ iff $A - A' \in Z^1_{dR}(M)$ is an equivalence relation that is usually called *gauge equivalence*. We denote the set of all gauge equivalent 1-forms (mod F) by $\mathcal{A}(M; F)$.

Note that two cohomologous closed 2-forms F and F' will generally differ by an exact 2-form $d\alpha$; hence, if $F = dA$ and $F' = dA'$ then:

$$F - F' = d(A - A') = d\alpha \neq 0. \tag{IV.41}$$

Since α is not closed A and A' cannot be gauge equivalent. That is, cohomologous 2-forms will not generally have gauge-equivalent potential 1-forms.

It is not true that if $A \sim A'$ then any linear combination $\alpha A + \beta A'$ will be an element of the same gauge equivalence class as A and A', since:

$$d(\alpha A + \beta A') = (\alpha + \beta)\, F, \tag{IV.42}$$

which is not generally equal to F. Hence, $\mathcal{A}(M; F)$ is not a vector space. However, the fact that the expression $A - A' \in Z^1_{dR}(M)$ is well-defined implies that it is an *affine* space that is modeled on the vector space $Z^1_{dR}(M)$. The fact that it is infinite-dimensional follows from the fact that at long as M has finite-dimensional de Rham cohomology in dimension one the dimension of $Z^1_{dR}(M)$ is the same as the dimension of $B^1_{dR}(M)$, and since each of its elements take the form $d\lambda$ for some smooth function λ, the dimension of $B^1_{dR}(M)$ is the same as the dimension of $C^\infty(M; \mathbb{R})$, which is denumerably or non-denumerably infinite depending upon whether M is compact or not, respectively.

Since gauge equivalence takes the local form $A - A' = d\lambda$, one can think of this as defining a *gauge transformation (of the second kind)* by the replacement $A \to A + d\lambda$. In order to define the gauge transformations of the first kind, one must express λ in the form $\lambda = \ln g$, which makes:

$$d\lambda = g^{-1}\, dg\,. \tag{IV.43}$$

Since multiplication by real numbers is commutative, one can express the gauge transformation of the second kind as the replacement of A with:

$$A' = g^{-1}Ag + g^{-1}\, dg. \tag{IV.44}$$

although the excess complexity in the expression is, of course, unnecessary for the Abelian case at hand. It does, however, become non-trivial for non-Abelian gauge theories.

c. Field equations in terms of a potential. If $F = dA$ then the first of equations (IV.30) becomes an identity. The constitutive law makes $\mathfrak{h} = \kappa(dA)$, and this puts the non-trivial differential equation of the set into the form:

$$\Box_\kappa A = (\delta \cdot \kappa \cdot d)A = \mathbf{J}. \tag{IV.45}$$

This means that the operator $\Box_\kappa$: $\Lambda^1 \to \Lambda_1$ plays the role of a "pre-metric d'Alembertian." Ordinarily – i.e., in the metric theory, which provides a Hodge * isomorphism in all dimensions, not just dimension two – we could say that this operator *equals* the d'Alembertian operator by the Lorentz choice of gauge for A, namely $\delta A = 0$. However, in the absence of a metric, we cannot generally define the codifferential operator for δA to make any sense.

If we represent A in a local coordinate chart as $A_\mu\, dx^\mu$, $\mathbf{J}$ as $J^\mu\, \partial_\mu$, and κ by means of functions $\kappa^{\mu\nu\alpha\beta}(x, F)$ then the field equation (IV.45) takes the local form:

$$\frac{\partial}{\partial x^\mu}\left(\kappa^{\mu\nu\alpha\beta}\, \frac{\partial A_\alpha}{\partial x^\beta} \right) = J^\nu, \tag{IV.46}$$

and, upon carrying out the partial derivatives, it becomes:

$$\tilde{\kappa}^{\mu\nu\alpha\beta}\, \frac{\partial^2 A_\alpha}{\partial x^\mu \partial x^\beta} + \frac{\partial \kappa^{\mu\nu\alpha\beta}}{\partial x^\mu}\, \frac{\partial A_\alpha}{\partial x^\beta} = J^\nu, \tag{IV.47}$$

with:

$$\tilde{\kappa}^{\mu\nu\alpha\beta} \equiv \kappa^{\mu\nu\alpha\beta} + \frac{\partial \kappa^{\mu\nu\alpha\beta}}{\partial F_{\kappa\lambda}}\, \frac{\partial A_\kappa}{\partial x^\lambda}\,. \tag{IV.48}$$

c. U(1) gauge structures [**9**, **10**]. Because the expression (IV.44) looks so formally similar to the way that the local representative of a connection 1-form A on a $U(1)$-principal bundle $P \to M$ would transform from one local trivialization over $U \subset M$ to another one over $V \subset M$ that overlaps the first, we briefly summarize the physical roots of the rest of that construction.

An element of the Abelian Lie group $U(1)$ can be represented by a complex number of the form $e^{i\theta}$. Hence, an element of its Abelian Lie algebra will take the form $i\theta$. Our group of gauge transformations of the second kind can then be represented by the group $C^\infty(U, U(1))$ of smooth maps from U to $U(1)$.

A $U(1)$-principal bundle $P \to M$ has a fiber over each point of M that looks like $U(1)$, which is topologically a circle. However, as usual, the diffeomorphism of each fiber with $U(1)$ is not canonical, and one finds that there are generally inequivalent $U(1)$-principal bundles over a given M. In particular, they are not all generally trivial – i.e., equivalent to the projection $M \times U(1) \to M$ – since that is equivalent to saying that there is a global section of the bundle.

The physical meaning of a local section $\phi\colon U \to P$ is that it represents a local choice of $U(1)$ gauge for an electromagnetic field F, which we now assume to be a 2-form on P. By means of ϕ, one can then pull F down to a 2-form $\phi^* F$ on U. If $F = dA$ then if one interprets the 1-form A on P as a $\mathfrak{u}(1)$-connection 1-form then this would make F the curvature 2-form that it defines (since the Lie algebra u(1) is Abelian). The 1-form A also pulls down to a 1-form $\phi^* A$ on U.

If $V \subset M$ is an open subset that overlaps U and supports a local section $\psi\colon V \to P$ then on the overlap $U \cap V$ there is a transition function $g\colon U \cap V \to U(1)$ that allows one to transform from one choice of local gauge to the other by way of:

$$\psi(x) = g(x)\phi(x). \tag{IV.49}$$

This is what one commonly calls a *gauge transformation of the first kind.*

One finds that the pull-down of A transforms to $A' = \psi^* A$ by way of:

$$A' = g^{-1}Ag + g^{-1}\,dg = A + d\lambda, \tag{IV.50}$$

and the curvature 2-form F transforms to:

$$F' = g^{-1}Fg = F. \tag{IV.51}$$

In the eyes of modern differential geometry [**11**], this implies that it is possible to interpret electromagnetism as involving basic constructions that are consistent with the geometry of a $U(1)$-principle bundle over spacetime that has been given a choice of connection. In the vocabulary of modern theoretical physics, one calls such a theory a $U(1)$ *gauge theory* of the interaction in question.

We pointed out that the triviality of a G-principal bundle $P \to M$ is equivalent to the existence of a global section of the fibration. The branch of topology called obstruction theory [**12**, **13**] gives the obstruction to the extension of a local section to a global one in terms of the non-vanishing of a certain cocycle called the *primary obstruction cocycle* whose dimension is one more than the dimension of the first non-vanishing homotopy group $\pi_k(G)$ of the fiber – namely, G – and whose coefficient group is $\pi_k(G)$. In the case of $G = U(1)$ the first, and only, non-vanishing homotopy group is $\pi_1(U(1)) = \pi_1(S^1) = \mathbb{Z}$, so the primary obstruction cocycle for a $U(1)$-principal bundle over a manifold M is a cocycle $c_1 \in H^2(M; \mathbb{Z})$ that one calls the *first Chern class* for P. The gist of the *Chern-Weil homomorphism* is that this integer cocycle can be represented in de Rham cohomology by the 2-form:

$$c_1 = \frac{1}{2\pi} F\,. \tag{IV.52}$$

As it turns out, the choice of $\mathfrak{u}(1)$-connection 1-form A is irrelevant, because any other possible choice A' would give a curvature F' that is cohomologous to F.

In the case of $M = S^2$, the issue of triviality leads to the consideration of Dirac monopoles, although not in the form that Dirac described. Suppose one has a a $U(1)$ principal bundle $P \to S^2$. Basically, one cannot cover the 2-sphere with a single local trivialization of P, but must cover it with two overlapping hemispheres U_N and U_S centered on the North and South pole. One assumes, moreover, that the overlap $U_N \cap U_S$ is homotopic to a circle.

We will call two local sections of the fibration over the two hemispheres ϕ_N and ϕ_S. There must be a transition function $g_{NS}: U_N \cap U_S \to U(1)$ that relates them be way of:

$$\phi_N = g_{NS}\,\phi_S\,. \tag{IV.53}$$

Since $U_N \cap U_S$ is homotopic to S^1, as is $U(1)$, the homotopy classes $[g_{NS}]$ are all elements of $\pi_1(S^1) = \mathbb{Z}$. Hence, there is a denumerable sequence of homotopically inequivalent transition functions, which also implies a corresponding sequence of homotopically inequivalent $U(1)$-principal bundles over S^2. In fact, the integers that index this sequence of bundles can be obtained by a simple integration of c_1 over S^2:

$$c_1[P] = \frac{1}{2\pi}\int_{S^2} F \,. \qquad \text{(IV.54)}$$

This integer $c_1[P]$ is called the *first Chern number* for P, and since it happens to equal to the *Euler-Poincaré characteristic* [19] of S^2 in this case, equation (IV.54) is another form of the *Gauss-Bonnet theorem*, which is what Chern was originally attempting to generalize when he was eventually led to define the characteristic classes that bear his name. For a trivial $U(1)$-principal bundle over S^2 the first Chern class will vanish, so a non-vanishing first Chern number is necessary and sufficient for the non-triviality of such a bundle.

The way that this relates to magnetic monopoles is that the right-hand side of (IV.54) is also proportional to the total magnetic flux though S^2 when one thinks of F as an electromagnetic field strength 2-form, hence, the total magnetic charge contained in it, assuming that S^2 bounds a closed 3-ball.

An interesting point to ponder, since magnetic monopoles have not been experimentally demonstrated, and even seem to contradict one's basic intuition about the nature of magnetic field sources, is whether it is actually necessary that one assume that the gauge transformation $A \mapsto A + d\lambda$ actually has to imply that one dealing with $U(1)$ indeed. Basically, it is only a statement about Lie *algebras*, and there is another Lie *group* that has a Lie algebra that is isomorphic to $\mathbb{R}$, namely, the additive group $(\mathbb{R}, +)$ of real numbers. The usual argument for going the route of $U(1)$ is purely reasoning by analogy with wave mechanics, on the assumption that gauge transformations of the first kind work the same way for potential 1-forms as they do for quantum wave functions. However, if this analogy were weak, and the actual gauge group was $(\mathbb{R}, +)$, not $U(1)$, that would account for the non-existence of magnetic monopoles, since $\mathbb{R}$ is contractible, so all of its homotopy groups vanish, and there is no obstruction to a global section of any $\mathbb{R}$-principal bundle; i.e., they are always trivial. Then again, the experimentally-verified validity of the Bohm-Aharonov effect suggests that $U(1)$ might be the correct choice, after all.

References

41. J.D. Jackson, *Classical Electrodynamics,* 2nd ed., Wiley, New York, 1976
42. F. Rohrlich, *Classical Charged Particles,* Addison-Wesley, Reading, MA, 1965.
43. A.O. Barut, *Electrodynamics and Classical Theory of Fields and Particles*, Dover, NY, 1980..
44. W. Thirring, *Classical Field Theory,* Springer, Berlin, 1978.

[19] The Euler-Poincaré characteristic $\chi[M]$ of a topological space is the alternating sum $\sum (-1)^k b_k$ of its Betti numbers b_k, which, in our case, will be the dimensions of the de Rham cohomology vector spaces in each dimension $k = 0, 1, \ldots, \dim(M)$. For a 2-sphere, the only non-vanishing Betti numbers are $b_0 = b_2 = 1$, so $\chi[S^2] = 2$. The fact that $\chi[S^2]$ is non-zero is also the reason that there are no everywhere non-zero vector fields on S^2, which is a special case of the *Poincaré-Hopf* theorem that a differentiable manifold M admits a global non-zero vector field iff $\chi[M] = 0$.

45. E. J. Post, *Formal Structure of Electromagnetics*, Dover, NY, 1997.
46. F. Hehl and Y. Obukhov, *Foundations of Classical Electrodynamics,* Birkhäuser, Boston, 2003.
47. L.D. Landau and E.M. Lifshitz, *Classical Field Theory,* Pergamon, Oxford, 1975.
48. D.H. Delphenich, "On the Axioms of Topological Electromagnetism," Ann. d. Phys. (Leipzig) **14** 347 (2005), and hep-th/0311256.
49. T. Eguchi, P. Gilkey, and A. Hanson, "Gravitation, Gauge Theories, and Differential Geometry," Phys Rep. **66** (1980), 213-393.
50. T. Frankel, *The Geometry of Physics: an introduction*, Cambridge University Press, Cambridge, 1997.
51. S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry,* Interscience, London, 1964.
52. N. Steenrod, *The Topology of Fiber Bundles,* Princeton Univ. Press, Princeton, 1951.
53. J. Milnor and J. Stasheff, *Characteristic Classes,* Princeton Univ. Press, Princeton, 1974.

Chapter V

Electromagnetic constitutive laws

Although the subject of electromagnetic constitutive laws seems to be best established for the response of macroscopic matter to electromagnetic fields, it seems to be gradually emerging from quantum electrodynamics that there is much more internal structure to the electromagnetic vacuum state than can be conveniently described by the constants ε_0 and μ_0 . Since the nature of that state is apparently beyond the limits of direct observation, one must hope that it is permeable to intuitive probes that are based on analogies with more established macroscopic phenomena. Therefore, we shall first discuss some of the issues that pertain to macroscopic matter and then summarize the most established aspects of the quantum electrodynamical vacuum state that suggest analogies with nonlinear optics.

1. Electromagnetic fields in macroscopic matter [1-4]. As pointed out previously, the application of an electromagnetic field F to a macroscopic medium will imply the application of forces and torques to the elementary charges and dipoles – both electric and magnetic – in the medium. The main issue for a given medium is the degree of freedom that those elementary charges and dipoles have in the medium. This, in turn, reverts to the question of how they interact.

Any model for macroscopic matter, such as one finds in condensed matter physics [**5**], is going to be based in some empirical approximation to the interaction of the elementary constituents of the medium. For instance, the ideal gas model assumes that the gas molecules all have zero volume and interact only by elastic collisions. Hence, the limits of the model for the medium are usually defined by the limits of the interaction model for its constituents. One often finds that the nature of phase transitions in such media is closely related to the transition from one interaction model to another. For instance, the ideal gas model begins to break down when the molecules get close enough for inter-atomic forces to take over, which suggests a high gas density, as one would expect during condensation from the gaseous phase to the liquid phase. Similarly, when a gas goes to the plasma state the outer electrons of the gas molecules are no longer bound to the molecules, so the interaction model for the electrons and gaseous ions must change accordingly.

The elementary charges in a macroscopic electromagnetic medium come in two basic forms: electrons and atomic ions. The electrons are free to move in response to applied electromagnetic fields to a greater or lesser degree that mostly determines whether the medium is regarded as a conductor, insulator, or something more exotic, such as a semiconductor. That is, the translational response of atomic electrons to applied electric and magnetic forces will take the form of an electric current that flows in the medium.

Atomic ions in a solid medium are often bound to each other into a crystal lattice, so they do not generally translate very far from their equilibrium positions, but they are still capable of exhibiting collective vibrational modes of response, as well as electronic

energy level transitions. Atomic ions can appear in non-solid media such as electrolytic solutions and in the plasma state.

Elementary electric dipoles are most commonly associated with rather complex molecules, such as one finds in plastic dielectrics, but also in the silicon dioxide that mostly constitutes many optical media, such as quartz and the various glasses. It is no coincidence that optically transparent media are not generally conductors, although we shall not go into the details here.

Elementary magnetic dipoles are mostly associated with the angular momenta, both orbital and intrinsic, of elementary charges. However, in the case of both nuclear and electronic magnetic moments, it is the intrinsic angular momentum – i.e., *spin* – that tends to dominate in macroscopic phenomena over the orbital angular momentum. In particular, the response of most magnetic media to applied magnetic fields is generally due to the degree of freedom that the electron spins have in the medium.

As mentioned above, the electromagnetic properties of macroscopic media tend to be partitioned according to the phase of the medium. This, in turn, is often related to the magnitudes of the field strengths involved. One simply has to consider the possible phase transitions that might come about when field strengths exceed critical limits. For instance, in most optical media, the propagation of waves involves the responses of the atomic electrons to the impinging photons. One can see that there will be various types of photons in that regard: photons that bring about level transitions and result in absorption or stimulated emission of photons, photons of sufficient energy to bring about the ionization of electrons, photons that get Compton scattered by the electrons, etc. Hence, it is clear the optical character of the medium will change dramatically when the energy of the photon – i.e., the field strengths of its electric and magnetic fields – exceeds the ionization limit or has the proper frequency to initiate stimulated emission, as in lasers, or perhaps when it equals a level transition that would cause absorption.

In this chapter, which is more phenomenological than the previous ones, we shall start with the linear electromagnetic media and then go on to the nonlinear ones. From that study, we shall comment upon some of the possible analogies with the quantum electrodynamical vacuum state. In all of the discussion, the primary objective will be to obtain possible forms for the electromagnetic constitutive law that shows up in the pre-metric Maxwell equations.

2. Linear constitutive laws [1-4]. Even though linearity is invariably an approximation in physics, nonetheless, it can also be such a useful approximation to a wide class of natural phenomena – which usually amounts to the realm of weak field strengths in the case of electromagnetism – that we still need to discuss the linear case. Indeed, one finds that the extension to nonlinearity is so fraught with complexity that it is only by starting with the firm intuitive foundation that one gains from the linear phenomena that one can even hope to choose a fruitful direction of mathematical generalization into the domain of nonlinear phenomena.

We shall first deal with the most general case of non-local linear constitutive laws, so that we can give the local laws their proper context. We will then discuss some of the classes of natural phenomena that pertain to the structure of local linear constitutive laws.

We shall also be considering only non-conducting media, so the applied electromagnetic field will not be assumed to produce an electric current in the medium.

a. Non-local linear laws. In its most general form, when an electromagnetic medium exhibits a linear response to an applied electromagnetic field strength $F \in \Lambda^2 M$ the resulting bivector field $\mathfrak{h} \in \Lambda_2 M$ is related to F by an linear operator κ: $\Lambda^2 M \rightarrow \Lambda_2 M$ that is not necessarily local. That is, it does not induce linear maps on the individual fibers of these vector bundles at each point of M. Rather, it can generally be represented as an integral operator:

$$\mathfrak{h}(x) = \int_M K(x, y) \wedge F(y), \tag{V.1}$$

in which K: $M \times M - \Delta \rightarrow \Lambda_2 M \otimes \Lambda^{n-2}\mathrm{M}$, $(x, y) \mapsto K(x, y) \in \Lambda_{2,x} \otimes \Lambda^{n-2}_y$ is a kernel function for the operator K that is not, however, defined on the diagonal Δ of $M \times M$; viz., the set of all (x, x). One way of characterizing the kernel function is to think of it as the *impulse-response* function for the medium. That is, use $F(y) = \delta(y)$ and note that the resulting $\mathfrak{h}(x)$ will equal $K(x, 0)$.

This non-locality takes two basic forms: temporal and spatial.

In the temporal case, one must consider the fact that a real-world medium will not respond to an applied field instantaneously, nor will its response cease immediately after the applied field itself vanishes. There will generally be a certain time delay before the medium begins to respond, as well as a certain decay envelope after the impulse ceases. One must also add causality constraints on the kernel to insure that $\mathfrak{h}$ is not responding to the future state of the field F.

Nonlocality in the spatial case is based on the notion that generally neighboring dipoles, whether electric or magnetic, are not completely independent of each other, but generally have some sort of interaction. Often, this interaction takes the form of essentially a "nearest-neighbor" interaction, as one finds in the Heisenberg ferromagnet. Spatial nonlocality does not involve a causality constraint in the same way as temporal nonlocality, but one might imagine that the state of a given dipole could not depend upon the state of dipoles that are not sufficiently close that electromagnetic waves could pass between them within a lightlike time interval.

If M is a compact region in $\mathbb{R}^4$ and we assume that K is translation-invariant, so it can be expressed in the form $K(x - y)$, then we can take the Fourier transform of (V.1) to get an algebraic equation in frequency-wave number space, which we express in component form for a choice of coordinates on $\mathbb{R}^4$:

$$\mathfrak{h}^{\mu\nu}(\omega, k_i) = \tfrac{1}{2} K^{\mu\nu\alpha\beta}(\omega, k_i)\, F_{ab}(\omega, k_i). \tag{V.2}$$

Hence, we are now dealing with a complex 2-form $F \in \Lambda^2(\mathbb{R}^{4*})$ on frequency-wave number space $\mathbb{R}^{4*}$, a complex bivector field $\mathfrak{h} \in \Lambda_2(\mathbb{R}^{4*})$, and a complex linear map K from one to the other, which is now local. In general, temporal nonlocality will manifest

itself as a frequency dependence of the components of K, and spatial nonlocality will imply wave-number dependence. The appearance of non-zero real parts to the components of K is related to the possibility that the medium might tend to absorb electromagnetic energy.

b. Local linear laws. From the preceding discussion, it should be clear that there is a fundamental approximation associated with passing from nonlocal linear constitutive laws to local linear ones, just as there is a fundamental approximation that is associated with the restriction to linear laws, in the first place. Basically, the approximation is that the temporal impulse response must involve a short time delay and a quick decay afterwards, while the spatial response must be confined to only a sufficiently small neighborhood of each elementary dipole.

In dealing with the general form of the component matrix for a local linear constitutive map κ, it is generally more convenient to ignore the specific nature of the elements of the vector spaces $\Lambda^2_x M$ and $\Lambda_{2,x} M$ as exterior products of other vector spaces and treat them as simply six-dimensional real vector spaces, at least when M is four-dimensional. Furthermore, since most of what we shall be discussing is local and linear in character, we shall consider only the vector spaces $A^2(\mathbb{R}^4)$ and $A_2(\mathbb{R}^4)$, which serve as typical fibers of the vector bundles $\Lambda^2 M$ and $\Lambda_2 M$.

