Chapter VII

The interaction of fields and currents

One can argue that there is always something solipsistic about any discussion of free field solutions to field equations, since the only way that one knows about the very existence of a field in a region of spacetime is by its interaction with other physical phenomena, often in the form of the motion of matter. However, this illustrates one of the subtle distinctions between linear field theories and nonlinear ones, since in a linear field theory as long as both fields that are present are solutions to the same field equations their combined field, which is obtained by summation, is also a solution, so the space of solutions is sufficiently rich to include combinations of fields. By contrast, if the field equations are nonlinear then there is no guarantee that the sum of two fields will still be a solution. However, this is not merely a technical nuisance, but also a statement about the nature of field interactions, since in the context of electromagnetic wave solutions linear superposition gives interference in the region of overlap without lasting changes to the fields, while nonlinear superposition can bring about lasting alterations to the form of the interacting waves, such as photon splitting.

Therefore, the purpose of this chapter will be to briefly explore the nature of the interactions between electromagnetic fields and electric currents, which can take the form of either the sources of electromagnetic fields or currents external to other fields. One sees that there are basically four possible pairings to consider: current $\rightarrow$ field, field $\rightarrow$ current, field $\rightarrow$ field, current $\rightarrow$ current, where the arrow suggests a cause-and-effect relationship.

In the first case, one must consider the way that source charges and currents generate electromagnetic fields. One sees that in the rest frame of the measurer/observer there are three distinct types of fields that couple to the successive terms in the kinematical state of a charge distribution that is moving relative to the measurement devices: Electrostatic fields couple to the relative position, magnetic fields couple to the relative velocity, and radiation fields couple to the relative acceleration.

Conversely, electromagnetic fields affect currents mostly by exerting forces on the currents. The simplest force to consider, beyond the electrostatic force, is the Lorentz force, which couples to the velocity. However, it is in attempting to account for the effects of radiation that one finds that the theory itself becomes questionable. The problem seems to be in the fact that the way that the acceleration of a moving charge in an electromagnetic field couples to that field is by way of its radiation reaction – or radiation damping – which actually produces a third-order system of ordinary differential equations for the motion of a point charge in an electromagnetic field, namely, the Lorentz-Dirac equation. This system admits various unphysical solutions that make one suspicious about whether the equations are more heuristic than definitive.

As mentioned above, the interaction of electromagnetic fields themselves has a very different character depending upon whether one considers linear constitutive laws or nonlinear ones. It is in the phenomena of nonlinear optics and quantum electrodynamics

that one can gain some intuition as to the best way to proceed into the nonlinear realm of electromagnetism.

Finally, one can consider that two currents can interact in various ways. The simplest way, beyond the electrostatic interaction of their charge distributions, is due to the fact that one current produces a magnetic field that exerts a force on the other current, and vice versa. The result is a mutual magnetostatic force of attraction and repulsion between currents that basically has the opposite sign to the electrostatic one; i.e., like currents attract and unlike ones repel. This sort of static "action-at-a-distance" is, of course, a non-relativistic oversimplification of a more involved interaction picture in which the currents can only propagate electromagnetic waves that effect the interaction in a more causal way.

1. Construction of fields from sources. The most elementary, if not the most fundamental, interaction between fields and sources is the one that amounts to the statement that a field source must generate a field, in some sense. Here again, one can distinguish the problem of the generation of static fields from the generation of dynamic fields since it will affect the type of boundary/initial-value problems that one can pose.

Generally, the association of a field ϕ with its source ρ follows from the solution of the inhomogeneous system of partial differential equations $D\phi = \rho$, where D is a differential operator. One is basically looking for a left inverse operator D^{-1} for D that makes $\phi = D^{-1}\rho$. Since a left inverse is not generally unique, one must then narrow down the set of possibilities by imposing boundary/initial-value conditions. Furthermore, the source ρ must satisfy the integrability condition that it lie in the image of D.

Hence, one can immediately distinguish the problem of generating fields in linear media from generating ones in nonlinear media.

In the case of linear partial differential equations, such as $\delta\mathfrak{h} = \mathbf{J}$, the association of a field with a source by the method of Green functions is preferred. When the manifold in which the fields live is an affine space, or possibly a more general homogeneous space, such as a sphere, the method of the Fourier transform is also very powerful.

a. Static fields. The linear differential operators that are most fundamental in electrostatics and magnetostatics are basically the exterior derivative operator d and the divergence operator δ. One can absorb the constitutive law into the divergence operator so that one has two first-order differential equations in a single field – viz., E or H. If one assumes the existence of a potential – whether local or global – then one can further reduce the system to a single second-order differential equation in the potential.

In the electrostatic case, the basic problem to be solved is the inhomogeneous partial differential equation $\delta\mathbf{D} = \rho$, for $\mathbf{D} = \varepsilon(E)$, along with the constraint on E that $dE = 0$. One can define a pair of equations for E:

$$dE = 0, \qquad \delta_\varepsilon E = \delta \cdot \varepsilon(E) = \rho. \tag{VII.1}$$

Hence, the solution E to this pair of equations will involve an operator G_ε: $\Lambda_0 \to \Lambda^1$ such that:

$$E = G_\varepsilon \rho = (\varepsilon^{-1} \cdot G_\delta)\rho, \tag{VII.2}$$

where G_δ is a right-inverse to d.

When ε is linear, so is δ_ε, as well as G_ε, and as an integral operator it will have a Green function $G_\varepsilon(x, y)$ for its kernel that satisfies the distributional equations:

$$dG_\varepsilon(x, y) = 0, \qquad \delta_\varepsilon G_\varepsilon(x, y) = -\delta(x, y). \tag{VII.3}$$

However, actually solving this equation for $G_\varepsilon(x, y)$ is simply a special case of solving the inhomogeneous problem by assuming that ρ represents essentially a point source of unit charge.

If the constitutive law ε is not only linear, but homogeneous, which naturally assumes that the spatial manifold Σ is an affine space, then we can use the Fourier transform to solve for $G_\varepsilon(x, y) = G_\varepsilon(\mathbf{x} - \mathbf{y})$. The Fourier-transformed equations that one obtains from (VII.3) are:

$$k \wedge \hat{G}_\varepsilon(k) = 0, \qquad i_k[\varepsilon(\hat{G}_\varepsilon(k))] = -i\,. \tag{VII.4}$$

The first one can be solved, up to multiplication by a scalar function $\alpha(k)$, in the form:

$$\hat{G}_\varepsilon(k) = \alpha(k)k. \tag{VII.5}$$

Substituting this in the second one gives:

$$\alpha(k)\,\varepsilon(k, k) = -i\,, \tag{VII.6}$$

which allows us to solve for $\alpha(k)$; in this expression $\varepsilon(k, k) = \varepsilon^{ij} k_i k_j$ represents the quadratic form that e defines on k. We then find our fundamental solution in k-space to be:

$$\hat{G}_\varepsilon(k) = \frac{-ik}{\varepsilon(k, k)}. \tag{VII.7}$$

The inverse Fourier transform of this is:

$$G_\varepsilon(x) = \frac{-i}{(2\pi)^3}\int_{\mathbb{R}^{3*}} \frac{k}{\varepsilon(k, k)} e^{ik(x)}\, \mathcal{V}_k\,. \tag{VII.8}$$

When ε is the Euclidian scalar product, we know that taking the inverse Fourier transform will give the usual Coulomb law expression for the Green function. However, since ε is symmetric, non-degenerate, and positive-definite, one can just as well use ε in place of the Euclidian scalar product as long as one regards the expression $\varepsilon^{ij} k_i k_j$ as simply the form that it takes in a non-orthonormal frame.

As far as physics is concerned, there is another problem with using ε as a metric: it still has the units of electric permittivity. Hence, we normalize it to a dimensionless metric:

$$g_\varepsilon = (\det \varepsilon)^{-1/3}\, \varepsilon, \tag{VII.9}$$

so we can set:

$$\varepsilon(k,k) = (\det \varepsilon)^{1/3}\, g_\varepsilon(k,k), \qquad k(\mathbf{x}) = g_\varepsilon(k, x), \qquad x \equiv \tilde{g}_\varepsilon(\mathbf{x}). \tag{VII.10}$$

This puts the *k*-space Green function (VII.7) into the form:

$$\hat{G}_\varepsilon(k) = \frac{1}{(\det \varepsilon)^{1/3}} \frac{-ik}{g_\varepsilon(k,k)}, \tag{VII.11}$$

and the integral (VII.8) into the form:

$$G_\varepsilon(x) = \frac{-i}{(2\pi)^3 (\det \varepsilon)^{1/3}} \int_{\mathbb{R}^{3*}} \frac{k}{g_\varepsilon(k,k)} e^{ig_\varepsilon(k,x)}\, \mathcal{V}_k . \tag{VII.12}$$

The integrand now takes the standard Euclidian form that makes the spatial Green function:

$$G_{\varepsilon i}(\mathbf{x}) = \frac{1}{4\pi(\det \varepsilon)^{1/3}} \frac{\tilde{g}_\varepsilon(\mathbf{x})}{[\tilde{g}_\varepsilon(\mathbf{x},\mathbf{x})]^{3/2}} = \frac{1}{4\pi} \frac{\tilde{\varepsilon}(\mathbf{x})}{[\tilde{g}_\varepsilon(\mathbf{x},\mathbf{x})]^{3/2}}; \tag{VII.13}$$

the tilde on the g_ε signifies that we are dealing the inverse metric on $T(\Sigma)$ to the one that g_ε defines on $T^*\Sigma$.

One immediately verifies that when the medium is isotropic, as well as linear and homogeneous, so one has:

$$\varepsilon^{ij} = \varepsilon\, \delta^{ij}, \qquad \varepsilon_{ij} = 1/\varepsilon\, \delta_{ij}, \qquad \det \varepsilon = \varepsilon^{3/2}, \tag{VII.14}$$

the form that the Green function takes is:

$$G_{\varepsilon i}(x) = \frac{1}{4\pi\varepsilon} \frac{x_i}{\| x \|^{3/2}}, \tag{VII.15}$$

in which one uses the Euclidian components δ^{ij} for the scalar product.

Given $G_\varepsilon(x)$, one obtains E by convolving G_ε with ρ:

$$\begin{aligned} E(\mathbf{x}) &= \int_{\operatorname{supp}\rho} G_\varepsilon(\mathbf{x}-\mathbf{y})\rho(\mathbf{y})\, \mathcal{V}_y \\ &= \varepsilon^{-1} \left[\frac{1}{4\pi} \int_{\operatorname{supp}\rho} \frac{\rho(\mathbf{y})}{[\tilde{g}_\varepsilon(\mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y})]^{3/2}} (\mathbf{x}-\mathbf{y})\, \mathcal{V}_y \right]. \end{aligned} \tag{VII.16}$$

Note that the expression inside the brackets is nothing but the vector field $\mathbf{D}(\mathbf{x})$.

In the case of a linear, isotropic, homogeneous electrostatic medium with a point charge Q at the origin for a source one then arrives at *Coulomb's law*:

$$\mathbf{D}(\mathbf{x}) = \frac{Q}{4\pi r^2}\hat{\mathbf{r}}, \qquad E(\mathbf{x}) = \frac{1}{4\pi\varepsilon}\frac{Q}{r^2}\hat{r}. \tag{VII.17}$$

with:

$$r^2 = x^2 + y^2 + z^2, \qquad \hat{\mathbf{r}} = \frac{x^i}{r}\partial_i. \tag{VII.18}$$

The basic potential equation for electrostatics is $\Delta_\varepsilon\phi = \rho$, with $\Delta_\varepsilon = \delta \cdot \varepsilon \cdot d$: $\Lambda^0 \to \Lambda_0$. When ε is linear, its symbol is the function:

$$\sigma[\Delta_\varepsilon; k] = i_k \cdot \varepsilon \cdot e_k = \varepsilon(k, k). \tag{VII.19}$$

Since ε is assumed to be positive definite, this function is always non-zero when k is non-zero.

When the medium is electrically homogeneous, the Fourier transform for the Green function for Δ_ε is then:

$$\hat{G}_{\Delta\varepsilon}(k) = \frac{-i}{\varepsilon(k,k)} = \frac{-i}{(\det\varepsilon)^{1/3}}\frac{1}{g_\varepsilon(k,k)}. \tag{VII.20}$$

Taking the inverse Fourier transform gives:

$$G_{\Delta\varepsilon}(\mathbf{x}) = \frac{1}{4\pi(\det\varepsilon)^{1/3}[\tilde{g}_\varepsilon(\mathbf{x},\mathbf{x})]^{1/2}}, \tag{VII.21}$$

and in a linear, homogeneous medium that is isotropic, as well, this takes the Coulomb form:

$$G_{\Delta\varepsilon}(\mathbf{x}) = \frac{1}{4\pi\varepsilon\,\|\mathbf{x}\|}. \tag{VII.22}$$

The solution to the inhomogeneous potential problem for a source charge density ρ in an electrostatically linear and homogeneous medium is then obtained by convolution:

$$\phi(\mathbf{x}) = \frac{1}{4\pi(\det\varepsilon)^{1/3}}\int_{\mathrm{supp}(\rho)}\frac{\rho(\mathbf{y})}{[\tilde{g}_\varepsilon(\mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y})]^{1/2}}\mathcal{V}_y. \tag{VII.23}$$

In magnetostatics, the corresponding equations for the magnetic field B are:

$$dB = 0, \qquad \delta\mathbf{H} = \mathbf{I}, \qquad \mathbf{H} = \tilde{\mu}(B), \tag{VII.24}$$

which can be consolidated into:

$$dB = 0, \qquad \delta\cdot\tilde{\mu}(B) = \mathbf{I}. \tag{VII.25}$$

The integrability condition for the source current **I** is then:

$$\delta \mathbf{I} = 0, \tag{VII.26}$$

which also represents conservation of charge.

Now, let us examine the Fourier-transforms of equations (VII.25) in the linear homogeneous case:

$$k \wedge \hat{B}(k) = 0, \qquad i_k \cdot \tilde{\mu}(\hat{B}) = -i\hat{\mathbf{I}}(k). \tag{VII.27}$$

One can solve the first one by means of:

$$\hat{B}(k) = k \wedge \hat{A}(k), \tag{VII.28}$$

in which $\hat{A}(k) \in \Lambda^1(\mathbb{R}^{3*})$ is some 1-form on $\mathbb{R}^{3*}$, which we shall see in a bit represents the Fourier transform of a covector potential for B. Hence, $\hat{B}(k)$ can only lie in a 2-dimensional subspace of $\Lambda^2_k(\mathbb{R}^{3*})$ at each $k \in \mathbb{R}^{3*}$, namely, the image of $\Lambda^1_k(\mathbb{R}^{3*})$ under e_k. This subspace gets mapped isomorphically to a two-dimensional subspace of $\Lambda_{2,k}(\mathbb{R}^{3*})$ under $\tilde{\mu}$.

When we attempt to solve the second equation in (VII.27) for $\hat{B}(k)$ we are immediately obstructed by the fact that the map i_k is not invertible. However, it does behave somewhat like a projection of the fibers of $\Lambda_2(\mathbb{R}^{3*})$ onto two-dimensional subspaces of the fibers of $\Lambda_1(\mathbb{R}^{3*})$ at each $k \in \mathbb{R}^{3*}$. Hence, if we restrict $\hat{\mathbf{I}}(k)$ to these subspaces we can define a left-inverse to i_k in the form of $e_{\mathbf{k}}$, where $\mathbf{k}$ is a vector field such that $k(\mathbf{k}) = 1$. This defines two problems: how to characterize the subspaces in $\Lambda_1(\mathbb{R}^{3*})$ and how to define the vector field $\mathbf{k}$.

The solution to the first problem follows immediately from the fact that $i_k \cdot i_k = 0$, namely, $\hat{\mathbf{I}}(k)$ must lie in the kernel of i_k. However, this follows naturally from the conservation of charge constraint on $\mathbf{I}$, namely, (VII.26), which leads to the Fourier-transformed constraint on $\hat{\mathbf{I}}(k)$ that it must lie in the hyperplane defined by:

$$0 = i_k \hat{\mathbf{I}} = k_i \hat{I}^i. \tag{VII.29}$$

In order to solve the second problem, we must confront another subtle distinction between the magnetostatic case and the previous electrostatic one that is also related to the fact that we are dealing with bivector fields and 2-forms. It is the fact that the magnetic constitutive law μ: $\Lambda_2 \rightarrow \Lambda^2$ not only "goes backwards," but it also most naturally defines a scalar product on either Λ_2 or Λ^2:

$$\mu(\mathbf{A}, \mathbf{B}) = \mu(\mathbf{A})(\mathbf{B}), \qquad \tilde{\mu}(A, B) \equiv \tilde{\mu}(A)(B). \tag{VII.30}$$

What we ultimately need, though, is a scalar product on Λ_1 or Λ^1. The solution to this problem is simply to use Poincaré duality to define the isomorphism:

$$\# \cdot \mu \cdot \#: \Lambda_1 \to \Lambda^1,$$

which then defines scalar products on the tangent and cotangent bundles, as desired. At the risk of confusion, we denote them by the same letters that we used for the scalar products that μ defines directly.

Furthermore, we need to normalize the resulting metric to be dimensionless:

$$g_\mu = (\det \mu)^{-1/3} \mu . \tag{VII.31}$$

In order to make $k(\mathbf{k}) = 1$, we then divide $\mathbf{k}$ by $g_\mu(k, k)$.

We find that the vector field $\mathbf{k}$ can be obtained from the covector field k by mapping k to a vector field using the normalized form of μ:

$$\mathbf{k} = g_\mu(k), \tag{VII.32}$$

which then makes:

$$k(\mathbf{k}) = g_\mu(k, k). \tag{VII.33}$$

We can then solve the second equation in (VII.27) for $\hat{B}(k)$:

$$\hat{B}(k) = -i\,\mu \cdot \frac{e_{\mathbf{k}}}{g_\mu(k,k)} \cdot \hat{\mathbf{I}}(k) = -i\mu\left[\frac{1}{g_\mu(k,k)}\mathbf{k} \wedge \hat{\mathbf{I}}(k)\right]. \tag{VII.34}$$

We then obtain the k-space Green function:

$$\hat{G}_\mu(k) = \frac{-i}{g_\mu(k,k)}\mu(\mathbf{k}), \tag{VII.35}$$

which is analogous to (VII.11).

Its inverse Fourier transform is:

$$G_\mu(\mathbf{x}) = \frac{1}{4\pi}\frac{1}{[\tilde{g}_\mu(\mathbf{x},\mathbf{x})]^{3/2}}\mu(\mathbf{x}), \tag{VII.36}$$

and the form that it takes in a magnetically isotropic medium is:

$$G_{\mu,i}(\mathbf{x}) = \frac{\mu}{4\pi}\frac{x_i}{\|\mathbf{x}\|^{3/2}}. \tag{VII.37}$$

If we express the integral operator whose kernel is $G_\mu(\mathbf{r})$ in the form:

:

$$B(\mathbf{x}) = \int_{\mathrm{supp}\,\mathbf{I}} G_\mu(\mathbf{x}-\mathbf{y}) \wedge \mathbf{I}(\mathbf{y})\, \mathcal{V}_y \tag{VII.38}$$

then the solution for B that corresponds to (VII.16) is:

$$B(\mathbf{x}) = \mu(\mathbf{H}) = \mu\left[\frac{1}{4\pi} \int_{\mathrm{supp}\,\mathbf{I}} \frac{1}{[\tilde{g}_\mu(\mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y})]^{3/2}} (\mathbf{x}-\mathbf{y}) \wedge \mathbf{I}(\mathbf{y})\, \mathcal{V}_y \right]. \tag{VII.39}$$

In the case of a magnetically isotropic medium, this solution takes the form of the *Biot-Savart law*:

$$B(\mathbf{x}) = \frac{\mu}{4\pi} \int_{\mathrm{supp}\,\mathbf{I}} \frac{1}{\| \mathbf{x}-\mathbf{y} \|^3} (x-y) \wedge I(\mathbf{y})\, \mathcal{V}_y\,, \tag{VII.40}$$

in which x, y, and I denote the covector fields that correspond to the vector fields **x**, **y**, and **I** under the isomorphism that is defined by the Euclidian scalar product.

As for the magnetostatic potential problem $\Delta_\mu A = \mathbf{I}$, the analogy with electrostatics is also imprecise, since the magnetic Laplacian $\Delta_\mu = \delta \cdot \tilde{\mu} \cdot d$: $\Lambda^1 \to \Lambda_1$ has a symbol:

$$\sigma[\Delta_\mu, k] = i_k \cdot \tilde{\mu} \cdot e_k\,, \tag{VII.41}$$

that is not invertible. Not only does i_k: $\Lambda_2(\mathbb{R}^{3*}) \to \Lambda_1(\mathbb{R}^{3*})$ define a projection onto a two-dimensional subspace of $\Lambda_1(\mathbb{R}^{3*})$, but the kernel of the map e_k: $\Lambda^1(\mathbb{R}^{3*}) \to \Lambda^2(\mathbb{R}^{3*})$ is one-dimensional, namely, the line $[k]$ in $\Lambda^1(\mathbb{R}^{3*})$ that is generated by k.

This is where a choice of gauge for the potential 1-form A is useful. If one imposes a generalized Coulomb condition on A*:*

$$\delta \cdot \mu(A) = 0 \tag{VII.42}$$

then the corresponding constraint on the Fourier transform $\hat{A}(k)$ of A is:

$$0 = i_k \mu(\hat{A}) = \mu(k, \hat{A}). \tag{VII.43}$$

That is, $\hat{A}$ is confined to the orthogonal complement to $[k]$. When $\sigma[\Delta_\mu, k]$ is restricted to this subspace of the domain space $\Lambda^1(\mathbb{R}^{3*})$ and its image in the range space $\Lambda_1(\mathbb{R}^{3*})$ one finds that it is invertible. Of course, this also means that the Fourier transform $\hat{I}$ of the current can only lie in the image, but that is a natural consequence of the conservation of charge, as before.

One finds that the inverse of $\sigma[\Delta_\mu, k]$ then takes the form:

$$\sigma[\Delta_\mu, k]^{-1} = -\hat{G}_\mu(k) = i_{\mathbf{k}} \cdot \mu \cdot e_{\mathbf{k}} = \frac{1}{g_\mu(k,k)}\mu\,; \tag{VII.44}$$

hence:

$$\hat{G}_\mu(k) = -\frac{1}{g_\mu(k,k)}\mu\,. \tag{VII.45}$$

In a magnetically homogeneous medium this takes the component form:

$$\hat{G}_{\Delta\varepsilon}(k) = \frac{\mu}{k^2}\delta_{ij}\,. \tag{VII.46}$$

The inverse Fourier transform of (VII.45) then becomes:

$$G_{\Delta\varepsilon}(\mathbf{x}) = \frac{1}{4\pi}\frac{1}{[\tilde{g}_\mu(\mathbf{x},\mathbf{x})]^{1/2}}\mu\,. \tag{VII.47}$$

In a magnetically isotropic medium, we then have:

$$G_{\Delta\varepsilon}(\mathbf{x}) = \frac{\mu}{4\pi\,\|\,\mathbf{x}\,\|}\delta_{ij}\,. \tag{VII.48}$$

The covector potential $A(\mathbf{x})$ that is produced by a source current density **I** is then:

$$A(\mathbf{x}) = \mu\left[\frac{1}{4\pi}\int_{\mathrm{supp}(\rho)}\frac{\mathbf{I}(\mathbf{y})}{[\tilde{g}_\mu(\mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y})]^{1/2}}\mathcal{V}_y\right]. \tag{VII.49}$$

One should compare the expressions that we have derived here with the corresponding ones in – say – Jackson [**1**] to see how much vector calculus takes for granted as a consequence of dealing with a three-dimensional Euclidian space. For instance, it regards vectors, covectors, bivectors, and 2-forms as essentially the same, since one has isomorphisms of all four spaces. As we saw, just dropping the assumption that the Euclidian structure is purely geometric in origin makes it necessary to reevaluate all of these isomorphisms.

b. Dynamic fields. In the case of dynamic fields, we must use the spacetime manifold M in place of the spatial manifold Σ and a source current density **J** on M in place of ρ or **I**.

However, the inhomogeneous problem that one poses for a given source current **J** is still essentially the same: one must solve the partial differential equations:

$$dF = 0, \qquad \delta_\kappa F = \delta \cdot \kappa(F) = \mathbf{J} \tag{VII.50}$$

for F, with the integrability condition for **J** that:

$$\delta \mathbf{J} = 0, \tag{VII.51}$$

which still represents conservation of charge.

On the surface of things, this problem is no different from the one that we defined above for magnetostatics. However, we shall see that the previous logic breaks down at a crucial point.

The Fourier-transformed equations that one obtains when κ is linear and homogeneous are:

$$k \wedge \hat{F} = 0, \qquad i_k \cdot \kappa(\hat{F}) = -i\hat{\mathbf{J}}. \tag{VII.52}$$

The solution to the first equation still takes the form:

$$\hat{F} = k \wedge \hat{A} \tag{VII.53}$$

for a suitable 1-form $\hat{A}$. Hence, $\hat{F}$ must lie in a three-dimensional subspace of the fiber of $\Lambda^2(\mathbb{R}^{4*})$ at each point, and it gets mapped to a three-dimensional subspace in $\Lambda_2(\mathbb{R}^{4*})$.

Once again, i_k not invertible unless we restrict ourselves to a hyperspace in $\Lambda_1(\mathbb{R}^{4*})$ that that takes the form of the kernel of i_k: $\Lambda_1(\mathbb{R}^{4*}) \rightarrow \Lambda_0(\mathbb{R}^{4*})$. This again comes from the integrability condition on **J** that charge be conserved:

$$0 = i_k \hat{\mathbf{J}}(k). \tag{VII.54}$$

One can then invert i_k when restricted to this subspace by means of $e_{\mathbf{k}}$ when **k** is a vector field with $k(\mathbf{k}) = 1$.

However, this is where things break down compared to the three-dimensional magnetostatic case. Basically, Poincaré duality no longer takes vector fields to 2-forms, but to 3-forms; similarly, it now takes bivector fields to 2-forms, not 1-forms. Hence, we can no longer use the # isomorphism to associate a scalar product on $\Lambda^1(\mathbb{R}^{4*})$ with the scalar product on $\Lambda^2(\mathbb{R}^{4*})$ that is defined by κ; and we will have to find some other way of obtaining **k**.

A further complication arises when one considers the inhomogeneous potential problem for a dynamic field, which is defined by substituting $F = dA$ in the field equation for F to give $\Box_\kappa A = \mathbf{J}$, where $\Box_\kappa = \delta \cdot \kappa \cdot d$: $\Lambda^1 \rightarrow \Lambda_1$ is the field operator for κ:

In the linear case, its symbol is:

$$\sigma[\Box_\kappa, k] = i_k \cdot \kappa \cdot e_k = Q(k, k). \tag{VII.55}$$

One finds that, in addition to the aforementioned restrictions, one must contend with the fact that if $Q(k, k)$ is to behave like a generalization of the Lorentzian dispersion law then it will vanish not just at the origin, but on an algebraic hypersurface in $\Lambda^1(\mathbb{R}^{4*})$.

Since this is the starting point for any discussion of electromagnetic wave motion and the ultimate appearance of a Lorentzian structure as a consequence of the pre-metric electromagnetic structure of spacetime – i.e., its constitutive law and field equations – we shall simply return to the topic in the next chapter of this book.

c. Electromagnetic radiation. We pause to observe that each of the successive terms in the kinematical state of a relatively moving electric charge distribution $\rho(t, x^i)$ seems to be associated with its own special field. The relative position of the points of the distribution generates the electrostatic field and the relative velocity generates a magnetic field; if the relative velocity is "constant in time" then this magnetic field is static, as well.

Since historically the rigorous geometrical treatment of the concept of acceleration for motion in manifolds that have no natural affine structure opened up the subtleties of general relativity, and the consequent representation of gravity as a sort of "fictitious force," it is not surprising that one of the most debatable extensions of Maxwellian electromagnetism regards the manner by which the relative acceleration of an electric charge distribution is associated with an electromagnetic field.

It is generally agreed that the electromagnetic field that is generated by a relative acceleration is a *radiation field*, which implies, among other things, that it will be a wavelike solution of the field equations. It is not, however, generally agreed that such a radiation field will be observed by a comoving measure-observer; e.g., a static electron observed in a freely-falling laboratory. Basically, the issue is that of the relativity of acceleration, which leads into the study of conformal Lorentzian geometry, since the transformations between frames that differ by a constant relative acceleration is related to a composition of translations and inversions. Although the former class of transformations is contained in Poincaré group, the latter class is not. Inversions are more projective-geometric in character, and relate to the conformal Lorentz group.

Although we shall make a more complete discussion of electromagnetic waves in the next chapter, including the manner by which a Lorentzian structure might emerge from the electromagnetic field equations, for now, we provisionally revert to the metric formulation of the Maxwell equations, namely:

$$dF = 0, \qquad \delta F = J, \tag{VII.56}$$

in which δ is the codifferential operator that is associated with the Hodge dual isomorphism *, which is defined using a Lorentzian metric g on the $T(M)$, and the 1-form J is associated with the source current density **J** by way of g; i.e.:

$$J = i_{\mathbf{J}} g = (g_{\mu\nu} J^{\nu})\, dx^{\mu}. \tag{VII.57}$$

If we take the codifferential of both sides of the first equation in (VII.56) and the exterior derivative of the second one then the result is:

$$\Box F = dJ, \tag{VII.58}$$

in which $\Box = \delta d + d\delta$ is the d'Alembertian operator for g.

Equation (VII.58) represents a forced linear wave equation for the 2-form F with a source term that consists of the 2-form dJ. Its solution will then have a different character from an electromagnetic field that has J for its source, and we refer to a solution F_{rad} to (VII.58) as a *radiation field*.