Hence, a basis for A_2 will consist of six linearly independent vectors $\{\mathbf{b}_I\,,\ I = 1, \ldots, 6\}$ (which are really *bi*vectors) and a basis for A^2 will consist of six linearly independent covectors$\{b^I\,,\ I = 1, \ldots, 6\}$ (which are really algebraic 2-forms). As in any other vector space, they will be reciprocal to each other iff $b^I(\mathbf{b}_J) = \delta^I_J$.

If $\{\mathbf{e}_\mu,\ \mu = 0, \ldots 3\}$ represents a basis for $\mathbb{R}^4$ and $\{\theta^\mu,\ \mu = 0, \ldots, 3\}$ its reciprocal basis for $\mathbb{R}^{4*}$ then one can define a basis for A_2 by means of:

$$\mathbf{b}_i = \mathbf{e}_0 \wedge \mathbf{e}_i\,, \qquad \mathbf{b}_{i+3} = \tfrac{1}{2}\,\varepsilon_{ijk}\,\mathbf{e}_j \wedge \mathbf{e}_k, \qquad i, j, k = 1, 2, 3 \tag{V.3}$$

and one for A^2 by means of:

$$b^i = \theta^0 \wedge \theta^i\,, \qquad b^{i+3} = \tfrac{1}{2}\,\varepsilon^{ijk}\,\theta^j \wedge \theta^k. \tag{V.4}$$

It is easy to see that these bases are themselves reciprocal.

For instance, in this basis we can represent the electromagnetic field strength F as:

$$F = E_i\, b^i + B_i\, b^{i+3}\,. \tag{V.5}$$

Because the bases seem to split evenly into three members that have a common exterior factor of $\mathbf{e}_0$ (θ^0, resp.) and three members that do not involve it, we shall find it convenient, as well as mathematically and physically significant, to represent A_2 as the direct sum $A_2^{\mathrm{Re}} \oplus A_2^{\mathrm{Im}}$ and A^2 as the direct sum $A^2_{\mathrm{Re}} \oplus A^2_{\mathrm{Im}}$ of two three-dimensional real vector spaces. Our notations "Re" and "Im" are suggestive of "real" and "imaginary,"

respectively, although we have more to say about why that is actually appropriate in Chapter XII.

When a basis has been chosen for both A_2 and A^2, any linear transformation L: $A^2 \to A_2$, $\alpha \mapsto L(\alpha)$ can be represented by a matrix L^{IJ} with respect to these bases that is defined by:

$$L(b^I) = L^{IJ}\, \mathbf{b}_J \,. \tag{V.6}$$

In block-matrix form, the matrix of a local linear constitutive map κ looks like:

$$\kappa^{IJ} = \left[\begin{array}{c|c} \alpha^{ij} & \beta^{ij} \\ \hline \gamma^{ij} & \eta^{ij} \end{array}\right]. \tag{V.7}$$

So far, we have said nothing concerning the symmetry of the indices I and J. Indeed, one can first regard them as having no specified symmetry. However, one can polarize κ^{IJ} into a sum $\kappa^{IJ}_+ + \kappa^{IJ}_-$ of a symmetric part κ^{IJ}_+ and an anti-symmetric part κ^{IJ}_- , where:

$$\kappa^{IJ}_+ = \tfrac{1}{2}(\kappa^{IJ} + \kappa^{JI}), \qquad\qquad \kappa^{IJ}_- = \tfrac{1}{2}(\kappa^{IJ} + \kappa^{JI}) \,. \tag{V.8}$$

However, κ^{IJ}_+ can be decomposed further, since the inverse $\#^{-1}$ of the Poincaré duality isomorphism also has a matrix that is symmetric:

$$\tilde{\#}^{IJ} = \left[\begin{array}{c|c} 0 & I \\ \hline I & 0 \end{array}\right]. \tag{V.9}$$

In the terminology of Hehl and Obukhov [**3**], the *principal part* of κ^{IJ} is then defined to be the part of κ^{IJ}_+ that does not include this contribution:

$${}^{(1)}\kappa^{IJ} = \kappa^{IJ}_+ - \tfrac{1}{6}\mathrm{Tr}(\kappa^I_J)\,\tilde{\#}^{IJ} = \left[\begin{array}{c|c} -\varepsilon^{ij} & \gamma^{ij} \\ \hline \overline{\gamma}^{ji} & \tilde{\mu}^{ij} \end{array}\right]. \tag{V.10}$$

In the second term, we have used $\kappa^I_J \equiv \#_{JK}\, \kappa^{KI}$.

The sub-matrix ε^{ij}, which is Hermitian, represents the *electric permittivities* of the medium, which generalize the dielectric constant; the reason for the minus sign will become clear shortly. The matrix $\tilde{\mu}^{ij}$, which is also Hermitian, represents the inverse of the matrix of *magnetic permeabilities* of the medium. The remaining off-diagonal matrices γ^{ij} and its Hermitian transpose $\overline{\gamma}^{ji}$ represent the contribution of the magnetic field B_i to the electric excitation D^i and that of the electric field E_i to the magnetic excitation H^i. We shall discuss possible origins for these contributions in a bit.

The anti-symmetric part of κ – namely, ${}^{(2)}\kappa^{IJ} = \kappa^{IJ}_{-}$ – is called the *skewon* part of κ and the remaining "trace-class" contribution ${}^{(3)}\kappa^{IJ} = \frac{1}{6}\mathrm{Tr}(\kappa^{I}_{J})\tilde{\#}^{IJ}$ is called its *axion* part, which represents a contribution from the volume element on $\mathbb{R}^4$.

3. **Examples of local linear media**. It will prove essential in what follows to have some specific examples of constitutive laws to consider, so we shall show how one accounts for some of the most commonly used electromagnetic media in the aforementioned formalism. One thing that should be self-evident is that when one is referring to the constancy of matrix components this unavoidably begs the question of "in what frame?" In most cases the answer is simply: the rest frame of the medium itself. Of course, in the case of massless media, such as the electromagnetic vacuum – whether classical or quantum – it is meaningless to speak of a rest frame for the medium, which was, of course, the problem with the former concept of the "ether," and one must then accept the more ambiguous answer: the frame of some measurer/observer.

a. Linear isotropic media. The most elementary electromagnetic media of all are the *linear isotropic media*, which include the classical vacuum itself. The term "isotropic" refers to invariance under spatial rotations in the chosen frame, such as the rest frame of the medium, when it is not massless. At the elementary level, this generally means that the electric and magnetic dipoles of the medium must be capable of aligning themselves to an imposed field just the same way at any point and in any direction.

Such media are characterized by the fact that κ^{IJ} agrees with its principal part, so it is symmetric, and its constituent submatrices are of the form:

$$\varepsilon^{ij} = \varepsilon\, \delta^{ij}, \qquad \tilde{\mu}^{ij} = (1/\mu)\, \delta^{ij}, \qquad \gamma^{ij} = 0. \tag{V.11}$$

The function ε is referred to as the *electric permittivity* of the medium. The function μ is its *magnetic permeability*. When these functions are constant in the chosen frame one calls the medium *homogeneous*, as well. The constant ε is usually referred to as the *dielectric constant* of the medium, in that case.

The classical electromagnetic vacuum is assumed to be linear, isotropic, and homogeneous; one denotes its dielectric constant by ε_0 and its magnetic permeability by μ_0 . We shall denote its constitutive map by κ_0, and it will play a recurring role in what follows as essentially an asymptotic state that one finds in the absence of electromagnetic fields.

b. Linear optical media. The next level of generality is defined by the *linear optical media*, which are basically like (V.11), except that one allows ε^{ij} to possibly be inhomogeneous or anistropic. In regular optical practice, the most common inhomogeneity that one treats involves matrix components that are piecewise constant and undergo jump discontinuities on the interfaces between dissimilar media. However, an important example of continuous inhomogeneity is defined by the variation of the optical properties of some transparent media under stress. Since the magnetic

permeability of an optical medium generally plays no role, one customarily thinks of them as dielectric media; i.e. $\tilde{\mu}^{ij} = (1/\mu_0)\ \delta^{ij}$, with μ_0 a constant that is usually set to unity.

When ε^{ij} is symmetric, but anisotropic, it is customary to define the *principal frame*, in which it becomes the diagonal matrix diag[ε_x , ε_y , ε_z]. The diagonal elements, which are the *principal permittivities*, are the eigenvalues of the matrix $\varepsilon^i_j = \delta_{jk}\ \varepsilon^{ik}$, and the vectors of the principal frame (in $\mathbb{R}^3$) are its normalized eigenvectors. Because ε^{ij} is symmetric, these eigenvalues exist and are real, and the eigenvectors associated with distinct eigenvalues must be orthogonal. Note that one implicitly introduces the Euclidian metric on the spatial manifold in order to speak of eigenvectors and eigenvalues.

There are three basic ways that the eigenvalues of ε^{ij} can equal or differ from each other:

1. When they are all equal, one calls the medium *isotropic*.
2. When two are equal, and the third one differs, the medium is called *uniaxial*. (The axis in question belongs to the distinct eigenvalue.)
3. When all three are unequal, the medium is called *biaxial*.

Generally, the symmetry type of a dielectric is traceable to the symmetry type of the crystal lattice of the material that it is composed of.

The complementary class to that of optical media is defined by *magnetic media*, for which $\varepsilon^{ij} = \varepsilon_0\ \delta^{ij}$, where ε_0 is regarded as constant, but $\tilde{\mu}^{ij}$ is possibly inhomogeneous or anisotropic. For magnetic media, one has an analogous notion of a principal frame for the symmetric matrix $\tilde{\mu}^{ij}$, whose eigenvalues are $1/\mu_x$, $1/\mu_y$, $1/\mu_z$, are then *principal permeabilities.* Clearly, the principal frame for $\tilde{\mu}^{ij}$ will be the same as the principal frame for its inverse matrix μ_{ij} , and the eigenvalues of the inverse matrix will be μ_x , μ_y , μ_z .

It is important to understand that the physical nature of electric and magnetic polarization in a medium suggests that there is no reason to believe that in the general case where ε^{ij} and $\tilde{\mu}^{ij}$ are both anisotropic they will have a common principal frame. Basically, this is equivalent to the condition that the matrices commute under multiplication, which is always the case when one of them is a scalar multiple of δ^{ij}, but not true in general. When both ε^{ij} and $\tilde{\mu}^{ij}$ have a common principal frame, we shall call such a medium *electromagnetically diagonalizable*. In such a frame, one has:

$$\kappa^{IJ} = \text{diag}[-\varepsilon_x , -\varepsilon_y , -\varepsilon_z, 1/\mu_x , 1/\mu_y , 1/\mu_z]. \qquad \text{(V.12)}$$

In all of the cases for which $\alpha^{ij} = 0$ one can summarize the constitutive relations by the conventional three-dimensional component equations:

$$D^i = -\varepsilon^{ij} E_j , \qquad B_i = \mu_{ij} H^j . \qquad \text{(V.13)}$$

c. Bi-isotropic media. An elementary example of a medium in which the off-diagonal matrices can be non-vanishing is that of bi-isotropic media (see Lindell [**4**]), which are like isotropic media in their diagonal matrices, but have:

$$\gamma^{ij} = \overline{\gamma}^{ji} = \gamma\,\delta^{ij}\,. \tag{V.14}$$

In this case, the off-diagonal matrices are symmetric, and the part of κ^{IJ} is $\gamma\tilde{\#}^{IJ}$.

As we shall see, these media include the case of greatest interest in quantum electrodynamics, namely, the Heisenberg-Euler media.

d. Lorentzian media. Recall that the crucial step in pre-metric electromagnetism is the replacement of the purely geometric tensor field $\imath_g \wedge \imath_g$ that represents the isomorphism of the vector space of 2-forms with the vector space of bivectors at each point of the spacetime manifold with a purely physical – indeed, as we have seen, largely phenomenological – tensor field that represents the response of the medium to the presence of electric and magnetic fields. It is then heuristically useful to examine the form that the isomorphism $\imath_g \wedge \imath_g$ takes when expressed in terms of the *I, J* indices in order to compare the effect that it has to that of an electromagnetic constitutive law.

One simply has to start with the expression for the raising of two indices:

$$\mathfrak{h}^{\mu\nu} = g^{\mu\alpha}\,g^{\nu\beta}\,F_{\alpha\beta} = \tfrac{1}{2}(g^{\mu\alpha}\,g^{\nu\beta} - g^{\mu\beta}\,g^{\nu\alpha})F_{\alpha\beta} \tag{V.15}$$

and expand:

$$\mathfrak{h}^{0i} = (g^{00}\,g^{ij} - g^{0i}\,g^{0j})\,F_{0j} + \tfrac{1}{2}(g^{0j}\,g^{ik} - g^{0k}\,g^{ij})\,F_{jk}\,, \tag{V.16a}$$

$$\mathfrak{h}^{ij} = \tfrac{1}{2}(g^{i0}\,g^{jk} - g^{0j}\,g^{ik})\,F_{0k} + \tfrac{1}{2}(g^{ik}\,g^{jl} - g^{il}\,g^{jk})\,F_{kl}\,. \tag{V.16b}$$

If we set:

$$F_{0i} = E_i\,, \qquad F_{ij} = \varepsilon_{ijk}B^i, \qquad \mathfrak{h}^{0i} = D^i, \qquad \mathfrak{h}^{ij} = \varepsilon^{ijk}H_k \tag{V.17}$$

then (V.16a, b) take the form:

$$D^i = -\varepsilon^{ij}\,E_j + \gamma^i_j B^j, \qquad H_i = \gamma^j_i\,E_j + \tilde{\mu}_{ij}\,B^j, \tag{V.18}$$

as long as we set:

$$\varepsilon^{ij} = g^{0i}\,g^{0j} - g^{00}\,g^{ij}, \qquad \gamma^i_j = \varepsilon_{klj}\,g^{0k}\,g^{il}, \qquad \tilde{\mu}_{ij} = \varepsilon_{nmi}\,\varepsilon_{klj}\,g^{nk}\,g^{ml}. \tag{V.19}$$

Of particular interest is the case of the Minkowski space metric:

$$g^{\mu\nu} = \eta^{\mu\nu} = \mathrm{diag}[+1, -1, -1, -1], \tag{V.20}$$

which makes:

$$\varepsilon^{ij} = \delta^{ij}, \qquad \gamma^i_j = 0, \qquad \tilde{\mu}_{ij} = \delta_{ij}\,. \tag{V.21}$$

Hence, we can put κ^{IJ} into the form:

$$\kappa^{IJ} = \left[\begin{array}{c|c} -\delta^{ij} & 0 \\ \hline 0 & \delta_{ij} \end{array}\right] \tag{V.22}$$

in this case; we now see the reason for the inclusion of the minus sign in the general form (V.10) for ${}^{(1)}\kappa^{IJ}$ above.

Note that one cannot generally solve (V.19) for the $g^{\mu\nu}$ when given the ε^{ij}, γ^{i}_{j}, $\tilde{\mu}_{ij}$ since the manifold Lor(4) of all $g^{\mu\nu}$ is 10-dimensional, whereas the manifold Cons(4) of all ε, μ, and γ, in which the first two are symmetric and the last one has no specified symmetry, is 6 + 6 + 9 = 21-dimensional. Therefore, not all possible combinations (ε, μ, γ) will be consistent with the mapping Lor(4) $\rightarrow$ Cons(4), $g \mapsto (\varepsilon, \mu, \gamma)$ that that is defined by (V.19), and one must specify 11 compatibility conditions, in order to define the image of that map as a submanifold of Cons(4).

Actually, there is an increasing body of literature, both mathematical and physical (see, e.g., [**6**]), that is based on the fact that the principal part of a linear electromagnetic constitutive law defines a fourth-rank totally covariant tensor field κ on the spacetime manifold of a symmetry type that is sometimes called an *algebraic curvature tensor* [**7**], which also defines an *area metric* – i.e., a metric on the vector bundle $\Lambda^2 M$. We shall return to discuss this latter aspect of k in Chapters XII and XIII, but for now, we simply point out that a key result is the *Gilkey decomposition* [**7**]:

$$\kappa_{\kappa\lambda\mu\nu} = \sum_{(i)=1}^{N} Z_{(i)} \left[{}^{(i)}g_{\kappa\mu}\, {}^{(i)}g_{\lambda\nu} - {}^{(i)}g_{\kappa\nu}\, {}^{(i)}g_{\lambda\mu} \right], \tag{V.23}$$

in which the $Z(i)$, $i = 1, \ldots, N$ are frame-invariant scalar functions and the ${}^{(i)}g_{\kappa\lambda}$ are metric tensor fields. This decomposition is not, however, unique, since, at the very least one can choose other scalar factors.

We shall refer to an electromagnetic medium whose constitutive law takes the form of (V.19) for some Lorentzian metric as a *Lorentzian* medium.

e. Almost-complex media [**8**]. If one left-multiplies κ^{IJ} in (V.22) by the matrix $\#_{IJ}$ that is inverse to (V.9) then one gets a matrix:

$$*^{I}_{J} = \#_{JK}\, \kappa^{KI} = \left[\begin{array}{c|c} 0 & I \\ \hline -I & 0 \end{array} \right] \tag{V.24}$$

that represents a linear isomorphism $*: \Lambda^2 \rightarrow \Lambda^2$ with the property that:

$$*^2 = -I; \tag{V.25}$$

In particular, it is the Hodge duality isomorphism that is associated with the Lorentzian metric, as it acts on 2-forms.

Now, a linear isomorphism of an even-dimensional real vector space with itself that has the property (V.25) defines a *complex structure* on that vector space, because multiplication by the imaginary i has that same effect on the vectors in a complex vector space. We shall have much more to say about the role of complex structures in pre-

metric electromagnetism in Chapter XII, but, for now, we confine ourselves to the aspects that pertain to constitutive laws.

Since an electromagnetic constitutive law κ is supposed to replace the * isomorphism that one derives from a Lorentzian structure, it is natural to *postulate* that it have the same property (V.25). However, in the simplest case of an isotropic medium (V.11), since:

$$\tilde{\kappa}^I_J = \left[\begin{array}{c|c} 0 & (1/\mu)I \\ \hline -\varepsilon I & 0 \end{array}\right], \qquad (\tilde{\kappa} \equiv \# \cdot \kappa) \tag{V.26}$$

one must have:

$$\tilde{\kappa}^2 = -\frac{\varepsilon}{\mu} I\,. \tag{V.27}$$

Hence, it probably more prudent to postulate that there is some function λ on the spacetime manifold M such that:

$$\tilde{\kappa}^2 = -\,\lambda^2\, I\,. \tag{V.28}$$

One can then define * by normalization:

$$* = (1/\lambda)\,\tilde{\kappa}\,. \tag{V.29}$$

The function λ can be shown to have the dimension of an admittance (i.e., 1/impedance), and for the vacuum it has the numerical value of 1/377 mhos.

An immediate consequence of the condition (V.28) is that it implies further restricting conditions on the submatrices that represent $\tilde{\kappa}$. If κ agrees with its principal part, so its matrix has the form (V.10), then:

$$\tilde{\kappa} = \left[\begin{array}{c|c} \gamma^T & \tilde{\mu} \\ \hline -\varepsilon & \gamma \end{array}\right]. \tag{V.30}$$

From the condition (V.28), a direct evaluation of $\tilde{\kappa}^2$ gives a set of four matrix equations that reduce to just two:

$$\varepsilon = \lambda^2 \mu + \mu\,\gamma^2, \qquad \mu\gamma^T = \gamma\mu\,. \tag{V.31}$$

Here, we see that this constraint is actually more restrictive than it sounds like it would be. If we assume that $\gamma = 0$ then we are left with the constraint that the matrix ε must be proportional to the matrix μ. Hence, the medium must have the same symmetry under spatial rotations for both its electric and magnetic properties; i.e., both must be isotropic, uniaxial, or biaxial, resp. Whereas this is certainly the case for the classical vacuum, and any other isotropic space, it is not generally true of most anisotropic optical media, since one assumes that they are magnetically isotropic and homogeneous.

For a bi-isotropic medium, the second equation in (V.31) is trivial, and the first says that one must have:

$$\varepsilon = (\lambda^2 + \gamma^2)\,\mu \tag{V.32}$$

for some λ. Once again, ε must be proportional to μ, but the factor of proportionality depends upon γ.

Nonetheless, we shall see that media for which κ satisfies (V.33), which we call *almost-complex media*, provide a considerable wealth of applications for the methods of complex projective geometry, which is also closely suggested by the use of 2-forms and bivector fields.

f. Purely axionic media. A *purely axionic* electromagnetic medium will have a constitutive tensor field of the form:

$$\kappa = \alpha \left[\begin{array}{c|c} 0 & I \\ \hline I & 0 \end{array}\right], \tag{V.33}$$

which makes:

$$* = \alpha I. \tag{V.34}$$

This sort of relationship between F and $H = *F$ is strongly analogous to a variation on Ohm's law for coupling current to voltage in a two-port electrical network that was obtained by Tellegen [**9**, **10**]. For the network that he devised, which he called a "gyrator" instead of the usual $V = IR$ rule, he obtained:

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = s \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}, \tag{V.34}$$

in which i_a, v_a, $a = 1, 2$ are the currents and voltages at the two ports, respectively, while s has the dimensions of a resistance.

Physical examples of electromagnetic media that exhibit such laws also exist. For instance, Lindell and Sihvola [**10**] have described such media as *perfect electromagnetic conductors*, and the emerging class of *metamaterials* [**11**] appears to offer the promise of implementing such a law in a tangible medium.

g. Magneto-electric materials [**12**]. As a contribution to an electromagnetic constitutive law, and not a law unto itself, an axion part represents one type of magneto-electric coupling, in addition to the contributions of skewon type. In the next section, we shall discuss how magneto-electric couplings can come about as a result of frame transformations, and some of the experimentally established physical that exhibit them, but, for now, we cite as an example of a *magneto-electric material* the much-studied anti-ferromagnetic material chromium sesquioxide (Cr_2O_3).

It gives a * isomorphism whose matrix has the block matrix form:

$$[*] = \left[\begin{array}{c|c} \varepsilon & 0 \\ \hline 0 & \tilde{\mu} \end{array}\right] + \alpha \left[\begin{array}{c|c} 0 & I \\ \hline -I & 0 \end{array}\right], \tag{V.35}$$

in which:

$$\varepsilon = \varepsilon_0 \begin{bmatrix} \varepsilon_\perp - \alpha_\perp^2/\mu_\perp & 0 & 0 \\ 0 & \varepsilon_\perp - \alpha_\perp^2/\mu_\perp & 0 \\ 0 & 0 & \varepsilon_\parallel - \alpha_\parallel^2/\mu_\parallel \end{bmatrix}, \qquad \tilde{\mu} = \frac{1}{\mu_0} \begin{bmatrix} 1/\mu_\perp & 0 & 0 \\ 0 & 1/\mu_\perp & 0 \\ 0 & 0 & 1/\mu_\parallel \end{bmatrix}, \tag{V.36}$$

$$\alpha = \frac{1}{3}\left(2\frac{\alpha_\perp}{\mu_\perp} + \frac{\alpha_\parallel}{\mu_\parallel} \right) \sqrt{\frac{\varepsilon_0}{\mu_0}}\,. \tag{V.37}$$

Hence, such a medium has no skewon contribution, and its magneto-electric coupling is solely due to the axionic part.

Experiments with Cr_2O_3 [**12**] have verified that $\alpha \neq 0$.

h. Fresnel-Fizeau effect [**1**, **2**]. It has long been established that magneto-electric couplings in an electromagnetic medium can appear in as a result of the propagation of electromagnetic waves in a massive electromagnetic medium that is in a state of motion relative to the measurer/observer. This effect of relative motion on such a medium was first studied by Fresnel and later by Fizeau; it was also discussed by Einstein in his early work on special relativity.

Basically, the effect of a relative velocity **v** on the medium is to couple magnetic fields to electric ones and vice versa in a manner that is analogous to the Lorentz force, at least up to first order. All that one needs to do is subject the **E** and **B** field to a Lorentz transformation that corresponds to a transformation to a frame that has a relative velocity of **v**. One finds (see Jackson [**13**], Landau and Lifschitz [**14**]) that the transformation of fields is, to first order:

$$\mathbf{E}' = \gamma\mathbf{E} + \gamma c\, \mathbf{v} \times \mathbf{B}, \qquad \mathbf{B}' = \gamma\mathbf{B} - \gamma c\, \mathbf{v} \times \mathbf{E}\,. \tag{V.38}$$

Hence, the effect of the relative velocity is to make:

$$\gamma^{ij} = 1/c\ \varepsilon^{ijk} v_k = \frac{1}{c} \begin{bmatrix} 0 & -v_z & v_y \\ v_z & 0 & -v_x \\ -v_y & v_x & 0 \end{bmatrix}. \tag{V.39}$$

Note that although κ^{IJ} remains symmetric the individual off-diagonal matrices γ^{ij} and $\overline{\gamma}^{ji}$ are anti-symmetric.

i. Plasmas [**15-18**]. One of the most important examples of an electromagnetic medium is defined by the plasma state. This state, which comes about from the ionization of gases as a result of high temperatures or applied electric fields, is characterized by an ensemble of (usually) two oppositely charged fluids that are collectively neutral. The positively-charged fluid, which is composed of the atomic or molecular ions, has a low mobility due to the higher masses of the ions, while the negatively-charged fluid is composed of free electrons, which will then have a high degree of mobility. Examples of plasmas are found at all levels of scale, including

possibly the scale of strongly-interacting particles. Commonly, one encounters them in the form of flames, glowing gases in discharge tubes, such as neon lights, the Earth's ionosphere, as well as the solar atmosphere, and even interstellar space seems to be composed of a very dilute plasma whose ions are mostly protons – i.e., hydrogen ions. Since the scope of plasma physics is so vast as to touch upon most of the established branches of physics, we shall confine our remarks to only those aspects of plasmas that affect the electromagnetic constitutive properties and later, the propagation of electromagnetic waves.

The key aspect of plasmas that affects their electromagnetic constitutive properties is the fact that high electron mobility implies high conductivity. Hence, in the usual Maxwell equations for electromagnetic fields in plasmas, one must add an electric current $\mathbf{i}_\sigma = \mathbf{i}(E)$ that comes about in response to the applied electric field. Customarily, one assumes that the response is linear but not necessarily isotropic or homogeneous, so the induced current obeys Ohm's law in the form:

$$i^i_\sigma = \sigma^{ij}(t, x)E_i \,, \tag{V.40}$$

in which the matrix:

$$\sigma^{ij} = \begin{bmatrix} \sigma_x & \sigma_{xy} & \sigma_{xz} \\ -\sigma_{xy} & \sigma_y & \sigma_{yz} \\ \sigma_{xz} & -\sigma_{yz} & \sigma_z \end{bmatrix} \tag{V.41}$$

is the electrical conductivity of the plasma and is generally complex.

For as *lossless* plasma one will have that σ is anti-Hermitian (i.e., $\sigma^\dagger = -\sigma$). In this case, the elements σ_{xy} and σ_{yz} will be real, while the others are imaginary.

Generally, the origin of the anisotropy in the conductivity of a plasma is the presence of a background magnetic field, which then introduces a preferred direction and an orbital motion to the charged particles, although the electrons will be dynamically affected by the magnetic field more than the ions.

The electric current due to the applied E can be absorbed into the dielectric constant of the plasma to give an effective dielectric constant:

$$\varepsilon^{ij} = \varepsilon_0\, \delta^{ij} + \frac{1}{i\omega}\,\sigma^{ij} \,. \tag{V.42}$$

Hence, ε^{ij} will have the same symmetry as σ^{ij}, except that for lossless plasmas, one will have that ε^{ij} is Hermitian (i.e., $\varepsilon^\dagger = \varepsilon$).

In the case of an isotropic plasma, such as when no background magnetic field is present, one has that $\sigma^{ij} = \sigma\delta^{ij}$, with:

$$\sigma = -\, i\frac{ne^2}{m_e} \,, \tag{V.43}$$

in which m_e is the electron mass and n is the number density of the electrons.

This makes the effective dielectric constant take the form:

$$\varepsilon(\omega) = \left(1 - \frac{ne^2}{m_e \omega}\right)\varepsilon_0 . \qquad \text{(V.44)}$$

One immediately sees that the dependence of the dielectric constant on the frequency of the electric field that is present is such that it will vanish when that frequency equals:

$$\omega_p(n) = \left(\frac{4\pi ne^2}{m_e}\right)^{1/2} = 2p(8980)\sqrt{n}\,, \qquad \text{(V.45)}$$

and is referred to as the *plasma frequency*.

The process of electric polarization in a medium with considerable charge mobility implies the formation of a counter-electric field that tends to oppose the applied field. For instance, in the close proximity of a charged electrode in a plasma one can expect that there will be a shielding layer with a characteristic thickness that one calls the *Debye screening length:*

$$\lambda_D = \left(\frac{\varepsilon_0 KT_e}{ne^2}\right)^{1/2} . \qquad \text{(V.46)}$$

The fact that this length depends upon the temperature T_e of the electrons, in addition to their number density n, derives from the fact that when the electrons have sufficiently high temperature they can escape the electrostatic well that is created by the screening effect.

Note that such a statistical argument is applicable only when the total number N of electrons in the shielding layer is appreciable. This also defines a characteristic number, in the form of the total number of free electrons in a "Debye sphere":

$$N_D = n\left(\frac{4}{3}\pi\lambda_D^3\right) = 1.38\times 106\left(\frac{T^3}{n}\right)^{1/2} \qquad (T \text{ in } {}^{o}\text{K}). \qquad \text{(V.47)}$$

Actually, since the number density of the free electrons is not constant, but is assumed to vary in response to the ambient electric and magnetic fields, one must add a further equation to the Maxwell equations that is based in the Boltzmann equation of physical kinetics [**18**] and expresses the balance law for the number density in terms of these ambient fields. This equation is often called the *Vlasov equation.*

4. Some physical phenomena due to magneto-electric couplings. The presence of magneto-electric couplings in an electromagnetic medium can give rise to various phenomena that are well-documented in the world of experimental physics and sometimes the most powerful tools that physics – especially astrophysics – has in determining the nature of the medium that electromagnetic waves are propagating through. Although we shall discuss the propagation of electromagnetic waves in a later

chapter, for now we shall mostly concentrate on the way that these effects manifest themselves in changes to the constitutive properties of the medium.

a. Faraday effect [**1**, **2**]. The way that electromagnetic waves propagate through an electromagnetic medium can be influenced by the presence of "external" electric and magnetic fields in the medium, in addition to those of the wave itself. For instance, in astrophysics, one generally has to deal with the presence of galactic magnetic fields that might affect the propagation of photons from the stars of the galaxy in question.

In particular, the presence of a magnetic field can bring about a rotation of the polarization planes (which is generated by the **E** and **k** vectors) within the tangent spaces. This is because an imposed magnetic field, which we assume to point in the positive *z*-direction, will affect both the dielectric and magnetic properties of the medium, and either effect is called the *Faraday effect.*

The dielectric Faraday effect changes the ε^{ij} matrix of an isotropic dielectric with permittivity ε to one of the form:

$$\varepsilon^{ij} = \begin{bmatrix} \varepsilon' & 0 & 0 \\ 0 & \varepsilon & i\varepsilon_{yz} \\ 0 & -i\varepsilon_{yz} & \varepsilon \end{bmatrix}. \tag{V.48}$$

Note, in particular, that the contribution is imaginary and the medium is no longer isotropic, but uniaxial.

Similarly, the magnetic Faraday effect changes the $\tilde{\mu}^{ij}$ matrix for an isotropic magnetic medium with a permeability of μ to one of the form:

$$\tilde{\mu}^{ij} = \begin{bmatrix} 1/\mu' & 0 & 0 \\ 0 & 1/\mu & i\mu_{yz} \\ 0 & -i\mu_{yz} & 1/\mu \end{bmatrix}. \tag{V.49}$$

b. Natural optical activity [20]. In Landau, Lifschitz, and Pitaevski [**1**], the phenomenon of natural activity is described as something that occurs in optical media with no center of symmetry and takes the form of allowing ε^{ij} to be a function of both ω and **k**.

To first order, this takes the form:

$$\varepsilon^{ij}(\omega, \mathbf{k}) = \varepsilon_0^{ij}(\omega) + i\gamma^{ijk} k_k . \tag{V.50}$$

Hence, the effect of natural optical activity is equivalent to the dielectric Faraday effect for a background magnetic field in an optically inactive medium.

However, in Post [**2**] the effect of natural optical activity is to make:

[20] Since the mathematical representations of this phenomenon that were given in Landau, et al [**1**] and Post [**2**] are inconsistent, we shall defer to the former reference.

$$\gamma^{ij} = -\overline{\gamma}^{ji} = i\gamma\delta^{ij} . \qquad \text{(V.51)}$$

That is, it affects the off-diagonal matrices in κ^{IJ}, not the off-diagonal terms in ε^{ij} itself.

5. Nonlinear constitutive laws [19-21]. Since the appearance of linearity in physics is invariably based in some simplifying empirical approximation, such Hooke's law, Ohm's law, or others, the consideration of nonlinear phenomena is also invariably the most promising mathematical horizon for further exploration in physics.

Generally, the transition from the linear regime to the nonlinear regime is indexed by some magnitude, such as displacement, temperature, or field strength. Quite often, a complicating factor that can drastically change the nature of a system's response from linear to nonlinear is the possibility of a phase change. For instance, in the case of Ohm's law (viz., $I = \Delta V/R$) one finds that current I flowing through a resistor R in response to an applied voltage difference ΔV will cause it to heat up, which increases the resistance. That, in its own right, would make the current through the resistor related to the voltage drop across the resistor in a nonlinear manner, but ultimately the heat will bring about complete vaporization of the resistor, as in fuses. This then represents the sort of catastrophic nonlinearity that one associates with phase transitions.

In the case of electromagnetic constitutive laws, since the elementary electric and magnetic dipoles are associated with more complicated systems, such as atoms and crystal lattices, it is not surprising that linear constitutive laws are just as much of an approximation as in any other linear law of nature. For instance, one can imagine that an intense laser beam in a transparent plastic fiber will have a similar effect to a high current in a resistor, and at a threshold level of intensity the fiber will melt or vaporize.

As with linear effects, one can distinguish between non-local nonlinear effects and local nonlinear effects. An example of a non-local nonlinear effect that is quite common in electromagnetics is *magnetic hysteresis*. The idea is that as one increases the magnitude B of $\mathbf{B}$ in most magnetic media, the resulting H field will involve a time lag in its response that depends upon the magnitude B in such a way that if one decreases the value of B back to its original value then the medium will respond to the same values of B differently. This situation is illustrated in Fig. 4.

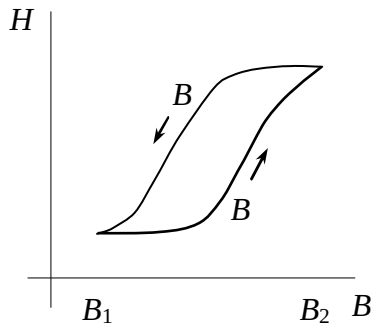


Figure 4. Magnetic hysteresis.

The same sort of situation occurs when an incandescent filament heats up at one rate as the current in it increases and then cools down at a different rate when the current decreases, so instead of the straight-line plot of Ohm's law one has something more like Fig. 4 for I vs. ΔV.

a. Electromagnetic susceptibilities. There is a regime between linearity and the onset of some phase transition that is sometimes referred to as "weak nonlinearity." It is characterized by the possibility of representing the diffeomorphism κ: $A^2 \to A_2$, $F \mapsto \mathfrak{h} = \kappa(F)$ by the first terms of a Taylor series in F:

$$\mathfrak{h}(F) = \mathfrak{h}_0 + d\kappa|_0(F) + \tfrac{1}{2} d^2\kappa|_0(F, F) + \ldots \tag{V.52}$$

which can be expressed in coordinates as:

$$\mathfrak{h}^{\mu\nu}(F_{\alpha\beta}) = \mathfrak{h}_0^{\mu\nu} + \left.\frac{\partial \mathfrak{h}^{\mu\nu}}{\partial F_{\alpha\beta}}\right|_{F=0} F_{\alpha\beta} + \frac{1}{2}\left.\frac{\partial^2 \mathfrak{h}^{\mu\nu}}{\partial F_{\alpha\beta}\partial F_{\gamma\delta}}\right|_{F=0} F_{\alpha\beta}F_{\gamma\delta} + \ldots \tag{V.52}$$

Actually, the leading term in this expansion, which we shall suggestively call the *zero-point field*, represents a possible source of nonlinearity, in the sense that it makes the linear relationship that is defined by the second term into an affine relationship. In the case of magnetism the presence of a non-zero constant term in **H** is usually attributed to *ferromagnetism*; in the case of electric fields, it is called *ferroelectricity*.

For the purposes of nonlinear optics, it is usually preferable to use a Taylor expansion in the electric polarization vector field **P**, which relates to **E** by way of:

$$\mathbf{D} = \varepsilon_0\mathbf{E} + 4\pi\mathbf{P}. \tag{V.53}$$

Similarly, in magnetic media one introduces the *magnetization* vector field **M** by way of:

$$\mu_0\mathbf{H} = \mathbf{B} + 4\pi\mathbf{M}. \tag{V.54}$$

In order to combine these into something more useful to pre-metric electromagnetism – i.e., a four-dimensional construction – we must use the classical vacuum isomorphism κ_0 to define a sort of reference medium, in order to define the subtraction above.

We call the bivector field:

$$\mathbf{Q}(F) = \frac{1}{4\pi}(\kappa - \kappa_0)(F) = P_i\, b^i + M_i\, b^{i+3}, \tag{V.55}$$

the *electromagnetic polarization* bivector field.

Now, we can expand $4\pi Q(F)$ in a Taylor series:

$$4\pi\mathbf{Q}(F) = \chi^{(0)} + \chi^{(1)}(F) + \chi^{(2)}(F, F) + \ldots \tag{V.56}$$

The successive terms $\chi^{(i)}$, $i = 0, 1, \ldots$, are the *electromagnetic susceptibilities*. To the zeroth and first order, they relate to κ by way of:

$$\chi^{(0)} = 4\pi\eta_0, \qquad \chi^{(1)} = 4\pi\, d(\kappa - \kappa_0)|_{F=0}, \quad \ldots, \chi^{(i)} = 4\pi\, d^{(i)}(\kappa - \kappa_0)|_{F=0}, \ldots, \qquad \text{(V.57)}$$

in which the $d^{(i)}$ refers to the i^{th} differential map. Therefore, the essential difference between the susceptibilities and the corresponding terms in the Taylor series for the constitutive law is in the linear terms.

b. Nonlinear optical effects. Although the field of nonlinear optics has expanded by now into a vast volume of literature (for some elementary treatments, however, cf. Mills [**20**] or Butcher and Cotter [**21**], as well as the synopsis in Delphenich [**22**]), most of which is considerably more applied and empirical than is relevant to our present theoretical discussion, there are some general phenomena that occur as a result of the nonlinearity itself that we shall find intuitively illuminating to summarize. Basically, the nonlinear effects are classified according to the level of susceptibility that first introduces the effect.

At second order, one first encounters a departure from the law of superposition that is definitive of linear fields or waves. In particular, when two interacting electromagnetic waves have frequencies ω_1 and ω_2, one can find that waves with their difference and sum frequencies appear in addition to the original frequencies. When they both have same frequency, the resulting waves can include a DC component and one with the double frequency. The former effect is called *optical rectification*, while the latter one is called *second-harmonic generation*.

In some cases, such as when one F_0 of two fields F_0 and F can be treated as a "background" field, in some sense, instead of dealing with – say – second-order effects in the combined field $F_0 + F$, one can absorb the background field into the second-order susceptibility:

$$\chi^{(2)}(F_0 + F, F_0 + F) = \chi^{(2)}(F_0, F_0) + 2\chi^{(2)}(F_0, F) + \chi^{(2)}(F, F) \qquad \text{(V. 58)}$$

to produce a first-order effect, in which the background field has "modulated" the first order susceptibility. For instance, one might consider, among other terms:

$$4\pi\mathbf{Q} = \chi^{(1)}(F) + \chi^{(2)}(F_0, F) = [\chi^{(1)} + 2\chi^{(2)}(F_0, .)](F). \qquad \text{(V.59)}$$

Hence, the background field has effectively altered the first order susceptibility. In the context of electric fields, which is of interest to optics, this is referred to as the *linear electro-optic* – or *Pockels – effect*. Similarly, the cubic term $3\chi^{(2)}(F_0, F_0, .)$ can also contribute to the linear term, which is referred to as the *quadratic electro-optic* – or *Kerr – effect* in the electric case.

Another cubic effect that gets considerable attention is due to the fact that in addition to the original frequencies and their sums and differences, one can also produce waves of other combination frequencies, such as $2\omega_1 - \omega_2$. This case is called *four-wave mixing*. However, one must always impose the constraint on the incoming and outgoing frequencies and wave numbers that they sum to zero, which is essentially a form of the conservation of energy-momentum.

When the background field is a magnetic field, there are magnetic phenomena that correspond to the electrical ones. For instance, there are linear and quadratic magneto-optical effects due to the modulation of the *electrical* susceptibilities by a background magnetic field. The conversion of circularly polarized waves into elliptically polarized ones under reflection is called the *magneto-optic Kerr effect*. The presence of non-zero magnetization in a medium can make an electrically isotropic medium behave like a uniaxial dielectric, with the associated birefringence [21], and this is referred to as the *Couton-Mouton effect*.

Many of the optical phenomena that nonlinear optics is concerned with are based in the idea that macroscopic media will generally have resonant frequencies, due to either the discrete nature of electron level transitions in atoms or the normal modes of mechanical vibrations in the crystal lattice. When the incoming frequency mixes with a resonant frequency to produce sum and difference frequencies, one gets *Raman scattering*; the difference frequency is the *Stokes* component of the scattered wave and the sum frequency is the *anti-Stokes* component. In *Brillouin scattering*, the acoustic modes of the lattice create an effective diffraction grating, producing dispersion.

The nonlinear wave equations of solitons have found considerable application in nonlinear optics, as well. In particular, the nonlinear Schrödinger equation relates to the phenomenon of *self-focusing* in cylindrical beams, while the sine-Gordon equation describes *self-transparency* in a nonlinear medium. In the former case, the index of refraction can vary with the beam intensity, which varies with radial distance from the center of the beam, and in the latter case, the frequency of the incoming wave couples to a resonant frequency that originates in an electronic level transition.

c. Nonlinear plasma phenomena [**15**]. The range over which the plasma state of matter manifests itself – indeed, over ninety percent of the matter in the universe is in that state – is sufficiently vast in scope that the breakdown of linear approximations in its mathematical modeling can also be quite broad-ranging.

Most of the nonlinear phenomena first manifest themselves in the form of nonlinear wave phenomena, which we will defer to a later discussion. However, as far as the actual constitutive properties of plasma are concerned, we can mention here that the electrical resistivity of plasmas can take on anomalous contributions due to the fact that the diffusion equation that one uses is fundamentally nonlinear, while treating it as a linear partial differential equation is only a limiting approximation.

6. Effective quantum constitutive laws. Beyond a doubt, the most challenging task for pre-metric electromagnetism is to see if some sort of formal extension of the scope of classical – i.e., linear – Maxwellian electromagnetism makes it possible to absorb some of the phenomenological successes of quantum electrodynamics into the domain of things that can be explained by the theoretical methodology that is sometimes called “classical.”

[21] Birefringence – or “double refraction” – means that the index of refraction depends upon the direction of the wave vector for the wave. As we shall see in Chapter VIII, it plays a fundamental role in pre-metric electromagnetic dispersion laws.

At this point, it is necessary to clarify that the use of the word "classical" to refer to physics or mathematics that was not actually being discussed by the classical physicists – say, the Copenhagen school of quantum physics – is incorrect, if not pejorative. The essential difference between what is currently dismissed as "classical" formalism and what is considered to be the more modern quantum formalism is essentially the difference between field theories and scattering theories.

That is, modern quantum electrodynamics (cf., for instance, Berestetskii, Lifschitz, and Pitaevski [**23**] or Jauch and Rohrlich [**24**]) does not even attempt to pose boundary-value problems in electrostatics or Cauchy problems in electrodynamics because it has long since been agreed that the details of physical processes at that level of resolution are beyond the limits of experimental measurement or observation, and therefore it would be pointless for theoretical physics to speculate on them. (Of course, this argument does not seem to apply to the nature of the early Big Bang, or physics at the Planck energy scale, or at the scale of supersymmetry breaking, all of which is many orders of magnitude beyond the limits of experiment, but nonetheless quite dominant in the topics that modern theoretical physics is concerned with!)

Rather, the methodology of quantum physics is based in the statistical interpretation of wave mechanics, combined with the notion that if the only things that experimental physics will ever "know" about the nature of physics at the atomic-to-subatomic scale must manifest themselves in the results of particle scattering experiments then there is no loss in theoretical scope associated with treating the scattering theory as if it were identical with the field theory. In fact, scattering theories follow from dynamical field theories as an asymptotic approximation whereby one does not address the time evolution of a set of fields in interaction during that interaction itself, but only the relationship of the asymptotic incoming fields at a time long before the interaction took place to the asymptotic fields at a time long afterwards (cf., e.g., Lax and Phillips [**25**]). Hence, in a sense, a scattering theory follows from a field theory in a manner that is similar to the way that static fields relate to dynamic fields.