In order to get a better physical intuition for the nature of the forcing term, we assume that J takes the form ρu, where ρ is an electric charge density function and $u = u_\mu \, dx^\mu$ is its associated covelocity 1-form; both ρ and u are assumed to have the same spatially compact support. One sees that:

$$dJ = d\rho \wedge u + \rho \, du. \tag{VII.59}$$

Hence, there are two essentially distinct sources of radiation fields: the non-constancy of ρ over its support and the "kinematical vorticity" du of its flow. In time + space form, du looks like:

$$du = -dt \wedge a - \omega = -dt \wedge a - \#_s \, \boldsymbol{\omega}. \tag{VII.60}$$

in which the 1-form a and the 2-form ω have the local component forms:

$$a = (u_{i,0} - u_{0,i}) \, dx^i, \qquad \omega = \tfrac{1}{2}(u_{i,j} - u_{j,i}) \, dx^i \wedge dx^j . \tag{VII.61}$$

The 1-form a amounts to the spatial acceleration of the distribution ρ, with a correction for the spatial gradient of the Fitzgerald-Lorentz factor, which reverts to the spatial gradient of the spatial speed $v = \| u_s \|$. The 2-form ω defines the spatial vorticity of the covelocity u, which is essentially the curl of the spatial velocity vector field $\mathbf{v}$.

If one has a Green function for the operator $\Box$, which we now represent as $G(x, y) \in \Lambda^2 \otimes \Lambda^2$, then the volume potential term in (VI.88) can be given the form:

$$F_{\text{rad}}(x) = \int_M G(x, y) \wedge dJ(y) = \int_{\text{supp}(J)} G(x, y) \wedge dJ(y). \tag{VII.62}$$

We can then substitute from (VII.59) and (VII.60) to express F_{rad} as the sum of three contributions:

$$\begin{aligned} F_{\text{rad}}(x) = & \int_{\text{supp}(J)} G(x, y) \wedge (d\rho \wedge u)(y) \\ & - \int_{\text{supp}(J)} G(x, y) \wedge (dt \wedge a)(y) - \int_{\text{supp}(J)} G(x, y) \wedge \omega(y). \end{aligned} \tag{VII.63}$$

It is important to point out that when one gets into the realm of interactions that must be moderated by the exchange of dynamic fields, such as photons, that move with finite speeds one must restrict the Green function accordingly. That is, the support of G must be *causal*, which usually means that it can only be a subset of the past light cone at each point. We shall not elaborate at the moment, as the theory of radiation, especially when one includes the quantum consideration, is a deep problem that requires further analysis.

Of the three contributions to F_{rad} in (VII.63), the ones that seem to get the most attention in practice are the first two.

The first one usually gets applied in the context of antenna theory, where it is not the spatial variation of the charge density ρ that is important, but the time variation. One sees that since G defines a linear operator from 2-forms to 2-forms, it is meaningful and useful to speak of eigenvalues and eigenforms. For instance, sinusoidal variations in the charge will produce sinusoidal variations in the radiation field.

The second contribution says that accelerating (or, of course, decelerating) charges generate radiation fields. For instance, *brehmstrahlung*, or "braking radiation," is an example of this situation. *Čerenkov radiation* is the form of braking radiation that is produced when the speed of a charge is greater than the speed of light in the medium.

It is the fact that accelerating charges, such as charges confined to circular orbits, must radiate energy, and therefore lose energy, that defined one of the first unavoidable flaws in the Bohr planetary model for atomic electrons, since this classical radiation model would suggest that a planetary electron would only lose energy continuously and spiral down into the nucleus itself. There were two problems with this picture in the eyes of experimental physics:

1. A continuous decay of energy would produce a continuous spectrum of frequencies for the radiated field, which contradicts the discrete nature of atomic spectra.

2. There is a non-zero ground state energy for atomic electrons that they reach before they are ever absorbed into the nucleus.

Although it is traditional to say that all discussion of the radiation of atomic electrons must necessarily fall within the purview of quantum electrodynamics, one must keep in mind that there is still a sizable gap between the field-theoretic formalism of Maxwellian electromagnetism and the formalism of the scattering approximation of quantum electrodynamics. Hence, there is still some validity to posing the problem of extending the field-theoretic formalism into the realm that is usually treated in the scattering approximation. Therefore, many of the generalizations that we shall consider in the name of pre-metric electromagnetism are guided by the hope that they will extend the field-theoretic formalism further into the quantum – i.e., atomic-to-subatomic – domain.

2. Lorentz force. In this section, we address the converse of the problem of the previous section: That is, we now consider the effect of an electromagnetic field on a current that is not its source. Of course, there is something unavoidably linear about assuming that an electromagnetic field can be "external" to an electric current, because the current will, of course, generate an electromagnetic field of its own. It is only linear superposition that allows one to treat the combined effect as a simple addition of 2-forms.

a. Lorentz force on one current. Suppose we have an electric current $\mathbf{J} = \rho\mathbf{u} = \rho\partial_t + \rho\mathbf{v}$ in an external electromagnetic field $F = dt \wedge E - \#_s\mathbf{B}$. We have already discussed the force density that the electric field E exerts on the charge density ρ, namely:

$$\rho E = \rho\, i_{\partial t}(dt \wedge E). \tag{VII.64}$$

We now address the force density that $\mathbf{B}$ exerts on $\rho\mathbf{v}$. In the conventional formulation of electromagnetism in terms of vector analysis, it takes the form of the *Lorentz force (density):*

$$\mathbf{f} = \rho\mathbf{v} \times \mathbf{B} = \mathbf{I} \times \mathbf{B}. \tag{VII.65}$$

in which we have introduced $\mathbf{I} = \rho\mathbf{v}$ as the spatial current density.

If we wish to put this into the language of exterior algebra then we need only represent f by a spatial 1-form, and we can rewrite (VII.65) in the form:

$$f = \rho\, \#_s(\mathbf{v} \wedge \mathbf{B}) = \#_s(\mathbf{I} \wedge \mathbf{B}). \tag{VII.66}$$

But:

$$\#_s(\mathbf{v} \wedge \mathbf{B}) = i_{\mathbf{v} \wedge \mathbf{B}}\mathcal{V} = -\, i_{\mathbf{v}}(\#_s\mathbf{B}) = -\, i_{\mathbf{u}}(\#_s\mathbf{B}). \tag{VII.67}$$

We find that the combined Coulomb force on the charge density and the Lorentz force on the spatial current density, namely:

$$f = \rho[E + \#_s(\mathbf{v} \wedge \mathbf{B})] = \rho[i_{\partial t}(dt \wedge E) - i_{\mathbf{v}}(\#_s\mathbf{B})], \tag{VII.68}$$

can be combined into the four-dimensional expression:

$$f = i_{\mathbf{J}}F = J^{\mu}F_{\mu\nu}\,dx^{\nu}. \tag{VII.69}$$

However, the f of (VII.69) describes only the spatial force density, so there is one extra piece to $i_{\mathbf{J}}F$ that is missing from (VII.68), namely, the temporal piece $-\,\rho\, E(\mathbf{v})\, dt$. Hence, we correct (VII.68) to:

$$f = \rho\,[-\, E(\mathbf{v})\, dt + E + \#_s(\mathbf{v} \wedge \mathbf{B})]. \tag{VII.70}$$

From a consideration of its basic units, the temporal contribution represents a power density.

Now, let us define the energy-momentum density associated with the vector field $\mathbf{v}$ to be:

$$p = \mu u, \tag{VII.71}$$

in which μ is the rest mass density, whose support is the same as for ρ and $\mathbf{v}$, while the 1-form u is a covelocity 1-form that relates to $\mathbf{v}$ by means of:

$$u(\mathbf{v}) = \text{const.} \tag{VII.72}$$

We define the proper-time derivative of something to be its Lie derivative along the flow of $\mathbf{v}$:

$$\frac{d}{d\tau} = \mathrm{L}_{\mathbf{v}} = i_{\mathbf{v}}d + di_{\mathbf{v}}, \tag{VII.73}$$

so:

$$\frac{dp}{d\tau} = \frac{d\mu}{d\tau}u + \mu\frac{du}{d\tau}. \tag{VII.74}$$

The first term vanishes iff mass is conserved along the flow of $\mathbf{v}$.

If we express **v** and u in time + space form as:

$$\mathbf{v} = v^0\partial_t + \mathbf{v}_s\,, \qquad u = u_0\, dt + u_s \tag{VII.75}$$

then:

$$\frac{du}{d\tau} = i_{\mathbf{v}}du + di_{\mathbf{v}}u = i_{\mathbf{v}}du = -\, a_s(\mathbf{v}_s) + [v^0 a_s + \#_s(\mathbf{v}_s \wedge \boldsymbol{\omega}_s)], \tag{VII.76}$$

where the term $di_{\mathbf{v}}u$ vanishes on account of (VII.72). The spatial vector field $\boldsymbol{\omega}_s$ and 1-form a_s that we introduced are defined by (VII.59) and (VII.60).

Hence, we can regard a_s as the acceleration, in the sense of the convected derivative of the covelocity, and $\boldsymbol{\omega}_s$ is the (kinematical) vorticity vector field.

Thus, if we consider the four-dimensional form of Newton's second law of motion – i.e., conservation of energy-momentum:

$$f = \frac{dp}{d\tau}\,, \tag{VII.77}$$

then we arrive at the following pair of equations:

$$\mu a_s(\mathbf{v}_s) = \rho E(\mathbf{v}_s), \tag{VII.78a}$$
$$\mu[v^0 a_s + \#_s(\mathbf{v}_s \wedge \boldsymbol{\omega}_s)] = \rho[v^0 E + \#_s(\mathbf{v}_s \wedge \mathbf{B})]. \tag{VII.78b}$$

The first one is a statement about power, i.e., the rate at which energy is being added or subtracted from the motion of the charge cloud. The second one is a generalization of the usual equation of motion defined by the combined Coulomb and Lorentz force that one encounters for pointlike charges, for which the vorticity $\boldsymbol{\omega}_s$ vanishes. It also vanishes for irrotational flows, such as ones for which $\mathbf{v}_s$ is spatially uniform.

When $E = 0$ one finds that the power equation says simply that:

$$a_s(\mathbf{v}_s) = 0. \tag{VII.79}$$

In the metric case, for which:

$$a_s(\mathbf{v}_s) = g(\mathbf{a}_s, \mathbf{v}_s) = \tfrac{1}{2}\,\frac{dv_s^2}{d\tau}\,, \tag{VII.80}$$

this says that the motion of a charge cloud in a magnetic field is uniform circular.

b. Radiation reaction. When an electromagnetic field exerts a force on a current – i.e., a moving charge distribution – that charge distribution will accelerate or decelerate. Consequently, since accelerating charges emit radiation fields, which carry energy and momentum, the charge distribution will also decelerate due to the fact that it is losing energy and momentum. This loss of energy-momentum due to radiation, or rather, its proper time derivative, is referred to as the *radiation reaction* or *radiation damping*.

A full accounting of the equations of motion for a charge distribution in an external electromagnetic field must then include not only the Lorentz force but the radiation reaction, as well. However, since a full accounting of the radiation reaction involves a

more detailed discussion of the problem of electromagnetic radiation, we shall defer that study to future research.

3. Interaction of fields. The main issue that concerns the interaction of more than one field in a region of space is linearity versus nonlinearity. In effect, the very assumption that there is interaction to begin with has a distinctly nonlinear sort of character, but there are still enough linear phenomena in nature to make a brief discussion of linear field interactions worthwhile. Furthermore, what starts off as a linear combination in the field space might very well project to a nonlinear interaction in the measurement space.

Generally, there are three basic regimes in the name of linearity, which are parameterized by some appropriate magnitude: the linear regime, which is characterized by weak magnitudes, strong nonlinearity, which is usually characterized by the onset of some phase transition at a critical magnitude, and the regime of weak nonlinearity, which is an intermediate sort of realm between linearity and the onset of a phase transition.

For instance, in the response of elastic materials to applied stresses, one starts out with Hooke's law for small strains, which is the linear regime. Between linear response and the onset of plastic deformation, one is in the regime of nonlinear elasticity, which is what we are calling weak nonlinearity. Once plastic deformation begins, a phase transition has taken place, and the concept of elastic deformation is no longer applicable.

Since the actual differential equations that we consider in pre-metric electromagnetism, namely, $dF = 0$, $\delta\mathfrak{H} = \mathbf{J}$, are both linear, the only possible source of nonlinearity is in the constitutive law $\mathfrak{H} = \kappa(F)$ that couples the excitation $\mathfrak{H}$ of a medium to the presence of an electromagnetic field F. We then see that for small field strengths the response of the medium is approximately linear, whereas for some high enough field strength, phase transitions can change the very nature of the medium. For instance, one can get melting of solids, ionization of gases, the onset of ferromagnetism, and the formation of particle-anti-particle pairs.

a. Linear constitutive laws. When one is concerned with weak enough field strengths to remain well within the linear response regime for a medium, one is dealing with a vast variety of possible phenomena. In particular, electronics, optics, and communications are mostly concerned with linear phenomena. Indeed, most practical applications are greatly complicated by the onset of nonlinearity. For instance, in the design of capacitors, the onset of a phase transition is clearly undesirable, since the phase transition that one has to contend with is the breakdown of the dielectric at high field strengths.

The equivalent statement to saying that one is in the linear response regime for a medium is the principle of superposition, which says, in effect, that if the constitutive map $\kappa\colon \Lambda^2 \to \Lambda_2$ is linear on each fiber – hence, linear on sections – then the combined operator $\delta\cdot\kappa\colon \Lambda^2 \to \Lambda_1$, $F \mapsto \delta\kappa(F)$ is linear, as well. Its kernel $\ker(\delta\kappa) = \{F \in \Lambda^2 \mid \delta\kappa(F) = 0\}$ is therefore a linear subspace of Λ^2, and since an element of $\ker(\delta\kappa)$ is a solution to the first-order partial differential equation $\delta\kappa(F) = 0$ this implies that linear combinations of solution fields will be solution fields.

One of the most far-reaching consequences of dealing with the linear regime of phenomena of any sort is the applicability of the methods of Fourier analysis when one is also dealing with a homogeneous medium. Indeed, if one has a nonlinear operator from a Hilbert space of "input" fields to another Hilbert space of "output" fields, there is nothing to stop one from performing Fourier transforms of the inputs and outputs. However, it is only in the case of a homogeneous linear operator that the relationship between the Fourier transformed fields is linear, as well. One can see how this is immensely useful in the problem of modulating and demodulating information into carrier signals. It is also the basis for most of the constructions of quantum field theory, in which one is chiefly concerned with constructing the Fourier transform of the integral kernel for a unitary map from a Hilbert space of "incoming" scattering states to a Hilbert space of "outgoing" ones that one thinks of as the scattering operator or *S* matrix for the interaction in question.

There is a subtle relationship between linear supposition and the optical phenomenon of interference, which leads to diffraction, which defines the most definitive difference between classical mechanics and wave mechanics. Basically, one is dealing with a linear superposition of *complex* wave functions whose effect in the eyes of the measurer/observer is what we might call "pseudo-nonlinear." This takes the form of a nonlinear projection of a linear combination. In the case of wave interference, the linear combination is formed in the complex vector space $\mathbb{C}$, while the nonlinear projection maps $\mathbb{C} - \{0\}$ onto the non-negative real axis $\mathbb{R}^+$ by way of the transformation from Cartesian to polar coordinates $\mathbb{C} \to \mathbb{R}^+$, $x + iy \mapsto \sqrt{x^2 + y^2} = \| z \|$.

Although this map is homogeneous of degree one with respect to real scalar multiplication, since $\lambda(x + iy)$ goes to $\lambda\sqrt{x^2 + y^2}$, nevertheless, the sum of two complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ goes to $\sqrt{(x_1 + x_2)^2 + (y_1 + y_2)^2} = \| z_1 + z_2 \|$, which differs from $\sqrt{x_1^2 + y_1^2} + \sqrt{x_2^2 + y_2^2} = \| z_1 \| + \| z_2 \|$. Indeed, since we are dealing with a norm on $\mathbb{C}$ the sum satisfies the triangle inequality $\| z_1 + z_2 \| \leq \| z_1 \| + \| z_2 \|$. Hence, the projection is nonlinear whenever strict integrability is obtained.

Another example of nonlinear superposition is given by the addition of velocities in special relativity, which originates in the fact that the projection of four-dimensional spacetime events into the rest space of a measure/observer is not a simple Cartesian projection of a four-tuple (v^0, v^1, v^2, v^3) of components onto a triple (v^1, v^2, v^3) of components, but more like the projection of homogeneous coordinates (v^0, v^1, v^2, v^3) for $\mathbb{R}\mathrm{P}^3$ onto inhomogeneous coordinates (V^1, V^2, V^3), which is a nonlinear projection, since $V^i = v^i/v^0$. Since we shall have much more to say about the role of projective geometry in special relativity and electromagnetism later in Chapter XI, we shall suspend our discussion of the subject for now.

b. Nonlinear constitutive laws. When κ is a nonlinear map on the fibers, $\ker(\delta \cdot \kappa)$ is not generally a linear subspace, but a more elaborate nonlinear subspace of the vector space Λ^2 that will generally be infinite-dimensional, as well. Hence, if F_1 and F_2 are both solutions of $\delta\kappa(F) = 0$ then their sum is not another solution, in general. One concludes

that the effect of superposing the two fields in a nonlinear medium will be to produce a field $\Sigma(F_1, F_2)$ that is generally distinct from $F_1 + F_2$, and satisfies $\delta\kappa(\Sigma(F_1, F_2)) = 0$.

Of course, since we dealing with a vector space we can always define a 2-form $\Delta(F_1, F_2)$ such that:

$$\Sigma(F_1, F_2) = F_1 + F_2 + \Delta(F_1, F_2). \tag{VII.81}$$

However, unless there is some convenient way of characterizing the interaction term $\Delta(F_1, F_2)$ one must realize that its utility is mostly in the weakly nonlinear regime.

Indeed, if we assume that $\Sigma(F_1, F_2)$ can be approximated by a Taylor series:

$$\Sigma(F_1, F_2) \approx F_1 + F_2 + d\Sigma|_0(F_1 + F_2) + \tfrac{1}{2} d^2\Sigma|_0(F_1 + F_2, F_1 + F_2) + \ldots \tag{VII.82}$$

then we see that:

$$\Delta(F_1, F_2) \approx d\Sigma|_0(F_1 + F_2) + \tfrac{1}{2} d^2\Sigma|_0(F_1 + F_2, F_1 + F_2) + \ldots \tag{VII.83}$$

Since $d\Sigma|_0$ is linear and $d^2\Sigma|_0$ is bilinear, we can expand this into:

$$\begin{aligned}\Delta(F_1, F_2) \approx\ & d\Sigma|_0(F_1) + d\Sigma|_0(F_2) \\ & + \tfrac{1}{2} d^2\Sigma|_0(F_1, F_1) + d^2\Sigma|_0(F_1, F_2) + \tfrac{1}{2} d^2\Sigma|_0(F_2, F_2) + \ldots\end{aligned} \tag{VII.84}$$

Of course, the problem now reverts to that of characterizing the nature of the successive derivatives in this expression in terms of something that pertains to the nonlinearity in κ. One way of doing this is to note that since the operator $\delta \cdot \kappa$ annihilates the left-hand side of (VII.81), one has:

$$\delta\kappa[F_1 + F_2 + \Delta(F_1, F_2)] = 0. \tag{VII.85}$$

If one expands κ in terms of nonlinear susceptibilities then presumably one might arrive at successive levels of perturbation to the law of linear superposition, but that also appears to be a major undertaking in terms of computational complexity. Of course, there might be useful parallels to the more established perturbational methods of quantum field theory that could prove heuristically probative.

4. Interaction of currents. Since all physical fields presumably have sources one can consider the interaction of fields as being, in a sense, dual to the problem of the interaction of the source currents themselves. From that viewpoint, one regards the fields of the sources as being essentially intermediaries for the interaction of the sources themselves.

We see that there are two basic levels of complexity associated with the interaction of sources, which correspond to the non-relativistic and relativistic approximations, respectively. Namely, when one is dealing with slowly-varying fields produced by closely-spaced sources, one can use the non-relativistic approximation of action-at-a-distance. By contrast, when the currents are rapidly varying or the sources are

sufficiently distant from each other, one must respect causality and regard the interaction of sources as only taking place by the intermediary of electromagnetic waves.

a. Force between two currents (linear case). In the case of static – or at least, slowly-varying – fields with neighboring sources $\mathbf{J}_1$ and $\mathbf{J}_2$, one can combine the coupling of – say – $\mathbf{J}_1$ to the field F_1 that it generates with the Lorentz force density $f_{12} = i_{\mathbf{J}_2} F_1$ that the field F_1 exerts on $\mathbf{J}_2$. Since there is nothing, at this point, to distinguish $\mathbf{J}_1$ from $\mathbf{J}_2$, one could just as well consider the force density $f_{21} = i_{\mathbf{J}_1} F_2$ that the field F_2, which is generated by $\mathbf{J}_2$, exerts on $\mathbf{J}_1$. By Newton's third law of motion, one expects that $f_{12} = -f_{21}$.

When the medium in which $\mathbf{J}_1$ and $\mathbf{J}_2$ are defined is linear in its response, one can use the method of Green functions to establish the map from each source to its corresponding field; i.e., one can obtain a linear function $F_i = F_i(\mathbf{J}_i)$, $i = 1, 2$. As a consequence, one can define f_{12} as an anti-symmetric bilinear functional:

$$f_{12}(\mathbf{J}_1, \mathbf{J}_2) = -f_{12}(\mathbf{J}_2, \mathbf{J}_1) = -f_{21}(\mathbf{J}_1, \mathbf{J}_2). \tag{VII.86}$$

For instance, in the electrostatic case, Coulomb's law (VII.17) associates an electrostatic field:

$$E_1(\mathbf{r}) = \frac{1}{4\pi\varepsilon} \frac{Q_1}{\| \mathbf{r} - \mathbf{r}_1 \|^2} d \| \mathbf{r} - \mathbf{r}_1 \| \tag{VII.87}$$

with a point charge Q_1 that is located at $\mathbf{r}_1 \in \mathbb{R}^3$.

The force that Q_1 exerts on a point charge Q_2 that is located at $\mathbf{r}_2$ is:

$$f_{12}(Q_1, Q_2) = Q_2 E_1(\mathbf{r}_2) = \frac{1}{4\pi\varepsilon} \frac{Q_1 Q_2}{\| \mathbf{r}_2 - \mathbf{r}_1 \|^2} d \| \mathbf{r}_2 - \mathbf{r}_1 \| = -f_{12}(Q_2, Q_1). \tag{VII.88}$$

(The sign change originates in the 1-form $d\| \mathbf{r}_2 - \mathbf{r}_1 \|$.)

Actually, the anti-symmetry in $f_{12}(Q_1, Q_2)$ is closely related to the fact that the charge distributions associated with Q_1 and Q_2 are pointlike. If one replaced them with more general charge distributions ρ_1 and ρ_2 then $f_{12}(\rho_1, \rho_2)$ would become a force density that was distributed over the support of ρ_2, which would take the form:

$$f_{12}(\rho_1, \rho_2)(y) = \rho_2(y) \int_{\text{supp}\, \rho_1} G_\varepsilon(x, y) \rho_1(x) \mathcal{V}_x = \int_{\text{supp}\, \rho_1} \rho_1(x) G_\varepsilon(x, y) \rho_2(y) \mathcal{V}_x, \tag{VII.89}$$

while $f_{12}(\rho_2, \rho_1)$ would be a force density that was distributed over the support of ρ_1 that would take the form:

$$f_{12}(\rho_2, \rho_1)(x) = \int_{\text{supp}\, \rho_2} \rho_1(x) G_\varepsilon(x, y) \rho_2(y) \mathcal{V}_y \tag{VII.90}$$

Even with an anti-symmetric kernel $G_\varepsilon(x, y)$, it makes no sense to speak of the symmetry properties of $f_{12}(\rho_1, \rho_2)(x)$ with respect to $f_{12}(\rho_2, \rho_1)(y)$ since they have their supports on two different regions of space. However, if one integrates each of them over

their respective supports then one obtains a total force of interaction between the two distributions:

$$f_{12}(\rho_1, \rho_2) = \int_{\mathrm{supp}\,\rho_1} \mathcal{V}_x \int_{\mathrm{supp}\,\rho_2} \rho_1(x) G_\varepsilon(x, y) \rho_2(y)\, \mathcal{V}_y\,, \qquad \text{(VII.91)}$$

which is anti-symmetric when $G_\varepsilon(x, y)$ is.

Of course, this process of reducing extended charge distributions to point distribution by integration is not entirely mathematically rigorous, since one obtains a 1-form that is not associated with any specific point of Σ in the general case. Furthermore, the integral does not transform properly under frame changes that involve transition functions that are not constant. However, one does see how the anti-symmetry of the integral kernel implies the anti-symmetry of the force between point charges as approximations to extended charges.

The situation with the magnetic forces of interaction between currents is similar, but complicated by various factors, mostly relating to the non-existence of point currents, except in the form of the transversal intersection of curves with surfaces, and the fact that the Green function for the interaction acts on a current density by the exterior product, not the scalar product.

The Lorentz force density f_{12} that a current $\mathbf{I}_1$ exerts at a point y of $\mathbf{I}_2$ is of the form:

$$\begin{aligned} f_{12}(\mathbf{I}_1, \mathbf{I}_2)(y) &= \#_s(\mathbf{I}_2(y) \wedge \mathbf{B}(y; \mathbf{I}_1) \\ &= \#_s\left[\mathbf{I}_2(y) \wedge \int_{\mathrm{supp}\,\mathbf{I}_1} G_\mu(x, y) \wedge \mathbf{I}_1(x)\, \mathcal{V}_x \right] \\ &= \#_s\left[\int_{\mathrm{supp}\,\mathbf{I}_1} \mathbf{I}_2(y) \wedge G_\mu(x, y) \wedge \mathbf{I}_1(x)\, \mathcal{V}_x \right], \end{aligned} \qquad \text{(VII.92)}$$

and conversely, the force density that $\mathbf{I}_2$ exerts on a point x of $\mathbf{I}_1$ is:

$$f_{12}(\mathbf{I}_2, \mathbf{I}_1)(x) = \#_s\left[\int_{\mathrm{supp}\,\mathbf{I}_2} \mathbf{I}_1(x) \wedge G_\mu(x, y) \wedge \mathbf{I}_2(y)\, \mathcal{V}_y \right]. \qquad \text{(VII.93)}$$

If one then integrates over y or x, resp., then one obtains a total force of the form:

$$f_{12}(\mathbf{I}_1, \mathbf{I}_2) = \#_s\left[\int_{\mathrm{supp}\,\mathbf{I}_1} \mathcal{V}_x \int_{\mathrm{supp}\,\mathbf{I}_2} \mathbf{I}_1(x) \wedge G_\mu(x, y) \wedge \mathbf{I}_2(y)\, \mathcal{V}_y \right] = -\, f_{12}(\mathbf{I}_2, \mathbf{I}_1) \qquad \text{(VII.94)}$$

that is anti-symmetric.

Since the only way of reducing currents to points is by reducing three-dimensions to two dimensions, one can then think of this expression as something like the interaction of magnetic charges, except that there is a significant difference: In the case of currents in parallel infinite wires, one knows, from elementary physics that like – i.e., parallel – currents *attract*, while unlike ones *repel*. Hence, the magnetic interaction of currents behaves more like the Newtonian interaction of masses than the electrostatic interaction of charges.

b. Causal interactions between currents. When the time variation of a current reaches the point that its second derivative is no longer negligible, or the spatial separation between two currents is appreciable, the approximation of action-at-a-distance breaks down and one sees that interactions between currents must be mediated by radiation fields and take into account the time lag that it takes for a photon to go from one source to the other.

Something else that must break down in the process is the anti-symmetry of the interaction, since one is essentially dealing with "signals" that travel from one place to another in a finite amount of time, so there will be a noticeable difference between a signal that is sent from point A at time t_A to a point B at t_B and one that is sent from B at time t_A and arrives at A at time t_B.

It is also reasonable to ask whether the response of the medium itself will depend upon the direction of travel, since there are such things as one-way mirrors. If the constitutive law is invariant under such a replacement of signal source with receiver then one calls the medium *reciprocal*.

Once again, in order to do justice to the causal interactions of currents, one must discuss the theory of electromagnetic radiation in greater detail. Hence, we content ourselves for the moment with only these cursory remarks.

References

1. Jackson, J.D., *Classical Electrodynamics,* 2nd ed., Wiley, New York, 1976.

(other references not cited)

2. Rohrlich, F., *Classical Charged Particles,* Addison-Wesley, Reading, MA, 1965.
3. Barut, A.O., *Electrodynamics and Classical Theory of Fields and Particles*, Dover, NY, 1980.
4. Thirring, W., *Classical Field Theory,* Springer, Berlin, 1978.
5. Landau, L.D., Lifshitz, E.M., *Classical Field Theory,* Pergamon, Oxford, 1975.
6. F. W. Hehl, Y. N. Obukhov. *Foundations of Classical Electrodynamics*, Birkhäuser, Boston, 2003.

Chapter VIII

Electromagnetic waves

In this chapter, we finally get to the most fundamental aspect of pre-metric electromagnetism, which is to account for the "emergence" of the Lorentzian structure on the spacetime manifold as a *possible consequence* of a more general set of possibilities. These possibilities follow from examining the dispersion law for the propagation of electromagnetic waves that one defines from the field equations and the constitutive law of the medium. In particular, the quadratic nature of the characteristic hypersurfaces for the Maxwell field equations, as they are usually presented for isotropic media, is a degenerate case of the quartic dispersion law that appears in anisotropic media.

It is essential to understand that the process of reduction by which we obtain the dispersion law implicitly involves not only a linearization of a potentially nonlinear constitutive law, as well as the fact that dispersion laws are generally subordinate to the basic assumption that one makes about the form of the wavelike solutions to the field equations. However, many of the common forms that wavelike solutions take produce the same dispersion equation. This is closely related to the fact that when one passes from a differential operator to its symbol one loses a considerable amount of information, since the (principal) symbol really only tells one about the highest order of derivatives, and a quasi-linear differential operator will have the same principal symbol as a linear one.

Nevertheless, one thing that all of the common forms for wavelike solution have in common is the fact that they involve some sort of amplitude function, which defines the spatial shape of the wave envelope, and a phase function, which defines the spatial shape of the momentary wave fronts themselves. Although the concept of a plane wave, for which the amplitude function is constant and the phase function is linear, is of limited utility in differentiable manifolds that are not affine spaces, one finds that the concepts of amplitude and phase function can be generalized to nonlinear manifolds.

The geometrical optics approximation that one makes into order to go from the propagation of electromagnetic waves to the behavior of null geodesics, which become light rays when one looks at their spatial projections, amounts to ignoring the contribution of the amplitude function to the wave and concentrating on the phase function as root of the essential information in the wave. As we shall see, this is equivalent to considering only the dispersion law that follows from passing to the symbol of the second-order differential operator that defines the electromagnetic field equations.