Nevertheless, one must respect the empirical successes of quantum electrodynamics, even if the theoretical formalism seems to be manifestly limiting in its scope. Here, it helps to remind oneself that the problems of the qualitative and quantitative description of natural phenomena are quite distinct from that of the mathematical modeling of those phenomena. Hence, one can accept that charged particles have certain properties that are entirely consistent with the formalism of quantum electrodynamics without taking the position that the established formalism is the only possible solution to the problem of mathematical modeling. For instance, one must address the unavoidable facts that – at least as far as the electromagnetic interaction is concerned – there is a minimum total electric charge e that forms the basis for all larger total charges, that charged particles all have anti-particles, that the most elementary charges have non-vanishing internal angular momentum or spin, and other established truths of electromagnetism at the elementary level.

One of the recurring themes of quantum electrodynamics is that the fundamental process by which photons of more than a critical energy – viz., $2m_ec^2$, or about 1 MeV – can split into an electron-positron pair plays a fundamental role in all of the other processes. Conversely, an electron-positron pair can annihilate each other to form a photon. One calls this process and its inverse *pair creation and annihilation*,

respectively, and it can lead to what one calls *vacuum polarization*, a phenomenon that accounts for some of the most definitive experimental confirmations of quantum electrodynamics, such as the anomalous magnetic moment of the electron and the associated Lamb shift of atomic electron energy levels.

Since we have been treating the electromagnetic polarization of a medium as a purely macroscopic process that eventually resolves to the alignment of elementary atomic dipoles there is always the question of whether one might take the phrase "vacuum polarization" literally and propose to regard it as the basis for replacing the elementary constants ε_0 and μ_0 – and thus also $c_0 = 1/\sqrt{\varepsilon_0\mu_0}$ – with more elaborate functions, most likely, functions of the electric and magnetic field strengths. There are various arguments for and against taking this step.

On the one hand, one might argue that is it not really the vacuum itself that is polarizing, but the photon. That is, electron-positron pairs originate in the splitting of photons, not from regions of space in which there is not even so much as a photon present. However, this argument is weakened by other considerations.

For one thing, one regularly uses vacuum polarization as a means of accounting for the difference between the "bare" charge and the "dressed," or renormalized, charge of a particle, namely, one assumes that the dressed charge comes about as a result of the polarization of the surrounding vacuum in the presence of the high electric field strength that one finds close to elementary charges. In that event, there is no photon present, only the static electric field of the charge distribution and the static magnetic field due to its spin.

Perhaps the most compelling argument for treating vacuum polarization more generally is based in the idea that nowadays the electromagnetic vacuum is not regarded as a region of space in which no fields are present, and which is associated with various empirical constants, such as ε_0 and μ_0. Rather, the vacuum is regarded as a state in a large state space of electromagnetic fields, which is not only infinite-dimensional to begin with, but, more to the point, one must clearly distinguish the vacuum state, in the sense of energetic ground state, from the "zero" state, if indeed the space in question is a linear space, and not some more general manifold. This opens the possibility that the vacuum state might even be non-unique, as with the phenomenon of spontaneous symmetry breaking, which gives ground states non-zero expectation values for their energies.

In the case of the electromagnetic field, one must intuitively replace the classical conception of that field as being a spatial distribution of simple harmonic oscillators with its conception as a spatial distribution of *quantum* harmonic oscillators. A consequence of this is that since quantum harmonic oscillators have a non-zero ground state energy the quantum electromagnetic field must also have a non-zero ground state configuration, which one calls the *zero-point field*. Note that this field is entirely distinct from the 2.7K cosmic microwave background radiation that is also ubiquitous to space.

The existence of a zero-point electromagnetic field has actually been confirmed directly by experiments in the form of the *Casimir effect* [**26**]. This effect amounts to the fact that an ideal parallel-plate capacitor in a vacuum will experience a slight, but measurable, force of attraction between its plates.

One sees that an immediate problem with the conception of the zero-point field as a spatial distribution of quantum harmonic oscillators is that if one adheres to the constraint

that *all* of them must be in their non-zero ground state then no matter how small that energy is, the fact that there are an infinitude of points in space will make the total energy of the vacuum state infinite, as well. Hence, one must accept that "most" of that infinitude of points must be in the *zero* energy state and that only the *total* energy of the field must exist in some ground state. Of course, this suggests that the vacuum ground state is anything but unique, since once generally imagines an infinite set of possible "vacuum fluctuations" of a largely stochastic nature.

One of the compelling experimental verifications of vacuum polarization in the neighborhood of elementary charges takes the form of *Delbrück scattering*. In that process, a photon can be scattered by the electrostatic field of an atomic nucleus, which is a classically non-existent possibility, due to the fact that the combined field of the photon and nucleus might exceed the threshold for electron-positron pair production, and the electron-positron pair then couples to the nuclear field electromagnetically.

Another non-classical possibility that follows from vacuum polarization is *photon-photon scattering*. This would imply that the superposition of the photon fields is nonlinear, since the only linear effect of combining electromagnetic waves is possible interference where they intersect each other, but no lasting effects on their subsequent propagation. Interestingly, this process has yet to be experimentally verified, although the minimum energy level for observing it seems only incrementally beyond the limits of laser technology.

Although it seems unavoidable that the proper context for the mathematical modeling of the electromagnetic vacuum must involve infinite-dimensional spaces, nevertheless, since experimental physics always involves a finite set of measurements, which can only span a finite-dimensional space, one must eventually come back to *effective models* for the vacuum state. In our case, these will be effective quantum constitutive laws that are derived from effective field theories for quantum electrodynamics, so we briefly discuss that notion.

a. Effective actions [**27**-**31**]. Although we shall have more to say about the variational formulation of electromagnetism later, at this point, since we are mostly concerned with simply stating electromagnetic constitutive laws for nonlinear field theories that have emerged from quantum electrodynamics, we shall assume a minimal familiarity with conventional Lagrangian field theory.

Without going into the details here, we simply say that the difference between a full quantum field theory and an effective quantum field theory goes back to some of the early problems in quantum electrodynamics [22] that all originated in the concept of the "Dirac Sea" as a model for the electromagnetic vacuum state. In that model, electrons were positive energy states and positrons were negative energy states, with an energy gap of $2m_e c^2$ between them that represented the difference between the non-zero rest energies of the electron and the positron. The problem was that in the eyes of classical physics there is no mirror symmetry between positive energy, which represents free particle states, and negative energy, which represents bound particle states. Furthermore, an energetic state is stable iff it is minimal; i.e., iff there are no lower energy states to decay

[22] An engrossing discussion of the early years of quantum electrodynamics can be found in Miller [**32**], along with translations of many of the key papers.

into. Hence, because the energy states of the Dirac Sea went down to negative infinity they would all be unstable.

The "solution" to this dilemma was to assume that all of the negative energy states were occupied, so that the Pauli exclusion principle would prevent the runaway decay of particles to negative infinity. However, this implies that the vacuum state would have to have infinite rest mass – hence, infinite rest energy – and infinite charge, which was clearly absurd.

The approach that Heisenberg took was a combination of the "exchange-particle concept" and the so-called "subtraction of infinities". The exchange-particle concept effectively replaced all consideration of fields and forces at the fundamental level with the consideration of particles that were exchanged during interactions. Basically, subtraction of infinities amounted to treating the unphysical infinities in the vacuum state as being passive to the process of interaction. That is, the only energy states that were actually affected by the process of interaction would be a finite number of low-energy states. The subtraction of infinities eventually turned into the more modern techniques for the regularization and renormalization of unphysical infinities.

A similar transition occurs between a complete quantum formulation of a field theory and an effective theory of that field. In the complete theory one starts with a classical action for the fields in question and "quantizes" it, either by regarding some of the fields as taking their values in an operator algebra or by forming an integral over an infinite-dimensional space of fields (or, at least, gauge equivalence classes of them) that gives the transition probability for the incoming scattering state to turn into the outgoing one. This then leads to unphysical infinities, such as mass and charge, which are then corrected by the processes of regularization of the integral and renormalization of the action.

Finally, one can often define an effective action for the process in question to be a correction to the original classical action that includes the effects of renormalization. In the language of the functional integral approach, one is performing a "loop expansion" of the scattering amplitude (really, the Green function). This is an asymptotic series expansion in powers of $\hbar$, in which the "tree level," which has no loops, represents the classical action. At the one-loop level, one must renormalize the first-order radiative corrections that come from the possibility of the creation and subsequent annihilation of one electron-positron pair from a photon. Similarly, succeeding levels of approximation will involve increasing numbers of loops on external and internal lines of the Feynman diagrams for the interaction.

One sometimes hears it said that what effective field theories amount to are low-energy theories that result from integrating out the higher-energy states. This is, of course, quite reminiscent of the Heisenberg approach to the subtraction of infinities from the Dirac Sea. They can be an invaluable tool in probing the quantum domain because what they provide is a strong sense of direction when desires to go beyond the classical theories into an infinite set of possible directions.

b. Heisenberg-Euler action. What Heisenberg and Euler were addressing in their seminal paper [**27**] was essentially the question of what happens to the classical action for an electromagnetic field F in vacuo, which is based on a Lagrangian of the form:

$$\mathcal{L}_{\text{em}} = \tfrac{1}{4} F_{\mu\nu} F^{\mu\nu} = \tfrac{1}{2} \hat{\kappa}(F, F) \tag{V.60}$$

when one includes the possibility that the vacuum might also polarize. That is, the electromagnetic field might produce any number of virtual electron-positron pairs. Although $\hat{\kappa}$ is basically distinct from the ultimate nonlinear constitutive law κ; it still defines a scalar product on 2-forms, as well as an isomorphism of 2-forms and bivectors.

Hence, the full quantum treatment of this situation must include contributions to the action that represent the fields of the fermions and their interactions with the given electromagnetic field. This must then be renormalized beyond the tree level, and one finds that the one-loop correction to $\mathcal{L}_{\text{em}}$ is the effective Lagrangian:

$$\mathcal{L}_{\text{HE}} = \frac{\alpha}{2\pi}\int_0^\infty d\eta \frac{e^{-\eta}}{\eta^3}\left\{(E_c^2 - \tfrac{1}{3}\eta^2\mathcal{F}) - i\eta^2\mathcal{G}\frac{\cos\left(\frac{\eta}{E_c}\sqrt{\mathcal{F}-i\mathcal{G}}\right)+c.c}{\cos\left(\frac{\eta}{E_c}\sqrt{\mathcal{F}-i\mathcal{G}}\right)-c.c}\right\}. \qquad \text{(V.61)}$$

In this expression, $\alpha = e^2/\hbar c$ = 1/137 is the fine structure constant; the fact that it appears to the first power is associated with the fact that this is a one-loop correction. The symbol $E_c = m_e^2c^3/e\hbar$ refers to the critical field strength for pair production; it equals either 1.3×10^{16} V/cm or 4.4×10^{13} G. The notations $\mathcal{F}$ and $\mathcal{G}$ refer to the fundamental Lorentz-invariant expressions that F defines:

$$\mathcal{F} = \tfrac{1}{2}F_{\mu\nu}F^{\mu\nu} = \hat{\kappa}(F, F), \qquad \mathcal{G} = \tfrac{1}{2}F_{\mu\nu}{}^*F^{\mu\nu} = \mathcal{V}(F, F). \qquad \text{(V.62)}$$

We have, as a consequence, that:

$$\mathcal{L}_{\text{em}} = \tfrac{1}{2}\mathcal{F}. \qquad \text{(V.63)}$$

The Lagrangian (V.61) is, of course, quite complicated to work with directly, and one most often encounters it in applications in its weak-field ($E < E_c$) Taylor series expansion:

$$\mathcal{L}_{\text{HE}} = \frac{\alpha}{360\pi E_c^2}(\mathcal{F}^2 + \frac{7}{4}\mathcal{G}^2). \qquad \text{(V.64)}$$

As Barcelo, Liberati, and Visser [**33**] point out, since this is a one-loop correction, it is absurd to carry the expansion to higher-order terms. One also needs to note that the numerical value of the leading scalar factor in c.g.s units is 10^{-42} cm^2/V^2 if one is to get some sense of how the correction term compares to the classical – i.e., tree-level – term (V.63).

In order to obtain an electromagnetic constitutive law from the combination $\mathcal{L} = \mathcal{L}_{\text{em}} + \mathcal{L}_{\text{HE}}$, one need only know at this point that in Lagrangian electromagnetics one has that the electromagnetic excitation 2-form $\mathfrak{H}$ is related to the field strength 2-form F by:

$$\mathfrak{H} = 2\frac{\partial \mathcal{L}}{\partial F} = 2\frac{\partial \mathcal{L}}{\partial \mathcal{F}}\frac{\partial \mathcal{F}}{\partial F} + 2\frac{\partial \mathcal{L}}{\partial \mathcal{G}}\frac{\partial \mathcal{G}}{\partial F} = \frac{\partial \mathcal{L}}{\partial \mathcal{F}}\hat{\kappa}(F) + \frac{\partial \mathcal{L}}{\partial \mathcal{G}}\#(F)$$
$$= \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}\hat{\kappa} + \frac{\partial \mathcal{L}}{\partial \mathcal{G}}\#\right)(F). \qquad \text{(V.65)}$$

Hence, we can make the general expression for the constitutive isomorphism:

$$\kappa = \frac{\partial \mathcal{L}}{\partial \mathcal{F}}\hat{\kappa} + \frac{\partial \mathcal{L}}{\partial \mathcal{G}}\#. \qquad \text{(V.66)}$$

One sees from this that if $\hat{\kappa}$ does not have an axion part to begin with then the origin of an axion contribution to κ will be in the functional dependency of the field Lagrangian on $\mathcal{G}$.

For the classical electromagnetic Lagrangian $\mathcal{L}_{\text{em}}$, one finds that:

$$\frac{\partial \mathcal{L}}{\partial \mathcal{F}} = 1, \qquad \frac{\partial \mathcal{L}}{\partial \mathcal{G}} = 0, \qquad \text{(V.67)}$$

and for the Lagrangian in question – viz., $\mathcal{L} = \mathcal{L}_{\text{em}} + \mathcal{L}_{\text{HE}}$ – one has:

$$\frac{\partial \mathcal{L}}{\partial \mathcal{F}} = 1 + \frac{\alpha}{360\pi}\frac{\mathcal{F}}{E_c^2}, \qquad \frac{\partial \mathcal{L}}{\partial \mathcal{G}} = \frac{7\alpha}{360\pi}\frac{\mathcal{G}}{E_c^2}. \qquad \text{(V.68)}$$

As E_c grows arbitrarily large, the Lagrangian $\mathcal{L}$ converges to the classical vacuum expression.

When the matrix $\hat{\kappa}^{IJ}$ is associated with the classical vacuum, we find that the constitutive matrix κ^{IJ} is of the bi-isotropic form:

$$\kappa^{IJ} = \left[\begin{array}{c|c} -\varepsilon\delta^{ij} & \gamma\delta^{ij} \\ \hline \gamma\delta^{i}_{j} & (1/\mu)\delta^{ij} \end{array}\right], \qquad \text{(V.69)}$$

in which:

$$\varepsilon = \left(1 + \frac{\alpha}{360\pi}\frac{\mathcal{F}}{E_c^2}\right)\varepsilon_0, \qquad \gamma = \frac{7\alpha}{360\pi}\frac{\mathcal{G}}{E_c^2}, \qquad 1/\mu = \left(1 + \frac{\alpha}{360\pi}\frac{\mathcal{F}}{E_c^2}\right)(1/\mu_0). \qquad \text{(V.70)}$$

We see that κ^{IJ} is symmetric, so its skewon part vanishes, and it has an axion part whose proportionality factor is γ.

As we shall see later, when we discuss complex geometry, the representation of κ by a 6×6 real matrix that is given by (V.69) is equivalent (by a rescaling of the field strengths to make $\varepsilon = 1/\mu$) to its representation by the 3×3 complex matrix $(\gamma + i\varepsilon)I$.

Hence, its effect on 2-forms, when regarded as elements of a complex three-dimensional vector space, is that of complex scalar multiplication. This also shows that a constitutive law of Heisenberg-Euler type is almost-complex iff γ vanishes, hence $\mathcal{G} = 0$.

c. Born-Infeld action. Born and Infeld [**29, 30**] were motivated by the fact that Coulomb's law leads to not only an infinite field strength at the origin, but also an infinite total self-energy for an electron. They postulated that if there were a maximum allowable field strength E_c then the infinite field strength predicted by Coulomb's law for a pointlike particle would be unphysical. They deduced a largely heuristic electromagnetic Lagrangian, which is also based in the Lorentz invariants $\mathcal{F}$ and $\mathcal{G}$, as before, and which has the property that the field strength of a pointlike charge origin will be E_c at the origin.

The Born-Infeld Lagrangian is:

$$\mathcal{L}_{\mathrm{BE}} = -2E_c^2\left(1-\frac{\alpha}{180\pi}\frac{\mathcal{F}}{E_c^2}-\frac{7\alpha}{720\pi}\frac{\mathcal{G}^2}{E_c^4}\right)^{1/2}. \tag{V.71}$$

In order to compute κ from this, we only need to find:

$$\frac{\partial\mathcal{L}}{\partial\mathcal{F}}=\left(1-\frac{\alpha}{180\pi}\frac{\mathcal{F}}{E_c^2}-\frac{7\alpha}{720\pi}\frac{\mathcal{G}^2}{E_c^4}\right)^{-1/2}, \tag{V.72a}$$

$$\frac{\partial\mathcal{L}}{\partial\mathcal{G}}=\left(1-\frac{\alpha}{180\pi}\frac{\mathcal{F}}{E_c^2}-\frac{7\alpha}{720\pi}\frac{\mathcal{G}^2}{E_c^4}\right)^{-1/2}\frac{7\alpha}{360}\frac{\mathcal{G}}{E_c^2}. \tag{V.72b}$$

Thus, in the limit as E_c grows arbitrarily large these expressions also converge to the classical vacuum expressions. If we expand the expression in parenthesis by means of the binomial theorem then we get:

$$\left(1-\frac{\alpha}{180\pi}\frac{\mathcal{F}}{E_c^2}-\frac{7\alpha}{720\pi}\frac{\mathcal{G}^2}{E_c^4}\right)^{-1/2}\approx 1+\frac{\alpha}{360\pi}\frac{\mathcal{F}}{E_c^2}, \tag{V.73}$$

and we see that the partial derivatives in (V72a,b) are approximately equal to the corresponding ones (V.68) for the Heisenberg-Euler Lagrangian.

Similarly, by direct inspection, one can see that the Born-Infeld vacuum is also of the bi-isotropic variety, as was the Heisenberg-Euler vacuum, and we see that the expressions in (V.70) now take the form:

$$\varepsilon=\left(1-\frac{\alpha}{180\pi}\frac{\mathcal{F}}{E_c^2}-\frac{7\alpha}{720\pi}\frac{\mathcal{G}^2}{E_c^4}\right)^{-1/2}\varepsilon_0, \tag{V.74a}$$

$$\gamma=\left(1-\frac{\alpha}{180\pi}\frac{\mathcal{F}}{E_c^2}-\frac{7\alpha}{720\pi}\frac{\mathcal{G}^2}{E_c^4}\right)^{-1/2}\frac{7\alpha}{360}\frac{\mathcal{G}}{E_c^2}, \tag{V.74b}$$

$$1/\mu = \left(1 - \frac{\alpha}{180\pi}\frac{\mathcal{F}}{E_c^2} - \frac{7\alpha}{720\pi}\frac{\mathcal{G}^2}{E_c^4}\right)^{-1/2} (1/\mu_0). \qquad \text{(V.74c)}$$

Therefore, the essential difference between the Heisenberg-Euler constitutive law and the Born-Infeld one amounts to the differences between precise values of the corresponding scalar multipliers.

References

54. L. D. Landau, E. M. Lifschitz, and L. P. Pitaevskii, *Electrodynamics of Continuous Media*, 2nd ed., Pergamon, Oxford, 1984.
55. E. J. Post, *Formal Structure of Electromagnetics*, Dover, NY, 1997.
56. F. W. Hehl and Y. N. Obukhov, *Foundations of Classical Electrodynamics*, Birkhäuser, Boston, 2003.
57. I. Lindell, *Differential Forms in Electromagnetics*, IEEE Press, NJ, 2004.
58. P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics*, Cambridge University Press, Cambridge, 1997.
59. E. Matagne, "Algebraic decomposition of the electromagnetic constitutive tensor. A step towards a pre-metric based gravitation," Ann. Phys. (Berlin), **17** (2008), 17-27.
60. P. D. Gilkey, *Geometrical Properties of Natural Operators Defined by the Riemann Curvature Tensor*, World Scientific, Singapore, 2001, pp. 41-44.
61. D. H. Delphenich, "On linear electromagnetic constitutive laws that define almost-complex structures," Ann. Phys. (Leipzig) **16** (2007), 207-217.
62. B. D. H. Tellegen, "The Gyrator, a new electric network element," Philips Research Reports, **3** (1948), 81-101; "The gyrator, an electrical network element," Philips Technical Reviews, **18** (1956/57), 120-124; reprinted in *An Anthology of Philips Research*, H. B. G. Casimir and S. Gradstein (eds.), Philips Gloeilampenfabriken, Eindhoven (1966), pp. 186-190.
63. A. H. Lindell and I. V. Sivola, "Perfect electromagnetic conductor," J. Electromag. Waves Appl. **11** (2005), 861-869; "Transformation method for problems involving perfect electromagnetic conductor structure," IEEE Trans. Ant. Prop., **53** (2005), 3005-3011.
64. A. H. Sihvola, "Metamaterials in electromagnetics," Metamaterials, **1** (2007), 2-11.
65. F. W. Hehl, Y. N. Obukhov, J.-P. Rivera, and H. Schmidt, "Relative nature of a magneto-electric modulus of Cr_2O_3 crystals: a new four-dimensional pseudoscalar and its measurement," (?).
66. J. D. Jackson, *Classical Electrodynamics*, 2nd ed., Wiley, New York, 1976.
67. L. D. Landau, E. M. Lifshitz, *Classical Field Theory*, Pergamon, Oxford, 1975.
68. F. F. Chen, *Introduction to Plasma Physics and Controlled Fusion, v. 1*," Plenum Press, New York, 1984.
69. W. P. Allis, S. J. Buchsbaum, and A> Bers, *Waves in Anisotropic Plasmas*, M.I.T. Press, Cambridge, MA, 1963.
70. V. L. Ginzburg, *The Propagation of Electromagnetic Waves in Plasmas*," Addison-Wesley, Reading, MA, 1964.

71. E. M. Lifschitz and L. P. Pitaevski, *Physical Kinetics*, Butterworth-Heinemann, Elsevier, Oxford, 1981.
72. J. Plebanski, *Lectures on Nonlinear Electrodynamics,* NORDITA Lectures, Copenhagen, 1970.
73. D. L. Mills, *Nonlinear Optics,* 2nd ed., Springer, Berlin, 1998.
74. P. N. Butcher and D. Cotter, *The Elements of Nonlinear Optics,* Cambridge Univ. Press, 1990.
75. D. H. Delphenich, "Nonlinear optical analogies in quantum electrodynamics," arXiv:hep-th/0610088.
76. V.B. Berestetskii, E.M. Lifschitz, and L.P. Pitaevskii, Quantum Electrodynamics, 2nd ed. (Elsevier, Amsterdam, 1984).
77. J. M. Jauch and F. Rohrlich, *The Theory of Photons and Electrons, 2nd ed.*, Springer, Berlin, 1976.
78. P. D. Lax and R. S. Phillips, *Scattering Theory*, Academic Press, New York, 1967.
79. V. M. Mostepanenko and N. N. Trunov, *The Casimir Effect and its Applications*, Clarendon Press, Oxford, 1997.
80. W. Heisenberg and H. Euler, "Folgerungen aus der Diracschen Theorie des Positrons," Zeit. f. Phys., **98** (1936), 714-732.
81. J. Schwinger, "On vacuum polarization and gauge invariance," Phys. Rev. **82** (1951), 664-679. Reprinted in *Selected Papers on Quantum Electrodynamics,* ed. J. Schwinger, Dover, NY, 1958.
82. M. Born and L. Infeld, "Foundations of a New Field Theory," Proc. Roy. Soc. A, **144** (1934), 425-451.
83. M. Born, "Théorie non-linéaire du champs électromagnétique, Ann. Inst. H.P., **7** (1937), 155-265.
84. W. Dittrich and H. Gies, *Probing the Quantum Vacuum,* Springer, Berlin, 2000.
85. A. I. Miller, *Early Quantum Electrodynamics: a source book,* Cambridge University Press, Cambridge, 1994.
86. C. Barcelo, S. Liberati, and M. Visser, "Bi-refringence versus bi-metricity," arXiv.org preprint gr-qc/0204017.

Chapter VI

Partial differential equations on manifolds

Since the most fundamental statements of electromagnetism, in any of its representations, involve systems of partial differential equations, and we are trying to maintain a consistent use of the topological and geometric benefits of differentiable manifolds to the greatest extent possible, it is inevitable that we must say something about the representation of such systems of partial differential equations on manifolds.

We hasten to point out that such a task can be quite formidable in the modern era, since, in effect, the first problems regarding partial differential equations on manifolds to be formally addressed by the mathematics community were not concerned with redefining the elementary methods that one applies to the equations of physics, but with more general problems, such as the nature of the eigenvalues of the Laplacian operator on Riemannian manifolds, the topological aspects of the Yang-Mills field equations, or the formal integrability of overdetermined systems of partial differential equations.

Certainly, these problems are not insignificant, even insofar as physics applications are concerned. However, many of the methods of mathematical physics that get used most widely by the physics community are more in the nature of applied mathematical techniques for solving specific problems, not general techniques for proving purely mathematical theorems. Hence, the purpose of this chapter will be to show how many of these elementary concepts can be represented in terms of things that pertain to differentiable manifolds and the elementary vector bundles that relate to them. In particular, we need to eventually address the linear differential operators on the vector bundles of k-forms and k-vector fields over the spatial manifold Σ or the spacetime manifold M that take the form of d and δ.

There are three basic ways of representing a system of partial differential equations on a manifold M: one can represent it by a differential operator $L: E \to F$ between vector bundles over M, by a hypersurface in a manifold $J^k(M, N)$ of k-jets of C^k maps from M to another manifold N, or by an exterior differential system on M. The choice of method is usually determined by the class of problems that one is ultimately addressing, so we shall begin by discussing each method at an elementary level.

Next, we discuss the nature of the most common types of problems that one encounters in the partial differential equations: boundary-value problems and initial-value problems. In the case of a linear system of partial differential equations, these problems can be reformulated in terms of a linear integral equation involving an integral operator whose kernel takes the form of a Green function, so we then discuss the extent to which the usual definitions can or cannot be carried over to corresponding construction on manifolds.

Finally, we briefly discuss the much more involved problem of what becomes of the Fourier transform on a differential manifold that is not necessarily an affine space, or even a homogeneous space. Indeed, this problem has been a predictable obstacle to the formulation of the foundations of quantum wave mechanics in the language of differentiable manifolds.

1. Differential operators on vector bundles [1, 2]. As far as field theories are concerned, the most convenient way of representing most of the fields of interest to physics is by means of sections ϕ: $M \to E$ of vector bundles E over the manifold M in which the fields are assumed to reside. Since the vector space V that represents the typical fiber of E – up to linear isomorphism – can be the tensor product of other vector spaces, one sees that any tensor or spinor field on M can be represented in such a manner. In particular, electromagnetism is often concerned with those vector bundles over the space manifold Σ or the spacetime manifold M whose sections are differential forms and multivector fields.

As pointed out above, the set $\Gamma(E)$ of sections of a vector bundle $E \to M$ can be given the structure of an infinite-dimensional vector space. Basically, one defines scalar combinations of sections pointwise. That is, if ϕ, $\psi \in \Gamma(E)$ are sections and $\alpha, \beta \in \mathbb{R}$ are scalars then the scalar combination $\alpha\phi + \beta\psi$ is defined to take each $x \in M$ to the vector:

$$(\alpha\phi + \beta\psi)(x) = \alpha\phi(x) + \beta\psi(x) \tag{VI.1}$$

in the vector space E_x .

If $E \to M$ and $F \to M$ are vector bundles over M then it is simple enough to define a linear operator L: $\Gamma(E) \to \Gamma(F)$ from the vector space of sections of the former bundle to the vector space of sections of the latter. Such an operator takes any linear combination $\alpha\phi + \beta\psi \in \Gamma(E)$ to the section:

$$L(\alpha\phi + \beta\psi) = \alpha L(\phi) + \beta L(\psi) \tag{VI.2}$$

in $\Gamma(F)$.

In order to give an operator $\mathcal{O}$: $\Gamma(E) \to \Gamma(F)$, whether linear or nonlinear, a local expression as a system of equations, one must choose a local trivialization $U \times E$ of E over a subset $U \subset M$. The best way to do this is to define the frame fields $\mathbf{e}_a$: $U \to E$, $x \mapsto \mathbf{e}_a(x)$, $a = 1, \ldots, N = \text{rank}(E)$ and $\mathbf{f}_b$: $U \to F$, $x \mapsto \mathbf{f}_b(x)$, $b = 1, \ldots, N' = \text{rank}(F)$; both fields consist of N (N', resp.) sections over U that are linearly independent at each point. Hence:

$$\mathcal{O}(\phi) = \mathcal{O}(\phi^a \mathbf{e}_a) = \mathcal{O}^b(\phi^a \mathbf{e}_a)\, \mathbf{f}_b \,, \tag{VI.3}$$

in which the ϕ^a and $\mathcal{O}^b(\phi^a \mathbf{e}_a)$ are smooth functions on U that represent the components of ϕ and $\mathcal{O}(\phi)$ with respect to the chosen frames. This gives rise to the following local system of N' equations in the N unknown functions ϕ^a:

$$\mathcal{O}^b = \mathcal{O}^b(\phi^a). \tag{VI.4}$$

When $\mathcal{O}$ is linear, one can replace the right-hand side of the latter expression with the column matrix $\mathcal{O}^b_a \phi^a$ of functions on U.

An operator $\mathcal{O}$: $\Gamma(E) \to \Gamma(F)$ is called *algebraic* iff it induces a corresponding map $\mathcal{O}_x$: $E_x \to F_x$ on each fiber of E. In that case, (VI.4) also works for the individual values:

$$\mathcal{O}^b(x) = \mathcal{O}^b(\phi^a(x)). \tag{VI.5}$$

If $\mathcal{O}$ is a linear algebraic operator then $\mathcal{O}_x$ will be a linear map at each $x \in M$. Some examples of linear algebraic operators that we have previously encountered include scalar multiplication, the identity operator, left-multiplication of k-forms by a 1-form, the interior product operator on either differential forms or multivector fields, and Poincaré duality.

Of the non-algebraic types of operators on sections the two that will be of interest to us will be *differential* and *integral* operators. We shall first discuss differential operators and then return to the discussion of integral operators.

Ordinarily – i.e., when M is a vector space – one tends to think of the action of a differential operator $\mathcal{D}$ of degree k on a smooth function $f \in C^\infty(M, \mathbb{R})$ as essentially a "functional" expression:

$$\mathcal{D}f(x) = F(x, f(x), df|_x, \ldots, d^k f|_x) \tag{VI.6}$$

in the points of M, the values of f at each point, and the values of its differentials $d^m f$ up to order k.

Actually, this way of looking at partial differential equations leads directly into the methods of jet manifolds, which we will discuss in the next section. However, for now, we point out that when M is not a manifold and the function f takes its values in a vector bundle E, instead of $\mathbb{R}$, one runs into the problem of whether the differential df of the section f: $U \to E$ transforms properly under changes of local frames in E.

In general, the only way to get around this is to introduce a connection on E and replace the differential d with the covariant differential ∇ that goes with the chosen connection. However, one advantage of using differential forms, i.e., sections of $\Lambda^k(M) \to M$ is the fact that the exterior derivative operator d transforms properly without the necessity of introducing a connection.

Hence, since we will be mostly concerned with the field equations of pre-metric electromagnetism, the differential operators of order k that will be of interest to us will consist of scalar combinations of compositions:

$$\mathcal{D} = A_1 \cdot d \cdot A_2 \cdot \ldots \cdot A_k \cdot d \cdot A_{k+1} \tag{VI.7}$$

of k+1 algebraic operators and the exterior derivative operator d. Note that if any of the algebraic operators besides A_1 and A_{k+1} are a scalar multiple of the identity operator then the resulting operator $\mathcal{D} = 0$.

This restricted class of differential operators on differential forms and multivector fields nevertheless includes the exterior derivative operator, the divergence operator $\delta = \#^{-1} \cdot d \cdot \#$, and the Lie derivative operator $\mathrm{L}_\mathbf{v} = i_\mathbf{v} d + d i_\mathbf{v}$, which are all first order, as well as the field operator $\Box_\kappa = \delta \cdot \kappa \cdot d$, which is second order.

a. Integrability. Suppose $\mathcal{D}$: $\Gamma(E) \to \Gamma(F)$ is a differential operator from sections of a vector bundle $E \to M$ to sections of a vector bundle $F \to M$. A typical problem for differential equations that $\mathcal{D}$ might define could take the form of the inhomogeneous problem: Given a section $\rho \in \Gamma(F)$, find a section $\phi \in \Gamma(E)$ such that:

$$\mathcal{D}\phi = \rho\,. \tag{VI.8}$$

The homogeneous problem is defined by choosing $\rho = 0$.

In general, the inhomogeneous problem is not going to have a solution, since it will usually be *overdetermined*; that is, the map $\mathcal{D}$ might not be a surjection. In such an event the image $\mathrm{im}(\mathcal{D}) = \mathcal{D}(\Gamma(E)) \subset \Gamma(F)$ is a proper subset; when $\mathcal{D}$ is linear, it is a proper linear subspace.

When $\rho \in \mathrm{im}(\mathcal{D})$ one says that the system of equations defined by (VI.8) is *integrable*. A set of *integrability conditions* – or *compatibility conditions* – usually takes the form of a set of equations that define the subset $\mathrm{im}(\mathcal{D})$. For instance, one might have another differential operator $\mathcal{D}'$: $\Gamma(F) \to \Gamma(G)$, and then characterize $\mathrm{im}(\mathcal{D})$ as the kernel of $\mathcal{D}'$. That is, $\rho \in \mathrm{im}(\mathcal{D})$ iff:

$$\mathcal{D}'\rho = 0. \tag{VI.9}$$

Of course, in the case of the exterior derivative operator this sort of condition introduces a subtlety in the form of the fact that whether or not $\ker(d) = \mathrm{im}(d)$ at some step in the sequence $\ldots \xrightarrow{d} \Lambda^k \xrightarrow{d} \Lambda^{k+1} \xrightarrow{d} \ldots$ depends upon the vanishing of de Rham cohomology in that dimension, which then defines a topological constraint on the manifold M; analogous remarks apply to the divergence operator δ.

b. Symbol of a differential operator. If $\mathcal{D}$: $\Gamma(E) \to \Gamma(F)$ is a first-order differential operator from sections of a vector bundle $E \to M$ to sections of a vector bundle $F \to M$ then its *symbol* is a bundle map $\sigma[\mathcal{D}]$: $T^*(M) \otimes \Gamma(E) \to \Gamma(F)$ that takes any $df \otimes \phi$ to $\mathcal{D}(f\phi) - f\mathcal{D}(\phi)$. Hence, it is a linear algebraic map between the fibers $T^*_x \otimes E_x$ and F_x at each point $x \in M$. When one fixes a covector field k the map $\sigma[\mathcal{D}, k]$ takes sections of E to sections of F linearly; hence, $\sigma[\mathcal{D}, k]$: $\Gamma(E) \to \Gamma(F)$.

In the case of the ordinary differential operator, which we denote by D to avoid confusion with the exterior derivative d, since $D(f\phi) = Df \otimes \phi + fD\phi$, the symbol of D is tensor multiplication by a covector field k:

$$\sigma[D, k](\phi) = k \otimes \phi. \tag{VII.10}$$

From this, by polarization, we find that:

$$\sigma[D_s, k](\phi) = k \odot \phi = \tfrac{1}{2}(k \otimes \phi + \phi \otimes k), \qquad \text{(VII.11a)}$$

$$\sigma[d, k](\phi) = k \wedge \phi = \tfrac{1}{2}(k \otimes \phi - \phi \otimes k). \qquad \text{(VII.11b)}$$

Here, the operator D_s represents symmetrized differentiation and d represents anti-symmetrized differentiation; i.e., the exterior derivative operator.

Hence, the linear algebraic operator on sections of E that corresponds to symmetrized or anti-symmetrized differentiation is symmetrized or anti-symmetrized tensor multiplication by k, respectively. By abuse of notation, we shall denote any of the three maps $\sigma[D, k]$, $\sigma[D_s, k]$, $\sigma[d, k]$ by e_k: $E \to F$ and let the context of its usage dictate the precise meaning implied.

Had we gone the route of replacing D with a covariant differential, we would have found that the symbol of the operator would not change, since the difference between ordinary and covariant differentiation, in any case, is an algebraic operator, whose symbol is then zero. Similarly, if a differential operator is quasi-linear – i.e., linear in its highest-order derivatives, but possibly nonlinear in its lower-level ones – then the symbol (really, the *principal* symbol) of that operator is the same as that of the linear operator that is defined by its highest-order term.

The symbol of the divergence operator δ is:

$$\sigma[\delta, k](\mathbf{\Phi}) = k(\mathbf{\Phi}) = i_k\mathbf{\Phi}, \qquad \text{(VII.12)}$$

i.e., interior multiplication by k.

In order to extend this conception of the symbol of a differential operator to operators of order higher than one, we confine ourselves to linear operators of the form (VII.7) and assume that:

i. The symbol of a composition of linear differential and algebraic operators is the composition of the symbols.

ii. The symbol of a linear algebraic operator (in such a sequence) is itself.

Hence, the symbol of an operator defined of the form (VII.7) is:

$$\sigma[\mathcal{D}, k] = A_1 \cdot e_k \cdot A_2 \cdot \ldots \cdot A_k \cdot e_k \cdot A_{k+1}\,. \qquad \text{(VI.13)}$$

In the case where any of the sequences $A \cdot e_k \cdot A'$ take the form of $i_k = \#^{-1} \cdot e_k \cdot \#$, which is the symbol of δ, it is generally simpler to replace them with i_k. For instance, the symbol of the field operator $\Box_\kappa = \delta \cdot \kappa \cdot d$, when κ is linear is most simply expressed as:

$$\sigma[\Box_\kappa, k] = i_k \cdot \kappa \cdot e_k = \kappa^{\mu\nu\alpha\beta} k_\nu k_\beta\,. \qquad \text{(VI.14)}$$

c. *Characteristic variety*. Since $\sigma[\mathcal{D}, k]$: $\Gamma(E) \to \Gamma(F)$ is a linear algebraic map, there is always the question of its invertibility to be addressed. As long as E and F both have finite rank, this amounts to the question of the integrability of the linear map $\sigma[\mathcal{D}, k]_x : E_x \to F_x$ at each $x \in M$. If the dimensions of both fibers are unequal then the map can never be invertible. When they are equal, invertibility can be characterized in a

frame-invariant way by considering the determinant of $\sigma[\mathcal{D}, k]_x$. This will depend not only upon x, but also upon the choice of $k \in T_x^*$, and one calls the set of all k such that:

$$\det\{\sigma[\mathcal{D}, k]_x\} = 0 \qquad \text{(VII.15)}$$

the *characteristic variety* of $\mathcal{D}$ at x; it will then be a hypersurface in T_x^*M. This has the immediate consequence that for differential operators of the type that we are considering, from (VI.13), we obtain:

$$\det\{\sigma[\mathcal{D}, k]\} = \det(A_1)\det(e_k)\det(A_2)\ \ldots\ \det(A_k)\det(e_k)\det(A_{k+1}). \qquad \text{(VI.16)}$$

However, one immediately sees that there is a problem, since all of the linear maps in the sequence (VI.13) must be invertible in order for the product to be non-vanishing, in general, and the map e_k is not invertible in any case. It is not injective, since its kernel in dimension m will be spanned by all simple m-forms that contain at least one exterior factor of k. It is not generally surjective, except for n–1 forms, since its image will also be spanned by simple m+1-forms that contain k as an exterior factor. Therefore, in order to arrive at a non-trivial characteristic variety, one must always restrict the domain of e_k to the quotient $\Lambda^m/\ker(e_k)$ and the range to $\mathrm{im}(e_k) = k \wedge \Lambda^m$ at each $x \in M$. For instance, in the case of e_k: $\Lambda^1 \to \Lambda^2$ for an n-dimensional manifold one must restrict e_k to any n–1-dimensional subspace of Λ^1 that is transverse to the line generated by k and the range to the linear subspace of Λ^2 that is spanned by all 2-forms of the form $k \wedge \alpha$ for some 1-form α.

Now, the determinant function on an $n\times n$ matrix represents a homogeneous polynomial of degree n in the components of the matrix. The number of times that the components of k appear in the terms of the *characteristic polynomial* that is defined by $\det\{\sigma[\mathcal{D}, k]\}$, with suitable restriction on the domains and ranges, will be equal to the order of the differential operator. Hence, in the case of a m^{th}-order differential operator on sections of a vector bundle of rank n, after restriction, the characteristic polynomial will be a homogeneous polynomial in k of degree mn.

In the cases of interest to us, when the relevant constitutive laws are linear:

i. The second-order field operator Δ_ε: $\Lambda^0 \to \Lambda_0$, with $\Delta_\varepsilon = \delta \cdot \varepsilon \cdot d$, has the symbol:

$$\sigma[\Delta_\varepsilon, k] = i_k \cdot \varepsilon \cdot e_k = \varepsilon(k, k) = \varepsilon^{ij} k_i k_j, \qquad \text{(VI.17)}$$

which is also the characteristic quadratic polynomial, since the matrix is 1×1.

ii. For three-dimensional space Σ, the second-order field operator Δ_μ: $\Lambda^1 \to \Lambda_1$, with $\Delta_\mu = \delta \cdot \tilde{\mu} \cdot d$, has the symbol:

$$\sigma[\Delta_\mu, k] = i_k \cdot \mu^{-1} \cdot e_k. \qquad \text{(VI.18)}$$

However, in order to derive the characteristic polynomial we must first note that since e_k is not invertible, for each choice of $k \in \Lambda^1$ we must restrict the map $\sigma[\Delta_\mu, k]$ to a two-dimensional subspace at each point of Σ that is transverse to k, and similarly, one must

restrict oneself to the two-dimensional subspaces of Λ_1 that define its image. If θ^a, θ^3 is an adapted coframe (a = 1, 2, $k = \kappa\theta^3$) and $\mathbf{e}_a$, $\mathbf{e}_3$ is its adapted reciprocal frame then the 2×2 matrix $\sigma[\Delta_\mu, k]^{ab}$ takes the form:

$$\sigma[\Delta_\mu, k]^{ab} = -\kappa^2 \varepsilon^a_c \tilde{\mu}^{cd} \varepsilon^d_b , \qquad (\varepsilon^a_b = \varepsilon_{ab}). \tag{VI.19}$$

The characteristic polynomial is actually a degenerate quartic in κ:

$$P[\Delta_\mu, k] = \kappa^4/\mu^2, \qquad (\mu = 1/\det[\tilde{\mu}]). \tag{VI.20}$$

iii. The second-order field operator $\Box_\kappa$: $\Lambda^1 \to \Lambda_1$, with $\Box_\kappa = \delta \cdot \kappa \cdot d$, has a symbol that is given by (VI.14). We shall devote a section of chapter VIII to the discussion of this case, since it is fundamental to geometrical optics, as well as the manner by which a Lorentzian structure "emerges" from the laws of pre-metric electromagnetism.

2. Jet manifolds [2-4]. If we return to the expression (VI.6) then we see that if we were to define a manifold that represented the space that is locally described by the components (x^μ, y^a, $y^a{}_\mu$, … , $y^a{}_{\mu_1 \cdots \mu_k}$) then we could regard a system of N partial differential equations of order k in the unknown functions y^a on the n-dimensional manifold M whose local coordinates are described by the functions x^μ as a level hypersurface of a function F on this manifold:

$$F(x^\mu, y^a, y^a{}_{,\mu}, \ldots, y^a{}_{,\mu_1,\cdots,\mu_k}) = \text{const.} \tag{VI.21}$$

For instance, the linear wave equation $\Box\psi = g^{\mu\nu}(x)\psi_{,\mu,\nu} = 0$ can be defined by the 0-hypersurface of the quadratic function:

$$F(x^\mu, \psi, \psi_\mu, \psi_{\mu\nu}) = g^{\mu\nu}(x)\psi_{\mu\nu} . \tag{VI.22}$$

The function F is often restricted by the requirement that one be able to solve (VI.21) for the highest-order derivatives:

$$y^a{}_{,\mu_1,\cdots,\mu_k} = F'(x^\mu, y^a, y^a{}_{,\mu}, \ldots, y^a{}_{,\mu_1,\cdots,\mu_{k-1}}). \tag{VI.23}$$

Courant and Hilbert [**5**] call this the *normal form* for a system of partial differential equations. This also makes the system *quasilinear*, since the only possible nonlinearity in the system (VI.23) must be in the derivatives of less than maximal order.