Eventually, we intend to show that the geometry of wave motion, at least for characteristic waves, is best addressed in the framework of projective geometry. In particular, both of the concepts of wave normal covector and group velocity vector are naturally defined in projective terms by means of inhomogeneous coordinates, as well as the relationship between them and the Fresnel surfaces for both. Although we shall return to a more general discussion of the role of projective geometry in electromagnetism in Chapter XII, we mention these aspects of it in this chapter for the sake of motivating that more general study.

1. Electromagnetic waves. For the sake of completeness, we rewrite the pre-metric Maxwell equations here:

$$dF = 0, \qquad \delta\mathfrak{h} = \mathbf{J}, \qquad \mathfrak{h} = \kappa(F). \tag{VIII.1}$$

In gauge form, they are:

$$F = dA, \qquad \delta\mathfrak{h} = \mathbf{J}, \qquad \mathfrak{h} = \kappa(F), \tag{VIII.2}$$

which can be combined into a single second-order equation:

$$\Box_\kappa A = \mathbf{J}, \tag{VIII.3}$$

in which:

$$\Box_\kappa = \delta \cdot \kappa \cdot d\colon \Lambda^1 \to \Lambda_1 \tag{VIII.4}$$

is the *electromagnetic field* operator.

In a local coordinate chart (U, x^μ) on M, when the constitutive law is both inhomogeneous and nonlinear, so κ has local components $\kappa^{\mu\nu\alpha\beta}(x, F)$, the operator $\Box_\kappa$ takes the local form:

$$\Box_\kappa^{\nu\alpha} = \tilde{\kappa}^{\mu\nu\alpha\beta}\frac{\partial^2}{\partial x^\mu \partial x^\beta} + \frac{\partial \kappa^{\mu\nu\alpha\beta}}{\partial x^\mu}\frac{\partial}{\partial x^\beta} \tag{VIII.5}$$

in which:

$$\tilde{\kappa}^{\mu\nu\alpha\beta} = \kappa^{\mu\nu\alpha\beta} + \frac{\partial \kappa^{\mu\nu\alpha\beta}}{\partial F_{\kappa\lambda}}\frac{\partial A_\kappa}{\partial x^\lambda}. \tag{VIII.6}$$

Although, as we shall see later when we discuss the dispersion law that follows from this operator, the operator $\Box_\kappa$ is something like a "pre-metric d'Alembertian," actually that is misleading. In particular, the field equation (VIII.3) admits solutions that represent static electric and magnetic fields, as well as time-varying solutions that are not wavelike in character. For instance, one can have fields that vary linearly or exponentially in time.

Hence, since the field equations (VIII.3) are more general than the usual wave equations of mathematics – whether linear or nonlinear – we must treat wavelike solutions as essentially subspaces of the (not necessarily linear) space of solutions A in Λ^1. Some of the common classes of solutions treated in conventional books on electromagnetic waves (e.g., [**1-6**]) are time-periodic electromagnetic fields, fields that are describable by the geometrical optics approximation, and waves as propagating discontinuities, so we show how these topics can be treated in the present formalism.

a. Time-periodic electromagnetic fields. Suppose the spacetime manifold M is space-time separable; i.e., expressible as a product $\mathbb{R} \times \Sigma$, with $\mathbb{R}$ playing the role of the time manifold and Σ, a three-dimensional manifold that plays the role of a spatial manifold. From the previous discussion of space-time splittings of $T(M)$ and its tensor

algebra, we then see that, in particular, all of the bundles of exterior differential forms on M will be decomposed into direct sums of temporal forms and spatial forms, and that the Poincaré duality isomorphism # permutes the temporal and spatial sub-bundles.

However, it is important to notice that the components of the spatial and temporal tensor fields are still functions on M, in general. Hence, when M itself can be decomposed one can also distinguish temporal and spatial functions, as well; that is, a *temporal function* is a function on $\mathbb{R}$ and a *spatial function* is a function on Σ. As a result, we can also distinguish *purely temporal* differential forms on M, which are forms on M that have components in $C^\infty(\mathbb{R})$ – for some adapted coordinate system on M – from *purely spatial* forms on M, which only have components in $C^\infty(\Sigma)$. However, one must clearly distinguish the purely spatial k-forms on M, which can locally have terms involving dt, from the k-forms on Σ, which cannot. We shall use the notation $\Lambda^k_{s'}(M)$ for the bundle of purely spatial k-forms on M, which must be clearly distinguished from the bundle $\Lambda^k_s(M)$ of spatial k-forms on M that is defined by the splitting $T(M) = T(\mathbb{R}) \oplus T(\Sigma)$. Locally, these k-forms look like:

purely temporal: $1/(k-1)!\ \alpha_{0\mu\ldots\nu}(t)\ dt \wedge dx^\mu \wedge \ldots \wedge dx^\nu$,

purely spatial ($a \in \Lambda^k_{s'}(M)$): $1/(k-1)!\ \alpha_{0\mu\ldots\nu}(x^1, \ldots, x^n)\ dt \wedge dx^\mu \wedge \ldots \wedge dx^\nu$,

$\alpha \in \Lambda^k\Sigma$: $1/k!\ \alpha_{\mu\ldots\nu}(x^1, \ldots, x^n)\ dx^\mu \wedge \ldots \wedge dx^\nu$.

In this section, we shall be concerned with a special class of differential forms on M that we call *stationary*. A stationary k-form $\alpha \in \Lambda^k(M)$ takes the form of $\alpha(t, x) = T(t)\alpha'(x)$ where T is a function on $\mathbb{R}$ that is the same for all such forms and α' is a purely spatial k-form on M.

For instance, if an electromagnetic field $F \in \Lambda^2(M)$ on a space-time separable manifold $M = \mathbb{R} \times \Sigma$ is stationary then it can be expressed in the form:

$$F(t, x) = T(t)f(x), \tag{VIII.7}$$

in which $T(t)$ is a smooth function on $\mathbb{R}$ and $f(x) \in \Lambda^k\Sigma$. In particular, the components f_{ij} of f in any coordinate system (t, x^i) on an open subset of M that is adapted to the product structure must be functions of only the x^i:

$$f(x) = \tfrac{1}{2} f_{ij}(x)\ dx^i \wedge dx^j. \tag{VIII.8}$$

From now on, we shall omit the explicit reference to x in the symbol f.

The exterior derivative of F then becomes:

$$dF = dT \wedge f + T\ df = T[d(\ln T) \wedge f + df] \equiv T[\varpi\, dt \wedge f + df], \tag{VIII.9}$$

in which we have introduced the function:

$$\varpi(t) = d(\ln T). \tag{VIII.10}$$

This integrates immediately to:

$$T(t) = \exp\left[\int_0^t \varpi(\tau)d\tau\right], \tag{VIII.11}$$

in which we have suppressed the integration constant by setting $t_0 = 0$.

Of course, in the usual case that is treated by geometrical optics, ϖ is a negative imaginary constant $-i\omega$ so the function $T(t)$ takes the form $e^{-i\omega t}$; from now on, we shall make that assumption. When T takes that form, a stationary field on M is then called *time-periodic*.

In order to make sense of the multiplication by the imaginary i, we could assume that the constitutive law defines an almost complex structure * on $\Lambda^2(M)$, but, as it turns out, in order get agreement with the stationary form of the Maxwell equations, we must not use that structure to define multiplication by i. For one thing, we shall need to multiply 1-forms and 3-forms by i, as well. Rather, we understand that if α is a k-form and $X_1, \ldots, X_k$ are vector fields on M then $i\alpha(X_1, \ldots, X_k)$ is the imaginary number that is obtained by multiplying the real number $\alpha(X_1, \ldots, X_k)$ by i. Hence, in this case we are passing to the complexification of $\Lambda^*(M)$.

For simplicity, we set $* = \# \cdot \kappa$; so the second Maxwell equation takes the form $d*F = 0$ in the absence of sources. Furthermore, we assume that $*^2 = -I$, so it does, in fact, define an almost-complex structure.

Since $*F$ will take the form $T(t)*f(x)$, we see that a straightforward replacement of f with $*f$ in (VIII.9), combined with aforementioned replacement of ϖ with the constant $-i\omega$, will give:

$$dF = T[-i\omega\, dt \wedge f + df]. \tag{VIII.12a}$$
$$d*F = T[-i\omega\, dt \wedge *f + d*f]. \tag{VIII.12b}$$

Maxwell's sourceless equations then imply that:

$$df = i\omega\, dt \wedge f, \qquad d*f = i\omega\, dt \wedge *f. \tag{VIII.13}$$

To see that these indeed give the customary E-B form, substitute:

$$F = dt \wedge E - *(dt \wedge B) = T[dt \wedge u - *(dt \wedge v)], \tag{VIII.14}$$

in which we have assumed that the 1-forms E and B on M take the form:

$$E(t, x) = T(t)u(x), \qquad B(t, x) = T(t)v(x), \tag{VIII.15}$$

so $u(x)$ and $v(x)$ are 1-forms on Σ. Hence, they will have the component form:

$$u(x) = u_i(x)\, dx^i, \qquad v(x) = v_i(x)\, dx^i, \tag{VIII.16}$$

in an adapted coordinate system.

From (VIII.14), we see that:

$$*F = T[dt \wedge v + *(dt \wedge u)], \tag{VIII.17}$$

Hence, between (VIII.14) and (VIII.17), we see that:

$$f = dt \wedge u + *_s v, \qquad *f = dt \wedge v - *_s u. \tag{VIII.18}$$

We have also replaced the expressions $*(dt \wedge u)$ and $*(dt \wedge v)$ with $-*_s u$ and $-*_s v$, respectively.

It is physically illustrative to see what sort of matrix the map $*_s : \Lambda^1\Sigma \to \Lambda^2\Sigma$ will have. We use a basis $\{b^i, i = 1, 2, 3\}$ for a fiber of $\Lambda^1\Sigma$ that is adapted to a basis for $\Lambda^1 M = \Lambda^1 L \oplus \Lambda^1\Sigma$, and the corresponding basis $*b^i$ for $\Lambda^2\Sigma$ that is a subset of the basis $\{dt \wedge b^i, *b^i\}$ for $\Lambda^2 M$. The linear map $*_s = -*e_{dt}$ then has a 3×3 matrix relative to these bases that is the composition of the matrices $[e_{dt}][*][P_{\mathrm{Im}}]$, since the components of $u = u_i\, b^i$ are row matrices. Here, the linear map P_{Im}: $\Lambda^2 M = \Lambda^2_{\mathrm{Re}} M \oplus \Lambda^2_{\mathrm{Im}} M$ is the canonical projection of any 2-form onto its imaginary or "magnetic" part. If one recalls the discussion of almost-complex media from Chapter V, and represents the 6×6 matrix [*] in the form of (V.34) then the product of matrices in question is:

$$[*_s] = -\begin{bmatrix} I & 0 \end{bmatrix} \left[\begin{array}{c|c} \gamma & -\varepsilon \\ \hline \tilde{\mu} & \gamma^T \end{array}\right] \begin{bmatrix} 0 \\ \hline I \end{bmatrix} = [\varepsilon]. \tag{VIII.19}$$

This somewhat surprising result, that spatial duality is due solely to the electric permittivity of the medium, is actually subordinate to the assumption that we are considering an almost-complex medium, since (VIII.35) suggests that in order for an electromagnetic constitutive law to define such a structure, one must satisfy rather severe restrictions on the form of the submatrices of κ. In particular, in the absence of electromagnetic couplings ($\gamma = 0$), one must have that ε is proportional to μ.

Exterior differentiation of the resulting expressions in (VIII.18) gives:

$$df = -dt \wedge du + d*_s v = i\omega dt \wedge f = i\omega dt \wedge *_s v, \tag{VIII.20a}$$
$$d*f = -dt \wedge dv - d*_s u = i\omega dt \wedge *f = -i\omega dt \wedge *_s u, \tag{VIII.20b}$$

which gives:

$$-dt \wedge du + d*_s v = dt \wedge (i\omega *_s v), \tag{VIII.21a}$$
$$dt \wedge dv + d*_s u = dt \wedge (i\omega *_s u). \tag{VIII.21b}$$

From (VIII.13), (VIII.18), and (VIII.20), and considering the temporal or spatial nature of the 3-forms in question, we deduce:

$$du = -i\omega *_s v, \qquad d*_s u = 0, \tag{VIII.22a}$$
$$dv = +i\omega *_s u, \qquad d*_s v = 0. \tag{VIII.22b}$$

If we apply the $*_s$ operator to both sides of each equation then we see that they take the vector calculus form:

$$\nabla\times\mathbf{u} = +\, i\omega\mathbf{v}, \qquad \nabla\cdot\mathbf{u} = 0, \tag{VIII.23a}$$

$$\nabla\times\mathbf{v} = -\, i\omega\mathbf{u}, \qquad \nabla\cdot\mathbf{v} = 0, \tag{VIII.23b}$$

that is more customary in geometrical optics (see Luneberg [**6**]).

b. Stationary potential 1-forms. Since Maxwell's first equation suggests that the 2-form F can be locally represented by dA for some non-unique potential 1-form A, we next look at what happens when this 1-form is also stationary:

$$A(t, x) = T(t)a(x). \tag{VIII.24}$$

A change of gauge $A \mapsto A + d\lambda$ must also contain T as a common factor in both terms. In particular, $\lambda(t, x)$ must take the form $T(t)l(x)$. As a consequence, a change of gauge for A implies a change of the form $a \mapsto a + dl - i\omega l\, dt$ for a.

One must be careful to note that this time, although we are assuming that the components of $a(x)$ are functions on Σ, nevertheless, we are allowing it to take the form:

$$a = \phi(x)\, dt + a_i(x)dx^i \equiv \phi\, dt + a_s\,; \tag{VIII.25}$$

hence, it is what we have been calling a purely spatial 1-form on M.

We then have:

$$da = -\, dt \wedge d\phi + da_s\,. \tag{VIII.26}$$

However, in optical problems it is customary to assume that the electrostatic field $d\phi$ is null.

This makes:

$$F = dA = T(-\, i\omega dt \wedge a + da) = T[-\, dt \wedge (d\phi + i\omega a_s) + da_s]. \tag{VIII.27}$$

Hence:

$$f = -\, dt \wedge (d\phi + i\omega a_s) + da_s\,. \tag{VIII.28}$$

By matching up temporal and spatial terms in (VIII.28) and the first of (VIII.18), we get:

$$u = -\, d\phi - i\omega a_s, \qquad *_s v = -\, da_s\,. \tag{VIII.29}$$

Since:

$$*f = -\, *[dt \wedge (d\phi + i\omega a_s)] + *da_s = *_s\, d\phi + i\omega *_s\, a_s + *da_s\,, \tag{VIII.30a}$$

$$d*f = d*_s\, d\phi + i\omega d*_s\, a_s + d*da_s\,, \tag{VIII.30b}$$

the remaining Maxwell equation for $*F$ gets converted into the form:

$$\begin{aligned} 0 &= d*f - i\omega dt \wedge *f \\ &= [d*da_s + \omega^2\, dt \wedge *_s\, a_s] + [d*_s\, d\phi - i\omega dt \wedge *_s\, d\phi\,] + i\omega d*_s\, a_s\,. \end{aligned} \tag{VIII.31}$$

With the choice of gauge:

$$d\phi = 0, \qquad d^*_s a_s = 0, \tag{VIII.32}$$

which is essentially the Coulomb gauge, equation (VIII.30) takes the form:

$$0 = d^* d a_s + \omega^2\, dt \wedge {}^*_s a_s. \tag{VIII.33}$$

In order to put this into the customary Helmholtz form, first apply the * operator to both sides:

$$0 = {}^* d^* d a_s - \omega^2\, a_s\,. \tag{VIII.34}$$

By following the sequence of operators d, $*$, d, $*$ acting on the 1-form $a_s \in \Lambda^1(\Sigma)$, one sees that $*d*d$ has the same effect as:

$$-\,{}^*_s d^*_s d = -\,\Delta + d^*_s d^*_s\,. \tag{VIII.35}$$

However, with the choice of gauge (VIII.32) the second term on the right-hand side will vanish, and we finally obtain:

$$0 = \Delta a_s + \omega^2 a_s\,. \tag{VIII.36}$$

c. Geometrical optics approximation. The form that electromagnetic fields take in geometrical optics is a generalization of the aforementioned time-periodic form. In particular, rather than factoring $F(t, x)$ into a purely temporal part and a purely spatial part by way of the time function $e^{-i\omega t}$, we now assume that M still takes the form $\mathbb{R} \times \Sigma$, but we factor F as:

$$F(t, x) = e^{-i\phi} f(x), \tag{VIII.37}$$

in which $\phi(t, x)$ is a smooth function on M that one calls the *phase function*. Its level hypersurfaces in M are then referred to as *isophases.*

The most common form that ϕ takes is the *plane-wave* form:

$$\phi(t, x) = k_\mu x^\mu = \omega t \pm k_i x^i\,, \tag{VIII.38}$$

which amounts to assuming that $M = \mathbb{R}^4$ and ϕ is linear in the coordinates x^μ. The isophases are then affine hyperplanes. By means of other choices of coordinate system on M one can similarly define cylindrical and spherical isophases.

One sees that time-periodic electromagnetic fields are the special case of (VIII.38) for which $k_i = 0$ for all $i = 1, 2, 3$. However, one then refers to the isophases as *isochrones*, or *simultaneity hypersurfaces*. Clearly, this notion demands a more relativistic treatment, but we shall only point the curious to the author's work [**7**], and the references that were cited therein. Furthermore, the spirit of pre-metric electromagnetism is that causality is a *consequence* of the manner by which electromagnetic waves propagate through spacetime, not a prerequisite for it.

The physical significance of an isophase is that it represents the time evolution of a *momentary wave surface* in the space Σ. That is, the projection of an isophase in *M* onto Σ will be, by definition, a momentary wave surface. Hence, if $f(x)$ represents the overall "shape" of the wave motion in Σ then each value $f(x)$ will propagate in *M* on a fixed isophase. In effect, an isophase is a higher-dimensional analogue of the world line of a point particle.

When *F* takes the form (VIII.1), one derives:

$$dF = e^{-i\phi}(-ik \wedge f + df), \tag{VIII.39}$$

in which we have set:

$$k = d\phi. \tag{VIII.40}$$

Similarly, since $*F = e^{-i\phi}*f$, one has:

$$d*F = e^{-i\phi}(-ik \wedge *f + d*f). \tag{VIII.41}$$

Maxwell's sourceless equations then give:

$$df = ik \wedge f, \qquad d*f = ik \wedge *f. \tag{VIII.42}$$

If we compare these equations with (VIII.13) then we see that we have essentially replaced $\omega\, dt$ with $k = \omega\, dt + k_i\, dx^i$.

Note that, from Frobenius, the existence of such a *k* as in (VIII.42) implies that when *f* – hence, $*f$ – is a decomposable (i.e., rank two) 2-form the exterior differential systems $f = 0$ and $*f = 0$ must be completely integrable.

Now, this is always the case for electromagnetic waves, due to the basic properties of wavelike solutions to the Maxwell equations:

$$0 = f \wedge f = f \wedge *f. \tag{VIII.43}$$

Hence, *M* will be doubly foliated by the two-dimensional leaves to each of these foliations. At each point of *M* one leaf will be tangent to the characteristic variety of the field operator and the other one will be normal to it, and their intersection will be tangent to a generator of the characteristic variety; we shall discuss this situation in more detail shortly. When projected into Σ, the leaf that is tangent will produce a momentary wave surface, while the leaf that is normal will produce the normal trajectory that is followed by a point on that surface in time.

This is the point at which geometrical optics makes an approximation. One assumes that the spatial variation of the amplitude 2-form *f* is sufficiently slow compared to the variation of the phase function θ in time, which is referred to as either the *high-frequency* limit or the *small-wavelength* limit. For wavelengths on the order of visible light and smaller, this approximation is generally more than sufficient, although for the purposes of radio waves, which can have wavelengths in meters, it is generally inadequate.

In order to make the statement of this approximation somewhat more precise, we introduce the following notation:

$$k = \omega\hat{k}, \qquad \hat{k} \equiv dt - n_i\, dx^i, \qquad n_i \equiv k_i / \omega. \tag{VIII.44}$$

The components n_i then amount to indices of refraction in the various spatial directions, as they are reciprocal to the phase velocities ω/k_i, which we shall discuss below.

This allows us to rewrite (VIII.41) in the form:

$$1/\omega\, df = i\hat{k} \wedge f, \qquad 1/\omega\, d{*}f = i\hat{k} \wedge {*}f. \tag{VIII.45}$$

If one takes the high-frequency limit of these expressions then the Maxwell equations reduce to the algebraic equations:

$$0 = \hat{k} \wedge f = \hat{k} \wedge {*}f, \tag{VIII.46}$$

along with the partial differential equation (VIII.40). One can interpret equations (VIII.46) as saying that $\hat{k}$ – or k, for that matter – is incident on both the annihilating plane of f and the annihilating plane of $*f$, which then implies that $\hat{k}$ generates their line of intersection.

Now, let us examine the 1+3 form of equations (VIII.46). Thus, we let $f = dt \wedge u + *_s v$ and $*f = dt \wedge v - *_s u$, as before. We further let $\hat{k} = dt - n$, which makes:

$$\hat{k} \wedge f = + dt \wedge (*_s v + n \wedge u) - n \wedge *_s v, \tag{VIII.47a}$$
$$\hat{k} \wedge {*}f = - dt \wedge (*_s u - n \wedge v) + n \wedge *_s u. \tag{VIII.47b}$$

By considering the vanishing of the temporal and spatial parts independently, (VIII.47) gives:

$$*_s v = - n \wedge u, \qquad n \wedge *_s v = 0, \tag{VIII.48a}$$
$$*_s u = + n \wedge v, \qquad n \wedge *_s u = 0. \tag{VIII.48b}$$

One sees that these equations are similar to (VIII.22a,b), except that now there is no differentiation involved, and we have absorbed the factor of ω into n.

Note that now the second equation in each set follows unavoidably from the first.

The effect of these equations can be better understood by relating them to the conventional forms that they take. By operating on both sides of the first equations with $*_s$, and taking advantage of the fact that $*_s(\mathbf{a} \wedge \mathbf{b}) = \mathbf{a} \times \mathbf{b}$ and $a \wedge *_s b = (\mathbf{a} \cdot \mathbf{b})V_s$, we can put (VIII.48a, b) into the vector form:

$$\mathbf{v} = + \mathbf{n} \times \mathbf{u}, \qquad \mathbf{n} \cdot \mathbf{v} = 0, \tag{VIII.49a}$$
$$\mathbf{u} = - \mathbf{n} \times \mathbf{v}, \qquad \mathbf{n} \cdot \mathbf{u} = 0. \tag{VIII.49b}$$

Hence, given **u** one can deduce **v**, and vice versa. Similarly, if one given **u** and **v**, which are basically the Cauchy data for the initial-value problem associated with the field equations, when expressed in space-time form, one can derive **n** by using the first equation in either (VIII.48a) or (VIII.48b). One simply takes the interior product of both sides of (VIII.48a) with **u**, taking into account that $u(\mathbf{u}) = u^2$, $n(\mathbf{u}) = 0$:

$$n = u^{-2} *_s(u \wedge v), \tag{VIII.50}$$

Since the vector form of this is:

$$\mathbf{n} = \| \mathbf{u} \|^{-2} \mathbf{u} \times \mathbf{v}, \tag{VIII.51}$$

we see that the 2-form $u \wedge v$ is essentially the Poincaré dual of the *Poynting vector* for the wave.

Furthermore, when one includes the space-time form of (VIII.43), namely:

$$u \wedge *_s v = u \wedge *_s u - v \wedge *_s v = 0, \tag{VIII.52}$$

which has the vector form:

$$\mathbf{u} \cdot \mathbf{v} = u^2 - v^2 = 0, \tag{VIII.53}$$

one sees that the set $\{\mathbf{u}, \mathbf{v}, \mathbf{k}_s\}$ represents a set of orthogonal, but not orthonormal, vectors at each point of Σ, relative to the spatial metric that is defined by $\langle u, v \rangle = \mathcal{V}_s(u \wedge *_s v)$.

We can also substitute the values of $*_s u$ and $*_s v$ that we obtained from (VIII.47a, b) into the definitions of f and $*f$ to deduce that:

$$f = \hat{k} \wedge u, \qquad\qquad *f = \hat{k} \wedge v. \tag{VIII.54}$$

This makes it clearer just what the nature of the 2-planes that are associated with f and $*f$ amounts to. The plane of f in a cotangent space is spanned by the orthogonal covectors k and u, while that of $*f$ is spanned by k and v. It is traditional to call the plane that is spanned by f the *polarization plane*.

d. Propagating discontinuities. Although the Fourier conception of waves as being linear superpositions of elementary plane waves is quite mathematically powerful and pervasive in its practical applications, there are nonetheless several disadvantages to this approach as far as physics is concerned.

For one thing, the very concept of a plane wave in $\mathbb{R}^3$ is unphysical, since such a wave ends up having infinite total momentum and infinite total kinetic energy, due to the non-compactness of its support. Of course, one almost always restricts one's consideration to plane waves in a compact region of $\mathbb{R}^3$, but, strictly speaking, such a spatial truncation of the wave must necessarily introduce higher-wave-number contributions to the spectral density of the wave that imply that it is not really a plane wave. Another disadvantage of planes waves is that they can only be defined globally in affine spaces, so their use in more general differentiable manifolds has a purely local character.

In the Hadamard approach to wave motion [**8-12**], a wave is defined to be a disturbance in a region of a medium that propagates through that medium. In particular, the region of the disturbance is assumed to be bounded by a surface $\mathcal{F}$ that one can call a *singular hypersurface* or the *wave front*. There might also be a bounding surface at the trailing edge of the wave, although generally one expects the shape of the wave envelope to decay smoothly from dissipative forces in the medium. Indeed, waves in even-

dimensional spaces (i.e., odd-dimensional spacetimes) will always be associated with such a decaying "tail."

Across this surface, one expects that there is a finite jump discontinuity in the field under consideration or one of its normal derivatives at some order. For instance, acoustic shock waves represent jump discontinuities in the covelocity 1-form u, which represents the exterior derivative of the velocity potential function. One also thinks of the sources of mechanical waves as being due to jump discontinuities in the acceleration vector field that originate in the fact that the driving force that produces them is approximately time-impulsive. That is, as a function of time, it looks like a Dirac delta function about some initial time point.

For electromagnetic waves, a time-impulsive source current $\mathbf{J}$ will produce a jump discontinuity in the bivector field $\mathfrak{h}$, and we assume that the set of all points at which $\mathfrak{h}$ is discontinuous is the wave front $\mathcal{F}$. Whether the jump discontinuity in $\mathfrak{h}$ also produces a jump discontinuity in F depends upon the linearity of κ: For the sake of progress, we assume for the rest of this section that F also has a jump discontinuity across $\mathcal{F}$. We denote the discontinuities in the fields in question by $[\mathbf{J}]$, $[F]$ and $[\mathfrak{h}]$, respectively. They then represent a smooth vector field, a smooth 2-form, and a smooth bivector field on $\mathcal{F}$, respectively.

In order to relate these discontinuities to the pre-metric Maxwell equations, one needs to first recognize that the most rigorous formalism in which to treat jump discontinuities analytically is in the language of distributions. We must then reformulate the pre-metric Maxwell equations in terms of distributions on differential forms and multivector fields.

Since we described several ways of defining continuous linear functionals on Λ^* and Λ_*, we need to first specify which definition that we are using. Because we are using the differential operators d and δ, the definition that is most convenient to our immediate purposes is the one that makes these operators adjoint to each other, namely:

$$\langle F, \mathbf{A}\rangle = \int_M F \wedge \#\mathbf{A} = \int_M F(\mathbf{A})\,\mathcal{V}, \tag{VIII.55a}$$

$$\langle \mathfrak{h}, \alpha\rangle = \int_M \#\mathfrak{h} \wedge \alpha = \int_M \alpha(\mathfrak{h})\,\mathcal{V}, \tag{VIII.55b}$$

$$\langle \mathbf{J}, \beta\rangle = \int_M \#\mathbf{J} \wedge \beta = \int_M \beta(\mathbf{J})\,\mathcal{V}. \tag{VIII.55c}$$

In these expressions, $\mathbf{A}$ is an arbitrary smooth bivector field of compact support, α is a smooth 2-form of compact support, and β is a smooth 1-form of compact support.

The Green formula for d and δ is obtained by integrating the product rule for $d(\alpha \wedge \#\mathbf{A})$:

$$\langle d\alpha, \mathbf{A}\rangle + (-1)^k \langle \alpha, \delta\mathbf{A}\rangle = \int_{\partial M} \alpha \wedge \#\mathbf{A}, \tag{VIII.56}$$

which vanishes when M has no boundary or a boundary that is disjoint from the support of $\mathbf{A}$. In that event, the operators d and δ are adjoint with respect to this bilinear pairing.

Hence, we can define the exterior product of the distribution F and the divergence of the distribution $\mathfrak{h}$ by way of:

$$\langle dF, \mathbf{A}\rangle = -\langle F, \delta\mathbf{A}\rangle, \qquad \langle \delta\mathfrak{h}, \alpha\rangle = -\langle \mathfrak{h}, d\alpha\rangle. \qquad \text{(VIII.57)}$$

Therefore, these definitions allow one to define the differentiation of fields with jump discontinuities by differentiating the test fields on which the distributions act.

Thus, in order to make sense of the pre-metric Maxwell equations as equations involving distributions, one then says that what they really mean is that for every smooth trivector field **A** of compact support and every smooth 3-form α of compact support, one has:

$$\langle dF, \mathbf{A}\rangle = 0, \qquad \langle \delta\mathfrak{h}, \alpha\rangle = \langle \mathbf{J}, \alpha\rangle, \qquad \mathfrak{h} = C(F), \qquad \text{(VIII.58)}$$

or:

$$\langle F, \delta\mathbf{A}\rangle = 0, \qquad \langle \mathfrak{h}, d\alpha\rangle = -\langle \mathbf{J}, \alpha\rangle, \qquad \mathfrak{h} = C(F), \qquad \text{(VIII.59)}$$

One can then consolidate these equations into equations for F alone or $\mathfrak{h}$ alone by using the third equation. In the sourceless case, this gives:

$$\langle F, \delta\mathbf{A}\rangle = 0, \qquad \langle C(F), d\alpha\rangle = 0, \qquad \text{(VIII.60)}$$

or:

$$\langle C^{-1}(\mathfrak{h}), \delta\mathbf{A}\rangle = 0, \qquad \langle \mathfrak{h}, d\alpha\rangle = 0, \qquad \text{(VIII.61)}$$

respectively.