From the implicit function theorem, the condition on F that makes this possible is that one must have:

$$\frac{\partial F}{\partial y^a{}_{,\mu_1,\cdots,\mu_k}} \neq 0 \qquad \text{(all } a, \mu_1, \ldots, \mu_k). \tag{VI.24}$$

Otherwise, one must deal with a singular system of equations whose order is reduced at some points. For instance, in the case of systems of first-order ordinary differential equations, which can be represented by vector fields on manifolds, the points at which the derivative vanishes will also be zeroes of the vector field, and therefore fixed points of its local flow.

i. Jets of functions and sections. The manifold that F is defined on in (VI.20) is called the manifold $J^k(M; N)$ of *k-jets* of C^k functions $f: M \to N$. The k-jet $j^k_x f$ of such a C^k function f at $x \in M$ is defined to be the equivalence class of all C^k functions that are defined in some neighborhood of x, which may vary with the function, and have the same values at x as functions, along with the same values of their first k derivatives. Hence, it is easy to see how this gives rise to local coordinate charts on $J^k(M; N)$ of the form $(x^\mu, y^a, y^a{}_\mu, \ldots, y^a{}_{\mu_1 \cdots \mu_k})$; note that we do not include commas in the subscripts, which will become significant shortly.

One immediately has three manifold projections for any k that are defined by:

$$\begin{aligned} &J^k(M; N) \to M, && j^k_x y \mapsto x, \\ &J^k(M; N) \to N, && j^k_x y \mapsto y, \\ &J^k(M; N) \to M \times N, && j^k_x y \mapsto (x, y). \end{aligned}$$

The first two are referred to as the *source* and *target* projections.

In general, these projections do not define fiber bundles, but only give $J^k(M; N)$ the structure of a *fibered manifold*, since they are surjective submersions; i.e., the projections have differential maps that have maximal rank at each point of $J^k(M; N)$.

Of the three possible types of sections for the projections above, the ones that are most fundamental for us are the sections of the first projection, which then take the form of differentiable maps s: $M \to J^k(M; N)$ that give the identity when composed with the projection. This basically means that the value $s(x)$ of s at each $x \in M$ is an element of the fiber $J^k_x(M; N)$. Locally, the values of $s(x)$ have the coordinates

$$s(x) = (x^\mu, y(x), y_\mu(x), \ldots, y_{\mu_1 \cdots \mu_k}(x)). \tag{VI.25}$$

Of particular interest are the *integrable* sections, for which these coordinates also locally satisfy:

$$y_\mu(x) = y_{,\mu}(x), \ldots, \; y_{\mu_1 \cdots \mu_k}(x) = y_{,\mu_1, \cdots, \mu_k}(x). \tag{VI.26}$$

The global way of characterizing integrable sections is that they represent *k-jet prolongations* of differentiable functions on M, which can be thought of as differentiable sections of the third projection above. One then notates the k-jet prolongation of a function $f(x)$ by $j^k f(x)$. It is basically defined by the function f and its first k derivatives. One is cautioned that not all sections of the projection $J^k(M; N) \to M$ are integrable.

ii. Contact form. Although the contact form can be defined for manifolds of k-jets when $k > 1$, we shall confine our attention to the case of $k = 1$ for the moment.

The integrable sections of $J^1(M; N) \to M$ have the property that the pull-backs $s^*\theta^a$ of a certain set of N one-forms θ^a on $J^1(M; N)$ by a section s vanish iff the section is integrable. The 1-forms θ^a are collectively called the *contact form* on $J^1(M; N)$ and can be locally represented in the form:

$$\theta^a = dy^a - y^a{}_\mu \, dx^\mu \, . \tag{VI.27}$$

One is cautioned that the coordinates $y^a{}_\mu$ are functions on an open subset of $J^1(M; N)$, not an open subset of M. However, pulling θ^a down to M by means of s turns dy^a into $y^a{}_{,\mu}\, dx^\mu$ and θ^a into:

$$s^*\theta^a \; = (y^a{}_{,\mu} \; - y^a{}_\mu) \, dx^\mu \qquad \text{(all } a\text{)}, \tag{VI.28}$$

in which the $y^a{}_\mu$ are now functions on M.

Hence, locally, s is integrable iff:

$$y^a{}_\mu = y^a{}_{,\mu} \qquad \text{(all } \mu, a\text{)}. \tag{VI.29}$$

iii. Differential equations. The notion of higher jet prolongations gives a particular concise way of explaining the usual process by which a differential equation (either ordinary or partial) of order k in a C^k function y on M can be converted into an equivalent system of k first-order differential equations. All that one is doing is introducing the higher jet coordinates of $J^k(M; \mathbb{R})$, such as y_μ, $y_{\mu\nu}$, …, and coupling them to the derivatives of y by means of first order differential equations $y_\mu = y_{,\mu}$, etc.

For instance, suppose that we have an nth-order ordinary differential equation in normal form:

$$\frac{d^n y}{d\tau^n} = F(\tau, y, \dot{y}, \cdots, \overset{(n-1)}{y}) \, . \tag{VI.30}$$

If one introduces supplementary variables $v = \dot{y}$, $\overset{(1)}{v} = \ddot{y}$, …, $\overset{(n-2)}{v} = \overset{(n-1)}{y}$ for the successive derivatives then (VI.30) can be converted into the equivalent system of n first-order equations:

$$\begin{cases} \dfrac{dy}{d\tau} = v, \\ \quad \vdots \\ \dfrac{d\overset{(n-2)}{v}}{d\tau} = \overset{(n-2)}{v} \, , \\ \dfrac{d\overset{(n-1)}{v}}{d\tau} = F(\tau, y, v, \cdots, \overset{(n-2)}{v}) \end{cases} \tag{VI.31}$$

Since the supplementary variables are clearly the coordinates of $J^{n-1}(\mathbb{R}, \mathbb{R})$ beyond τ and y, and the first n–1 equations in (VI.27) amounts to the vanishing of the contact form, one sees that the n^{th}-order differential equation (VI.26) is equivalent to a first-order system on $J^{n-1}(\mathbb{R}, \mathbb{R})$, as well as a hypersurface in $J^n(\mathbb{R}, \mathbb{R})$. One can thus conclude that any non-singular system of n^{th}-order differential equations (ordinary or partial, linear or nonlinear) in normal form is equivalent to a first-order system of quasilinear equations.

Of particular interest to us are k-jets of C^k sections of vector bundles $E \to M$. In principle, the definitions of the k-jet $j^k_x\phi$ of a C^k section ϕ: $M \to E$ at $x \in M$ and its k-jet prolongation $j^k\phi$ are analogous to those for more general functions between these manifolds. The difference is in the fact that the fiber structure on E translates into a fiber structure on the target projection $J^1(E) \to E$; in fact, it is a vector bundle. One can actually regard the case of general C^k functions f: $M \to N$ as a special case of C^k sections by regarding each f as a section of the trivial fiber bundle $M \times N \to M$.

The way that the k-jet formulation of systems of partial differential equations relates to their formulation in terms of differential operators is quite simple to explain if one considers the case of k-jets of sections of vector bundles. If one wishes to represent a k^{th} order differential operator (linear or not) $\mathcal{D}$: $E \to F$, by which we really mean $\mathcal{D}$ acts on sections, as a hypersurface in a manifold of jets, one need only define a C^k fiber-preserving map $\mathcal{O}$: $J^kE \to F$ such that if s: $M \to E$ is a section then $\mathcal{D}s$ takes the form $\mathcal{D}s = \mathcal{O} \cdot j^k s$. Hence, $\mathcal{O}$ plays the same role in this case that F did in the more elementary case that we first considered.

3. Exterior differential systems [6-9]. A third way of representing systems of partial differential equations that can be of advantage to some classes of problems is the method of exterior differential equations. Much of that methodology goes back to the seminal works of Cartan [**6**] and Kähler [**7**].

In general, a *differential system of rank k* on a manifold M is a vector sub-bundle $D(M) \subset T(M)$ of constant rank k. That is, one associates a k-plane in T_xM with each $x \in M$. For instance, a line field on M is a differential system of rank 1 and a field of hyperplanes is a differential system of *corank* 1.

An *integral submanifold* of a differential system of rank k on M is a submanifold σ: $N \to M$ such that tangent space to the submanifold – i.e., $d\sigma|_x(T_xN)$ – is contained $D_{\sigma(x)}M$ for each $x \in N$. When the dimension of N equals k, one calls such an integral submanifold *maximal*. The differential system $D(M)$ is called *integrable* iff there is an integral submanifold through each of its points and *completely integrable* iff there is a maximal integral submanifold through point. In the latter case, the images of the integral submanifolds partition M into what one calls a *foliation* of dimension k (or codimension $n - k$, n being the dimension of M); the integral submanifolds are then called *leaves*.

The necessary and sufficient condition for complete integrability is given by *Frobenius's theorem*, which can be stated in various forms. The one that pertains to the present definition of a differential system is that $D(M)$ is completely integrable iff the

vector space $\mathfrak{X}(D)$ of all sections of $D(M) \to M$ is closed under the Lie bracket of vector fields. This also means that $\mathfrak{X}(D)$ is a Lie subalgebra of $\mathfrak{X}(M)$; one also calls $D(M)$ *involutory* when this is true.

The way that an exterior differential system differs from a more general one is in the representation of the sub-bundle $D(M)$. Any k-form $\alpha \in \Lambda^k M$ defines an n-k-dimensional linear subspace of T_xM at each $x \in M$ by way of the set D_xM of all tangent vectors with the property that for any k vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ in D_xM one must always have:

$$a(\mathbf{v}_1, \ldots, \mathbf{v}_k) = 0. \tag{VI.32}$$

More concisely, one says that D_xM is the solution to the exterior algebraic equation:

$$\alpha = 0. \tag{VI.33}$$

An exterior differential system on M is then a differential system on M that is the simultaneous solution of a set of exterior algebraic equations:

$$0 = \alpha_i , \qquad i = 1, \ldots, p. \tag{VI.34}$$

in which the α_i are p differential forms of varying degree.

Of particular interest are the *Pfaffian systems*, for which all of the differential forms in the systems are 1-forms. Since 1-forms are obstructed in the same way as vector fields from being globally non-zero, one sees that in the general case – at least when M is compact – one will be dealing with differential systems with singularities.

The complete integrability of an exterior differential system of the form (VI.34) is given by a variant form of Frobenius: The exterior differential system (VI.34) is completely integrable iff:

$$0 = \alpha_i \wedge d\alpha_i \qquad \text{(all } i\text{)} \tag{VI.35}$$

which, in turn, is equivalent to the condition that there exist 1-forms η^i_j such that [23]:

$$d\alpha_j = \eta^i_j \wedge \alpha_i . \tag{VI.36}$$

The actual representation of a system of partial differential equations by an equivalent exterior differential system is not generally uniquely defined. For one thing, it is often more convenient to put higher-order partial differential equations into the form of systems of first-order equations. Furthermore, the manifold on which the exterior differential system is ultimately defined will generally have to be adapted to the nature of the problem.

---

[23] One can also say that the "ideal" in the exterior algebra $\Lambda^* M$ that is generated by the set $\{\alpha_1 , \ldots, \alpha_p\}$, namely, the vector space spanned by all finite linear combinations of expressions of the form $\beta \wedge \alpha_i$ where β is arbitrary, is closed under the exterior derivative. This is yet another way of stating Frobenius that is favored by the school of Chern, et al. [**9**].

In the case of the pre-metric vacuum Maxwell equations in the form $dF = 0$, $d{*}F = 0$ on a manifold M, and with $* = \# \cdot \kappa$; one might think on first glance that they already represent an exterior differential system. However, their solutions are 2-forms, not submanifolds of M, so if one desires to define an exterior differential system whose integral submanifolds are 2-forms then it is better to realize that since a 2-form is a smooth map F: $M \to \Lambda^2 M$ it is also a four-dimensional submanifold of the ten-dimensional manifold $\Lambda^2 M$. Hence, if a solution is to be an integral submanifold, one should define the exterior differential system on $\Lambda^2 M$, not on M; in effect, one must "lift" the system from M to $\Lambda^2 M$.

Since a local coordinate system on $\Lambda^2 M$ – i.e., a local trivialization – has the form $(x^\mu, F_{\mu\nu})$, by differentiation of the local expressions $F = \frac{1}{2} F_{\mu\nu}\, dx^\mu \wedge dx^\nu$, $*F = \frac{1}{2} *F_{\mu\nu}\, dx^\mu \wedge dx^\nu$, one sees that the exterior differential system on $\Lambda^2 M$ that represents the vacuum Maxwell equations takes the form:

$$\Theta^1 = \Theta^2 = 0, \tag{VI.37}$$

where [24]:

$$\Theta^1 = dF_{\mu\nu} \wedge dx^\mu \wedge dx^\nu, \qquad \Theta^2 = d{*}F_{\mu\nu} \wedge dx^\mu \wedge dx^\nu. \tag{VI.38}$$

Be careful to note that in these expressions the $F_{\mu\nu}$ and $*F_{\mu\nu}$ represent functions on $\Lambda^2 M$, not functions on M. However, if F is a section of $\Lambda^2 M \to M$ then the pull-back of Θ^1 by F becomes dF itself, and similarly one pulls Θ^2 back to $d{*}F$ by the section $*F$.

Often, one defines exterior differential systems on jet manifolds. For instance, the methods of prolongations of exterior differential systems to integrable ones that were first suggested by Cartan and later proved by Kuranishi [**10**] essentially amount to higher jet prolongations in which the supplementary variables that one is extending to are the higher jet manifold coordinates. We have already encountered one exterior differential system on a jet manifold in the form of the integrability condition for a section of $J^1(M, M) \to M$, as it is expressed in terms of the contact form θ^a, namely, the exterior differential system that is defined by the vanishing of those 1-forms. The integral submanifolds are then the integrable sections.

4. Boundary-value problems. [**11-13**]. When one is confronted with a linear differential operator L: $C^\infty(M) \to C^\infty(M)$ on smooth functions on a manifold, a natural problem to pose is that of solving the overdetermined linear differential equation:

$$Lf = \rho. \tag{VI.39}$$

As we pointed out above, the first question to address is that of integrability, which simply amounts to the statement that unless ρ is in the image $\mathrm{im}(L) = L(C^\infty(M))$ of the map L there can be no solution to begin with.

The second question to address, once one has identified the subspace of $C^\infty(M)$ in which the solutions exist, is whether a solution is unique. Of course, in the case of

[24] Actually, one can generally expand $d{*}F_{\mu\nu}$ into an expression in dx^κ and $dF_{\alpha\beta}$, but, for now, we pass over that fact for the sake of brevity.

differential operators this is generally impossible since even for the most elementary case of the operator d/dx acting on $C^\infty(\mathbb{R})$, smooth functions that differ by a constant function will map to the same derivative. Hence, one can hope to find, at best, a section of a surjection.

The issue then becomes one of how to characterize the nature of a particular section, and this is, of course, where one introduces initial/boundary conditions on the function f that would single it out from an infinite class of possibilities. Hence, we shall assume that M has boundary ∂M, so the boundary conditions on f – and possibly its derivatives – will be defined on ∂M. Therefore, the initial/boundary-value problem that is defined by (VI.39) for an integrable ρ is: Find that unique function $f \in C^\infty(M)$ that satisfies equation (VI.39) and has specified functions $\phi, \phi', \ldots \in C^\infty(\partial M)$ for the restriction to ∂M of f and its derivatives up to some order.

As long as L is linear, one will expect a solution of (VI.39) for a well-posed initial/boundary value problem to take the form of a linear operator L^{-1}: $\mathrm{im}(L) \to C^\infty(M)$, where our notation is suggestive only of the operator being a *right inverse* to L, so $LL^{-1} = I$; i.e.:

$$LL^{-1}\rho = \rho \tag{VI.40}$$

when the boundary conditions have been imposed on $L^{-1}\rho$.

One can also represent this situation in the form:

$$f = L^{-1}\rho \qquad (\mathrm{mod}\ C^\infty(\partial M)). \tag{VI.41}$$

The actual distinction between an initial-value problem and a boundary-values problem does not become clear until one looks at the specific nature of L – e.g., elliptic, hyperbolic, parabolic – and how that affects the nature of a "well-posed" problem; i.e., one that has a unique solution that depends continuously upon the given data. For instance, elliptic problems, such as might be defined by Poisson's equation, are fairly restrictive as far as what sort of boundary data one can specify.

The *Dirichlet problem* is defined by specifying only the boundary values of the function f, while the *Neumann problem* is defined by specifying only the boundary values of its normal derivative – viz., its derivative in the normal direction $\mathbf{n}$:

$$f_n = \mathbf{n}f = n^i \frac{\partial f}{\partial x^i} \tag{VI.42}$$

(Of course, we have to assume that ∂M is orientable in order to define $\mathbf{n} = \#^{-1}\mathcal{V}_\Sigma$, where $\mathcal{V}_\Sigma$ is the volume element on ∂M.)

One can also envision *mixed* or *Robin* boundary-value problems, at least when ∂M is composed of more than one connected component. In such problems, one might define f on some components and its normal derivative on others.

5. Initial-value problems [4, 5, 11]. Often, the problems that one poses in the case of dynamical fields take the form of initial-value – or *Cauchy* – problems. If one were dealing with a system of n^{th}-order ordinary differential equations then an initial-value problem would take the form of finding that particular integral curve that has a given kinematical state, in the sense of the position, and subsequent time derivatives up to order $n - 1$ at the initial time t_0 . In the case of partial differential equations, the initial point becomes an initial hypersurface, the initial position becomes the initial values of the solution function or field on the initial hypersurface, and the various subsequent time derivatives become corresponding normal derivatives of the solution.

One usually finds that only the normal derivative can be specified independently of the initial function, at least when it has derivatives of all orders involved, since that would determine the derivatives in all of the directions are tangent to the initial hypersurface.

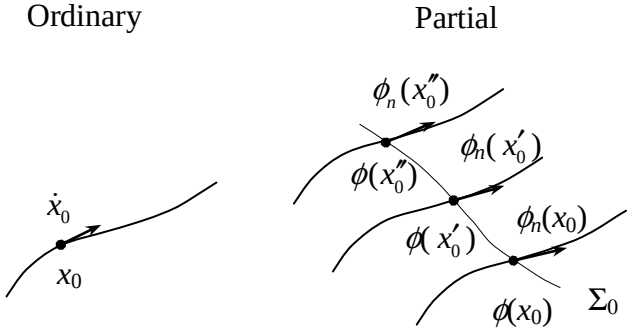


Figure 5. Initial-value problems for ordinary and partial differential equations.

a. Existence and uniqueness. It was Cauchy who first made significant progress towards the solution of the problem that bears his name, and later the proof was extended by Sonja Kowalevski [25] (see Courant and Hilbert [**5**] for the references). In essence, the Cauchy-Kowalevski theorem says that if a system of partial differential equations, in normal form, is given by an analytic function F on $J^{n-1}(M, \mathbb{R}^m)$ and the Cauchy (i.e., initial) data is analytic on some initial hypersurface Σ_0 then a unique solution to the Cauchy problem exists in a neighborhood of each point of Σ_0 for a sufficiently small time interval into the future.

One must be careful to distinguish positive time evolution from negative time evolution in the context of partial differential equations since in many cases the two are quite distinct. One can say that the "flows" of partial differential equations are then due to one-parameter *semigroups* of transformations, not one-parameter groups, as is the case for ordinary differential equations; this is especially true of diffusion equations.

Although the Cauchy-Kowalevski theorem is sufficiently general as to make no mention of the type of system that one dealing with – e.g., elliptic, parabolic, or hyperbolic – one finds that generally the Cauchy problem is not well-posed for elliptic

[25] Since there are a number of spellings for her name that are used in the literature, we follow the argument of John [**6**] in using this spelling, on the grounds that she herself used that spelling in her papers.

systems [26]. As it turns out, the very assumption that one can put the partial differential equation into normal form includes the constraint that initial hypersurface is not "characteristic," in a sense that follows from the signature type of the principal symbol. Furthermore, the restriction to analytic initial data is more severe than it sounds, since an important class of initial-value problems in the theory of waves concerns initial data that have discontinuities in their derivatives at some order, such as shock waves and acceleration waves. We shall discuss the Hadamard theory of the Cauchy problem for acceleration waves in chapter VIII.

The Cauchy-Kowalevski theorem was extended into the context of exterior differential systems by Cartan [**7**] and Kähler [**8**], although its very statement requires a considerable amount of preliminary definitions and constructions. We shall not elaborate here, but only refer the curious to Choquet-Bruhat [**9**] or Bryant, Chern, et al. [**10**]

b. Method of characteristics: first-order case. The simplest Cauchy problem is posed for a first-order partial differential equation in a real-valued function $y(x)$ on a manifold M. In local coordinates, the partial differential equation can be expressed in the form:

$$F(x^\mu, y, y_\mu) = 0, \tag{VI.43}$$

and for a function y that is a solution to this equation, one will have:

$$y_\mu = \frac{\partial y}{\partial x^\mu} \tag{VI.44}$$

in addition to (VI.43).

The initial-value problem for this class of partial differential equations is completely solvable by Cauchy's method of characteristics. This method reduces the solution of the initial-value problem for the partial differential equation (VI.43) to the solution of an initial-value problem for a system of first-order ordinary differential equations on the manifold $J^1(M; \mathbb{R})$, whose local coordinates take the form (x^μ, y, y_μ).

In order to define the characteristic equations for a first-order partial differential equation of the form (VI.43) we first regard the function F as a differentiable function on $J^1(M; \mathbb{R})$, so the partial differential equation in question represents a hypersurface in $J^1(M; \mathbb{R})$. Hence, F defines a 1-form on $J^1(M; \mathbb{R})$, namely:

$$dF = \frac{\partial F}{\partial x^\mu} dx^\mu + \frac{\partial F}{\partial y} dy + \frac{\partial F}{\partial y_\mu} dy_\mu . \tag{VI.45}$$

For an integrable section s: $M \to J^1(M; \mathbb{R})$ this pulls down to:

[26] One can, however, think of equipotential hypersurfaces as "evolving" along the field lines.

$$d(s^*F) = s^*dF = \left(\frac{\partial F}{\partial x^\mu} + y_{,\mu} \frac{\partial F}{\partial y} \right) dx^\mu + \frac{\partial F}{\partial y_\mu} dy_\mu . \qquad \text{(VI.46)}$$

In order to turn (VI.45) into a vector field on $J^1(M; \mathbb{R})$, we first look at the 2-form that the contact form $\theta = dy - y_\mu\, dx^\mu$ defines, namely:

$$d\theta = dx^\mu \wedge dy_\mu . \qquad \text{(VI.47)}$$

If one has a vector field:

$$X = X^\mu \frac{\partial}{\partial x^\mu} + X^y \frac{\partial}{\partial y} + X_\mu \frac{\partial}{\partial y_\mu} \qquad \text{(VI.48)}$$

on $J^1(M; \mathbb{R})$ then one can define a 1-form using X and $d\theta$, namely:

$$i_X d\theta = X^\mu\, dy_\mu - X_\mu\, dx^\mu . \qquad \text{(VI.49)}$$

If one requires, moreover, that θ itself annihilate X, so:

$$X^y = y_{,\mu} X^\mu , \qquad \text{(VI.50)}$$

then the combined algebraic equations:

$$i_X d\theta = dF, \qquad \theta(X) = 0, \qquad s^*\theta = 0 \qquad \text{(VI.51)}$$

define X uniquely, and its local components are then:

$$X^\mu = \frac{\partial F}{\partial y_\mu} , \qquad X^y = y_{,\mu} X^\mu , \qquad X_\mu = -\left(\frac{\partial F}{\partial x^\mu} + y_{,\mu} \frac{\partial F}{\partial y} \right) . \qquad \text{(VI.52)}$$

This vector field on $J^1(M; \mathbb{R})$ is then the *characteristic vector field* for the partial differential equation that F defines. The system of $2n + 1$ ordinary differential equations that X defines is then the system of *characteristic equations*:

$$\frac{dx^\mu}{d\tau} = \frac{\partial F}{\partial y_\mu} , \qquad \frac{dy}{d\tau} = y_{,\mu} X^\mu , \qquad \frac{dy_\mu}{d\tau} = -\left(\frac{\partial F}{\partial x^\mu} + y_{,\mu} \frac{\partial F}{\partial y} \right) . \qquad \text{(VI.53)}$$

To solve the initial-value problem that is defined by giving the values $y(x_0)$ of the function y on some initial hypersurface Σ_0 in M and requiring that y be a solution to the partial differential equation (VI.43) that is defined on all of M, one first converts it into an initial-value problem for the system of ordinary differential equations (VI.53) by means of the 1-jet prolongation of y on Σ_0. Hence, each pair $(x_0, y(x_0))$ turns into a 1-jet $(x_0, y(x_0), y_{,\mu}(x_0))$ and if one projects the unique integral curve $(x^\mu(\tau, x_0), y(\tau, x_0), y_{,\mu}(\tau, x_0))$ to

X that passes through this point back down to M then one obtains a curve $x^\mu(\tau) = x^\mu(\tau, x_0)$ in M that is transverse to the initial hypersurface Σ_0 at x_0 for each choice of x_0.

The solution $y(x)$ is then obtained by the requirement that y be constant on each of the projected integral curves. Hence, in a coordinate system (τ, x_0^i) that is adapted to Σ_0 one will have:

$$y(\tau, x_0^i) = y(x_0^i) \tag{VI.54}$$

for all τ in the range of values for which the integral curve is defined.

One must observe that for the case of an F that does not depend upon y, numerous reductions of scope follow. The manifold $J^1(M; \mathbb{R})$ reduces to T^*M, $F = F(x^i, y_i)$ becomes a function on T^*M, θ becomes its canonical 1-form, so $d\theta = -\, y_i\, dx^i$ is the canonical symplectic form, and the characteristic equations of F become:

$$\frac{dx^i}{d\tau} = \frac{\partial F}{\partial y_i}, \qquad \frac{dy_i}{d\tau} = -\frac{\partial F}{\partial x^i}, \tag{VI.55}$$

which are Hamilton's equations is one regards F as a Hamiltonian function.

Note that a section s: $M \to T^*M$ can be regarded as either a covector field on M or a 1-form. It is integrable relative to d iff:

$$s = dS \tag{VI.56}$$

for some differentiable function S on M; i.e., iff it is an exact 1-form.

The partial differential equation:

$$F(x^i, y_{,i}) = 0 \tag{VI.57}$$

that one obtains for any section s: $M \to T^*M$ is then the (stationary) *Hamilton-Jacobi equation* that F defines.

Although it may seem like a restriction in scope to eliminate the functional dependency of F on y, actually, it is not. One simply adds y, which we now call τ, as an extra dimension to M – i.e., one goes to $M\times\mathbb{R}$ – so $T^*(M\times\mathbb{R})$ takes on a fiber dimension y_0 that refers to partial differentiation with respect to τ.

In order to obtain the usual time-varying Hamilton-Jacobi equation, one must replace $F(x^i, y_i)$ with $y_0 + F(x^i, y_i)$. For an integrable section, $y_0 = \partial S/\partial\tau$ and $y_i = \partial S/\partial x^i$, so the Hamilton-Jacobi equation takes the time-varying form:

$$\frac{\partial S}{\partial \tau} + F\left(x^i, \frac{\partial S}{\partial x^i}\right) = 0. \tag{VI.58}$$

One must be aware that although the method of characteristics gives a complete solution to the Cauchy problem for a first-order partial differential equation in a real-valued function, it does not extend to the case of a system of first-order equations. This is because, as we mentioned above, any n^{th}-order partial differential equation can be

converted into an equivalent system of first-order equations by introducing the successive derivatives up to order $n - 1$ as auxiliary variables.

c. Method of characteristics: second-order case. A second-order partial differential equation in a C^2 function y on M takes the form:

$$F(x^\mu, y, y_{,\mu}, y_{,\mu,\nu}) = 0. \tag{VI.59}$$

Hence, F is now a differentiable function on the manifold $J^2(M; \mathbb{R})$ of 2-jets of C^2 functions on M. A 2-jet is simply the next level of differentiation from a 1-jet, namely, an equivalence class of functions that are defined on some neighborhood of each point x M that have common values as functions at x, along with their first and second derivatives. A local coordinate system about a point $j^2_x y \in J^2(M; \mathbb{R})$ then has the form $(x^\mu, y, y_\mu, y_{\mu\nu})$, in which $y_{\mu\nu} = y_{\nu\mu}$, since they must behave like second partial derivatives.

When one considers jet prolongations beyond the first one, one must be very careful concerning a certain subtlety: the first prolongation of a section s of $J^1(M; \mathbb{R}) \to M$ does not have to be a section of $J^2(M; \mathbb{R}) \to M$; i.e., a 1-jet does not always prolong to a 2-jet. This is because if s takes the local form $s(x) = (x^\mu, y(x), y_\mu(x))$ then its first prolongation will take the form $j^1 s(x) = (x^\mu, y(x), y_\mu(x), y_{\mu,\nu}(x))$. However, unless s is an integral section to begin with, so $y_\mu(x) = y_{,\mu}(x)$ the resulting functions $y_{\mu,\nu}(x))$ will not generally be symmetric in their lower indices. Hence, the manifold $J^1(J^1(M; \mathbb{R}))$ is higher-dimensional than $J^2(M; \mathbb{R})$, which can be embedded as a submanifold in the latter manifold.

The Cauchy problem for a second-order partial differential equation of the form (VI.59) for an initial (orientable) hypersurface Σ_0 in M then amounts to defining not only the values of y on Σ_0, but also its normal derivative y_n .

Although it would be nice if there were a characteristic vector field on $J^2(M; \mathbb{R})$ that would allow one to convert the initial-value problem for the partial differential equation (VI.55) into an initial-value problem for a system of ordinary differential equations, alas, such is not the case. However, as we shall see in the next chapter, there is a process that is almost as convenient: One first converts the second-order partial differential equation into a – generally, nonlinear – first-order partial differential equation, which one also calls the "characteristic equation," and which usually captures the essentials of the second-order equation to a lesser extent. One then solves this first-order partial differential equation by the method of characteristics, and, to avoid confusion, one calls the characteristic equations for the first-order equation "bicharacteristic" equations for the second-order one.

6. Distributions on differential forms [15-17]. In order to lead into a proper treatment of solving boundary-value problems for systems of linear differential equations – whether ordinary or partial – one must generalize from differential operators that act on

functions or differential forms to differential operators that act on distributions, which are also sometimes called "generalized functions." This is largely due to the fact that in order to define the Green function for an integral operator, which we shall do in the next section, one must unavoidably introduce the Dirac delta function, which is not a function, at all, but a distribution. Consequently, the Green function itself becomes a distribution, as well, and the defining equation makes sense only in the distributional sense.

a. Continuous linear functionals. The infinite-dimensional vector spaces $\Lambda^k M$ and $\Lambda^k_C M$, which consists of smooth *k*-forms with compact support, can be given topologies, although we shall not elaborate here [27], except to say that one can make sense of the notion of a *null sequence* of *k*-forms; viz., a sequence α_i, $i = 1, 2, \ldots$ of *k*-forms that converges uniformly to 0. One refers to the elements of $\Lambda^k_C M$ as *test k-forms.*

A linear functional *T* on either vector space is called *continuous* if the sequence of real numbers $T[\alpha_i]$ converges to 0 for every null sequence α_i. A *distribution* on either space $\Lambda^k M$ and $\Lambda^k_C M$ is a continuous, linear functional on that space. Hence, a distribution is an element of the dual spaces to $\Lambda^k M$ and $\Lambda^k_C M$, as topological vector spaces, which we denote by $(\Lambda^k)'$ and $(\Lambda^k_C)'$, resp. In the case of distributions on test *k*-forms, the term that de Rham used was "currents," and they defined the foundations for his formulation of his famous theorem.

This latter fact gives us our first example of a distribution on *k*-forms, regardless of whether their support is compact or not, namely, any *k*-chain $c_k \in C_k(M; \mathbb{R})$ defines a distribution by way of:

$$c_k[\alpha] = \int_{c_k} \alpha \,. \tag{VI.60}$$

Hence, one has a linear map [.]: $C_k(M; \mathbb{R}) \rightarrow (\Lambda^k M)'$ that takes the *k*-chain c_k to the distribution c_k[.]. Since we have not given $C_k(M; \mathbb{R})$ a topology, it is meaningless to speak of continuity for this map, but we shall not need that condition, anyway. However, we can say something about injectivity, which amounts to the issue of whether the kernel of the map [.] is trivial or not. An element of the kernel is a *k*-chain such that any *k*-form will integrate to zero over it. However, that can only be the 0-chain, since we are not allowing degenerate *k*-chains of dimension less than *k*, which might have "measure zero." Hence, the map in question is injective. It is not surjective, but the image of $C_k(M; \mathbb{R})$ is dense in $(\Lambda^k M)'$ (see de Rham [**15**]).

Any $(n - k)$-form α defines a distribution on test *k*-forms by way of:

$$\alpha[\beta] = \int_M \alpha \wedge \beta \,. \tag{VI.61}$$

[27] The details can be found in de Rham [**15**] or Hörmander [**16**]. For a general discussion of distributions, one might confer Friedlander [**17**].

Now, we have a linear map $[.]: \Lambda^{n-k} \to (\Lambda^k_C)'$, $\alpha \mapsto \alpha[.]$. It is continuous and injective since $\alpha \wedge \beta = 0$ for all β iff $\alpha = 0$.

Finally, since the evaluation $\alpha(\mathbf{A})$ of a k-form α on a k-vector field $\mathbf{A}$ produces a function, when α has compact support, one can use this to define yet another distribution on $\Lambda^k_C M$:

$$\mathbf{A}[\alpha] = \int_M \alpha(\mathbf{A})\, \mathcal{V} . \tag{VI.62}$$

b. Operations on distributions. Many of the operations that that one performs on differential forms can be performed on distributions, as well. For instance, one can clearly form scalar combinations of distributions.

One can form the exterior product of a distribution T that acts on $k+l$-forms with an l-form α to form a distribution $T \wedge \alpha$ that acts on k-forms:

$$(T \wedge \alpha)[\beta] = T[\alpha \wedge \beta]. \tag{VI.63}$$

We can thus define left-multiplication of distributions by α as a linear map $l_\alpha : (\Lambda^{k+l})' \to (\Lambda^k)'$, which then works more like the operator of taking the interior product by α does on $(k+l)$-vector fields.

The tensor product $T_1 \otimes T_2$, of two distributions can be defined as a distribution on $\Lambda^k \otimes \Lambda^l$:

$$(T_1 \otimes T_2)[\alpha \otimes \beta] = T_1(\alpha)\, T_2(\beta). \tag{VI.64}$$

Similarly, the exterior product of two distributions $T_1 \in (\Lambda^k)'$ and $T_2 \in (\Lambda^l)'$ can be defined as a distribution $(T_1 \wedge T_2)$ on Λ^{k+l}, but the anti-symmetrization of (VI.63) is not always immediate. Thus, we can define the exterior algebra $(\Lambda^*)'$ of distributions on differential forms as being generated by $(\Lambda^1)'$.

The interior product of a distribution $T \in (\Lambda^k)'$ with a vector field $\mathbf{v}$ can be defined predictably:

$$(i_\mathbf{v} T)[\alpha] = T[i_\mathbf{v} \alpha]. \tag{VI.65}$$

Hence, it is a linear map $i_\mathbf{v} : (\Lambda^{k-1})' \to (\Lambda^k)'$, so it behaves more like $e_\mathbf{v}$ does on Λ_{k-1} .

The divergence of a distribution T can be defined to be adjoint to the exterior derivative of the differential forms that it acts on:

$$\delta T[\alpha] = -(-1)^{n-k}\, T[d\alpha]. \tag{VI.66}$$

The choice of sign comes from the product rule for the exterior derivative if one uses $T = \beta[.]$, where β is an $(n-k)$-form:

$$d(\beta \wedge \alpha) = d\beta \wedge \alpha + (-1)^{n-k}\, \beta \wedge d\alpha \tag{VI.67}$$

which vanishes since $\beta \wedge \alpha$ is an n-form on an n-dimensional manifold.

We can also define distributions on the vector spaces Λ_k or $\Lambda_{k,C}$ by similar means to the foregoing. An $(n-k)$-chain c_{n-k} defines a distribution on Λ_k by way of:

$$c_{n-k}[\mathbf{A}] = \int_{c_{n-k}} \#\,\mathbf{A}\,. \tag{VI.68}$$

A k-form α defines a distribution on $\Lambda_{k,C}$ by way of:

$$\alpha[\mathbf{A}] = \int_M \alpha(\mathbf{A})\,\mathcal{V} = \int_M \alpha \wedge \#\,\mathbf{A}\,. \tag{VI.69}$$

We denote the topological vector spaces of distributions on Λ_k and $\Lambda_{k,C}$ by $(\Lambda_k)'$ and $(\Lambda_{k,C})'$, predictably.

The Poincaré duality isomorphism #: $\Lambda_k \to \Lambda^{n-k}$ can be transposed to a corresponding isomorphism #′: $(\Lambda^{n-k})' \to (\Lambda_k)'$ in the obvious way:

$$\#'T[\mathbf{A}] = T[\#\mathbf{A}]. \tag{VI.70}$$

The exterior derivative operator d: $(\Lambda_k)' \to (\Lambda_{k+1})'$ is basically the transpose of the divergence operator on Λ_{k+1}:

$$d\tau[\mathbf{A}] = -(-1)^{n-k}\tau[\delta\mathbf{A}]. \tag{VI.71}$$

In order to verify this, it is simplest to use $\tau[\mathbf{A}]$ in the form $\int_M \alpha \wedge \#\,\mathbf{A}$ and apply the product rule for exterior derivatives, as before.

c. Vector-valued distributions. It will prove necessary for us to extend our notion of a real-valued distribution on differential forms to that of a vector-valued distribution on differential forms. We define a *vector-valued distribution* on differential forms to be a continuous linear functional on Λ^k or Λ^k_C that takes its values in a topological vector space V, such as $\mathbb{R}^m$, a specified fiber E_x of a vector bundle $E \to M$, or the vector space $\Gamma(E)$ of sections of that vector bundle.

One example of a vector-valued distribution on k-forms is the *evaluation functional* $\delta_x[]$ for $x \in M$, which takes any k-form α to its value at x:

$$\delta_x[\alpha] = \alpha_x\,. \tag{VI.72}$$

Hence, the vector space V is the fiber Λ^k_x of the vector bundle $\Lambda^k M \to M$.

Another elementary example of a vector-valued distribution on k-forms is the *identity map*:

$$I[\alpha] = \alpha\,. \tag{VI.73}$$

The vector space V in this case is Λ^k itself.

A broad class of vector-valued distributions on differential forms is defined by *integral operators*, which take the form K: $\Lambda^k_C \to \Lambda^l_C$:

$$K[\alpha(y)] = \beta(x) = \int_M K(x,y) \wedge \alpha(y)\,. \tag{VI.74}$$

in which the integration is over all $y \in M$. (The fact that the notation for the independent variable changes is irrelevant to the fact that both x and y range over all of M.)

The *kernel* $K(x, y)$ of the operator then takes its values in the vector space $\Lambda_x^l \otimes \Lambda_y^{n-k}$ for each pair $(x, y) \in M \times M$; the integral operator K can then be regarded as a *two-point distribution.* In local coordinate terms, the kernel will be expressible in the form:

$$K(x, y) = \frac{1}{l!(n-k)!} K_{i_1 \cdots i_l, j_1 \cdots j_{n-k}}(x, y) dx^{i_1} \wedge \cdots \wedge dx^{i_l} \otimes dy^{j_1} \wedge \cdots \wedge dy^{j_{n-k}} . \qquad \text{(VI.75)}$$

The kernel will be called *decomposable* iff it takes the form $K(x, y) = \alpha(x) \otimes \beta(y)$, which implies that the local component functions take the form:

$$K_{i_1 \cdots i_l, j_1 \cdots j_{n-k}}(x, y) = K'_{i_1 \cdots i_k}(x) K''_{j_1 \cdots j_l}(y) . \qquad \text{(VI.76)}$$

It is commonplace to represent both the evaluation functional $\delta_x[.]$ and the identity operator $I[.]$ in terms of integral operators, even though both operators are purely algebraic – hence, local – and not integral operators, which are global in character:

$$\delta_x[\alpha] = \int_M \delta(x, y) \wedge \alpha(y) = \alpha(x), \qquad \text{(VI.77a)}$$

$$I[\alpha(x)] = \int_M I(x, y) \wedge \alpha(y) = \alpha(x). \qquad \text{(VI.77b)}$$

The fictitious kernel for the former distribution is the *Dirac delta function* $\delta(x, y)$; the equally fictitious kernel $I(x, y)$ is referred to as the *reproducing kernel.*

d. Fredholm theory. What Fredholm was originally concerned with was the solution of three basic classes of integral equations. When expressed in terms of an integral operator K that is the right inverse to a differential operator D they take the forms:

$$K\rho = f, \qquad \text{(VI.78a)}$$

$$(K - \lambda I)\rho = f, \qquad \text{(VI.78b)}$$

$$(K - \lambda I)\rho = 0. \qquad \text{(VI.78c)}$$

One then looks for solutions in the form of ρ.

Hence, we see that what he was concerned with were the solutions to the general inhomogeneous differential equation $Df = \rho$ and two equations that grew out of the eigenvalue equation for K:

$$K\rho = \lambda\rho . \qquad \text{(VI.79)}$$

The non-vanishing eigenvalues λ of K are easily seen to be inverses to the eigenvalues of D since one must have $DK = I$.

As for the zero eigenvalues, the corresponding eigenfunctions ρ belong to the kernel of the operator K. Hence, since a linear operator is injective iff it has a vanishing kernel,

one deduces the *Fredholm alternative:* Either (VI.78c) has a non-vanishing solution for $\lambda = 0$ or (VI.78a) has a unique solution for any f in the image of K.

7. Fundamental solutions [12-14, 18-20]. As we observed above, solving systems of linear differential equations, whether ordinary or partial, amounts to finding a right inverse to the linear differential operator that satisfies some boundary/initial-value problem that makes the solution exist uniquely. Such a right inverse will then represent a linear integral operator on the vector space of functions or sections that the differential operator acts on. In this section, we shall discuss the nature of this construction when one is concerned with functions and sections on more general manifolds than vector spaces.

a. General definitions. When the integral operator in question is the right-inverse operator D^{-1} that is associated with the differential operator D, the two-point function $\gamma(x, y)$ that represents the kernel of the integral operator is called a *fundamental solution*. Since the operator equation is $DD^{-1} = I$, the distributional equation that is associated with it is:

$$D\gamma_y(x, y) = -\delta(x, y). \tag{VI.80}$$

The subscript y indicates the independent variables that the differentiation affects.

Let us illustrate the representation of the solution of a boundary-value problem for a differential equation by fundamental solutions in the simplest possible case of solving:

$$\frac{df(x)}{dx} = \rho(x), \qquad f(0) = f_0 . \tag{VI.81}$$

A simple quadrature gives the solution as:

$$f(x) = f_0 + \int_0^x \rho(y)dy . \tag{VI.82}$$

Note that the only definitive restrictions on f and ρ are that they be differentiable and integrable, respectively.

At this point, we can introduce the fundamental solution:

$$G(x, y) = \begin{cases} 1 & x \le y, \\ 0 & \text{otherwise}, \end{cases} \tag{VI.83}$$

whose graph looks like:

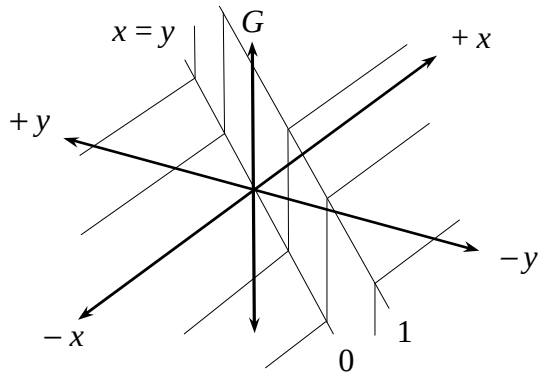


Figure 6. The fundamental solution for *d/dx*.

Note that both of the partial derivatives of $G(x, y)$ are zero everywhere except for the diagonal subset $\Delta = \{(x, y) \mid x = y\}$, where they are undefined and essentially infinite:

$$\frac{\partial G}{\partial x} = \frac{\partial G}{\partial y} = \begin{cases} 0 & x \neq y \\ \text{"}\infty\text{"} & x = y. \end{cases} \tag{VI.84}$$

(Our quotation marks around ∞ are there to suggest that it is more proper to say that the functions are simply undefined on the diagonal.)