If F has a jump discontinuity across a hypersurface $\phi(x) = 0$ then we let F_- represent the restriction of F to the half-space $\phi(x) < 0$ and F_+ represents its restriction to $\phi(x) > 0$. By the notation $[F] = F_+ - F_-$, we intend to denote a smooth 2-form on the hypersurface itself that represents the actual jump itself.

e. Polarization. Upon closer inspection of most treatments of the polarization of electromagnetic fields [**1**-**6**], one sees that the fact that one is using complex electric and magnetic field vectors is largely irrelevant to the nature of the discussion. In particular, one does not make use of either the electromagnetic constitutive laws of the medium or the Maxwell equations. Indeed, the essence of the construction amounts to the description of how the orbit of $U(1)$ on Euclidian $\mathbb{C}^3$, by way of scalar multiplication by the factor $e^{-i\omega t}$, projects onto the real $\mathbb{R}^3$ subspace.

Clearly, since $U(1)$ is a circle as a real manifold and its action on $\mathbb{C}^3$ is faithful, the image of $U(1)$ is diffeomorphic to a circle. If $\mathbf{a} = \mathbf{a}_R + i\mathbf{a}_I \in \mathbb{C}^3$ is an arbitrary non-zero vector then the orbit of $\mathbf{a}$ under the action of $e^{-i\omega t}$ takes the form:

$$\begin{aligned} e^{-i\omega t}\mathbf{a} &= \cos(t\omega)\,\mathbf{a} - i\sin(\omega t)\,\mathbf{a} \\ &= [\cos(t\omega)\,\mathbf{a}_R + \sin(t\omega)\,\mathbf{a}_I] - i\,[\sin(t\omega)\,\mathbf{a}_R - \cos(t\omega)\,\mathbf{a}_I] \end{aligned} \qquad \text{(VIII.62)}$$

The projection of $e^{-i\omega t}\mathbf{a}$ onto the real $\mathbb{R}^3$ subspace is then:

$$\mathrm{Re}[e^{-i\omega t}\mathbf{a}] = \cos(t\omega)\,\mathbf{a}_R + \sin(t\omega)\,\mathbf{a}_I\,. \qquad \text{(VIII.63)}$$

Although this is a loop in $\mathbb{R}^3$, one can perform a projection of $\mathbb{R}^3$ onto $\mathbb{R}^2$ (and possibly just $\mathbb{R}$) by taking the vectors $\mathbf{a}_R$ and $\mathbf{a}_I$ to $(a_R, 0)$ and $(0, a_I)$, respectively, unless $\mathbf{a}_R$ and $\mathbf{a}_I$ are collinear as real vectors, in which case, they go to $(a_R, 0)$ and $(a_I, 0)$, respectively. Hence, the image of $\mathrm{Re}[e^{-i\omega t}\mathbf{a}]$ in $\mathbb{R}^2$ takes the form $(a_R \cos(t\omega), a_I \sin(t\omega))$, in the non-degenerate case, and $(a \cos \omega t, 0)$, in the degenerate case, with $a = \max\{a_R, a_I\}$. In general, this curve takes the form of an ellipse with the given vectors as semi-major and semi-minor axes, depending upon the relative magnitudes of the vectors. As we have defined things, the vectors $(a_R, 0)$ and $(0, a_I)$ will be orthogonal even when $\mathbf{a}_R$ and $\mathbf{a}_I$ are not, which differs from the standard treatment of polarization. However, one can still distinguish the same three polarization classes, as usual.

One can then classify the type of ellipse in terms of the relationship between the two real 3-vectors $\mathbf{a}_R$ and $\mathbf{a}_I$. In the generic case, they are non-parallel and unequal in length; this is referred to as *elliptical* polarization. When they are equal in length, this is called *circular* polarization. The degenerate case, in which the ellipse flattens into a line segment, is called *linear* polarization.

Since a circle has two orientations – viz., clockwise and counter-clockwise – and the action of $U(1)$ preserves orientation, we can also speak of the sense in which the orbit of $\mathbf{a} \in \mathbb{C}^3$ is traversed. Rather than using the word "orientation" to describe this situation, one uses the word *helicity*. That quantity will have the value +1 or –1, depending upon the convention that one chooses for the orientation of circles in $\mathbb{R}^3$.

2. Characteristics. For the purposes of this section, in order to avoid confusion we shall refer to the constitutive map by the notation C, instead of κ:

a. Characteristic polynomial. In order to find the symbol of the second order differential operator $\Box_C$ we need to replace C with its linearization DC if it is nonlinear to begin with; from now on, we simply assume that C is linear. We then find that the symbol of the field operator $\Box_C$ is:

$$\sigma[\Box_C, k] = i_k \cdot C \cdot e_k\,. \qquad \text{(VIII.64).}$$

From (VIII.5), we see that the components of the linear operator $\sigma[\Box_C, k]$ are:

$$\sigma^{\mu\nu}[\Box_C, k] = -\,C^{\mu\kappa\lambda\nu}\,k_\kappa k_\lambda\,. \qquad \text{(VIII.65)}$$

It is illuminating to compute the 4×4 matrix $\sigma^{\mu\nu}[\Box_C, k]$ explicitly by regarding (VIII.64) as the product of the 4×6 matrix $\left[i_k\right]_I^\mu$, the 6×6 matrix C^{IJ}, and the 6×4 matrix

$[e_k]^\nu_J$, respectively, when we choose the coframe θ^μ for $\Lambda^1 M$, the coframe$\{b^i, \#\mathbf{b}_i\}$ for $\Lambda^2 M$, and the reciprocal frames $\{\mathbf{b}_i, \#b^i\}$ and $\mathbf{e}_\mu$ for $\Lambda_1 M$ and $\Lambda_2 M$, respectively. As usual, we set $b^i = \theta^0 \wedge \theta^i$ and $\mathbf{b}_i = \mathbf{e}_0 \wedge \mathbf{e}_i$. By direct computation, one finds that:

$$[i_k]^\mu_I = \left[\begin{array}{c|cc} -k_i & 0 & 0 \\ \hline \omega\delta^i_j & \multicolumn{2}{c}{-ad(k)^i_j} \end{array}\right], \qquad ad(k)^i_j \equiv \varepsilon^{ijk} k_k = \begin{bmatrix} 0 & -k_2 & k_3 \\ k_2 & 0 & -k_1 \\ -k_3 & k_1 & 0 \end{bmatrix}. \quad \text{(VIII.66)}$$

The matrix $[e_k]^\nu_J$ is simply the transpose of the matrix $[i_k]^\mu_I$ since the maps are adjoint to each other. If we give C^{IJ} the usual form that we discussed in Chapter V then the multiplication of the matrices gives:

$$\sigma^{\mu\nu}[\Box_C, k] = \left[\begin{array}{c|c} \varepsilon(k,k) & -\omega\varepsilon^{im}k_m - k_m\gamma^m_n ad(k)^{in} \\ \hline -\omega\varepsilon^{jm}k_m + k_m\hat{\gamma}^m_n ad(k)^{jn} & \omega^2\varepsilon^{ij} + \omega[\gamma\cdot ad(k) - ad(k)\cdot\hat{\gamma}]^{ij} - ad(k)^{im}\tilde{\mu}_{mn}ad(k)^{nj} \end{array}\right] \quad \text{(VIII.67)}$$

In general, the bundle map $\sigma[\Box_C, k]$: $T^*(M) \to T(M)$ does not have to be invertible, and this depends upon the choice of k. As we discussed in the chapter on partial differential equations, the definition:

$$P[k] \equiv \det \sigma[\Box_C, k] \quad \text{(VIII.68)}$$

then defines a polynomial in k that one calls the *characteristic polynomial* of the differential operator $\Box_C$. It vanishes iff $\sigma[\Box_C, k]$ is not invertible, and the zero locus of $P[k]$ is called the *characteristic hypersurface (or variety)* for $\Box_C$; we shall then call a k that lies in this hypersurface *characteristic*. The algebraic equation in k that is thus defined:

$$P[k] = 0 \quad \text{(VIII.69)}$$

is then what we call the *dispersion law* for the wave medium in question.

From (VIII.68), one can see that since k appears twice in $\sigma[\Box_C, k]$ this implies that the polynomial $P[k]$ will have a degree that is equal to $2n$. It is, moreover, homogeneous in k of degree $2n$. However, there are some traditional reductions that get applied to the degree.

First, one generally deals with the case of time-invariant constitutive laws on space-time separable manifolds, so M takes the form $\mathbb{R} \times \Sigma$, where $\mathbb{R}$ plays the role of the time manifold and Σ is the spatial manifold. This reduces our map $\sigma[\Box_C, k]$ to a map from $T^*\Sigma$ to $T(\Sigma)$, and the degree of $P[k]$ to $2(n-1)$; basically, one considers only the lower

right-hand sub-matrix in (VIII.67). For instance, when $n = 4$ the polynomial in k that one considers is homogeneous and sextic.

Second, as we are dealing with the case of electromagnetic waves, the map e_k: $\Lambda^1(\Sigma) \rightarrow \Lambda^2(\Sigma)$, $\phi \mapsto k \wedge \phi$ is not invertible for any k, since its kernel is one-dimensional, namely, all ϕ that take the form λk for some scalar $\lambda \in \mathbb{R}$. This is related to the fact that electromagnetic waves have no longitudinal modes of vibration, but are confined to the *Poynting plane,* which is 2-plane in T^*M that is spanned by E and H. Hence, the characteristic polynomial reduces to a homogeneous quartic polynomial in k in the electromagnetic case:

$$P[k] = P^{\kappa\lambda\mu\nu}\, k_\lambda\, k_\kappa\, k_\mu\, k_\nu\,. \tag{VIII.70}$$

By the fact that any polynomial of degree d in the coordinates k_μ of $\mathbb{R}^{n*}$ is associated with a completely symmetric covariant tensor of rank d, one can define a fourth-rank completely symmetric covariant tensor field on M that is associated with the polynomial $P[k]$. Locally, it looks like:

$$P = P^{\kappa\lambda\mu\nu}\, \partial_\kappa \partial_\lambda \partial_\mu \partial_\nu\,. \tag{VIII.71}$$

In the book by Hehl and Obukhov [**13**], this tensor field is referred to as the *Tamm-Rubilar tensor,* since it represents an enlargement of the scope of a tensor that was first defined by Tamm [**14**] in the early Twentieth Century that was defined by Rubilar [**15**]. In particular, the latter author derived an expression for the components $P^{\kappa\lambda\mu\nu}$ in terms of the components $\kappa^{\kappa\lambda\mu\nu}$ of the constitutive law:

$$P^{\kappa\lambda\mu\nu} = \frac{1}{4!}\, \varepsilon_{\alpha\beta\gamma\delta}\, \varepsilon_{\rho\sigma\tau\eta}\, \kappa^{\alpha\beta\rho(\kappa}\, \kappa^{\lambda\,|\gamma\delta|\mu}\, \kappa^{\nu)\sigma\tau\eta}\,. \tag{VIII.72}$$

One sees that, like the polynomial that spawned it, the tensor field is also homogeneous of degree four.

b. Propagation of discontinuities. One of the many good reasons for representing waves as propagating discontinuities, besides their generality and independence of linear space structures, is the fact that if the Cauchy problem for a second-order partial differential equation is well-posed then a second-order jump discontinuity in the solution – i.e., an acceleration wave – can only be defined across a characteristic hypersurface. Hence, the second-order jump discontinuities can only represent wavelike solutions.

This result was something that Hadamard showed in his seminal work [**8**] on the basis of compatibility relations for the wave function ψ that followed from the assumption that the function $[\psi] = \psi_+ - \psi_-$, which is defined on S, is continuously differentiable to at least second order. The only potentially unresolved aspect of the behavior of ψ across S is then in its normal – i.e., time – derivatives. Now, the first time derivative on S is assumed to be continuous, as part of the Cauchy problem, so the issue is whether one is free to specify the second time derivative of ψ on S or does it follow automatically by

specifying the Cauchy data and demanding that ψ satisfy the wave equation in question, viz., $\Box_C \psi = 0$, everywhere.

It is in the process of resolving the latter question that one finds yet another way of passing to essentially the symbol of $\Box_C$ and obtaining the characteristic equation. We exhibit this process in the form:

Theorem:

If F is a weak solution of the sourceless pre-metric Maxwell equations that has a jump discontinuity $[F]$ *across the hypersurface* $\phi(x) = 0$ *then that hypersurface is characteristic. In particular:*

$$d\phi \wedge [F] = 0, \qquad i_{d\phi}[\mathfrak{h}] = 0, \qquad [\mathfrak{h}] = C([F]), \tag{VIII.73}$$

so:

$$F[d\phi] = 0. \tag{VIII.74}$$

Proof:

Suppose $c = c_- + c_+$ is a 4-cycle that intersects $\phi(x) = 0$ such that c_- lies in the half-space $\phi(x) < 0$, while c_+ lies in the half-space $\phi(x) > 0$; hence, one also has $\partial c_- = -\partial c_+$. This situation is depicted in Fig. 7.

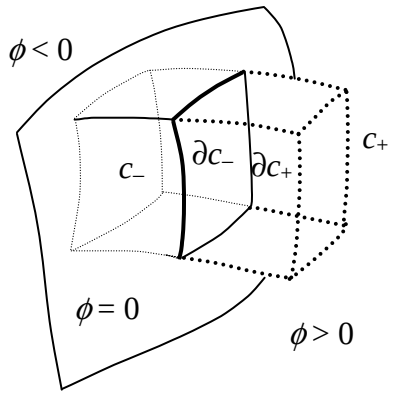

Figure 7. A typical 4-cycle that intersects the hypersurface $\phi(x) = 0$.

On c_- and c_+ the adjointness relations take the form:

$$<dF_\pm, \mathbf{G}> - <F_\pm, \delta\mathbf{G}> = \int_{\partial c_\pm} F \wedge \#\mathbf{G},$$
$$<\delta\mathfrak{h}_\pm, \alpha> - <\mathfrak{h}_\pm, d\alpha> = \int_{\partial c_\pm} \#\mathfrak{h} \wedge \alpha.$$

Since $F_\pm$ and $\mathfrak{h}_\pm$ are smooth solutions of the sourceless pre-metric Maxwell equations, the left-hand sides vanish. As for the right-hand sides, by addition, we see that for every

smooth bivector field **G** of compact support and every smooth 1-form α of compact support one must have:

$$0 = \int_{\partial c_-} F_- \wedge \#\mathbf{G} + \int_{\partial c_+} F_+ \wedge \#\mathbf{G} = \int_{\partial c_+} [F] \wedge \#\mathbf{G},$$
$$0 = \int_{\partial c_-} \#\mathfrak{h}_- \wedge \alpha - \int_{\partial c_+} \#\mathfrak{h}_- \wedge \alpha = \int_{\partial c_+} \#[\mathfrak{h}] \wedge \alpha .$$

In particular, they are true when $\#\mathbf{G} = \alpha = d\phi$.

Since they are also true for any choice of c, one has:

$$d\phi \wedge [F] = 0, \qquad [\#\mathfrak{h}] \wedge d\phi = 0, \qquad [\mathfrak{h}] = C([F]),$$

which is equivalent to (VIII.73).

By solving the first equation as $[F] = d\phi \wedge [A]$, where $[A]$ is a 1-form on the hypersurface, one can combine the equations into:

$$(i_{d\phi} \cdot C \cdot e_{d\phi})[A] = 0,$$

which is the same as:

$$\sigma[\Box_C, d\phi][A] = 0.$$

The condition for this to admit a non-trivial solution $[A]$ is the vanishing of $\det(\sigma[\Box_C, d\phi])$, which gives (VIII.74).

3. Examples of dispersion laws. In order to derive specific dispersion laws for specific constitutive laws, one finds that it is often easiest to start with the block matrix (VIII.67) and evaluate the determinant when one restricts the submatrices of the constitutive law accordingly.

a. Isotropic media. In the case of an *isotropic* medium, for which $\varepsilon_{ij} = \varepsilon(x)\delta_{ij}$, $\mu_{ij} = \mu(x)\delta_{ij}$, $\gamma^i_j = \hat{\gamma}^j_i = 0$, one finds that (VIII.67) takes the form:

$$\sigma^{\mu\nu}[\Box_C, k] = \left[\begin{array}{c|c} \varepsilon k^2 & -\varepsilon\omega k^i \\ \hline -\varepsilon\omega k^j & \varepsilon\omega^2\delta^{ij} - (1/\mu)\delta_{mn}ad(k)^{im}ad(k)^{nj} \end{array}\right]$$
$$= \varepsilon\left[\begin{array}{c|c} \kappa^2 & -\omega k^i \\ \hline -\omega k^j & \omega^2\delta^{ij} - (1/\varepsilon\mu)k^i k^j \end{array}\right], \qquad \text{(VIII.75)}$$

in which $\kappa^2 = \delta^{ij} k_i k_j$ takes the form of the square of the Euclidian spatial norm of the spatial wave number covector $k_i\, dx^i$.

Taking the determinant gives:

$$\det(\sigma^{\mu\nu}[\Box_C, k]) = \varepsilon^4 (\omega^2 - c^2\kappa^2)^2, \qquad \text{(VIII.76)}$$

in which $c^2 = 1/\varepsilon\mu$ then becomes the speed of propagation.

The vanishing of the determinant then reduces to the vanishing of a quadratic polynomial – viz.:

$$g^{\mu\nu} k_\mu k_\nu = \omega^2 - c^2 \delta^{ij} k_i k_j , \tag{VIII.86}$$

which involves the constitutive properties of the spacetime only by way of c.

Since the resulting quadratic form on k that this equation defines has a normal hyperbolic signature type (+1, –1, …, –1), one sees that this is where the light cones finally originate, as well as the unit proper time hyperboloids and mass shells.

b. Anisotropic optical media - Fresnel analysis. In traditional geometrical optics [**3-6**], one deals with electromagnetic constitutive laws κ of a very particular form, as we discussed in Chapter V. Mostly, one assumes that the γ and $\hat{\gamma}$ matrices vanish, while the magnetic properties of the medium are linear, homogeneous, and isotropic, so $\mu_{ij} = \mu\delta_{ij}$, and the dielectric tensor ε_{ij} is symmetric. One can then reduce the matrix (VIII.67) to:

$$\sigma^{\mu\nu}[\Box_C, k] = \left[\begin{array}{c|c} \varepsilon(k,k) & -\omega\varepsilon^{im}k_m \\ \hline -\omega\varepsilon^{jm}k_m & \omega^2\varepsilon^{ij} - ad(k)^i_m\, ad(k)^{mj} \end{array}\right] \tag{VIII.87}$$

Next, one factors out the ω^2, while keeping in mind that $n_i = k_i/\omega$, and obtains:

$$\sigma^{\mu\nu}[\Box_C, k] = \left[\begin{array}{c|c} \varepsilon(n,n) & -\varepsilon^{im}n_m \\ \hline -\varepsilon^{jm}n_m & \varepsilon^{ij} - ad(n)^i_m\, ad(n)^{mj} \end{array}\right] \tag{VIII.88}$$

In the usual formulation, the matrix that one obtains from the algebraic form of the Maxwell equations – viz., $\mathbf{v} = c\mathbf{n} \times \mathbf{u}$, $\mathcal{E}(\mathbf{u}) = -c\mathbf{n} \times \mathbf{v}$ – is:

$$\sigma^{ij} = \varepsilon^{ij} - (n^2\delta^{ij} - n^i n^j) \tag{VIII.89}$$

However, we see that this matrix is precisely the spatial sub-matrix in (VIII.87).

We can then think of the polynomial in the components of the covector:

$$\Phi[n] = \det \sigma^{ij} \tag{VIII.90}$$

as a polynomial in the inhomogeneous coordinates of $\mathbb{R}\mathrm{P}^{3*}$, which we call the *Fresnel polynomial*.

If one further considers ε^{ij} in its principal frame, which then gives ε^{ij} the form diag[ε_x, ε_y, ε_z], then when one takes the determinant of σ^{ij} one obtains a characteristic equation in the form:

$$\begin{aligned} 0 = \Phi[n] = {} & n^2[\varepsilon^x n_x^2 + \varepsilon^y n_y^2 + \varepsilon^y n_y^2] \\ & -[n_x^2\varepsilon^x(\varepsilon^y + \varepsilon^z) + n_y^2\varepsilon^y(\varepsilon^x + \varepsilon^z) + n_z^2\varepsilon^z(\varepsilon^x + \varepsilon^y)] + \varepsilon^x\varepsilon^y\varepsilon^z , \end{aligned} \tag{VIII.91}$$

which can be put into the more elegant forms:

$$\frac{n_x^2}{n^2-\varepsilon^x}+\frac{n_y^2}{n^2-\varepsilon^y}+\frac{n_z^2}{n^2-\varepsilon^z}=\frac{1}{n^2} \tag{VIII.92}$$

or:

$$\frac{n_x^2}{\frac{1}{v_p^2}-\frac{1}{v_x^2}}+\frac{n_y^2}{\frac{1}{v_p^2}-\frac{1}{v_y^2}}+\frac{n_z^2}{\frac{1}{v_p^2}-\frac{1}{v_z^2}}, \tag{VIII.93}$$

in which $v_p = \omega/\kappa = 1/n$ is the phase velocity of the wave whose normal is n_i and we have introduced the principal velocities $v^i = 1/\sqrt{\varepsilon^i}$ of the medium.

However, the form (VIII.91) makes it more explicit that one is dealing with a quartic polynomial in the inhomogeneous coordinates n_i of the wave normal n.

The quartic hypersurface in $\mathbb{R}P^{3*}$ that is defined by either (VIII.91), (VIII.92), or (VIII.93) is called the *Fresnel normal hypersurface.*

The general case of ε^x, ε^y, ε^z all distinct corresponds to the case of a *biaxial* medium. If we assume, without loss of generality that $\varepsilon^x < \varepsilon^y < \varepsilon^z$ then one octant of the Fresnel normal hypersurface can be depicted as in Fig. 8, along with its intersections with the coordinate planes.

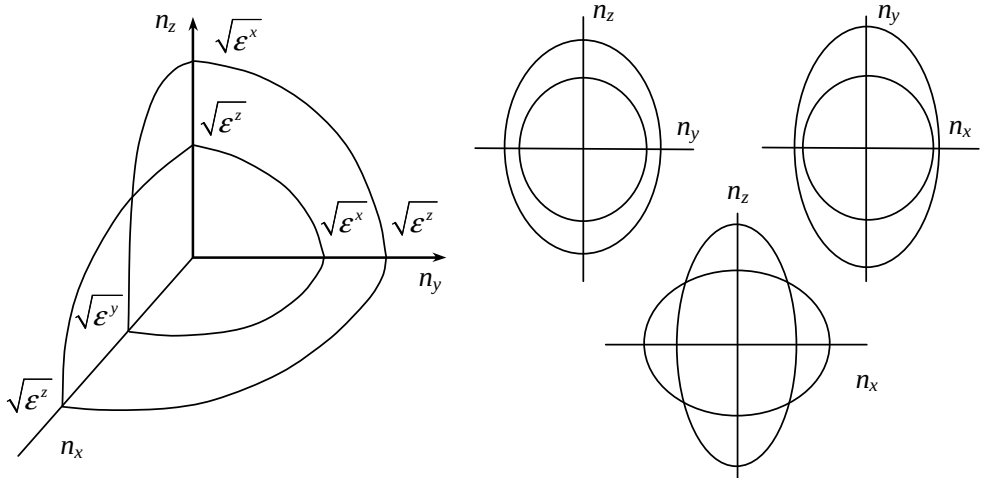

Figure 8. The general Fresnel quartic (biaxial media).

As one can see from the figure, it is not unheard of for the Fresnel quartic to be self-intersecting and the existence of such a singularity corresponds to the possibility of *conical refraction*. That is, one incoming ray can produce a cone of outgoing ones.

A possible property of $P[k]$ that is of considerable interest in electromagnetism is *birefringence*. In optics, this is associated with a certain type of anisotropy that is found in "uniaxial" media (see Landau, et al. [**3**]). For such media, two of the principal values of ε^{ij} – say, ε^x and ε^y – are equal. One refers to the two equal values as the *ordinary* values, while the third one is the *extraordinary* value, and one denotes them by ε^o and ε^e, respectively.

Birefringence refers to the fact that when $P[k_\mu]$ is quartic, if one fixes the spatial components k_i of k then the remaining polynomial $P[\omega, k_i]$ is quadratic in ω^2, and its roots can be shown to be real. This then implies that for any spatial direction of propagation there will generally be two distinct positive values of ω, and therefore two distinct values of the phase velocity ω/κ, where $\kappa^2 = \delta^{ij} k_i\, k_j$. This leads to double refraction of a given light ray. It is customary to refer to the resulting waves of the pair as the *ordinary* and *extraordinary* waves. This phenomenon can be observed by placing a slab of calcite over a page of print, which produces double images of the letters. The fainter image is then due to the extraordinary waves.

In terms of the Fresnel quartic, the effect of a uniaxial dielectic is to cause the quartic polynomial to factor into a product of quadratic ones:

$$(n^2 - \varepsilon^o)[\varepsilon^e n_z^2 + \varepsilon^0 (n_x^2 + n_y^2) - \varepsilon^0 \varepsilon^e] = 0. \qquad \text{(VIII.94)}$$

The factorization of the polynomial then implies that the quartic consists of the union of a sphere and an ellipsoid that intersect at their North and South poles. The sphere is indicative of an isotropic medium, so one sees why the corresponding principal value of ε^{ij} would be referred to as ordinary.

There are two possibilities for the extraordinary ellipsoid corresponding to whether it is outside or inside the ordinary sphere. In the former case, one says that the medium is *positive*, while in the latter case one refers to it as *negative*. We depict these two possibilities in Fig. 9 by means of the x–z sections of the surfaces.

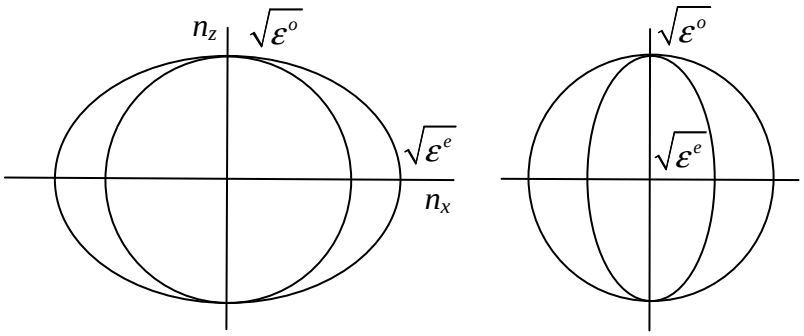

Figure 9. Fresnel quartic (uniaxial media).

b. Skewon contributions [**16**]. Although a purely skewonic medium would give a vanishing dispersion polynomial, which would imply the absence of wave modes, nonetheless, as a contribution to the principal part of a constitutive tensor field, it can have an effect.

According to the analysis that was given to the subject in [**16**, loc. cit.], one must distinguish between real and imaginary skewon contributions. Furthermore, one can classify skewons in terms of a basic decomposition as having "electric Faraday,"

"magnetic Faraday," and "magneto-electric optical activity" type. (see, loc. cit. for all definitions involved).

A real skewon contribution amounts to a source of resistivity and energy absorption for waves. For a skewon of "electric Faraday" type, the Fresnel wave surface can take on the character of a torus. That is, it has changed its topology, as well as its shape. When a real skewon has the "magneto-electric optical activity" type, the Fresnel surface can become two intersecting torii.

In the case of imaginary skewon contributions, one has to distinguish not only types of skewons, but also the ones that are "small," "medium," and "large" in magnitude when compared to the principal part. Generally, they are associated with natural optical activity and the Faraday effects. For the large ones of "magneto-electric optical activity type," the Fresnel surface will become merely an oblate spheroid, while for small ones it will be two concentric spheroids. For those of "magnetic Faraday" type, the large ones give intersecting hyperboloids, the small ones give intersecting spheroids, and in between one finds intersecting paraboloids.

c. Axion contributions. A direct verification using (VIII.67) shows that a axion part to a constitutive law does not contribute to the dispersion polynomial. That is, it does not affect the propagation of waves *in the geometrical optics approximation.* Perhaps that is why sometimes the vanishing of the axion part is a basic axion of electromagnetic constitutive laws that is referred to as the "Post constraint," since it was Post [**17**] who advocated that axiom.

However, as Itin [**18**] observes, if one goes beyond that approximation and considers waves for which the amplitude function is not effectively constant then the first derivatives of that amplitude can couple to the axionic part in a non-trivial way. In particular, one finds that dispersion polynomial is no longer a homogeneous polynomial in the wave covector, but an inhomogeneous one.

d. Bi-metric media. There often exists a factorization of the quartic polynomial $P[k]$ into a product of quadratic polynomials (which is called *bi-metricity* by Barcello, Liberati, and Visser [**19**]):

$$P[k] = g(k, k)\,\bar{g}(k, k)\,. \tag{VIII.95}$$

Hence, the Tamm-Rubilar tensor field that is associated with $P[k]$ takes the form of symmetrized tensor product of two Lorentzian metrics on $T^*(M)$:

$$P = g \odot \tilde{g}\,. \tag{VIII.96}$$

The components of P are then obtained from those of g and $\tilde{g}$ by way of:

$$P^{\kappa\lambda\mu\nu} = \tfrac{1}{6}(g^{\kappa\lambda}\tilde{g}^{\mu\nu} + g^{\mu\lambda}\tilde{g}^{\kappa\nu} + g^{\nu\lambda}\tilde{g}^{\mu\kappa} + g^{\kappa\mu}\tilde{g}^{\lambda\nu} + g^{\kappa\nu}\tilde{g}^{\mu\lambda} + g^{\mu\nu}\tilde{g}^{\kappa\lambda})\,. \tag{VIII.97}$$

Although fourth-degree polynomials in more than one real variable do not always have to factorize into products of quadratics, nonetheless, in the case of electromagnetic waves, it is widely known that as long as the Lagrangian for the electromagnetic field F

depends only upon the Lorentz invariants $F \wedge F$ and $F \wedge {}^*F$, the characteristic polynomial *will* factorize, even in the nonlinear case. (See [**15**].)

e. One-loop effective vacua. From the analysis that was given in Barcello, Liberati, and Visser [**19**], one sees that the dispersion laws for both the Heisenberg-Euler and Born-Infeld media take the general bimetric form:

$$(\eta_{\mu\nu} + \varepsilon_1 T_{\mu\nu})(\eta_{\mu\nu} + \varepsilon_2 T_{\mu\nu}) = 0. \qquad \text{(VIII.98)}$$

In this equation, the components $T_{\mu\nu}$ refer to the Faraday stress-energy-momentum tensor field for the background electromagnetic field, which we will discuss more thoroughly in Chapter X, while the scalar factors ε_1 and ε_2 depend upon the electromagnetic field strengths and would generally have units of 1/energy density. Hence, in this case the geometry of spacetime is perturbed quite directly by the presence of a background electromagnetic field.

f. Plasmas. In plasmas, not only the constitutive laws can vary, but so can the field equations themselves, depending upon what degree of interaction (usually collisions) one is assuming. Furthermore, plasmas are capable of propagating not only electromagnetic waves, but also mechanical – i.e., acoustic – ones. Rather than derive the dispersion laws, which would take us quite far afield, we simply summarize some of them that one deals with in the linear case (cf., [**20**]).

In general, waves in plasmas can be either acoustic or electromagnetic, and can be concerned with the basic oscillations of either the electrons or the ions. Furthermore, as we already saw in Chapter V, the presence or absence of a background magnetic field $\mathbf{B}_0$ can affect whether the dielectric – hence, optical – properties of the plasma are isotropic or anisotropic. They can also exhibit linear or nonlinear wave behavior, depending upon their amplitudes, even though often the actual number densities or temperatures involved can be relatively casual.

Electron waves can be electrostatic or electromagnetic and the dispersion laws for electrostatic electron waves are:

$$\omega^2 - 3/2 v_{th}^2 k^2 = \omega_p^2, \qquad (\mathbf{B}_0 = 0 \text{ or } \mathbf{k} \parallel \mathbf{B}_0) \qquad \text{(VIII.99a)}$$

$$\omega^2 = \omega_h^2 \equiv \omega_p^2 + \omega_c^2 \qquad (\mathbf{k} \perp \mathbf{B}_0) \qquad \text{(VIII.99b)}$$

respectively. In these expressions, $v_{th} = (2KT_e/m)^{1/2}$ is the thermal velocity of the electrons, while $\omega_p = (n_0 e^2/\varepsilon_0 m)^{1/2}$ is the *plasma frequency* of the electrons when their equilibrium number density is n_0, and ω_h is called the *upper hybrid frequency.* The plasma oscillations come about due to the fact that the equilibrium configuration of the electrons is stable, so any perturbation of an electron from equilibrium will produce a counter-electric field that tends to restore the equilibrium, bur produces a characteristic oscillation about the equilibrium configuration.

Note that the second relation (VIII.99b) does not give an actual dispersion law, so no traveling wave actually propagates, only a stationary oscillation.

For electromagnetic electron waves, one has:

$$\omega^2 - c^2 k^2 = \omega_p^2, \qquad (\mathbf{B}_0 = 0) \qquad \text{(VIII.100a)}$$

$$= \omega_p^2, \qquad (\mathbf{k} \perp \mathbf{B}_0, \mathbf{E}_1 \parallel \mathbf{B}_0) \qquad \text{(VIII.100b)}$$

$$= \omega_p^2 \left(\frac{\omega^2 - \omega_p^2}{\omega^2 - \omega_h^2} \right), \quad (\mathbf{k} \perp \mathbf{B}_0, \mathbf{E}_1 \perp \mathbf{B}_0) \qquad \text{(VIII.100c)}$$

$$= \omega_p^2 \left(\frac{\omega}{\omega - \omega_c} \right), \qquad (\mathbf{k} \parallel \mathbf{B}_0) \qquad \text{(VIII.100d)}$$

$$= \omega_p^2 \left(\frac{\omega}{\omega + \omega_c} \right), \qquad (\mathbf{k} \parallel \mathbf{B}_0), \qquad \text{(VIII.100e)}$$

in which $\omega_c = eB_0/m$ is the cyclotron frequency of the electron in the magnetic field.

One sees that in the absence of a background magnetic field, the dispersion law has the same "Klein-Gordon" form as in the electrostatic case, but with a different speed of propagation, and in which the plasma frequency plays a role that is analogous to the Compton frequency of a massive wave in relativistic quantum wave mechanics. In particular, the electromagnetic waves are not obtained from the vanishing of the characteristic polynomial – i.e., its zero locus – but from a non-zero locus. Whether this analogy between plasma waves and matter waves proves to be a useful tool in understanding quantum physics clearly deserves more attention.

The presence of a background field, the dielectric tensor becomes anisotropic, which leads to a bi-metric situation. When the background magnetic field is perpendicular to the wave vector, a wave that obeys (VIII.100b) is referred to as an *ordinary* wave, while (VIII.100c) describes the *extraordinary* wave, although the convention is reversed from the usual optical terminology. When $\mathbf{B}_0$ is parallel to the wave vector, one has two types of waves: (VIII.100d) describes the *whistler* mode and (VIII.100e) describes *L* waves.

Ion waves also come in the two types according to electrostatic and electromagnetic, and these can be further classified according to the relationship between the wave vector **k** and the background magnetic field.

For the electrostatic waves, one has:

$$\omega^2 - v_s^2 k^2 = 0, \qquad (\mathbf{B}_0 = 0 \text{ or } \mathbf{k} \parallel \mathbf{B}_0) \qquad \text{(VIII.101a)}$$

$$\omega^2 - v_s^2 k^2 = \Omega_c^2, \qquad (\mathbf{k} \perp \mathbf{B}_0) \qquad \text{(VIII.101b)}$$

$$\omega^2 = \omega_l^2. \qquad \text{(VIII.101c)}$$

Now, the speed $v_s = [(\gamma_e KT_e + \gamma_i KT_i)/M]^{1/2}$ is the speed of sound in the plasma, when the number fraction of the electrons is γ_e and that of the ions is γ_i, since the first type of wave is acoustic in character. $\Omega_c = eB_0/M$ is the cyclotron frequency of the ions in the magnetic field, and one calls the waves described by (VIII.101b) *electrostatic ion cyclotron waves.* The last law (VIII.101c) does not describe a wave, but only a state of oscillation at the *lower hybrid frequency* $\omega_l = (\Omega_c \omega_c)^{1/2}$.

As for electromagnetic ion waves, one first finds that there are no such things in the absence of $\mathbf{B}_0$, but when it is non-vanishing, there are two types:

$$\omega^2 - v_A^2 k^2 = 0, \qquad (\mathbf{k} \parallel \mathbf{B}_0) \qquad \text{(VIII.102a)}$$

$$\omega^2 - c^2 \left(\frac{v_s^2 + v_A^2}{c^2 + v_A^2} \right) k^2 = 0. \qquad (\mathbf{k} \perp \mathbf{B}_0) \qquad \text{(VIII.102b)}$$

The first ones are called *Alfvén waves*, which are hydromagnetic in character and $v_A = B_0 / \sqrt{\mu_0 n_0 M}$ is then their speed of propagation, while the second ones are called *magnetosonic waves*, and one sees that their speed of propagation is more involved.

One notes the following recurring aspects of these dispersion laws:

1. The ubiquitous role that seems to be played by laws of the Klein-Gordon type. $\omega^2 - v^2 k^2 = \omega_0^2$ for suitable choices of the parameters v and ω_0.

2. The fact that longitudinal electromagnetic wave modes are possible in plasmas, as well as transverse ones.

These dispersion laws define certain characteristic frequencies namely *cutoffs* and *resonances*. A cutoff frequency is defined by a frequency at which k (and therefore the index of refraction $n = k/\omega$, as we shall see)) vanishes. This is equivalent to saying that the phase velocity v_p goes infinite. Conversely, a resonance is a frequency at which either k or n goes infinite (i.e., v_p goes to zero).

In the case of the extraordinary wave (VIII.100c), one has a resonance at the hybrid frequency and the cutoff frequencies are the roots of:

$$\omega^2 \mp \omega_c \omega - \omega_p^2 = 0, \qquad \text{(VIII.103)}$$

namely:

$$\omega_R = \tfrac{1}{2}\left[\omega_c + \sqrt{\omega_c^2 + 4\omega_p^2} \right], \qquad \omega_L = \tfrac{1}{2}\left[-\omega_c + \sqrt{\omega_c^2 + 4\omega_p^2} \right]. \qquad \text{(VIII.104)}$$

The propagation of transverse electromagnetic waves in the Earth's ionosphere is limited by a cutoff frequency on the order of 10 MHz. Hence, radio waves of lower frequency will not penetrate the ionosphere, but only reflect back. Although this makes them useless as a means of communicating with spacecraft outside the ionosphere, nonetheless, it makes it possible for shortwave radio transmissions of low power to communicate with stations over the Earth's horizon from the transmitting antenna by means of successive reflections.

An important aspect of plasma oscillations is that they are damped by what amounts to the interaction of the plasma particles with passing waves. In effect, there is a resonance in this coupling when the particles have a velocity equal to the phase velocity $v_\phi = \omega/k$ of the wave. When the velocities v are far from v_ϕ, there is essentially no energy exchanged between the wave and the particle. As v approaches v_ϕ from below, a particle gains energy from the wave and as it approaches v_ϕ from above it loses energy to the wave. When the velocities agree there is no energy lost or gained and the particle is pushed along by the wave like a surfboard.

If one assumes that the number density function $f(x, v)$ for the velocities of the particles is Maxwellian then this tends to favor slow particles, which implies a damping of the wave itself due to the energy that it is losing to the particles and this damping,

which is not associated with any actual collisions between particles, is called *Landau damping*.

If one linearizes the Vlasov equation for *f*, which is the form that the Boltzmann equation takes for plasma dynamics, around a Maxwellian equilibrium distribution $f_0(x, v)$ then the resulting dispersion law for plasma oscillations is:

$$\omega(k) = \omega_p \left(1 + i \frac{\pi}{2} \frac{\omega_p^2}{k^2} \left[\frac{\partial f_0}{\partial v} \right]_{v=v_\phi} \right). \qquad \text{(VIII.105)}$$

The presence of damping in this expression derives from the fact that the imaginary part of $\omega(k)$ is negative.

Nonlinear wave phenomena are almost too numerous to mention. We briefly mention some of them just to give an impression of how wide-ranging they are.

1. One finds that "drift waves," in plasmas, whose amplitudes should grow exponentially in time, actually seem to reach a saturation limit, which is a symptom of nonlinearity in the dynamics of the wave.

2. Another common symptom of nonlinearity in the dynamics of waves is the changing of the shape of waves over time, although sometimes that can be accounted for by dispersion in a linear model. For instance, the breaking of surface wave on the ocean as it approaches the shore is due to the fact that the speed of propagation depends upon the depth of the water, which is different for the leading and trailing edges.

3. Waves in fluid media of sufficient amplitude often give rise to turbulence, and plasma waves are no exception.

4. The Landau damping that first appears in the linear approximation for the diffusion equation (viz., the Vlasov equation) eventually takes on a nonlinear form when one gives the Vlasov equation its full nonlinear treatment.

5. One has interactions between the plasma waves and the particles of the plasmas, such as particle trapping and plasma echoes, as well as interactions between the waves, which are somewhat analogous to photon-photon scattering in quantum electrodynamics.

6. One finds that the same nonlinear Schrödinger equation that played a role in nonlinear optics also plays a role in nonlinear plasma waves, while the Koortweg-deVries (KdV) equations, which originally described one-dimensional waves in shallow water, also appears.

4. Speed of wave propagation. Although the speed *c* of electromagnetic wave propagation in vacuo is treated as a "fundamental constant" in the eyes of special relativity, as well as quantum electrodynamics, nonetheless, the very fact that vacuum polarization seems to be fundamental to almost all quantum electrodynamical effects suggests that it is better to regard *c* as a *derived* constant, namely:

$$c = 1/\sqrt{\varepsilon_0 \mu_0}\ , \qquad \text{(VIII.106)}$$

and to recognize that the constancy of *c* would only be a consequence of the constancy of ε_0 and μ_0, or at least their product.

When one takes vacuum polarization into account, one must consider the possibility that both of these vacuum parameters are functions of the electromagnetic field strengths that are present in the region of space under consideration. Hence, one might recover the classical definition of c as an asymptotic limit in the absence of fields:

$$c = \lim_{F\to 0} \frac{1}{\sqrt{\varepsilon_0(F)\mu_0(F)}}. \qquad \text{(VIII.107)}$$

In full generality, however, since the parameters ε_0 and μ_0 are part of the constitutive law of the medium in question we should like to regard the speed of propagation of waves in a given medium as being a property of the medium that is derived from more fundamental assumptions about that constitutive law. In fact, it is only in the case of isotropic media that one can give any meaning to the notion of a unique speed of propagation at each point, and only with the further restriction of homogeneity that one can speak of constancy. In the general inhomogeneous, anisotropic medium the speed of propagation depends upon both position and direction, as well as possibly time and non-local considerations.

In order to derive the speed of propagation of electromagnetic waves in a medium from the constitutive law, one first has to recognize that since there is more than one way of defining a "wave" in the first place there is also more than one way of defining its speed of propagation [1]. The two that we shall consider are the phase velocity and the group velocity.

First, assume that the tangent and cotangent bundle of the spacetime manifold have been given a specific choice of space-time splitting. The *phase velocity* of propagation of a wave is a spatial vector field that is associated with the wave covector $k = \omega dt - k_i\, dx^i$, namely:

$$\mathbf{v}_p = (v_p^1, v_p^2, v_p^3), \qquad v_p^i = \frac{\omega}{k_i}, \qquad i = 1, 2, 3. \qquad \text{(VIII.108)}$$

Hence, $\mathbf{v}_p$ depends only upon k and the choice of space-time splitting, and not on the constitutive law. This basically accounts for the choice of the term "phase" in the definition, since it effectively amounts to the velocity that is associated with the isophase foliation defined by k itself when one chooses a space-time splitting.

One must note that there is something geometrically unnatural about taking the inverses of the components of vectors, since the operation is certainly not invariant under changes of frames. However, if we form the triple of components:

$$n = (n_1, n_2, n_3), \qquad n_i = -\frac{k_i}{\omega} \qquad \text{(VIII.109)}$$

then we see that the map from $(\omega, -k_i)$ to (n_1, n_2, n_3) makes perfect *projective* geometrical sense as the projection of the homogeneous coordinates for a chart on the

[1] For a thorough treatment of the other definitions that we do not use, see the book by Brillouin [**21**].

projective space $\mathbb{R}P^3$ onto the inhomogeneous coordinates, as we pointed out in chapter II. Hence, the *normal covector n* at each point $x \in M$ that is so defined is really the line [*k*] through the origin in T_x^*M that is generated by the covector *k*. It is important to note that this line is actually defined independently of any space-time splitting. Hence, the three-dimensional "rest space" that is most appropriate to geometrical optics is really a projective space, not an affine one. One then considers the notion of the *projectivized cotangent bundle PT*M*, whose fibers are the projective spaces PT_x^*M obtained from the cotangent spaces T_x^*M .

In order to define the *group velocity*, one must also consider the dispersion law *P*[*k*] = const. that follows from the constitutive law and field equations. Although in elementary physics, one usually assumes that the dispersion law has been solved for ω as a function of k_i , so one can define:

$$v_g^i = \frac{\partial \omega}{\partial k_i}, \tag{VIII.110}$$

since this implies that $\partial P/\partial \omega$ is non-vanishing, one can also define:

$$v_g^i = -\frac{\partial P / \partial k_i}{\partial P / \partial \omega}. \tag{VIII.111}$$

This definition has an interpretation in the language of projective geometry that is analogous to the previous interpretation of the normal covector. As we shall see below, the partial derivatives in the quotient are all components of a velocity vector field **v** on T^*M that represents part of the characteristic vector field X_P that is associated with *P*[*k*]:

$$v^\mu = \frac{1}{4}\frac{\partial P}{\partial k_\mu}, \tag{VIII.112}$$

which makes:

$$v_g^i = -\frac{v^i}{v^0}. \tag{VIII.113}$$

Hence, we see that the triple of group velocity components (v_g^1, v_g^2, v_g^3) is better regarded as the inhomogeneous coordinates of a line [**v**] in the *projectivized tangent bundle PT*(*M*), whose fibers are then the projective spaces PT_xM associated with all lines through the origins of the tangent spaces, than as a section of the tangent bundle itself. Since we shall have much more to say about the role of projective geometry in spacetime structure in chapter XI, we suspend our discussion on that point.

We shall point out that there is a duality defined by *P*[*k*] that links *k* with **v**, such that the dispersion law *P*[*k*] = 0 becomes *k*(**v**) = 0, while the image of this in the projective spaces is $n(\mathbf{v}_g) = 1$. Indeed, this sort of duality depends only upon the assumption of the homogeneity of the function *P*[*k*] and the invertibility of its Hessian, and not the assumption that it is a quadratic polynomial, as well.

Naturally, it is illuminating to see how the foregoing constructions work in the familiar case of a quadratic – i.e., Lorentzian – dispersion law:

$$P[k] = \tfrac{1}{2}(\omega^2 - c^2 k^2) = \tfrac{1}{2}\omega_0^2, \qquad k^2 = g^{ij} k_i k_j, \tag{VIII.114}$$

in which ω_0 is a constant that may or may not be zero and c has its usual meaning, although it is mostly being used to convert the spatial dimension to the time dimension.

This makes:

$$v^0 = \omega, \qquad v^i = -c^2 g^{ij} k_j, \tag{VIII.115}$$

so:

$$v_g^i = \frac{c^2 k^i}{\omega}, \tag{VIII.116}$$

and in the Euclidian case, in which $k_i = k^i$, one has:

$$v_p^i \, v_g^i = c^2. \tag{VIII.117}$$

One thus sees that generally the phase velocity and the group velocity are quite distinct from each other.

Let us compare the two velocities that one obtains in the lightlike case of $\omega_0 = 0$ with the one that one obtains in the timelike case in which $\omega_0 > 0$. In the former case, one can say that the dispersion law is simply:

$$\omega(k) = ck. \tag{VIII.118}$$

This makes:

$$v_p^i = \frac{ck}{k_i}, \qquad v_g^i = \frac{ck^i}{k}. \tag{VIII.119}$$

When one computes the spatial norms of these vectors relative to g_{ij}, which is inverse to g^{ij}, one gets:

$$v_p = v_g = c. \tag{VIII.120}$$

Thus, in this case either the phase velocity or the group velocity represents the most commonly-used way of referring to the speed of propagation of electromagnetic waves.

One also verifies that:

$$n_i v_g^i = \frac{k_i}{\omega}\frac{c^2 k^i}{\omega} = \left(\frac{ck}{\omega}\right)^2 = 1. \tag{VIII.121}$$

When $\omega_0 > 0$, one can rewrite the dispersion law as:

$$\omega = ck\sqrt{1 + \left(\frac{\omega_0}{ck}\right)^2}. \tag{VIII.122}$$

This dispersion law approaches the lightlike one as *k* grows indefinitely large.

This time, we get:

$$v_p^i = \frac{ck}{k_i}\sqrt{1+\left(\frac{\omega_0}{ck}\right)^2}\,, \quad v_g^i = \frac{ck^i}{k\sqrt{1+(\omega_0/ck)^2}}\,, \tag{VIII.123}$$

so:

$$v_p = c\sqrt{1+\left(\frac{\omega_0}{ck}\right)^2} = \frac{\omega}{k}\,, \qquad v_g = \frac{c}{\sqrt{1+(\omega_0/ck)^2}}\,. \tag{VIII.124}$$

From the facts that v_g goes to zero as *k* goes to zero and that it goes to *c* as *k* grows indefinitely large, we then see that the group velocity behaves more like the conventional velocity of a massive particle than the phase velocity, which grows indefinitely large as *k* goes to zero.

An interesting consequence of the first of these equations is that:

$$\sqrt{1-\frac{v_g^2}{c^2}} = \frac{\omega_0}{\omega}\,, \tag{VIII.125}$$

which shows that the Fitzgerald-Lorentz factor in special relativity can just as well be regarded in terms of the wave covector as in terms of the phase velocity vector.

References

1. J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill, New York, 1941.
2. J. D. Jackson, *Classical Electrodynamics*, 2nd ed., Wiley, New York, 1976.
3. L.D. Landau, E.M. Lifschitz, and L.P. Pitaevskii, *Electrodynamics of Continuous Media*, 2nd ed., Pergamon, Oxford, 1984.
4. M. Kline and I. W. Kay, *Electromagnetic Theory and Geometrical Optics*, Wiley-Interscience, New York, 1965.
5. M. Born and E. Wolf, *Principles of Optics*, Pergamon, Oxford, 1980.
6. R. K. Luneburg, *Mathematical Theory of Optics*, The University of California Press, Berkeley, 1964.
7. D. Delphenich, "Proper Time Foliations of Lorentz Manifolds," arXiv.org preprint gr-qc/0211066.
8. J. Hadamard, *Leçons sur la propagation des ondes et les équations de l'hydrodynamique*, Chelsea, NY, 1949.
9. R. Courant and P. Lax, "The propagation of discontinuities in wave motion," Proc. Nat. Acad. Sci. **42** (1956), 872-876.
10. R. Lewis, "Discontinuous initial value problems for symmetric hyperbolic linear differential equations," J. Math. Mech. **7** (1958), 571-592.
11. H. Bremmer, "The jumps of discontinuous solutions of the wave equation," Comm. Pure Appl. Math. **4** (1951), 419-426.

12. E. T. Copson, "The transport of discontinuities in an electromagnetic field," Comm. Pure Appl. Math. **4** (1951), 427-433.
13. F. Hehl and Y. Obukhov, *Foundations of Classical Electrodynamics*, Birkhäuser, Boston, 2003.
14. I. M. Tamm, "Relativistic crystal optics and its relation to the geometry of a bi-quadratic form," Zhurn. Ross. Fiz.-Khim. Ob. **57**, no. 3-4 (1925), 209-224 (in Russian); also reprinted in *I. E. Tamm, Collected Papers* (Nauka, Moscow, 1975), vol. 1, pp. 33-61 (in Russian); "Electrodynamics of an anisotropic medium in special relativity theory," ibid., pp. 19-31; an abbreviated version of this article appeared in German as: L. Mandelstam and J. Tamm, Math. Ann. **95** (1925), 154-160.
15. Y. N. Obukhov and G. F. Rubilar, "Fresnel Analysis of the wave propagation in nonlinear electrodynamics," arXiv.org preprint gr-qc/0204028.
16. F. W. Hehl and Y. N. Obukhov, "On possible skewon effects on light propagation," arXiv.org preprint, physics/0409155, v. 2 (2008).
17. E. J. Post, *Formal Structure of Electromagnetics*, Dover, Mineola, N.Y., 1997.
18. Y. Itin, "Wave propagation in axion electrodynamics," Gen. Rel. Grav., **40** (2008), 1219-1238.
19. C. Barcelo, S. Liberati, and M. Visser, "Bi-refringence versus bi-metricity," arXiv.org preprint gr-qc/0204017.
20. F. F. Chen, *Introduction to Plasma Physics and Controlled Fusion, v. 1,"* 2nd ed., Plenum Press,
21. L. Brillouin, *Wave Propagation and Group Velocity*, Academic Press, New York, 1960.

CHAPTER IX

Geometrical optics

When one has obtained a dispersion polynomial $P[k]$, which is a differentiable function on the cotangent bundle T^*M to the spacetime manifold M, one can take advantage of its natural symplectic structure to obtain a characteristic vector field on T^*M and a bicharacteristic flow whose integral curves project to generalized null geodesics in M. One obtains supplementary contributions to the geodesic equations, in addition to the generalized Levi-Civita contribution that one obtains for quadratic dispersion polynomials, which are based in the non-quadratic nature of the dispersion law. Hence, if the origin of the non-quadratic nature of the dispersion law is vacuum polarization due to high field strengths then one is justified in thinking of the departure of the null geodesics from the curves that one obtains in the quadratic case as a form of "quantum fluctuations about classical extremals."

Furthermore, since the derivation of a dispersion law follows only from the geometrical optics approximation, the deduction of null geodesics as the bicharacteristics of the field equations, which is so fundamental to the geometry of spacetime in the eyes of general relativity, is nevertheless a high-frequency (short-wavelength) approximation to a more involved geometrical picture that pertains to wave mechanics. Hence, one must also regard the diffraction effects that geometrical optics overlooks as being another source of quantum fluctuations.

Since the role of projective geometry in pre-metric electromagnetism is so fundamental and pervasive, we shall return to address it more thoroughly in Chapter XII. However, since the topic appears naturally in the context of geometrical optics, we shall nonetheless mention it briefly before then. At the root of the introduction of projective geometry is the fact that light rays have no preferred parameterization as curves so one must think in terms of tangent lines to the null geodesics, not tangent vectors. Similar considerations also apply to the wave covector, namely, it defines the same tangent hyperplane as any other covector that differs by a non-zero scalar multiple. Hence, one must pass from the tangent and cotangent bundles to their projectivizations in order to obtain the three-dimensional "rest spaces" in which geometrical optics takes place, rather than the usual space-time decompositions that conventional relativity considers.

We must also address the geometric nature of Huygens's principle in the context of pre-metric electromagnetism. As it turns out, the basic shift in emphasis that is required is from the arc-length functional that one obtains from a metric to the elapsed-time functional that is defined by the wave covector and the dispersion law. One finds that the resulting propagation of phase– i.e., the envelope of a family of elementary hypersurfaces that all represent the same elapsed time along a null geodesic – has a natural interpretation in terms of contact geometry, while the usual propagation of amplitude by Green function techniques is only appropriate to linear wave equations. However, one can consider more general "transport equations" for the propagation of amplitude.

1. The generalized eikonal equation. So far, by means of the geometrical optics approximation, we have reduced the pre-metric Maxwell equations from a set of first-order partial differential equations in the time-varying electric and magnetic fields to an algebraic equation involving the wave covector $k = \omega\, dt - k_i\, dx^i$ and a first-order linear partial differential equation for the phase function ϕ.

$$P[k] = 0, \qquad k = d\phi. \tag{IX.1}$$

Hence, one can combine them to obtain the generalized eikonal equation in the form:

$$P[d\phi] = 0, \tag{IX.2}$$

which will be nonlinear first-order partial differential equation for ϕ.

In order to obtain the spatial form of the eikonal equation one must first observe that, from what we said above, the three-dimensional "space" in question is really the *projective* space $\mathbb{R}\mathrm{P}^{3*}$, in the form of the projective cotangent spaces, not the *vector* space $\mathbb{R}^3$, as represented by hyperplanes in the cotangent spaces. Hence, we must first look at the projection of the dispersion law $P[k] = 0$ onto PT^*M.

We accomplish this by factoring k into $\omega(dt - n)$ with $n = n_i\, dx^i$, $n_i = k_i / \omega$. Since $P[k]$ is a homogeneous quartic polynomial in k, this gives:

$$0 = \omega^4 P[n], \tag{IX.3}$$

with:

$$P[n] = P_0 + P_1[n] + P_2[n] + P_3[n] + P_4[n], \tag{IX.4}$$

in which $P_k[n]$ refers to a homogeneous polynomial in n of degree k.

However, since P is also quadratic in ω^2 one expects that the odd-degree polynomials will vanish, which leaves only:

$$0 = P_0 + P_2[n] + P_4[n], \tag{IX.5}$$

as a spatial dispersion law.

We must also replace the requirement that the spacetime 1-form k be exact with the requirement that spatial 1-form $n = n_i\, dx^i$ be exact:

$$n = d\theta, \tag{IX.6}$$

where $\theta(x^i)$ is then a spatial phase function on $\mathbb{R}\mathrm{P}^3$.

By combining the dispersion law $\Phi[n] = 0$ with this differential expression for the wave normal n one obtains:

$$0 = P_0 + P_2[d\theta] + P_4[d\theta] \tag{IX.7}$$

for the spatial form of the generalized eikonal equation.

This equation represents a nonlinear first order partial differential equation for the spatial phase function θ, which takes the form of the time-invariant Hamilton-Jacobi equation when one regards the dispersion polynomial $P[k]$ as a Hamiltonian function on the cotangent bundle of Σ.

It takes the component form:

$$0 = P^{0000} + 6P^{00ij}\frac{\partial\theta}{\partial x^i}\frac{\partial\theta}{\partial x^j} + P^{ijkl}\frac{\partial\theta}{\partial x^i}\frac{\partial\theta}{\partial x^j}\frac{\partial\theta}{\partial x^k}\frac{\partial\theta}{\partial x^l} \ . \qquad \text{(IX.8)}$$

At this point, it helps to look at the form that equation (IX.8) takes when P is the quadratic form η that defines Minkowski space:

$$0 = \eta(dt, dt) + 2\eta(dt, d\theta) + \eta(d\theta, d\theta), \qquad \text{(IX.9)}$$

which then reduces to:

$$0 = 1 - c^2\delta^{ij}\,\theta_{,i}\,\theta_{,j} \qquad \text{(IX.10)}$$

by orthogonality.

Finally, this gives the partial differential equation for θ:

$$\delta^{ij}\frac{\partial\theta}{\partial x^i}\frac{\partial\theta}{\partial x^j} = n^2. \qquad \text{(IX.11)}$$

Of course, (IX.8) is still considerably more complicated than (IX.11), so we consider the fact that for a broad class of electromagnetic constitutive laws, including most of the popular nonlinear ones, the dispersion relation factorizes into a product of quadratic relations, as in (VIII.72). We see that it is not necessary to formulate a single quartic eikonal equation because it is sufficient to note that a polynomial $P[k]$ of the form (VIII.72) vanishes iff either $g(k, k)$ or $\bar{g}(k,k)$ vanishes, independently of each other. Hence, one simply obtains the union of the solution sets for quadratic eikonal equations of the form (IX.7), when δ^{ij} is replaced with either $-\,g^{ij}$ or $-\bar{g}^{ij}$, and there will be different functions for n^2.

A subtle point that one must consider in the foregoing is that since $n = 1/\omega k_s$, with $k_s = \phi_{,i}\,dx^i$, unless one assumes that ω is spatially homogeneous, one has no guarantee that the resulting 1-form is exact, even when k_s is. A necessary integrability condition for this is that n must be closed, which gives:

$$dk_s = d\omega \wedge n = d\ln\omega \wedge k_s \ . \qquad \text{(IX.12)}$$

From Frobenius, this is also the condition for the exterior differential system $k_s = 0$ to be completely integrable.