Let us apply $\partial G/\partial y$ to $f(y)$ as a distribution:

$$\begin{aligned} \partial G/\partial y\,[f] &= \int_0^1 \frac{\partial G}{\partial y} f(y)\,dy = \int_0^1 \frac{\partial (Gf)}{\partial y}\,dy - \int_0^1 G(x, y)\frac{df}{dy}\,dy \\ &= \int_0^1 d(Gf) - \int_0^1 G(x, y)\rho(y)dy\,. \end{aligned} \tag{VI.85}$$

The first integral vanishes since it is equal to $G(x, 1)f(1) - G(x, 0)f(0)$ and the second integral equals:

$$-f(x) = -\int_0^1 \delta(x, y)\, f(y)dy\,. \tag{VI.86}$$

Hence, as a distributional equation, we have:

$$\frac{\partial G(x, y)}{\partial y} = -\,\delta(x, y). \tag{VI.87}$$

One finds that generally fundamental solutions are undefined on the diagonal, and not merely discontinuous, as in the present case.

The fundamental solution $\gamma(x, y)$ can also be used to construct solutions to boundary-value problems for (VI.39). One starts with *Green's formula* as the definition of a *self-adjoint* linear operator L: $C^\infty(M) \to C^\infty(M)$:

$$\int_M (uLv - vLu)\mathcal{V} = \int_{\partial M} \left(uv_n - vu_n\right)\mathcal{V}_{\partial M} \,. \tag{VI.88}$$

Then, one applies this to the particular case in which the function v is the fundamental solution for L and u is a solution to the inhomogeneous equation $Lu = \rho$. (VI.88) then takes the form:

$$u(x) = \int_M \gamma(x,y)\rho(y)\mathcal{V}_y + \int_{\partial M} \frac{\partial \gamma(x,y)}{\partial n_y} u(y)\mathcal{V}_{\partial M} - \int_{\partial M} \gamma(x,y)\frac{\partial u(y)}{\partial n_y}\mathcal{V}_{\partial M} \,. \tag{VI.89}$$

The first integral on the right-hand side is called the *volume potential*, and it represents the contribution to u that is due to the presence of a non-zero source ρ. The second one is called the *single-layer surface potential*, and one sees that it is driven by the boundary values of u, while it is the normal derivative of γ that serves as the integral kernel. The final term is referred to as the *double-layer surface potential*, and one sees that it is driven by the normal derivatives of u on ∂M.

In the homogeneous case, where $\rho = 0$, one can think of the boundary-value functions f and f_n as the "source" of the field f inside of the boundary ∂M.

In the case where L is the Laplacian operator Δ, if one poses the Dirichlet problem for u then this implies the boundary condition on γ:

$$\gamma|_{\partial M} = 0, \qquad \frac{\partial \gamma(x,y)}{\partial n_y} = G(x,y), \tag{VI.90}$$

in which $\partial/\partial n_y$ refers to the normal derivative in terms of the y variables. The function $G(x, y)$ is then referred to as the *Green function* [28] for this problem.

If one poses the Neumann problem then the corresponding boundary conditions for γ are:

$$\left.\frac{\partial \gamma(x,y)}{\partial n_y}\right|_{\partial M} = 0, \qquad \gamma(x,y) = N(x,y), \tag{VI.91}$$

and one refers to the function $N(x, y)$ as the *Neumann function* for the problem.

The reader should be advised that it quite commonplace for the term "Green function" to be used as a generic reference to all integral operator kernels. In particular, the usage of that term in quantum field theory is not specific to the Dirichlet problem, and indeed it is entirely possible for a Green function to represent something that is not even a differential operator, such as a pseudo-differential operator.

The method of Green functions can still be used in the case of hyperbolic linear second-order partial differential equations. In particular, the construction of a solution to the Cauchy problem for a linear forced wave equation $\Box u = \rho$ by means of (VI.89) is still

[28] As pointed out by Jackson [**18**], Rohrlich [**19**], and others, the popular phrases "the Green's function" and "a Green's function" are ungrammatical, since one should not mix articles with possessive adjectives. After all, one does not say "the Laplace's equation" or "a Bessel's function."

valid. The main difference is that one must use "fractional hyperbolic potentials" in order to define the Green function, and this has the consequence that the character of the resulting wave solution depends upon whether the dimension of M is even or odd. One can also compute the Green function for $\Box$ by means of the Fourier transform, which we shall discuss below.

b. Topological integral operators [29]. Since we have already seen that the linear differential operators d and δ have topological significance, by way of de Rham cohomology and homology, respectively, it should come as no surprise that their operators also have topological significance.

A *chain map* f: $C_*(M) \to C_*(N)$ from one chain complex $C_*(M)$ to another one $C_*(N)$ is a linear map that commutes with the boundary operator ∂. That is, if c_k is a k-chain in M then the boundary of $f(c_k)$ is the image of c_k under f; in other words, a chain map takes boundaries to boundaries.

A *chain homotopy* from one chain map f to another chain map g is a linear operator H: $C_*(M) \to C_{*+1}(N)$, such that:

$$\partial H + H\partial = f - g. \tag{VI.92}$$

When this is applied to a k-cycle z_k the result is:

$$\partial H z_k = f(z_k) - g(z_k). \tag{VI.93}$$

Since f and g are chain maps the images $f(z_k)$ and $g(z_k)$ will be cycles in some dimension and (VI.92) then says that they will also be homologous. Hence, f and g induce the same map in homology since this means that $f[z_k] = g[z_k]$.

In particular, one might consider *chain contractions*, which are chain homotopies from the identity map to the zero map, which makes:

$$\partial H + H\partial = I. \tag{VI.94}$$

When applied to a specific k-chain c_k this says:

$$\partial H c_k + H\partial c_k = c_k\,. \tag{VI.95}$$

In particular, when applied to a k-cycle z_k, the result is:

$$\partial H z_k = z_k\,. \tag{VI.96}$$

However, here we see the limitation on whether such an H can even exist, since (VI.96) says that the k-cycle z_k must also be a k-boundary, namely, the boundary of the k+1-chain Hz_k ; that is, z_k must be homologous to zero. If H is to be defined on all k-cycles then one must have that they are all k-boundaries; i.e., $H_k(\Sigma)$ must be trivial.

[29] For the basic definitions in homology, see Vick, Rotman, or Greenberg, as cited in Chap. II. The representation in terms of differential forms is given in de Rham[**15**] or Bott and Tu, also cited in Chap. II.

We can represent this situation in de Rham homology, which pertains directly to static electric and magnetic fields. Now, we represent a chain contraction by a map H: $\Lambda_k \rightarrow \Lambda_{k+1}$ such that:

$$\delta H + H\delta = I. \tag{VI.97}$$

When applied to a k-cycle $\mathbf{A}$ this becomes:

$$\delta H\mathbf{A} = \mathbf{A}, \tag{VI.98}$$

which is to say that H is a right-inverse for the linear differential operator δ.

A local representation for the operator H can be defined about any point $x_0 \in M$ that is associated with a coordinate chart (U, x^μ), which is associated with a *radius vector field* $\mathbf{r}$:

$$\mathbf{r}(x) = x^\mu(x)\frac{\partial}{\partial x^\mu}. \tag{VI.99}$$

This vector field vanishes at the point x_0, which corresponds to the origin in $\mathbb{R}^n$ under the coordinate map.

By means of this vector field, one can define a differentiable curve $x(\lambda)$ from x_0 to any other $x \in U$ by means of $\lambda\mathbf{r}(x)$, where $\lambda \in [0, 1]$, and we abbreviate the expression $\lambda\mathbf{r}(x)$ to simply λx. If we are given an arbitrary k-vector field $\mathbf{A} \in \Lambda_k$ then we can define a k+1-vector field along this curve by way of $e_\mathbf{r}\mathbf{A}(\lambda\mathbf{r}) = \mathbf{r} \wedge \mathbf{A}(\lambda\mathbf{r})$ and a k+1-vector-valued 1-form $\lambda^{k-1}\, \mathbf{r} \wedge \mathbf{A}(\lambda\mathbf{r})\, d\lambda$. By integration along the curve, this gives a k+1-vector at x:

$$H\mathbf{A}(x) = \int_0^1 \lambda^{k-1}\, \mathbf{r} \wedge \mathbf{A}(\lambda x)\, d\lambda. \tag{VI.100}$$

As for the exterior derivative operator d: $\Lambda^k \rightarrow \Lambda^{k+1}$, a cochain contraction operator for it will take the form H': $\Lambda^{k+1} \rightarrow \Lambda^k$ such that:

$$dH' + H'd = I. \tag{VI.101}$$

Since the operator d is the transpose to the operator δ, the operator H' is the transpose to the operator H above. All that is actually necessary for the construction of H is to replace the operator $e_\mathbf{r}$ with its transpose $i_\mathbf{r}$. Hence, if $\alpha \in \Lambda^{k+1}(U)$ then we can define a k-form by means of $i_\mathbf{r}\alpha$, and for every curve of the form $x(l)$ we can then define a 1-form with values in $\Lambda^k(U)$ out of $\lambda^k\, i_\mathbf{r}\alpha_{\lambda x}\, d\lambda$. We then obtain our desired k-form $H'\alpha$ by integration:

$$H'\alpha_x = \int_0^1 \lambda^k i_\mathbf{r}\alpha_{\lambda x}\, d\lambda. \tag{VI.102}$$

Although we have constructed integral operators that represent right-inverses to the divergence and exterior derivative operators, nevertheless, we can see that they are not presented in a form that would suggest a Green function for a kernel. For one thing, we

are not integrating over *M* itself. In order to define such Green functions, one generally needs to look at specific problems, such as radially-symmetric solutions, which we shall do in the next chapter, since we are getting back into the realm of traditional electromagnetism.

8. Fourier transforms [14, 16, 17, 21, 22]. Obviously, since entire books have been written on just the subject of Fourier analysis, not to mention its applications to physics, we cannot hope to make a very far-reaching discussion of it in a single section of a chapter. However, since the ultimate objective of this book is to examine the aspects of electromagnetism that have a pre-metric character to them, we cannot avoid some discussion of Fourier analysis as it relates to such considerations.

a. Linear theory. Although sometimes Fourier transforms are defined in terms of scalar products, nevertheless, upon closer inspection one finds that as long as one regards a wave vector *k* as a *covector* the real issue is not one of scalar products but *bilinear pairings* of vectors and covectors. The scalar product then makes its first appearance in the context of the inner product on the function space in question.

Just to make the discussion more specific, let us first examine the usual Fourier transform for a complex function *f* on an *n*-dimensional vector space *V*:

$$\hat{f}(k) = \mathcal{F}f[k] = \int_V e^{-ik(x)} f(x)\, \mathcal{V}\,. \tag{VI.103}$$

In this definition, $k \in V^*$, so $\hat{f}$ – or $\mathcal{F}f$ – is a complex function on V^*, and $\mathcal{V}$ is the volume element for *V*.

Following Duistermaat [**21**], we restrict ourselves to complex functions in the *Schwartz spaces* $\mathcal{S}$ and $\mathcal{S}'$, which consist of complex functions on *V* and V^*, respectively, that satisfy the constraint that the local functions $x^\alpha(\partial^\beta f / \partial x^\beta)$ [$k_\alpha(\partial^\beta \hat{f} / \partial k_\beta)$, resp.] are bounded for all multi-indices α, β. (A *multi-index* is a notation for abbreviating elaborate algebraic expressions, as one often encounters in multivariable analysis. In the former expression, x^α means a product $x^{\alpha_1}x^{\alpha_2}\cdots x^{\alpha_k}$ while $\partial^\beta f / \partial x^\beta$ refers to the mixed partial derivative of $f(x^1, \ldots, x^n)$ with respect to the multi-indexed variables, and analogously for the expressions in *k* and $\hat{f}$.) The motivation for this constraint is to be found in the extension of Fourier transforms to distributions, namely, that it is not enough to be dealing with smooth functions of compact support, but one must restrict oneself to "tempered" distributions.

One then finds that the Fourier transform can be regarded as a linear isomorphism $\mathcal{F}\colon \mathcal{S} \to \mathcal{S}', f \mapsto \hat{f}$, whose inverse transform is:

$$\mathcal{F}^{-1}\hat{f}[x] = \frac{1}{(2\pi)^n}\int_{V^*} e^{ik(x)}\hat{f}(k)\,\mathcal{V}_k\,; \tag{VI.104}$$

this time $\mathcal{V}_k$ is the volume element on V^*.

There are various ways of interpreting the preceding constructions in terms of other mathematical concepts. For instance, one can think of the set $B = \{e^{-ik(x)} \mid k \in V^*\}$ as an uncountable basis for the vector space $\mathcal{S}$, and similarly, the set $B' = \{e^{ik(x)} \mid x \in V\}$ defines an uncountable basis for $\mathcal{S}'$. If one restricts the function f to some compact subset of V then the former basis can be reduced to a countable one.

The integrals:

$$<f, g> = \int_V f(x)\overline{g(x)}\mathcal{V}, \qquad < \hat{f}, \hat{g} > = \frac{1}{(2\pi)^n}\int_{V^*} \hat{f}(k)\overline{\hat{g}(k)}\mathcal{V}_k \qquad \text{(VI.105)}$$

then define inner products on $\mathcal{S}$ and $\mathcal{S}'$ that make them into Hilbert spaces. One of the forms that *Parseval's identity* takes is to say that the map $\mathcal{F}$ is also an isometry; i.e.:

$$<f, g> = < \hat{f}, \hat{g} > . \qquad \text{(VI.106)}$$

The integrands in the right-hand sides of (VI.103) and (VI.104) then represent the orthogonal projections of f ($\hat{f}$, resp.) onto the orthonormal bases that are defined by the sets B and B', while the integrals themselves represent the orthogonal decompositions of f ($\hat{f}$, resp.) with respect to the bases.

In this way of regarding Fourier transforms, the function $\hat{f}$ is considered to be the *spectral density* of the function f. In effect, it gives the probability density function (when suitably normalized) for the spectrum of wave numbers that contribute to the orthogonal decomposition of f.

One can also regard $\mathcal{F}$ as an integral operator whose kernel $K(k, x)$ is the complex function K on $V^* \times V$:

$$K(k, x) = e^{-ik(x)} . \qquad \text{(VI.107)}$$

The kernel of $\mathcal{F}^{-1}$ is then simply $\overline{K(k, x)}$.

One finds that the Fourier transform of $\partial f / \partial x^i$ is $ik_i \hat{f}$. Hence, one can think of the linear operator on $\mathcal{S}'$ that is defined by multiplication by ik_i as the Fourier transform of the operator of partial differentiation by x^i on $\mathcal{S}$.

This can be extended to linear differential operators with constant coefficients. Suppose the operator in question is $D = a^\alpha \partial^\alpha / \partial x^\alpha$ and it has order m. It Fourier transform – or *symbol* – is then the generally inhomogeneous polynomial of degree m in k:

$$\mathcal{F}[D](k) = a^\alpha (ik_\alpha). \qquad \text{(VI.108)}$$

The homogeneous polynomial $\sigma[D, k]$ of degree m in k that represents the highest-degree terms in the symbol is called the *principal symbol.* It agrees with the more abstract definition that we cited above, except for the factor of i before k.

One can find the Green function for the linear differential operator D when it has constant coefficients by means of the Fourier transform. However, one must take into account that the Fourier transform of the distributional equation $DG = -\delta$ is $\mathcal{F}[D](k)\hat{G}(k) = 1$. Hence, one can solve for the Fourier transform of the desired Green function:

$$\hat{G}(k) = \frac{1}{\mathcal{F}[D](k)}. \tag{VI.109}$$

If one then takes the inverse Fourier transform, one obtains the Green function in the form:

$$G(x, y) = G(\mathbf{x} - \mathbf{y}) = \frac{1}{(2\pi)^n}\int_{V^*} \frac{e^{ik(\mathbf{x}-\mathbf{y})}}{F[D](k)}\mathcal{V}_k. \tag{VI.110}$$

One sees that the assumption that the coefficients of D are constant in space, which already implies some sort of affine structure in order to define homogeneity, manifests itself in the fact that the function G must also be translation invariant. This is yet another fundamental limitation on the use of the Fourier transform in the eyes of the extension from vector spaces to manifolds.

Clearly, there will also be analytical problems with the convergence of the integral when the symbol of D has real roots k. Customarily, physics deals with such poles in the integrand of (VI.110) by analytically continuing the integrand to a complex function and replacing the real integrations by complex ones along contours that avoid the poles. However, this generally alters the character of the resulting Green function accordingly.

Since the linear differential operator with constant coefficients D is associated with a polynomial $F[D](k)$ of finite degree, one can envision a generalization of the integral operator G that involves replacing $F[D](k)$ with a more general function $\sigma[P, k]$ on V^*; at the very least, one might extend to non-constant coefficients. One might also assume that $\sigma[P, k]$ is assumed to be analytic in k, which is like generalizing from finite degree polynomials to infinite degree ones, and if it is smooth, then one is generalizing to formal power series expansions in k.

The resulting integral operator P: $\mathcal{S} \to \mathcal{S}$ that is associated with the symbol $\sigma[P, k]$ and kernel (VI.110) is then called a *pseudo-differential operator*. It then takes the general form:

$$Pf(x) = \frac{1}{(2\pi)^n}\int_{V\times V^*} \frac{e^{ik(x-y)}}{\sigma[P,k]} f(y)\mathcal{V}_y\mathcal{V}_k. \tag{VI.111}$$

Actually, the concept of a Fourier integral operator can be generalized in such a way that a pseudo-differential operator is a special case of it (cf., Duistermaat [**21**]).

b. Nonlinear extensions. The first point at which one might wish to generalize the foregoing discussion is to adapt it to the demands of spatial or spacetime manifolds that are not vector spaces. However, one finds that it is not enough to simply replace V with a more general n-dimensional differentiable manifold M.

For one thing, although it is conceivable that one could generalize $\mathcal{S}$ to a Schwartz space of complex functions on M, nonetheless, when M is not a vector space it is absurd to speak of its dual space. However, one does have vector spaces at each point of M in the form of tangent and cotangent spaces. This means that wave covector k must be associated with some specified point $x \in M$, or perhaps a covector field on M, as opposed to its definition in the linear case, which is independent of $x \in V$.

When one considers the expression $k(x) = k_i x^i$ one sees that when x does not belong to a linear space one must consider a more general *phase function* $\theta: M \to \mathbb{R}$. We assume that θ is analytic, or at least locally approximated by a finite-degree Taylor polynomial about each point of M:

$$\theta(x^i) = \theta_0 + d\theta|_0(x^i) + \mathcal{O}^2(x^i). \tag{VI.112}$$

Hence, since the initial phase θ_0 is essentially arbitrary, and can be neglected, we see that our previous phase function $k(x)$ represents the leading term of this series; i.e.:

$$k = d\theta. \tag{VI.113}$$

If we substitute $e^{-i\theta(x)}$ for the exponential in the Fourier transform then we obtain:

$$\mathcal{F}'f[k] = \int_M e^{-ik(x)} \left[e^{-i\mathcal{O}^2(x)} f(x) \right] \mathcal{V} = \mathcal{F}f'[k], \tag{VI.114}$$

where:

$$f'(x) = e^{-i\mathcal{O}^2(x)} f(x). \tag{VI.115}$$

That is, we have deformed the function that is being transformed with the higher-degree terms of the phase function.

Another reasonable assumption concerning the generalization of Fourier transforms from vector spaces to manifolds is that the space $V \times V^*$ will undoubtedly be replaced by the cotangent bundle T^*M, which looks like the latter vector space in any local trivialization.

Considerable progress has been made towards using the geometry of the cotangent bundle, especial its symplectic structure, to facilitate the generalization of *Fourier integral operators*, which take the form:

$$Af(x) = \int_{V \times V^*} e^{i\phi(x,y,k)} a(x,y,k) f(y) \mathcal{V}_y \mathcal{V}_k \tag{VI.116}$$

in the case where M is a vector space V. The function ϕ is a generalized phase function, which is assumed to be homogeneous of degree one in k, while the function a is a generalized amplitude function.

Of particular interest is the way that such expressions relate to asymptotic expansions, such as when a is a sum of terms a_j, $j = 0, 1, \ldots$ that are homogeneous of positive degree $\mu - j$, where μ is then called the *order* of the operator A, and go to zero asymptotically as some parameter, such as wave number or frequency, goes arbitrarily large. Such

asymptotic expansions play a fundamental role in not only the extension of geometrical optics into diffraction phenomena, but quantum wave mechanics, as well.

Of course, we are rapidly going beyond the scope of the present discussion, so we simply refer the curious reader to the literature cited at the beginning of this section.

References

87. R. Palais, "Differential operators on vector bundles," in *Seminar on the Atiyah-Singer Index Theorem*, ed. R. Palais, Princeton University Press, Princeton, 1965.
88. D. C. Spencer, "Overdetermined systems of linear partial differential equations," Bull. Amer. Math. Soc. **75** (1969), 179-239.
89. D. Saunders, *Geometry of Jet Bundles*, Cambridge University Press, Cambridge, 1989.
90. V. I. Arnol'd, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer, Berlin, 1983.
91. R. Courant and D. Hilbert, *Methods of Mathematical Physics*, v. 2, Interscience, NY, 1962.
92. F. John, *Partial Differential Equations*, Springer, Berlin, 1982.
93. E. Cartan, *Les systèmes différentielle extérieurs et leur applications géométriques*, Hermann, Paris, 1971.
94. E. Kähler, *Einführung in die Theorie der Systeme von Differentialgleichungen*, Teubner, Leipzig, 1934.
95. Y. Choquet-Bruhat, *Géométrie differentielle et systèmes différentieles exterieure*, Dunod, Paris, 1964.
96. R. L. Bryant, S.-S. Chern, *et al.*, *Exterior Differential Systems*, Springer, Berlin, 1991.
97. M. Kuranishi, "On E. Cartan's prolongation theorem of exterior differential systems," Am. J. Math. **79** (1957), 1-47.
98. G. F. D. Duff, *Partial Differential Equations*, Univ. of Toronto Press, Toronto, 1950.
99. O. D. Kellogg, *Foundations of Potential Theory*, Ungar, NY, 1929.
100. I. Stakgold, *Boundary-Value Problems of Mathematical Physics*, MacMillan, NY, 1967.
101. G. Rham de, *Differentiable Manifolds*, Springer, Berlin, 1984.
102. L. Hörmander, *Linear Partial Differential Operators*, Springer, Berlin, 1969.
103. F. G. Friedlander, *Introduction to the theory of distributions*, Cambridge University Press, Cambridge, 1982.
104. J. D. Jackson, *Classical Electrodynamics*, 2nd ed., Wiley, New York, 1976.
105. F. Rohrlich, *Classical Charged Particles*, Addison-Wesley, Reading, MA, 1965.
106. W. Thirring, *Classical Field Theory*, Springer, Berlin, 1978.
107. J. J. Duistermaat, *Fourier Integral Operators*, lecture notes, Courant Institute of Mathematical Sciences, New York University, New York, 1973.
108. V. Guillemin and S. Sternberg, *Geometric Asymptotics*, Mathematical Surveys, no. 14, Am. Math. Soc., Providence, 1977.