2. Bicharacteristics – null geodesics. Although the generalized eikonal equation (IX.2) is usually nonlinear, due to the polynomial character of the dispersion law, nevertheless, it is still a first-order partial differential equation in a single function ϕ.

Hence, one can solve it by Cauchy's method of characteristics. However, since we are already using the word "characteristic" in the context of the second-order system of partial differential equations in the potential 1-form A as a way of reducing it to a first-order equation, it is traditional to refer to the characteristic curves of the eikonal equation as the *bicharacteristics* of the original second-order system.

Consider the case of electromagnetic waves whose dispersion law is (VIII.69), which we rewrite in the form:

$$P(x)[k] = 0. \tag{IX.13}$$

The bicharacteristic equations are simply the Hamilton equations that one obtains by treating $F = 1/4\ P$ as a Hamiltonian:

$$\frac{dx^\mu}{d\tau} = \frac{\partial F}{\partial k^\mu} = \gamma^{\mu\nu}(x,\, k)\, k_\nu, \tag{IX.14a}$$

$$\frac{dk_\mu}{d\tau} = -\frac{\partial F}{\partial x^\mu} = -\frac{1}{4}\frac{\partial \gamma^{\kappa\lambda}}{\partial x^\mu} k_\kappa k_\lambda, \tag{IX.14b}$$

in which we have defined:

$$\gamma^{\mu\nu}(x,\, k) = P^{\mu\nu\kappa\lambda}(x)\, k_\kappa k_\lambda\,. \tag{IX.15}$$

Although the tensor field γ appears to define a second-rank covariant tensor field on M, actually, the fact that the local components are functions on T^*M shows that we are really "lifting" T^*M to the *vertical* sub-bundle $V(T^*M)$ of the tangent bundle $T(T^*M))$ to T^*M. That is, under the differential of the projection $T^*M \to M$ the tangent vectors in each $V_{(x,k)}(T^*M)$ project to 0 in each T_xM.

As usual, there is no natural complementary "horizontal" bundle to $V(T^*M)$ in $T(T^*M)$. However, from (IX.14a), we *do* have a natural algebraic correspondence between covectors on M and tangent vectors on M.

b. Dimension-codimension duality. If $k_x \in T_x^*M$ is a covector on M then one can associate k with the characteristic vector $X_P(k_x)$ on T_x^*M that is defined by P and obtain a tangent vector in $T_k T_x^* M$. Under the projection π: $T^*M \to M$ the vertical part of $X_P(k_x)$ will vanish and one obtains a tangent vector $d\pi|_{(x,\ k)}\ X_P(k_x)$ in T_xM. Hence, by composition, we obtain a map $i_P : T^*M \to T(M)$, $k_x \mapsto \mathbf{v}(k_x) = d\pi|_{(x,\ k)}\ X_P(k_x)$ In local form, it simply looks like:

$$v^\nu(x,\, k) = P^{\mu\nu\kappa\lambda}(x)\, k_\kappa k_\lambda\, k_\mu\ = \gamma^{\nu\mu}(x,\, k)\, k_\mu\,. \tag{IX.16}$$

If this system of homogeneous cubic equations is to replace the system of linear equations $v^\nu = g^{\nu\mu}\, v_\mu$ that one customarily derives from a Lorentzian structure g then just as one usually *postulates* the invertibility of the linear system, we must also postulate that the cubic system is invertible, as well. A necessary, but only locally sufficient, condition for this invertibility is given by the inverse function theorem, namely, the invertibility of the matrix:

$$\frac{\partial v^{\nu}}{\partial k_{\mu}} = \frac{\partial^2 P}{\partial k_{\mu} \partial k_{\nu}} = 3\gamma^{\mu\nu}(x, k) \tag{IX.17}$$

for every $(x, k) \in T^*M$. Hence, if we only assume that the algebraic correspondence $\mathbf{v} = \mathbf{v}(k_x)$ satisfies (IX.16) then the correspondence might possibly be neither one-to-one nor onto; for example, one might have folds and cusps.

Another complicating factor associated with the cubic case is the fact that now, since the polynomial $P^{\mu\nu\kappa\lambda}(x)\ k_{\kappa}\ k_{\lambda}\ k_{\mu}$ is homogeneous of degree three in k, the inverse algebraic relationship:

$$k_{\mu} = k_{\mu}(x, \mathbf{v}) \tag{IX.18}$$

is homogeneous of degree 1/3 in $\mathbf{v}$. Hence, whereas in the case of a quadratic dispersion law the map i_g and its inverse were both homogeneous of degree one, we now have a situation where the inverse map to i_P is not even a polynomial map anymore.

Moreover, in the Lorentzian case of a quadratic P the corresponding function $\bar{P}[\mathbf{v}] = P[k(\mathbf{v})]$ on $T(M)$ would also be quadratic. In other words, the light cones in the cotangent spaces would become light cones in the tangent spaces. However, that situation no longer obtains in the present case, since the function $\bar{P}$ does not have the same degree of homogeneity as the function P; in particular, it will have degree of homogeneity 4/3. One might consider $\bar{P}[\mathbf{v}] = (P[k(\mathbf{v})])^3$, which will have homogeneity degree four, but one is cautioned that this does not imply that $\bar{P}$ represents a homogeneous quartic *polynomial*, since there are many non-polynomial functions, such as rational functions, which are quotients of polynomials, that can still be homogeneous of degree four.

The fact that the function $\bar{P}[\mathbf{v}]$ on $T(M)$ is still homogeneous, but not generally a polynomial, implies that the hypersurface $\bar{P}[\mathbf{v}] = 0$ projects to a surface in $PT(M)$, which we will call the *Fresnel ray surface*, to be consistent with the literature on geometrical optics.

Due to the homogeneity of P, one finds that the normal surface in PT^*M and the ray surface in $PT(M)$ are related by a simple relationship that is true for any degree of homogeneity, even though it is usually established for only the quadratic case (cf., Kommerel [**1**]). One starts with the fact that Euler's formula for homogeneous functions of degree r, when combined with the dispersion law, gives us that:

$$0 = rP[k] = k_{\mu} \frac{\partial P}{\partial k_{\mu}} = k_{\mu} v^{\mu} \tag{IX.19}$$

for any r.

When one expresses k as $\omega dt - k_i\, dx^i$, this becomes $\omega v^0 = k_i\, v^i$ or:

$$n_i V^i = 1. \tag{IX.20}$$

In the case of optical media, one obtains an equations that are quite analogous to (VIII.102), (VIII.103), or (VIII.104) for the inhomogeneous coordinates (V^1, V^2, V^3) of the ray. These coordinates are obtained from the components v^{μ} of any velocity vector by

the usual process of regarding them as homogeneous coordinates for the ray and setting $V^i = v^i/v^0$. One also finds that the principal values ε^i of ε^{ij} get replaced by their reciprocals, since it is the inverse matrix ε_{ij} that now figures. The optical equation that corresponds to (VIII.104) is then:

$$\frac{(V^x)^2}{\frac{1}{v_r^2}-\frac{1}{v_x^2}}+\frac{(V^y)^2}{\frac{1}{v_r^2}-\frac{1}{v_y^2}}+\frac{(V^z)^2}{\frac{1}{v_r^2}-\frac{1}{v_z^2}}=0. \qquad \text{(IX.21)}$$

In this equation, we are now using the *ray velocity* v_r, which relates to the flow of energy in the direction of the Poynting vector, as opposed to the phase velocity v_p, which only relates to the normal to the wave front. The two are traditionally related by the relation:

$$\frac{v_p}{v_r}=\hat{n}(\hat{\mathbf{V}}), \qquad \text{(IX.22)}$$

in which $n = (1/v_p)\,\hat{n}$ and $\mathbf{V} = v_r\,\hat{\mathbf{V}}$, so $\hat{n}$ and $\hat{\mathbf{V}}$ are the unit covector and vector in those directions, relative to the Euclidian metric, which then gives $\hat{n}(\hat{\mathbf{V}})$ the interpretation of the cosine of the angle between the vectors **n** and **V**. In particular, this means that the wave normal does not have to be collinear with the ray velocity. Of course, in the pre-metric formulation of electromagnetism the very introduction of a metric onto the fibers of PT^*M and $PT(M)$ must be a consequence of the dispersion law for waves in the medium in question.

One can think of the "classical" – i.e., correspondence principle – limit of the quartic theory as being the degenerate case in which the metric $\gamma^{\mu\nu}(x, k)$ degenerates to $g^{\mu\nu}(x)$, so the cubic map i_P goes to the linear map i_g, because the dispersion polynomial $P[k]$ degenerates to the square of a quadratic. Hence, in general, the components of the metric γ we are using will depend on the covector k that we start with, as well as the point of M. This situation is sometimes referred to as a "rainbow metric" (cf., e.g., [**2**]).

It is important to understand that the linear map $d\pi \cdot X_P$ is only "generically" injective, since there is always the possibility that it might have projective singularities – i.e., points of T^*M at which X_P goes vertical, – which then depends upon the nature of P. One then has a second source of singularity in the association that did not appear in the linear isomorphism i_g: $T^*M \to T(M)$ that one obtained from a Lorentzian structure g.

As we have seen, one can use the polynomial $P[k]$ to define a corresponding map i_P : $\Lambda^1 M \to \Lambda_1 M$, $k(x) \mapsto \mathbf{v}(k(x))$. In the invertible case, this association between covector fields and vector fields defines what one might call *dimension-codimension duality*, since the differential system that is defined by a covector field has a dimension that is complementary to the dimension of the differential system that is defined by a vector field. One can eventually see that this is quite similar to the spirit of wave-particle duality if one understands that the isophase hypersurfaces that are defined by an integrable k represent the motion of wave surfaces and the integral curves of **v** represents the paths of the points of the wave surfaces when regarded as pointlike particles.

This situation also points out that there is an essential physical difference between the map i_P and the map i_g in terms of the units and interpretation of the covectors k_μ and v_μ that result from applying the inverses of these two maps to the same velocity vector v^μ. One sees that i_g^{-1} does not change the units of velocity in mapping it to covelocity, while i_P^{-1} essentially inverts the units of velocity (times time) into the corresponding units of frequency and wave number. The best way to get around this is to express both velocity and frequency-wave number in dimensionless units by rescaling in terms of some characteristic value of each.

c. Null geodesics. If we "solve" (IX.14a) for $k_\mu = \gamma_{\mu\nu} v^\nu$, where $\gamma_{\mu\nu}$ is the inverse matrix of $\gamma^{\mu\nu}$, and substitute in (IX.14b) then, after various straightforward manipulations of the expressions, we ultimately obtain a first-order system of differential equations for v^μ whenever $k(x)$ has been chosen:

$$\frac{dv^\mu}{d\tau} + \Gamma^\mu_{\kappa\lambda}(k) v^\kappa v^\lambda = --B^\mu_{\kappa\lambda}(k) v^\kappa v^\lambda + C^{\mu\kappa}_\lambda(k) k_\kappa v^\lambda , \tag{IX.23}$$

in which we have defined:

$$\Gamma^\mu_{\kappa\lambda}(k) = \tfrac{1}{2}\gamma^{\mu\nu}(\gamma_{\nu\lambda,\kappa} + \gamma_{\nu\kappa,\lambda} - \gamma_{\kappa\lambda,\nu}) \tag{IX.24a}$$

$$B^\mu_{\kappa\lambda}(k) = \tfrac{1}{4}\gamma^{\mu\nu}\frac{\partial\gamma_{\kappa\lambda}}{\partial x^\nu}, \tag{IX.24b}$$

$$C^{\mu\nu}_\lambda(k) = \gamma^{\mu\nu}\frac{\partial\gamma_{\nu\lambda}}{\partial k_\kappa}. \tag{IX.24c}$$

We recognize that for each choice of k the expressions $\Gamma^\mu_{\kappa\lambda}(k)$ take the form of the components of the Levi-Civita connection for the metric $\gamma_{\mu\nu}$. The differential equations for **v** then amount to perturbations of the conventional geodesic equations for the metric $\gamma_{\mu\nu}$ by the contributions on the right-hand side of (IX.23). Furthermore, the fact that k is characteristic suggests that we are justified in calling the geodesics thus obtained *null geodesics*. However, we must emphasize that this means that the "spectrum" of rainbow metrics $\gamma_{\mu\nu}(x, k)$ that we obtain as k varies over the characteristic quartic is associated with a spectrum of connections and a spectrum of null geodesics, as well.

The expressions $B^\mu_{\kappa\lambda}(k)$ and $C^{\mu\nu}_\lambda(k)$ essentially embody the perturbations to geodesic motion, in the Levi-Civita sense, that originates in the possibility that the quartic form $P[k]$ is not the square of a quadratic form of Lorentzian type. In order to see this, when $P[k] = (Q[k])^2$, with $Q[k] = g^{\mu\nu} k_\mu k_\nu$ one must first replace the dispersion law with its square root:

$$g^{\mu\nu}(x) k_\mu k_\nu = 0. \tag{IX.25}$$

One then finds that the bicharacteristic equations that result from using $Q[k]$ in place of $P[k]$ are the conventional geodesic equations for the Levi-Civita connection that is

defined by the metric $g^{\mu\nu}(x)$. Hence, $\Gamma^{\mu}_{\kappa\lambda}$ would no longer depend upon k, while $B^{\mu}_{\kappa\lambda}(k)$ and $C^{\mu\nu}_{\lambda}(k)$ would vanish.

It is tempting to identify the spectrum of geodesics that originate in the quartic nature of $P[k]$ as being, in some sense, "quantum fluctuations about the classical extremals;" for instance, one might have zitterbewegung in mind. However, one must remember that the quartic nature of $P[k]$ has more to do with the symmetry of constitutive laws of the medium, while quantum fluctuations are more commonly associated with going beyond the geometrical optics approximation that is implicit in the present discussion. We shall return to this issue at the end of this chapter in our discussion of diffraction, but for now we simply point out that quartic dispersion laws can also constitute quantum corrections.

As an example of the foregoing, let us look at what happens to the geodesic equations in the bi-metric case where:

$$P[k] = \tfrac{1}{4} g(k, k)\, \bar{g}(k, k)\,. \tag{IX.26}$$

The bicharacteristic equations are then:

$$\frac{dx^{\mu}}{d\tau} = \tfrac{1}{2}(\bar{\kappa}^2 g^{\mu\nu} + \kappa^2 \bar{g}^{\mu\nu})k_{\nu}\,, \quad \frac{dk_{\mu}}{d\tau} = -\tfrac{1}{4}(\bar{\kappa}^2 g^{\kappa\lambda}_{,\mu} + \kappa^2 \bar{g}^{\kappa\lambda}_{,\mu})k_{\kappa}k_{\lambda}\,, \tag{IX.27}$$

in which we have set:

$$\kappa^2 = g(k, k), \qquad \bar{\kappa}^2 = \bar{g}(k, k)\,. \tag{IX.28}$$

Although one can derive the corresponding geodesic equations that follow from setting:

$$\gamma^{\mu\nu} = \tfrac{1}{2}(\bar{\kappa}^2 g^{\mu\nu} + \kappa^2 \bar{g}^{\mu\nu})\,, \tag{IX.29}$$

nevertheless, since the process of inverting this matrix is not linear, a better way to understand the nature of the geodesic equations is by expressing the characteristic vector field in the form:

$$X_P = \bar{\kappa}^2 X_g + \kappa^2 X_{\bar{g}}\,, \tag{IX.30}$$

with:

$$X_g = \left(\tfrac{1}{2} g^{\mu\nu} k_{\nu}\right)\frac{\partial}{\partial x^{\mu}} - \left(\tfrac{1}{4} g^{\kappa\lambda}_{,\mu} k_{\kappa} k_{\lambda}\right)\frac{\partial}{\partial k_{\mu}}, \tag{IX.31}$$

and an analogous expression for $X_{\bar{g}}$.

One sees that, in a sense, there are two competing dynamical systems that represent the geodesic vector fields X_g and $X_{\bar{g}}$ on T^*M for the individual Lorentzian metrics g and $\bar{g}$ coupled in a linear combination by means of κ^2 and $\bar{\kappa}^2$ in a symmetric fashion. As a consequence, whenever the covector field k makes either coefficient vanish, the characteristic vector field is proportional to the geodesic vector field of the other metric.

3. Parallel translation. Since the bicharacteristic equations are apparently related to geodesic equations for the Levi-Civita connection that one obtains from the "metric" $\gamma^{\mu\nu}(x, k)$, and indeed agree with them when the dispersion polynomial $P[k]$ is quadratic, we naturally ought to examine how other geometric objects that just the velocity vectors to a geodesic congruence are translated along such curves.

a. Translation of k. First, let us return to the basic bicharacteristic equations (IX.14a, b). Once again, we express the first one as $v^\mu = \gamma^{\mu\nu} k_\nu$, with the usual notation. When one "inverts" the matrix $\gamma^{\mu\nu}$ this says $k_\mu = \gamma_{\mu\nu} v^n$, although when $P[k]$ is not quadratic the inverse only makes sense when one chooses a wave covector field k. When one substitutes this expression for one of the k's in (IX.14b), that equation takes the form:

$$0 = \frac{dk_\mu}{ds} + \frac{1}{r}\gamma_{\kappa p}\gamma^{\kappa\lambda}_{,\mu} v^\rho k_\lambda = v^\kappa \left[\frac{\partial k_\mu}{\partial x^\rho} + \frac{1}{r}\gamma_{\rho\kappa}\gamma^{\rho\lambda}_{,\mu} k_\lambda \right]. \tag{IX.32}$$

We have also generalized the degree of homogeneity of $P[k]$ in k to be r, although this will only equal 2 or 4 for our purposes.

By further manipulations, which include replacing $\gamma_{\rho\kappa}\gamma^{\rho\lambda}_{,\mu}$ with $-\gamma_{\rho\kappa,\mu}\gamma^{\rho\lambda}$ and observing that:

$$\gamma^{\rho\lambda}(\gamma_{\mu\rho,\kappa} - \gamma_{\kappa\mu,\rho}) v^\kappa k_\lambda = (\gamma_{\mu\rho,\kappa} - \gamma_{\kappa\mu,\rho}) v^\kappa v^\rho = 0, \tag{IX.33}$$

which follows from the conflict of symmetries in the indices κ and ρ, one finds that equation (IX.14b) actually says:

$$0 = v^\kappa \left[\frac{\partial k_\mu}{\partial x^\kappa} - \frac{2}{r}\Gamma^\lambda_{\kappa\mu} k_\lambda \right], \tag{IX.34}$$

with the same notation as above for the Christoffel symbols.

When $P[k]$ is quadratic ($r = 2$) the form that (IX.34) takes is precisely the equation of parallel translation when one uses the Levi-Civita connection for $\gamma^{\mu\nu}$, which then represents the Lorentzian structure. However, when $P[k]$ is a homogeneous quartic polynomial one sees that the connection that gives parallel translation associates infinitesimal Lorentz transformations that have half the magnitude of those in the quadratic case. One can also express this in the form:

$$\nabla_{\mathbf{v}} k_\mu = \begin{cases} 0 & (r = 2), \\ -\frac{1}{2}\Gamma^\lambda_{\kappa\mu} v^k k_\lambda & (r = 4). \end{cases} \tag{IX.35}$$

In either case, we conclude that the wave covector field k is parallel-translated along the null geodesics in manner that is even simpler than the way that the velocity vectors are subjected to. Of course, if one substitutes $k_\mu = \gamma_{\mu\nu} v^\nu$ in (IX.34), one sees that it is the differentiation of $\gamma_{\mu\nu}$ by k that can make the resulting equations of parallel translation for $\mathbf{v}$ more complicated than the ones for k.

*b. Translation of F and *F.* A immediate property of the bicharacteristic flow on M that is easy to verify is the fact that when the 2-forms F and $^*F$ – as well as f and $^*f$ – are associated with electromagnetic waves in the geometrical optics approximation in the geometrical optics approximation, they are constant along the flow of the geodesic vector field **v** that is associated with a given geodesic covector field k. This follows from the fact that since:

$$F = k \wedge E \qquad {}^*F = k \wedge B, \tag{IX.36}$$

in that approximation, one must have:

$$i_{\mathbf{v}}F = k(\mathbf{v})\, E - E(\mathbf{v})k = 0, \tag{IX.37a}$$
$$i_{\mathbf{v}}{}^*F = k(\mathbf{v})\, B - B(\mathbf{v})k = 0, \tag{IX.37b}$$

as **v** is transverse to the hyperplanes defined by E and B, for a suitable choice of E, B. We choose the E and B such that the triple $\{k, E, B\}$ is linearly independent and spans the annihilating hyperplane of **v** in each cotangent space; i.e., $E(\mathbf{v}) = B(\mathbf{v}) = 0$.

From this, and the sourceless field equations, we find that the Lie derivatives of F and $^*F$ along the flow of **v** are:

$$\mathrm{L}_{\mathbf{v}}F = i_{\mathbf{v}}dF + di_{\mathbf{v}}F = 0, \tag{IX.38a}$$
$$\mathrm{L}_{\mathbf{v}}{}^*F = i_{\mathbf{v}}d{}^*F + di_{\mathbf{v}}{}^*F = 0, \tag{IX.38b}$$

Hence, F and $^*F$ are, in a sense, convected by the flow of **v**.

Now, if we go back to the equations:

$$di_{\mathbf{v}}F = di_{\mathbf{v}}{}^*F = 0, \tag{IX.39}$$

and expand them in local components then we get, for the first one:

$$v^\mu\, dF_{\mu\nu} \wedge dx^\nu + F_{\mu\nu}\, dv^\mu \wedge dx^\nu = 0, \tag{IX.40}$$

The first term in this becomes:

$$v^\mu\, dF_{\mu\nu} \wedge dx^\nu = \tfrac{1}{2} v^\mu\, F_{\kappa\nu,\,\mu}\, dx^\kappa \wedge dx^\nu, \tag{IX.41}$$

in which we have used the fact that $dF = 0$ in order to obtain this.

In the second term of (IX.40) we note that if $v^\mu = \gamma^{\mu\lambda} k_\lambda$ then we can have two possible forms of the resulting expression depending upon whether $r = 2$ or 4.

i. Quadratic case ($r = 2$). In this case, $\gamma^{\mu\lambda}$ is independent of k and we obtain:

$$dv^\mu = k_\lambda\, \gamma^{\mu\lambda}{}_{,\kappa}\, dx^\kappa, \tag{IX.42}$$

which makes:

$$F_{\mu\nu}\, dv^\mu \wedge dx^\nu = \tfrac{1}{2}(F_{\mu\nu}\, k_\lambda\, \gamma^{\mu\lambda}{}_{,\kappa} - F_{\mu\kappa}\, k_\lambda\, \gamma^{\mu\lambda}{}_{,\nu})\, dx^\kappa \wedge dx^\nu. \tag{IX.43}$$

Substituting for $k_\lambda = \gamma_{\lambda\rho} v^\rho$ and rearranging puts the expression in parentheses into the form:

$$\gamma_{\lambda\rho}\,\gamma^{\mu\lambda}{}_{,\kappa}\,v^\rho F_{\mu\nu} + \gamma_{\lambda\rho}\,\gamma^{\mu\lambda}{}_{,\nu}\,v^\rho F_{\kappa\mu} = -\,\gamma^{\mu\lambda}\,\gamma_{\lambda\rho,\,\kappa} v^\rho F_{\mu\nu} - \gamma^{\mu\lambda}\,\gamma_{\lambda\rho,\,\nu}\,v^\rho F_{\kappa\mu}\ . \qquad \text{(IX.44)}$$

One can verify directly that if one substitutes $\Gamma^\mu_{\rho\kappa}$ for $\gamma^{\mu\lambda}\,\gamma_{\lambda\rho,\,\kappa}$ and similarly in the second term of the final expression in (IX.44) then the extra term that this introduces into the computations is symmetric in k and v, and therefore does not contribute to the components of the 2-form in (IX.43).

Ultimately, we find that when F has the form in question, **v** is the velocity vector field of a null geodesic congruence, and $P[k]$ is homogeneous of degree 2, one must have:

$$0 = v^\kappa\left[\frac{\partial F_{\mu\nu}}{\partial x^\kappa} - \Gamma^\lambda_{\kappa\mu}F_{\lambda\nu} - \Gamma^\lambda_{\kappa\nu}F_{\mu\lambda}\right] \equiv \nabla_{\mathbf{v}} F_{\mu\nu}\,. \qquad \text{(IX.45)}$$

That is, the electromagnetic field strength 2-form is parallel-translated by the Levi-Civita connection of the Lorentzian metric that is defined by $\gamma^{\mu\nu}$.

We could also use the property of the operator $\nabla_{\mathbf{v}}$ that it is a derivation with respect to the tensor product to see that:

$$\nabla_{\mathbf{v}} F_{\mu\nu} = \tfrac{1}{2}\nabla_{\mathbf{v}}(k_\mu E_\nu - k_\nu E_\mu) = \tfrac{1}{2}(k_\mu \nabla_{\mathbf{v}} E_\nu - k_\nu \nabla_{\mathbf{v}} E_\mu), \qquad \text{(IX.46)}$$

when one keeps in mind that k is parallel-translated along **v**.

In order for $\nabla_{\mathbf{v}} F_{\mu\nu}$ to vanish, we must have:

$$\nabla_{\mathbf{v}} E_\mu = \alpha k_\mu \qquad \text{(IX.47)}$$

for some real scalar α.

Hence, the 1-form E does not have to be parallel-translated, precisely, but its covariant derivative must be parallel to the wave covector field. This has the effect of allowing for a rotation of the polarization plane in the cotangent spaces, which is spanned the 1-forms k and E.

One finds that analogous results are arrived at for $*F = k \wedge B$; viz., $*F$ is parallel-translated along the null geodesics by means of the Levi-Civita connection, while B must have a covariant derivative that is parallel to k:

$$\nabla_{\mathbf{v}} *F_{\mu\nu} = 0, \qquad \nabla_{\mathbf{v}} B_\mu = \alpha k_\mu\,. \qquad \text{(IX.48)}$$

The fact that the proportionality constant is the same in either case follows from duality; in effect, E and B must remain perpendicular at all times.

The picture that emerges is that the 3-coframe $\{k, E, B\}$ moves along a null geodesic in such a manner that k remains parallel to itself while the 2-coframe $\{E, B\}$ can rotate in its plane. One sees that this generalizes the argument of Kline and Kay [**6**] in the first Appendix to Chapter V, in which they established, by means of vector calculus, that the 3-frame in question, or rather, its spatial projection, is parallel-translated along a null geodesic when the constitutive properties of the space are isotropic, but not necessarily

homogeneous, and the spatial metric that defines the connection is conformal to the Euclidian one, with a conformal factor that is the square of the index of refraction.

ii. Quartic case ($r = 4$). We return to (IX.42), only this time, we allow $\gamma^{\mu\lambda}$ to also be a differentiable function of k, so:

$$dv^{\mu} = \gamma^{\mu\lambda}{}_{,\kappa}\, k_{\lambda}\, dx^{\kappa} + \left[k_{\lambda} \frac{\partial \gamma^{\mu\lambda}}{\partial k_{\kappa}} + \gamma^{\mu\kappa} \right] dk_{\kappa}, \tag{IX.49}$$

before one pulls the dk_k down to M.

The term in brackets can be simplified by routine calculations that involve the definition of $\gamma^{\mu\lambda}$ and Euler's formula, and one gets:

$$dv^{\mu} = \gamma^{\mu\lambda}{}_{,\kappa}\, k_{\lambda}\, dx^{\kappa} + 3\gamma^{\mu\kappa} dk_{\kappa}, \tag{IX.50}$$

and after pulling the dk_k down to M by means of a section $k_k = k_k(x)$ – so $dk_{\kappa} = k_{\kappa,\lambda}\, dx^{\lambda}$ – one obtains the local 1-forms on M:

$$dv^{\mu} = \left[\gamma^{\mu\lambda}{}_{,\kappa} k_{\lambda} + 3\gamma^{\mu\lambda} k_{\lambda,\kappa} \right] dx^{\kappa}. \tag{IX.51}$$

Since this differs from the expression in (IX.42) only by the second term in brackets, we see that (IX.42) becomes:

$$\begin{aligned} &F_{\mu\nu}\, dv^{\mu} \wedge dx^{\nu} \\ &\qquad = (\text{R.H.S. of IX.42}) + \tfrac{3}{2}\, \gamma^{\mu\lambda}\, (k_{\lambda,\kappa} F_{\mu\nu} - k_{\lambda,\nu} F_{\mu\kappa})\, dv^{\kappa} \wedge dx^{\nu}. \end{aligned} \tag{IX.52}$$

However, from (IX.34) we can substitute $1/2\Gamma^{\rho}_{\lambda\kappa} k_{\rho}$ for $k_{\lambda,\kappa}$, which makes the components of the supplementary term in (IX.52) become:

$$3\gamma^{\kappa\lambda}\, (k_{\lambda,\mu}\, F_{\kappa\nu} - k_{\lambda,\nu} F_{\kappa\mu}) = -\tfrac{3}{2}\, \gamma^{\kappa\lambda}\, (\Gamma^{\rho}_{\lambda\mu} F_{\kappa\nu} - \Gamma^{\rho}_{\lambda\nu} F_{\kappa\mu})\, k_{\rho}\,. \tag{IX.53}$$

The ultimate effect of the extra term is then to replace (IX.45) with:

$$\nabla_{\mathbf{v}} F_{\mu\nu} = \tfrac{3}{2} (\gamma^{\kappa\lambda}\, \Gamma^{\rho}_{\lambda\mu} k_{\rho} F_{\kappa\nu} + \gamma^{\kappa\lambda}\, \Gamma^{\rho}_{\lambda\nu} k_{\rho} F_{\mu\kappa})\,. \tag{IX.54}$$

Hence, this is not parallel translation of F by means of the Levi-Civita connection of $\gamma^{\mu\nu}$, anymore.

In order to make more intuitive sense of the right-hand side of (IX.54), we introduce the notations:

$$\Gamma^{\kappa\rho}_{\mu} \equiv \gamma^{\kappa\lambda}\, \Gamma^{\rho}_{\lambda\mu}, \qquad \tilde{\omega}^{\kappa}_{\mu}(k) = \Gamma^{\kappa\rho}_{\mu} k_{\rho}, \tag{IX.55}$$

and re-write (IX.54) as:

$$\nabla_{\mathbf{v}} F_{\mu\nu} = \tfrac{3}{2}[\,\tilde{\omega}^{\kappa}_{\mu}(k)\,F_{\kappa\nu} + \tilde{\omega}^{\kappa}_{\nu}(k)\,F_{\mu\kappa}]\,. \qquad \text{(IX.54′)}$$

Replacing $F_{\mu\nu}$ with $k_\mu E_\nu - k_\nu E_\mu$ puts this into the form:

$$\begin{aligned}\nabla_{\mathbf{v}} F_{\mu\nu} &= \tfrac{3}{2}[\,\tilde{\omega}^{\kappa}_{\mu}(k)\,k_\kappa E_\nu - \tilde{\omega}^{\kappa}_{\nu}(k)\,k_\kappa E_\mu + k_\mu\,\tilde{\omega}^{\kappa}_{\nu}(k)\,E_\kappa - k_\nu\,\tilde{\omega}^{\kappa}_{\mu}(k)\,E_\kappa\,] \\ &= \tfrac{3}{2}[\,\tilde{\omega}(k)\,k \wedge E + k \wedge \tilde{\omega}(k)\,E]_{\mu\nu}\,, \qquad \text{(IX.56)}\end{aligned}$$

with some self-explanatory abbreviations.

If we apply $\nabla_{\mathbf{v}}$ to $k \wedge E$ and keep in mind (IX.35) then we also get:

$$\nabla_{\mathbf{v}}(k \wedge E) = \nabla_{\mathbf{v}} k \wedge E + k \wedge \nabla_{\mathbf{v}} E = -\tfrac{1}{2}\,\omega(\mathbf{v}) k \wedge E + k \wedge \nabla_{\mathbf{v}} E. \qquad \text{(IX.57)}$$

Hence, in order for this to be consistent with (IX.56), one must have:

$$\nabla_{\mathbf{v}} E_\mu = \alpha k_\mu + \tfrac{3}{2}\,\tilde{\omega}^{\kappa}_{\mu}(k) E_\kappa\,, \qquad [3\,\tilde{\omega}^{\kappa}_{\mu}(k) + \omega^{\kappa}_{\mu}(\mathbf{v})]\,k_\kappa = \beta E_\mu \qquad \text{(IX.58)}$$

for appropriate choices of real scalars α and β.

If we compare the first expression with the corresponding quadratic expression (IX.47) then we see that the net effect of the extension to a quartic dispersion polynomial is the addition of the second term on the right-hand side. The condition expressed by the second equation in (IX.58) gives a restriction on the connection itself, and if it is to be true for any conceivable E one sees that the proportionality constant β must vanish:

$$[3\,\tilde{\omega}^{\kappa}_{\mu}(k) + \omega^{\kappa}_{\mu}(\mathbf{v})]\,k_\kappa = 0. \qquad \text{(IX.59)}$$

This really just says that the effect of the infinitesimal transformation $3\,\tilde{\omega}^{\kappa}_{\mu}(k) + \omega^{\kappa}_{\mu}(\mathbf{v})$ on cotangent vectors should leave k fixed in any event.

The results for $^*F$ and B are obtained by analogous computations that essentially replace F with $^*F$ and E with B:

$$\nabla_{\mathbf{v}} {}^*F_{\mu\nu} = \tfrac{3}{2}[\,\tilde{\omega}(k)\,k \wedge B + k \wedge \tilde{\omega}(k)\,B]_{\mu\nu}, \qquad \text{(IX.60)}$$

$$\nabla_{\mathbf{v}} B_\mu = \alpha k_\mu + \tfrac{3}{2}\,\omega^{\kappa}_{\mu}(k) B_\kappa\,, \qquad [3\,\tilde{\omega}^{\kappa}_{\mu}(k) + \omega^{\kappa}_{\mu}(\mathbf{v})]\,k_\kappa = 0. \qquad \text{(IX.61)}$$

4. Huygens's principle. Huygens's principle is sufficiently fundamental and far-reaching in scope that it is presented in many forms in the literature (cf., [**3-11**]). The one that we shall first focus on is a modernization of the form that Huygens described graphically in his original treatise: Suppose we are given an initial hypersurface x_0: $N \to M$, $u \mapsto x_0(u)$, where N is an n–1-dimensional parameter manifold – e.g., hyperplane, sphere, ellipsoid, torus – and M is n-dimensional and path-connected. The time-t evolute of that initial hypersurface when it is in a state of wave motion takes the form of the

envelope of the (n–1)-parameter family of elementary hypersurfaces that emanate from each point of $x_0(u)$ and have the same value of t when the hypersurfaces are level surfaces of the function t.

a. The projectivized tangent bundle. The first obstacle to reconciling the traditional literature is the fact that most of it is set in a spatial manifold Σ of dimension two or three, not a spacetime manifold M of dimension four. Hence, as usual, one must proceed carefully in order to correctly account for the transition from spacetime to space. Indeed, we again find that the physically fundamental process that seems to assert itself is the projection of homogeneous coordinates for a projective space onto inhomogeneous coordinates, not merely the Cartesian projection that omits the temporal component.

In particular, it is the projection of any cotangent space T_x^*M onto its projectivized form PT_x^*M that first suggests this fact, since we are looking at the projection of the components $(\omega, -k_i)$ of any covector at $x \in M$ to the coordinates $(n_i = -k_i/\omega)$. One must then correctly account for the fact that although one regularly forms 1-forms such as:

$$n = n_i\, dx^i \tag{IX.62}$$

nevertheless, this is somewhat naïve since a local coordinate chart for PT^*U over an open subset $U \subset M$ would look like (t, x^i, n_i), in which the n_i are inhomogeneous coordinates for the real projective space $\mathbb{R}\mathrm{P}^3$ that models the fiber of PT^*U at each point of U, not components in the vector space $\mathbb{R}^3$.

It so happens that when M takes the form of the product manifold $\mathbb{R}\times\Sigma$ there is an accidental local equivalence of $PT^*M = PT^*(\mathbb{R}\times\Sigma)$ with $J^1(\Sigma; \mathbb{R})$, which is the manifold of 1-jets of differentiable functions on Σ; hence, such a construction has a distinctly non-relativistic sort of character. However, the use of functions of the form $t(x, y, z)$ in geometrical optics is so commonplace in the classical literature that we must unavoidably comment on this situation.

Locally, the manifold $J^1(\Sigma; \mathbb{R})$ looks like (x^i, f, f_i) and it projects onto Σ, $\mathbb{R}$, and $\Sigma\times\mathbb{R}$ in the predictable ways:

$$\begin{aligned} J^1(\Sigma; \mathbb{R}) &\to \Sigma, & j^1_x f &\mapsto x, \\ J^1(\Sigma; \mathbb{R}) &\to \mathbb{R}, & j^1_x f &\mapsto f, \\ J^1(\Sigma; \mathbb{R}) &\to \Sigma\times\mathbb{R}, & j^1_x f &\mapsto (x, f), \end{aligned}$$

It is not a fiber bundle under these projections, but a fibered manifold whose fibers locally look like the $\mathbb{R}\times\mathbb{R}^3$, $\mathbb{R}^3\times\mathbb{R}^3$, and $\mathbb{R}^3$, respectively.

The manifold $J^1(\Sigma; \mathbb{R})$ has a canonical 1-form $\theta = df - f_i\, dx^i$ that makes it a contact manifold. The significance of θ is based on the fact that when one looks at sections s: Σ

$\to J^1(\Sigma; \mathbb{R})$, $x \mapsto (x^i, f(x), f_i(x))$, such a section is integrable iff it is the 1-jet prolongation $j^1 f$ of the differentiable function f. Locally, this means that:

$$f_i = \frac{\partial f}{\partial x^i}, \tag{IX.63}$$

which is equivalent to requiring that:

$$0 = s^* \theta = df - f_i(x)\, dx^i. \tag{IX.64}$$

The contact structure on $J^1(\Sigma; \mathbb{R})$ then assigns each point of that manifold with the hyperplane $\theta = 0$ in its tangent space.

The manifold PT^*M also has a contact structure whose canonical 1-form $[\theta]$ we shall write in the form:

$$[\theta] = dt - n_i\, dx^i, \tag{IX.65}$$

when the local coordinate charts take the form (t, x^i, n_i).

We note that under the projection $T^*M \to PT^*M$, $k \mapsto [k]$, the canonical 1-form $k_\mu\, dx^\mu = \omega\, dt - k_i\, dx^i$ on T^*M projects to $\omega(dt - n_i\, dx^i)$ on PT^*M, which is proportional to the canonical 1-form, as long as one regards t as a function on Σ, not a coordinate of M. Hence, the vanishing of one of the canonical 1-forms is equivalent to the vanishing of the other.

Since $n_i = k_i / \omega$, we see that an integrable section $[k]$ of $PT^*M \to M$, i.e., an integrable line field on M, must satisfy the condition:

$$n_i = -\frac{\partial \phi / \partial x^i}{\partial \phi / \partial t} = \frac{\partial t}{\partial x^i} \tag{IX.66}$$

for some differentiable function ϕ on M and a differentiable function t on Σ. Note that changing the function ϕ to some other function ϕ' does not actually change the components n_i, since a change of ϕ is equivalent to a re-parameterization of $\mathbb{R}$.

The condition (IX.66) then leads to:

$$dt = n_i\, dx^i \tag{IX.67}$$

quite naturally.

We also find that if F is a differentiable function on T^*M then as long as it is homogeneous of some degree r in the fiber – i.e., $F[\lambda k] = \lambda^r F[k]$ – it will project to a function $[F]$ on PT^*M. Furthermore, this means that the characteristic vector field X_F on T^*M might be associated with a characteristic line field $[X_F]$ on PT^*M.

One any contact manifold (see Arnol'd [**11**]), the line in question is defined by the intersection of the contact hyperplane and the hyperplane that is annihilated by the 1-form $d[F]$. Hence, it will be a solution – up to a scalar multiple – of the equation:

$$i_{[X]} d\theta = d[F], \tag{IX.68}$$

in which we have abbreviated slightly.

Locally, this becomes the system:

$$[X]_t = 0 = \frac{\partial [F]}{\partial t}, \qquad [X]^i = \frac{\partial [F]}{\partial x^i}, \quad [X]_i = -\frac{\partial [F]}{\partial n_i}. \tag{IX.69}$$

Hence, we see immediately that not every homogeneous function on T^*M is projectable to something on PT^*M that will give a characteristic line field, but only the ones that are "time-invariant," in the sense that their first partial derivative with respect to t vanishes. Of course, when F refers to the dispersion law for electromagnetic wave propagation, this is a common assumption, since it amounts to assuming that the optical properties – or more generally, constitutive properties – of the medium are constant in time.

However, if F is homogeneous and time-invariant then the characteristic line field $[X_F]$ on PT^*M gives rise to a congruence of geodesics – up to parameterization – by integrating the system of ordinary differential equations that make $[X_F]$ a velocity vector field:

$$\frac{dx^i}{ds} = \frac{\partial [F]}{\partial n_i}, \qquad \frac{dn_i}{ds} = -\frac{\partial [F]}{\partial n_i}, \tag{IX.70}$$

in which the curve parameter is s, here.

Of course, when $M = \mathbb{R}\times\Sigma$, we see that essentially the same system of equations will come about by looking at the symplectic structure on $T^*\Sigma$ and the characteristic vector field that is defined by the analogous differentiable function $[F]$ on $T^*\Sigma$.

Furthermore, if we locally equate $PT^*(\mathbb{R}\times U)$ with $J^1(U; \mathbb{R})$ for some $U \subset \Sigma$, we see that we are also dealing with the system of bicharacteristic equations that follow from the first-order partial differential equation defined by:

$$[F][n] = 0, \qquad n = dt, \tag{IX.71}$$

which is the spatial version of the eikonal equation that one obtains from F, as we discussed in Section 1; however, we are now referring to the spatial phase function as t.

It is interesting that classical mechanics starts with objects in $J^1(\mathbb{R}; \Sigma)$ – viz., 1-jets of differentiable curves in Σ – while classical geometrical optics should model wave motion in terms of the dual objects in $J^1(\Sigma; \mathbb{R})$, especially since quantum physics seems to favor the representation of matter by waves rather than points. One can even define a form of *conjugacy* between curves in Σ and functions on Σ that is defined by the possibility that under composition of the curve γ: $\mathbb{R} \to \Sigma$ with the function f: $\Sigma \to \mathbb{R}$, one might obtain the identity map on $\mathbb{R}$:

$$f(\gamma(t)) = t \qquad \text{(for all } t \in \mathbb{R}\text{)}. \tag{IX.72}$$

Of course, given γ the choice of f is not unique, nor conversely, since one can see that a given curve in Σ can be embedded in many hypersurfaces, just as a given hypersurface admits many embedded curves.

Since the manifolds $J^1(\mathbb{R}; \Sigma)$ and $J^1(\Sigma; \mathbb{R})$ are topologically the same as $\mathbb{R} \times T(M)$ and $T^*M \times \mathbb{R}$, respectively, one sees that this sort of conjugacy is another form of dimension-codimension duality.

b. Elapsed-time functional. Now that we have made these prefatory remarks on the contact manifolds that we are concerned with, we see that some of the constructions of geometrical optics that lead to the principles of Huygens and Fermat follow quite naturally with no further restricting conditions.

For instance, given the spatial 1-form $n = n_i\, dx^i$ on Σ, one sees that its integral along any curve segment γ: $[0, 1] \to \Sigma$:

$$\Delta t[\gamma] = \int_\gamma n \tag{IX.73}$$

defines what we can call the *elapsed-time functional*, since in the case where n is exact and takes the form dt that is precisely what the integral will represent.

So far, our elapsed-time functional is a singular 1-cochain with coefficients in $\mathbb{R}$. If one is to resolve it to a two-point function on Σ that will represent a *characteristic function* for our geodesic flow on PT^*M, we have two basic possibilities:

1. If $\Delta t[\gamma]$ takes on the same values for any two homologous curves γ – i.e., ones that all have the same endpoints $\partial\gamma = y - x$ – then one can define:

$$\Delta t[x; y] = \Delta t[\gamma] \tag{IX.74}$$

unambiguously.

2. If there is some well-defined set of curves from x to y that all give the same value of $\Delta t[\gamma]$ then (IX.74) makes sense as long as γ is an element of that set.

In the first event, we must have the 1-form n is exact in order that the integral over curve reverts to the integral over its boundary, à la Stokes. The 1-cochain Δt then becomes a 1-coboundary.

In the second event, we see that we can take advantage of the fact that when one is given n: $\Sigma \to J^1(\Sigma; \mathbb{R})$, $x \mapsto (x, t(x), n_i(x))$, one can take advantage of the fact that then we have been given a spatial dispersion function $[F]$ on $J^1(\Sigma; \mathbb{R})$ there is a unique path from $n(x)$ to $n(y)$ in $J^1(\Sigma; \mathbb{R})$ – at least when they are sufficiently close – that is defined by the geodesic flow of $[F]$. The projection of this path onto Σ then gives the curve γ that we can integrate n over in order to make sense of (IX.74).

What makes $\Delta t[x; y]$ a characteristic function on $\Sigma \times \Sigma$ is the property that

$$n(x) = -\frac{\partial}{\partial x}\Delta t[x; y], \qquad n(y) = \frac{\partial}{\partial y}\Delta t[x; y]. \tag{IX.75}$$

Later, when we have discussed contact transformations, we shall see that this property also makes it a "generating functional" for the contact transformation that takes $(x, n(x))$ to $(y, n(y))$.

c. Elementary hypersurfaces. Since our elapsed-time functional is really just a generalization of the arc-length functional that one obtains in the case where $P[k]$ degenerates to the square of a quadratic polynomial, we can – at least for a sufficiently small elapsed time t – define a *geodesic sphere* about a given $x \in \Sigma$ of radius t to be the set $S_x(t)$ of all $y \in$ M such that $\Delta t[x, y]$ exists and is equal to t.

In the elementary geometrical case of Euclidian spatial geometry, the vanishing of the curvature of the Euclidian metric guarantees that geodesic spheres will exist about each point and have arbitrarily large radius. As long one considers only a constant speed of propagation for electromagnetic waves the elapsed time along any curve segment will be proportional to its arc length. One can then define a differentiable one-parameter family of geodesic spheres about each point for every $t > 0$.

However, when the curvature of the spatial manifold Σ is non-vanishing there will generally be a finite "radius of injectivity" for the exponential map that the chosen connection defines; that is, there will be pairs of points that are connected by geodesics of unequal length, such as non-antipodal points on a sphere.

Since the role of the dispersion polynomial $P[k]$ has been somewhat obscured by now, it is important to see that as long as one restricts oneself to only those 1-forms n on Σ that satisfy the requirement:

$$[P][n] = 0 \tag{IX.76}$$

one will find that one is dealing with 1-forms $k = \omega(dt - n)$ on $M = \mathbb{R}\times\Sigma$, where ω is non-zero, but arbitrary, that satisfy the requirement:

$$P[k] = 0. \tag{IX.77}$$

This means that elementary waves propagate in the characteristic hypersurfaces.

Hence, in the Minkowskian case, for instance, one sees that the elementary hypersurfaces in Σ are expanding spheres of radius t about each point x that lie on the light cone at the corresponding point (t, x) in Minkowski space.

d. The propagation of phase. The traditional formulation of Huygens's principle allows one to construct the time evolution of a momentary wave surface within its characteristic hypersurface from one moment to another by means of these elementary wave surfaces.

More precisely, let $\mathcal{F}_0 \subset \Sigma$ be a momentary wave surface at time $t = 0$; i.e., the image of an embedded two-dimensional submanifold. If one wishes to obtain the momentary wave front $\mathcal{F}_t$ at some later value of t then, as long as a geodesic sphere of radius t exists

about each $x_0 \in \mathcal{F}_0$, if we represent such a sphere by an embedding $\imath(x_0)$: $S^2 \to \Sigma$ then one can define a smooth two-parameter family of geodesic spheres of radius t by means of $\mathcal{F}_0 \times S^2 \to \Sigma$, $(x_0, y) \mapsto \imath(x_0)(y)$.

Of course, the image of $\mathcal{F}_0 \times S^2$ in Σ under this map will generally be a three-dimensional region. In order to obtain the evolute of the initial momentary wave surface $\mathcal{F}_0$ for this particular time interval, we need to consider only the boundary of that region – viz., the envelope of the family. A point y of this envelope $\mathcal{F}_t$ is characterized by the fact that if one moves along a curve through y whose tangent vector field is tangent to the envelope then the value of $\Delta t[x_0, y] = t$ does not change in an infinitesimal neighborhood of y. Hence, $y \in \mathcal{F}_t$ iff there is an $x_0 \in \mathcal{F}_0$ such that:

$$\Delta t[x_0, y] = t\,, \tag{IX.78a}$$

$$\frac{\partial}{\partial x_0}\Delta t[x_0, y] = n(x_0) = 0. \tag{IX.78b}$$

Between these equations, we obtain three local component equations for the three spatial coordinates of y as a function of x_0 . For instance, in the conventional case where Σ is Euclidian $\mathbb{R}^3$, when one uses:

$$\Delta t[x_0, y] = [\,(y^1 - x_0^1)^2 + (y^2 - x_0^2)^2 + (y^3 - x_0^3)^2\,]^{1/2}/c, \tag{IX.79}$$

with the points of $\mathcal{F}_0$ being parameterized by $x_0^i = x_0^i(u, v)$ the equations take the form:

$$(y^1 - x_0^1)^2 + (y^2 - x_0^2)^2 + (y^3 - x_0^3)^2 = (ct)^2, \tag{IX.80a}$$

$$\delta_{ij}(y^i - x_0^i)\frac{\partial x_0^j}{\partial u} = 0, \tag{IX.80b}$$

$$\delta_{ij}(y^i - x_0^i)\frac{\partial x_0^j}{\partial v} = 0. \tag{IX.80c}$$

The last two of these equations basically say that the line segment in $\mathbb{R}^3$ that goes from x_0 to y will always be orthogonal to the surface $\mathcal{F}_0$ and the first one says that it will have a length equal to ct. Since this defines a unique line segment (up to orientation) through each point of $\mathcal{F}_0$ there will be two well-defined endpoints $\pm y$ and thus two well-defined evolutes of $\mathcal{F}_0$, one of them a "forward" evolute $\mathcal{F}_{+y}$ and the other one $\mathcal{F}_{-y}$ a "backward" evolute. It was precisely this ambiguity in the solution that worried Huygens most. As it turned out, when one looks at the propagation of *amplitude*, the ambiguity is resolved. We illustrate the construction of the phase evolute by means of the envelope of geodesic phase spheres in Fig. 10.

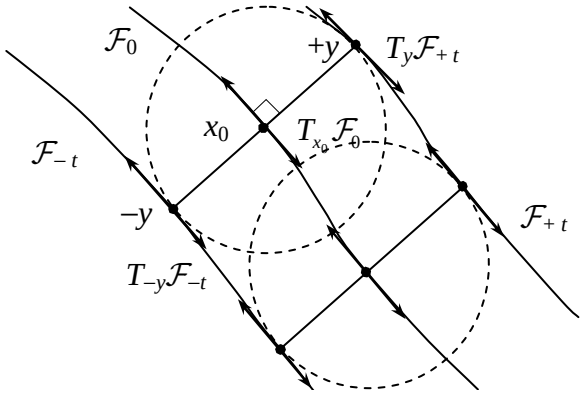

Figure 10. The construction of the evolved momentary wave surfaces by means of Huygens's principle.

e. Contact transformations. If we examine the construction of the evolved momentary wave surfaces then we see that we are not only mapping each point $x_0 \in \mathcal{F}_0$ to a pair of points on $\mathcal{F}_{+t}$ and $\mathcal{F}_{-t}$, but we also defining a map from the tangent space $T_{x_0}\mathcal{F}_0$ to the tangent spaces $T_{-y}\mathcal{F}_{-t}$ and $T_{+y}\mathcal{F}_{+t}$.

Now, since the tangent planes to any momentary wave surface $\mathcal{F}_t$ are hyperplanes in the tangent spaces to Σ at each point of $\mathcal{F}_t$, and a hyperplane in any tangent space $T_x\Sigma$ is associated with a unique 1-form n_x (up to non-zero scalar multiplication), we see that a momentary wave surface $\mathcal{F}_t$, along with its tangent spaces, defines a section $[n]$: $\mathcal{F}_t \to PT^*\Sigma$, $x \mapsto [n_x]$ of the projectivized cotangent bundle $PT^*\Sigma$ of Σ, that is, the bundle of lines through the origin in each fiber of $T^*\Sigma$. We can then extend this to a section n: $\mathbb{R}\times\mathcal{F}_t \to PT^*M$ by extending n to $dt - n$ at each (t, x).

What we now find is that since we have a canonical congruence of geodesics on PT^*M whenever $[F]$ has been chosen, it is actually unnecessary to go the route of first defining a family of geodesic spheres parameterized by the points of $\mathcal{F}_0$ and then going to the envelope of this family since, as long as a unique curve goes through each $n(x_0)$ in PT^*M over the points of $\mathcal{F}_0$ one can map the point x_0 *directly* to a unique point $y = x_0(t)$ for each t in the domain of definition for the local flow of X_F about $n(x_0)$.

One then sees that a basic property of the time evolution of momentary wave surfaces is that it must consist of a one-parameter family of *contact transformations* [1] whose parameter is t, in this case More precisely, a contact transformation is a diffeomorphism of PT^*M that preserves the hyperplanes in $T(PT^*M)$ that are annihilated by θ; note that this is not the same thing as preserving θ, as much as preserving θ up to multiplication by

[1] In addition to the references to Lie and Vessiot that were given in the Introduction to this book, one might also confer Eisenhart [**10**] or Arnol'd [**11**] on the subject of contact transformations..

a non-zero scalar function [2]. Hence, it will take a tangent plane at one point to a tangent plane at another point.

An immediate advantage of the use of contact transformations is that for a given t, one only produces *one* evolved momentary wave surface, namely, the one at $+t$, and not its somewhat unphysical twin at $-t$.

The elapsed-time function Δt: $\Sigma \times \Sigma \to \mathbb{R}$, $(x, y) \mapsto \Delta t[x, y]$ has a fundamental property in the eyes of contact transformations, namely that its partial derivatives at x and y define the 1-forms n_x and n_y when one is integrating n:

$$n_x = -\frac{\partial(\Delta t)}{\partial x} = -\frac{\partial(\Delta t)}{\partial x^i} dx^i, \qquad n_y = \frac{\partial(\Delta t)}{\partial y} = \frac{\partial(\Delta t)}{\partial y^i} dy^i; \tag{IX.81}$$

they also define the tangent planes to the momentary wave surfaces at each point. One then calls the function Δt a *generating function* for the contact transformation, since a given contact element (x, n_x) will get mapped to the corresponding (y, n_y) that one obtains from (IX.79).

f. The propagation of amplitude – general case. So far, we have only discussed Huygens's principle as it relates to the propagation of momentary wave surfaces; i.e., the propagation of the phase function for the wave. However, in elementary optics when one first encounters explanations for the phenomena of interference and diffraction one hears Huygens's principle applied to the propagation of wave amplitudes, or equivalently, intensities.

In that context, one finds that what one is really dealing with is the fact that elementary solutions of wave equations involve both phase functions and amplitude functions. We recall that the first step in the geometrical optics approximation was to consider only electromagnetic waves of the form $F(t, x^i) = e^{i\phi(t,x^i)} f(x^i)$, in which $\phi \in C^\infty(M)$ is a phase function and $f \in \Lambda^2_s M$, so it takes the form $1/2\, f_{\mu\nu}(x^i)\, dx^\mu \wedge dx^\nu$. This then makes $*F(t, x^i) = e^{i\phi(t,x^i)} * f(x^i)$.

Note that in order to make rigorous sense of the decomposition of F and $*F$, we have to explain how complex scalars act on real 2-forms. As it turns out, and we shall discuss this in greater depth in a later chapter, if the * isomorphism behaves like a Hodge * for a Lorentzian metric, then one has $*^2 = -I$, which means it defines an *almost-complex structure* on the real vector bundle $\Lambda^2 M$. One can then define the action of complex scalars by treating multiplication by * as the same thing as multiplication by i and:

$$(\alpha + i\beta)F = \alpha F + \beta *F. \tag{IX.82}$$

Hence:

$$e^{i\phi} F = \cos\phi\, F + \sin\phi\, *F. \tag{IX.83}$$

[2] According to Arnol'd [**11**], people who confuse the two conditions are "bad people," and pure mathematics could always use another moral principle.

One sees that the time evolution of the spatial amplitude f – or $*f$, for that matter – is confined to the orbit that it makes in the vector space $\Lambda^2 M$ under this action. Recall that this was precisely the issue that we discussed in the section of Chapter VIII on polarization, so we are again looking at duality rotations.

By exterior differentiation, the sourceless Maxwell equations $dF = d*F = 0$ give:

$$df + ik \wedge f = 0, \qquad d*f + ik \wedge *f = 0. \tag{IX.81}$$

The second step in the geometrical optics approximation is when one introduces the approximation that makes df and $d*f$ effectively zero. This would be like assuming that the spatial amplitudes f and $*f$ are constant, which would make the right-hand sides of (IX.83a, b, c) all vanish. Hence, one immediate way of going beyond the geometrical optics approximation, before one gets to the asymptotic series methods of diffraction theory, is to consider spatial 2-forms f and $*f$ that satisfy the differential equations (IX.81) instead of the algebraic equations that follow from omitting their differential terms.

One finds that the previous results that F and $*F$ are Lie-transported along the flow of $\mathbf{v}$ imply corresponding conditions on f and $*f$:

$$0 = \mathrm{L}_{\mathbf{v}} F = \mathrm{L}_{\mathbf{v}}(e^{i\phi} f) = e^{i\phi}(ik \wedge f + \mathrm{L}_{\mathbf{v}} f), \tag{IX.82a}$$
$$0 = \mathrm{L}_{\mathbf{v}} *F = \mathrm{L}_{\mathbf{v}}(e^{i\phi} *f) = e^{i\phi}(ik \wedge *f + \mathrm{L}_{\mathbf{v}} *f), \tag{IX.82b}$$

which gives:

$$\mathrm{L}_{\mathbf{v}} f = - ik \wedge f, \qquad \mathrm{L}_{\mathbf{v}} *f = - ik \wedge *f. \tag{IX.83}$$

Hence, the spatial 2-forms are no longer convected by the flow of $\mathbf{v}$.

d. The propagation of amplitude – linear case. Usually the propagation of amplitude in geometrical optics is addressed in the case of the linear wave equation. Since elementary wave solutions and fundamental solutions – in the Green function sense of the term – are closely-related concepts, one also finds that the basic construction associated with the propagation of amplitude functions from an initial wave surface $\mathcal{F}_0$ to some later surface $\mathcal{F}_t$ is essentially a graphical way of representing a linear operator [3] $K_{(0,t)} : \mathcal{L}^1(\mathcal{F}_0) \to \mathcal{L}^1(\mathcal{F}_t)$, $\psi \mapsto K_{(0,t)}\,\psi$ as an integral operator with a kernel $K(x, y)$:

$$(K_{(0,t)}\,\psi)(x) = \int_{\mathcal{F}_0} K(x, y)\psi(y)\,\mathcal{V}_y \,. \tag{IX.84}$$

Hence, this sort of construction is only useful for linear wave equations.

Since the kernel $K_{(0,t)} : \mathcal{F}_0 \times \mathcal{F}_t \to \mathbb{R}$ of the integral operator is a two-point function on a subset of $\Sigma \times \Sigma$, one naturally wonders how it relates to the elapsed-time $\Delta t(x, y)$.

[3] Here, we are using the notation $\mathcal{L}^1(N)$ to mean Lebesgue-integrable real functions on the manifold N. The Lebesgue measure on any manifold starts off on $\mathbb{R}^n$ in the coordinate charts and eventually defines the volume element $\mathcal{V}$ more globally.

One finds that, in practice, the kernel function is usually a function of the elapsed-time function, which usually takes the form of the geodesic distance function r: $\Sigma\times\Sigma \rightarrow \mathbb{R}$, $(x, y) \mapsto r(x, y)$ between pairs of points in Σ, as long as $r = ct$.

For example, consider the fundamental solutions $K(x, y)$ of the stationary linear wave equation, which takes the form of *Helmholtz's equation:*

$$(\Delta + k^2)\psi = 0, \tag{IX.85}$$

which one generally uses in the construction of interference patterns.

They take the form of expanding spherical wave solutions:

$$K(y - x) = \frac{e^{ik(y-x)}}{4\pi r(y - x)}, \tag{IX.86}$$

which implies that one must be dealing with a manifold Σ that has an affine structure, such as $\mathbb{R}^3$ if one is to make sense of the expression $y - x$. The distance function r is then the conventional Euclidian one.

It is important to note that $K(x - y)$ is not defined on the diagonal of $\Sigma\times\Sigma$ (i.e., all points of the form (x, x)). For this reason, fundamental solutions are often referred to as *fundamental singularities* (of the pole type).

The propagation of amplitude functions from a closed (= compact, without boundary) initial surface $\mathcal{F}_0$ to $\mathcal{F}_t$ is then given by *Helmholtz's theorem*: *Let:*

$$(K_{(0,t)}\,\psi)(y) = \int_{\mathcal{F}_0}\left[\frac{\partial K(y-x)}{\partial n}\psi(y) - K(y-x)\frac{\partial \psi(y)}{\partial n}\right]\mathcal{V}_y\,. \tag{IX.87}$$

with the kernel $K(y - x)$ defined as in (IX.86). One then has that:

$$(K_{(0,t)}\,\psi)(y) = \begin{cases}\psi(x) & x \text{ outside } \mathcal{F}_0,\\ 0 & x \text{ inside } \mathcal{F}_0.\end{cases} \tag{IX.88}$$

The fact that this integral vanishes for points inside the surface $\mathcal{F}_0$ accounts for the fact that the secondary solution $\mathcal{F}_{-t}$ to the propagation of the momentary wave surfaces is essentially physically irrelevant. That is, since the propagated wave surface $\mathcal{F}_{-t}$ lies inside the initial surface $\mathcal{F}_0$ the propagated amplitude of ψ will vanish on that surface.

This is the resolution of the dilemma that Huygens was worried about, since the existence of a secondary, inward-propagating wave front would not be physically significant if the amplitude of the wave on it was identically zero.

In order to restore the time evolution, one needs only to recall the separation of variables that produced the stationary solution:

$$\Psi(t, x) = T(t)\psi(x) = e^{-i\omega t}\psi(x). \qquad \text{(IX.89)}$$

However, there is an important subtlety associated with this reduction: By assuming that the time variation is governed by a single frequency ω, one is considering only *monochromatic* solutions of the full linear wave equation. Of course, as long as one is only concerned with the linear case, one can express the more general solutions in terms of a Fourier integral over all frequencies. However, the reduction does have the effect of converting a hyperbolic Cauchy problem into an elliptic boundary value problem. Hence, the values of ψ and its normal derivative on $\mathcal{F}_0$ are no longer independent of each and cannot be specified independently, as they would be in the hyperbolic Cauchy problem.

The reduction to a monochromatic stationary wave also implies that the solution space has been reduced from infinite-dimensional to finite-dimensional by essentially projecting onto the subspace that is singled out by a choice of ω. One can see that if each ω is associated with a distinct finite-dimensional vector space of solutions to the Helmholtz equation then clearly the generalized Cartesian product of this one-parameter family of vector spaces gives an infinite-dimensional vector space.

5. Diffraction. The foregoing discussion has been subordinate to the geometrical optics approximation, in which one regards the wave number (or frequency) of the wave in question as sufficiently large in comparison to the gradients of the electromagnetic field amplitudes that one can ignore those gradients and consider only a set of algebraic equations for the field amplitudes that amount to the passage from the field operator $\Box_\kappa$ to its symbol. As a consequence, one finds that the usual machinery of spacetime geometry – e.g., its Lorentzian metric, its geodesics, etc – are natural constructions that follow from the dispersion law, which gives to the characteristic hypersurfaces in the cotangent spaces. We have also seen that even within the geometrical optics approximation there is room for expansion in the geometry of spacetime when one considers birefringence and nonlinearity.

It is only natural to wonder how the geometry of spacetime would be further affected by eliminating, or at least weakening, the approximation that gave us this geometrical machinery. In particular, the system of linear, first-order, partial differential equations for f that we obtained in (VIII.35) would not reduce to algebraic equations if the differential contributions df and d^*f were no longer negligible compared to ω.

So far, the most definitive progress towards bridging the gap between wave optics and geometrical optics has been in the theory of diffraction. Since the original intent of wave mechanics was to make the transition from wave mechanics to classical mechanics analogous to the transition from wave optics to geometrical optics, there is also a close parallel between the methods of diffraction optics and the loop expansions of quantum field theory.

The starting point for most optical diffraction theory [4] is essentially the Cauchy problem for the Helmholtz equation for some C^2 function u on the spatial manifold Σ. As we saw, such a function gives the shape of stationary or time-harmonic waves. The

[4] The standard references on the subject of diffraction theory include Kline and Kay [**6**], Born and Wolf [**7**], Luneburg [**8**], and Baker and Copson [**9**].

Cauchy data u_0, $\partial u/\partial n$ are defined on some compact, possibly bounded, surface S in space, such as a rectangular slit or circular aperture in an opaque screen. Although it is usually emphasized that the waves propagating up to the slit or aperture are plane waves, actually, if the Cauchy problem is well-posed then that condition is irrelevant to the time evolution of the wave fronts past the opening. One must also keep in mind that since one is dealing with the elliptic equation that follows from the hyperbolic wave equation that the spatial Cauchy data are no longer independent of each other, but must satisfy a compatibility condition.

Since the Helmholtz operator $\Delta + k^2$ is linear and self-adjoint, one then solves the Cauchy problem for the Helmholtz equation by way of:

$$u(y) = -\frac{1}{4\pi}\int_S \left[u_0(x)\frac{\partial}{\partial n}\left(\frac{e^{ik(y-x)}}{\| y-x \|}\right) - \frac{\partial u_0(x)}{\partial n}\left(\frac{e^{ik(y-x)}}{\| y-x \|}\right)\right] \mathcal{V}_x . \qquad \text{(IX.90)}$$

However, like most integrals, this integral cannot generally be evaluated directly. Hence, one either goes the route of numerical integration or series approximations. The series approximation that is customarily employed is particularly difficult to fully comprehend, namely, the *asymptotic series* approximation (see, e.g., [**6-8**, **12-15**]). The justification for the use of such a series follows from the assumption that the Cauchy data was rapidly oscillating on the initial surface S, which amounts to the large ω approximation. In physics, this usually implies that the methods of diffraction theory are not useful until at least the microwave part of the electromagnetic wavelength spectrum, since the spatial dimensions of S can be quite appreciable for radio wave lengths, which might be in meters, or even kilometers.

Unlike the usual power series expansions of elementary calculus, an asymptotic series expansion is not presumed to converge in general; that is, it starts out as a *formal* power series. Furthermore, the series is also assumed to depend upon some parameter, such as k, in such a way that the series becomes an increasingly accurate approximation in the asymptotic limit as k grows arbitrarily large. One also frequently encounters the possibility that going to higher-order terms in the series might improve the approximation up some order, but then reduce the accuracy as the order goes beyond that point.

For the function $u(x)$, the expansion takes the form:

$$u(x) = e^{ik\psi(x)}\sum_{n=0}^{\infty}\frac{U_n(x)}{(ik)^n} . \qquad \text{(IX.91)}$$

Note that we have implicitly factored our phase function $\theta(x)$ into the product $k\psi(x)$, which assumes that we actually can define some characteristic wave number k, such as the Euclidian norm of the spatial wave number 1-form.

In order to get the functions $U_n(x)$ one can then substitute (IX.91) into (IX.90) and obtain a sequence of recurrence relations. However, in practice, one performs other transformations on the basic integral (IX.90) in order to obtain more manageable results. A further approximation that is often imposed is to assume that the dimensions of S are large compared to the wavelength of the waves passing though it, although diffraction

usually does not become noticeable until the dimensions of S are comparable to several wavelengths.

Generally, one expects that the series will start with the geometrical optics field $u^*(x) = e^{ik\psi(x)}U_0(x)$. The subsequent contributions to the series then represent successive orders of correction to the geometrical optics approximation. One can then regard the field $u - u^*$, which represents the terms of the asymptotic series beyond zeroth order, as the *diffracted field.*

The methods of asymptotic approximations get applied to quantum wave mechanics by using $1/\hbar$ in place of k as the large parameter. Presumably, one retrieves classical mechanics in the asymptotic limit as $\hbar$ goes to zero, which is like saying that the wave associated with a moving mass, such as an electron, ceases to diffract when passing though an opening. In fact, it has long since been established that low-energy electrons diffract in a wavelike manner when passing through openings whose dimensions are comparable to atomic spacings in crystal lattices, or about 1 Å, which was seen as a definitive proof of the de Broglie hypothesis of matter waves.

When one goes to first order in the asymptotic expansion, one obtains the WKB [5] approximation of quantum mechanics. This approximation has the advantage of being precise for the hydrogen spectrum, which also follows from the largely heuristic Bohr-Sommerfeld rules. In the eyes of quantum field theory, which uses asymptotic expansions for the evaluation of momentum-space Green functions for particle scattering operators, the powers of $\hbar$ represent the number of loops in the Feynman diagram for the scattering process, so the WKB approximation is regarded as a "one-loop" approximation to the Green function, while the classical scattering process is described by the zero-loop or "tree-level" diagrams.

From a geometrical standpoint, the most important question is that of what happens to the geodesics of spacetime – or even just space – as one goes to successive orders of diffraction. That is, can one still think in terms of transverse trajectories to the momentary wave surfaces, and, if so, how do the diffracted trajectories differ from the classical ones that represent the geometrical optics approximation? Here, one must clearly distinguish between the effect of diffraction on the propagation of phase, which is what the geodesics relate to, and its effect on the propagation of amplitude, which is what produces the diffraction patterns by way of interference patterns. We illustrate the situation that we have in mind in Fig. 11.

[5] The acronym WKB stands for Wentzel-Kramers-Brillouin. Sometimes one finds it referred to as the WKBJ approximation, where the J stands for Jeffreys.

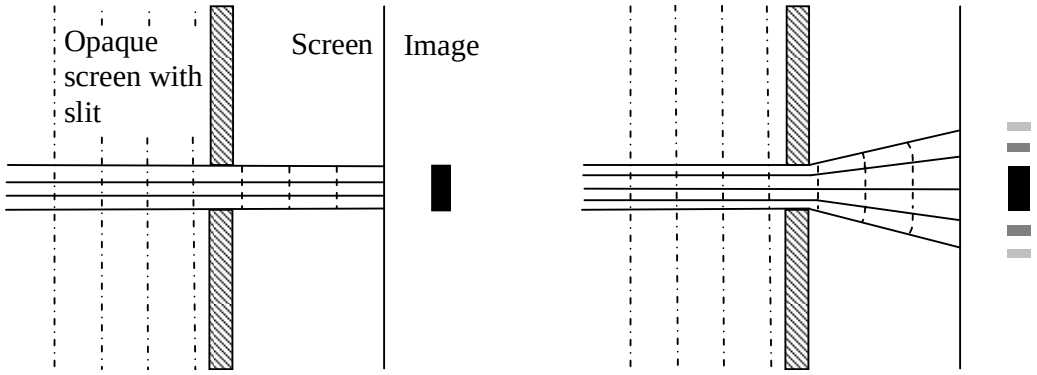

Figure 11. Geodesics in the geometrical optics approximation and in the diffracted case.

We shall conclude our discussion of the topic of diffraction with only these foregoing cursory observations, and the idea that since the spirit of pre-metric electromagnetism is to exhibit the manner by which spacetime geometry emerges from the way that electromagnetic waves propagate in that manifold, it appears that the most promising direction of further exploration in the eyes of understanding quantum physics is the examination of how diffraction affects the structure of geodesics.

References

1. K. Kommerel, *Vorlesungen über Analytische Geometrie des Raumes*, K. F. Koehler Verlag, Leipzig, 1940.
2. M. Visser, "Emergent rainbow spacetimes: two pedagogical examples," Time and Matter 2007 Conference, Lake Bled, Slovenia, arXiv.org: 0712.0810.
3. M. Herzberger, *Strahlenoptik*, Springer, Berlin, 1931.
4. C. Carathéodory, *Geometrische Optik*, Springer, Berlin, 1937.
5. J. L. Synge, *Geometrical optics; an introduction to Hamilton's method*, Cambridge University Press, 1962.
6. M. Kline and I. W. Kay, *Electromagnetic Theory and Geometrical Optics*, Wiley-Interscience, New York, 1965.
7. M. Born and E. Wolf, *Principles of Optics*, Pergamon, Oxford, 1980.
8. R. K. Luneburg, *Mathematical Theory of Optics*, The University of California Press, Berkeley, 1964.
9. B. Baker and E. Copson, *The Mathematical Theory of Huygens' Principle*, Oxford University Press, Oxford, 1950.
10. L. P. Eisenhart, "Contact geometry," Ann. Math. (2) **30** (1928/29), 211-249.
11. V. I. Arnol'd, "Contact Geometry and Wave Propagation," L'Enseignement Mathématique, 1989.
12. A. Erdelyi, *Aymptotic Expansions*, Dover, New York, 1956.
13. V. Guillemin and S. Sternberg, *Geometric Asymptotics*, Mathematical Surveys, no. 14, Am. Math. Soc., Providence, R. I., 1977.
14. V. P. Maslov and M. V. Fedoriuk, *Semi-classical Approximation in Quantum Mechanics*, D. Reidel, Dordrecht, 1981.
15. J. J. Duistermaat, *Fourier Integral Operators*, Lecture notes, Courant Institute, New York University, 1973.

Chapter X

The calculus of variations and electromagnetism

One of the most established, accepted, and far-reaching foundations for theoretical physics – whether one is concerned with mechanics or field theories – takes the form of Hamilton's least action principle. In essence, it says that what characterizes the state of a system that represents its "natural" state is the fact that that the state in question is an extremum of some performance index on the state space that one calls the "action functional." It is a principle that resonates with even the most intuitive and elementary concepts in natural philosophy, such as the Panglossian belief that we live in the best of all possible worlds. Of course, to the scientist that is not a sufficient justification, since Ptolemy's conception of the place of the Earth in the cosmos had a high degree of intuitive appeal in its day, but one still prefers to believe that the most fundamental laws of nature should have that sort of basis in intuitive concepts that one sees manifested in all common phenomena.

To be sure, there is a considerable degree of ambiguity in the concept of "action" that probably accounts for the generality of its applications, just as Newton's second law of motion needs to be made more specific as far as the nature of the forces are concerned if one is to use it in practice. One finds that although the most common definition of action in physics gives it the units of energy-time, nevertheless, one also finds it used to describe things with the units of length, as in the geodesic problem, or time, as in Fermat's principle, which is the basis for Hamiltonian optics. Indeed, in more general optimization problems, the performance index can take almost any imaginable form.

One must also be advised that one of the recurring themes of Twentieth Century quantum physics, besides the ubiquitous role of vacuum polarization and the existence of non-zero ground states, is the idea that the classical least-action principle may represent essentially a first-order approximation to something more involved in nature. That is, in addition to considering the extremal states, which are analogous to the equilibrium or ground states of thermodynamics, one must consider the non-extremal states in the neighborhood of the extremals, as well. This is then analogous to the consideration of fluctuations about equilibrium states in non-equilibrium thermodynamics, which are then referred to in the quantum context as "quantum fluctuations about vacuum ground states." One can even consider global topological issues, such as phase transitions under large perturbations.

As an analogy, it is conceptually useful to view the calculus of variations as being something like "the calculus of infinity variables." Indeed, if the action functional represents a "differentiable" function on an infinite-dimensional "differentiable manifold" of system states then its first variation functional would represent its differential map and the extremal states would be the critical points of the function. Although there is a considerable body of mathematical literature that explore this possibility, generally under the mantle of "global analysis," one finds that even some pure mathematicians agree that this approach is not generally the most useful when it comes to the application of the calculus of variations to more tangible problems than the ones that global analysis poses. Hence, although we shall occasionally employ the

analogy as a heuristic device, we shall present the actual *calculus* in terms of local expressions.

Since the most natural mathematical formalism for the definitions of the calculus of variations seems to be the methods of jet manifolds [1] that we have already introduced to a limited extent, we shall add to the previous definitions with ones that are more specific to the present discussion. Although it might seem more conceptually logical to start with variational mechanics and then generalize to variational field theory, we shall present the topics in the opposite order, since our primary object under scrutiny is the pre-metric formulation of electromagnetic field theory, and we introduce mechanics only in the context of coupling the energy of an electromagnetic field to an "external" current.

One will also observe that the introduction of a metric into the field theory is generally unnecessary from the variational perspective since the only possible role it would play is in the association of field strengths with excitations. Since we have repeatedly emphasized that this sort of duality is best defined by the electromagnetic constitutive law of the medium, it is reassuring that we shall find that the introduction of a spacetime metric does not seem to be unavoidable until we discuss the notion of kinetic energy for point particles. Indeed, even in the case of extended matter, one must introduce a mechanical constitutive law in order to effect the association, and not necessarily a metric.

In the first section of this chapter, we shall discuss the energetics of electromagnetic fields. After that, we formulate the calculus of variations, first for sections of vector bundles in general and then for differential forms and multivector fields, more specifically. We then specialize the general expressions to the cases of electrostatics, magnetostatics, and electromagnetic fields, with particular attention to the case of electromagnetic waves. In that section, we also discuss how one makes a variational formulation of the problem of a charge distribution moving in an external electromagnetic field. Finally, we discuss the variational formulation of geometrical optics using Fermat's principle.

1. Electromagnetic energy. As a prelude to the discussion of the action functionals for electromagnetic fields, we first discuss the way that one associates potential energy with distributions of field sources, as well as the fields themselves, in the cases of electostatics, magnetostatics, and electromagnetic fields, more generally.

a. Electrostatic energy. One immediately recognizes that for static fields the only kind of energy that would be relevant is potential energy, or – more precisely – work. However, one has a choice of two directions to follow in this regard: the work done configuring a distribution of electric charges, electric dipoles, or magnetic dipoles, and the work done establishing the field that they generate as sources.

Of course, one expects that the total work done in one case should equal the total work done in the other, but the difference is that one generally desires that the total work

[1] For a good general treatment of the geometry of jet manifolds, one might confer Saunders [**1**]. Its application to physical field theories has been discussed to a considerable extent by Sardanashvily [**2**] and his colleagues.

done establishing the field be distributed over space as an energy density. Hence, there are more subtle issues to address in that case.

Since we are talking about work, we must immediately restrict our scope by requiring that the path functional in question be path-independent; i.e., the force in question must be conservative. This suggests that we are considering mostly "elementary" systems of charges or dipoles, since it is commonplace for dissipative forces to exist in the "complex" systems found in macroscopic matter, such as the heating of magnetic armatures exposed to time-varying magnetic fields. Indeed, a more comprehensive discussion of the non-conservative case brings one unavoidably to the issues of thermodynamics, which is beyond the scope of our present analysis.

The most elementary case in which one can define the potential energy of an electrostatic system consists of one point charge Q in an external field $\mathbf{E}$, which we assume to be conservative. The force of interaction that Q experiences at a point $x \in \Sigma$ is then $Q\mathbf{E}(x)$, with the usual caveat about the validity of the test-charge/external-field approximation; i.e., one also assumes that the field $\mathbf{E}$ is independent of Q.

The differential change in potential energy when Q goes from x to another point $x + dx$ – along any path $l(s)$ – is then:

$$dU(x) = \text{“}\mathbf{F}(x)\cdot d\mathbf{l}\text{”} = QE_i(x)\, dx^i = QE_i(s)\, v^i(s)\, ds\,. \qquad \text{(X.1)}$$

in which $v^i(s) = dx^i/ds|_s$ is the velocity of the parameterization for the chosen curve.

The reason for the quotation marks is to emphasize the fact that the introduction of the dot product is unnecessary if one regards the force as a 1-form to begin with. Instead of a scalar product of vectors one is then dealing with either the components F_i of a 1-form and the elements of a local coframe field dx^i or the bilinear pairing of a covector field $F_i(s)\, dx^i$ along a curve with a vector field $v^i(s)\, \partial/\partial x_i$ along it.

By integration along any curve γ from x to y, the change in potential energy of the charge-field system is then:

$$\Delta U[x, y] = Q\int_\gamma E(x(s)) = Q\int_0^1 E_i(s)v^i(s)\, ds = Q(U(y) - U(x)). \qquad \text{(X.2)}$$

If one brings Q in "from infinity" to y and expects that $U(y)$ is a finite number then this change in potential energy is finite iff $U(x)$ approaches a finite value as x goes to infinity. (Customarily, one lets this asymptotic value be zero, although any value of a potential function is ambiguous up to an arbitrary additive constant.) It is worth pointing out that if the external field is "constant" in space – assuming that the concept of constancy means something for the space in question – then the potential energy of the charge/field system is infinite no matter where you put Q, since it is proportional to the distance from any point "to infinity."

The second most elementary electrostatic system consists of two point charges Q_1 and Q_2 separated by a distance d. The essential difference between this case and the previous one is due to the fact that one cannot always regard the system as Q_1 in the external field due to Q_2, or vice versa, unless the test-charge/external-field approximation applies; i.e., the presence of Q_1 does not change the field of Q_2, or conversely. Hence, this assumption is more justifiable in the case of linear electrostatics than in the nonlinear case.

From the anti-symmetry of the electrostatic force $\mathbf{F}_{12}(d)$ between the two charges, combined with the change in sign when the line element $d\mathbf{l} = \mathbf{v}(s)\, ds$ for one direction changes to the opposite direction, one has the symmetry of the differential element of work done while changing their separation distance:

$$dF_{12} = \text{“}\mathbf{F}_{12} \cdot d\mathbf{l}\text{”} = F_i(l)\, dl^i = F_i(s)\, v^i(s)\, ds\,. \tag{X.3}$$

If one brings them together at a distance d by starting with one of them “at infinity” then the total work done is:

$$\Delta U_{12}[\infty, d] = \lim_{d_0 \to \infty} [d_0, d] = \int_{\infty}^{d} F_i(l) dl^i = \int_{\infty}^{d} F_i(s) v^i(s)\, ds\,. \tag{X.4}$$

If one uses the Coulomb expression for the force between them then the integration is immediate and one finds that the total potential energy associated with the configuration of Q_1 and Q_2 at a separation distance d is:

$$\Delta U_{12}[\infty, d] = \frac{1}{4\pi\varepsilon_0} \frac{Q_1 Q_2}{d^2}\,. \tag{X.5}$$

One sees that, in principle, the potential energy of the system grows without bound as the separation distance grows vanishingly small.

Implicit in all of this is the assumption that ε_0 is actually a constant and not a function of possibly Q_1, Q_2, and d, or simply the strength of the combined $\mathbf{E}$ field at each point. Hence, the aforementioned calculation has a distinctly linear field-theoretic character to it. In the nonlinear realm the total potential energy would involve a more complicated integration.

In order to extend the elementary case of two point charges at a distance d to N point charges Q_i, $i = 1, \ldots, N$ at points $x_i \in \Sigma$, we essentially use an inductive argument that assumes that if Q_k is any chosen one of the set of charges and $E(x_k)$ represents the field of the remaining $N - 1$ charges at the point x_k then the total potential energy of the system when one brings all of the charges in from infinity is simply:

$$U_{\text{tot}} = Q_k \int_{\infty}^{x_k} E(x(s)) = Q_k\, \phi_k(x_k). \tag{X.6}$$

Of course, this is based on the assumption that any of the possible choices for Q_k will produce the same value of the total energy by this process, which represents a consistency requirement.

When linear superposition is in effect, such as in the case of the Coulomb field, the total potential energy then takes the form of one-half the sum of all pairwise contributions over all pairs of charges:

$$U_{\text{tot}} = \tfrac{1}{2} \sum_{i=1}^{N} Q_i \phi_i(x_i)\,. \tag{X.7}$$

For each i, $\phi_i(x_i)$ represents the electric potential at x_i due to the remaining $N - 1$ charges, excluding Q_i ; the factor of ½ accounts for the fact the sum counts each pair of charges twice.

In order to extend this to a continuous distribution of charges that is described by a charge density $\rho(x)$, one basically replaces the finite charge Q_k with the infinitesimal charge $\rho(x)\mathcal{V}$, the electric potential at x due to the rest of the charge distribution by $\phi(x)$, and the summation by an integration over the support of ρ:

$$U_{\text{tot}} = \tfrac{1}{2}\int_{\text{supp}\,\rho} \rho\phi\,\mathcal{V}\,. \tag{X.8}$$

If one wishes to deduce the manner in which the total potential energy gets distributed through space as an energy density, one now imagines that the region described by supp ρ is dielectric in character. Hence, the electric field $E = -\,d\phi$ produces a response in the form of the electric excitation **D**, which couples to the charge density ρ by way of the fundamental equation:

$$\delta\mathbf{D} = \rho, \tag{X.9}$$

which we also express in the form:

$$d\#\mathbf{D} = \rho\mathcal{V}. \tag{X.10}$$

Hence, the integrand in (X.8) becomes:

$$\rho\phi\mathcal{V} = \phi\, d\#\mathbf{D} = d(\phi\#\mathrm{V}) - d\phi\wedge\#\mathbf{D} = d(\phi\#\mathrm{V}) + E\wedge\#\mathbf{D}\,. \tag{X.11}$$

This makes:

$$U_{\text{tot}} = \tfrac{1}{2}\int_{\partial(\text{supp}\,\rho)} \phi\,\#\mathbf{D} + \tfrac{1}{2}\int_{\text{supp}\,\rho} E\wedge\#\,\mathbf{D} = \tfrac{1}{2}\int_{\text{supp}\,\rho} E\wedge\#\,\mathbf{D}\,, \tag{X.12}$$

if we assume that the support of ρ has vanishing boundary.

From the form of this resulting expression, we see that we can define the energy density in supp ρ to be the 3-form:

$$w_E = \tfrac{1}{2}E\wedge\#\mathbf{D} = \tfrac{1}{2}E(\mathbf{D})\,\mathcal{V}. \tag{X.13}$$

In terms of local components, we have that:

$$E(\mathbf{D}) = E_i D^i = \varepsilon^{ij}E_i\,E_j = \tilde{\varepsilon}_{ij}\,D^i\,D^j\,. \tag{X.14}$$

We can also express this as a Euclidian scalar product:

$$E(\mathbf{D}) = \varepsilon(E, E) = \tilde{\varepsilon}\,(\mathbf{D}, \mathbf{D}). \tag{X.15}$$

Of course, the symbol $\tilde{\varepsilon}$ represents the scalar product on the tangent spaces that is inverse to the one that ε defines on the cotangent spaces.

Although we have defined w_E only on the support of ρ, where, by definition, ρ is non-vanishing, nevertheless, it is conventional to simply generalize the resulting expression for energy density (X.13) to the case of the classical vacuum, in which ρ vanishes, so one has:

$$w_E = \tfrac{1}{2}\,\varepsilon_0 E^2\,\mathcal{V}. \tag{X.16}$$

Apparently, the notion of energy being distributed throughout space according to the square of the field strength, and not simply a total value that was associated with the charge configuration was not always regarded as indisputably meaningful. According to Stratton [**3**], in a 1929 text on electromagnetic theory by Mayer and Weaver the authors argued that distributing the energy of a charge configuration throughout space made as much sense as distributing the beauty of a painting across the canvas! However, Stratton counters that certainly one distributes the energy of deformation in an elastic medium throughout the medium according to the distribution of strain, but he also admits that the continuum-mechanical analogy for electromagnetism was losing favor as a consequence of relativity.

Perhaps the best way to reconcile the distribution of energy with the notion of work done is to recall that each point is associated with an electric dipole moment by way of ε and it takes work to create an electric dipole. Of course, the formation of an electric dipole *in vacuo* only seems relevant for electric fields whose field strength approaches the critical value. However, the fact that ε_0 is not identically zero suggests that even the electric dipoles that eventually polarize into electron-positron pairs have a non-zero vacuum ground state, just as the space of electromagnetic fields does in the form of the zero-point field(s), and for sub-critical field strengths the change in the electric dipole moment is negligible.

b. Magnetostatic energy. The process of deriving the energy density of a static magnetic field is similar to the foregoing, although there is a fundamental difference: Since a current must be distributed over a chain that is at least 1-dimensional, instead of zero-dimensional, one cannot start with the work done moving a point in from infinity. Indeed, since a steady-state current can exist only in a closed circuit – i.e., a 1-cycle – one must first deal with the work done bringing a 1-cycle c_1 in from infinity when it carries a current vector field **I** when it moves in an external static magnetic field **B** distributed throughout space Σ.

First, look at the work done by the Lorentz force $f = \#(\mathbf{I} \wedge \mathbf{B})$ on the current in c_1 when one displaces it by a finite amount. We represent this displacement by the 2-chain $[-l, 0] \times c_1$, by which we really mean the embedding of the formal sum of squares formed from the products $[-l, 0] \times [0, 1]_a$, where a ranges through the intervals that define c_1. We represent the situation in Fig. 12:

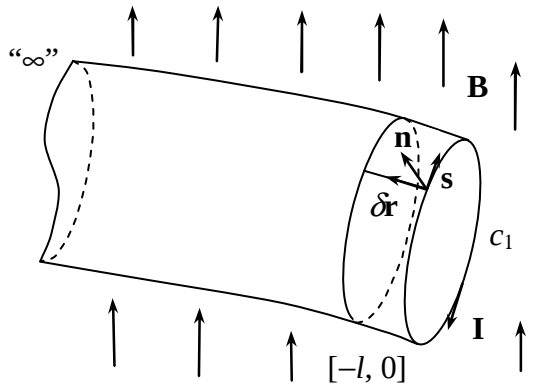

Figure 12. Bringing a current loop in from infinity in an external magnetic field.

First one must recall that actually f is a linear force density on c_1, not a force. Hence, the scalar function $\delta W = f(\delta\mathbf{r})$ on $[-l, 0] \times c_1$, which represents the linear work density due to the displacement $\delta\mathbf{r}$, must be integrated over c_1, as well as $[-l, 0]$. Hence, we must multiply it by the area element $\#\mathbf{n} = i_{\mathbf{n}}\mathcal{V}$. Now:

$$f(\delta\mathbf{r}) = \#(\mathbf{I} \wedge \mathbf{B})(\delta\mathbf{r}) = \mathcal{V}(\mathbf{B} \wedge \mathbf{I} \wedge \delta\mathbf{r}). \tag{X.17}$$

We define the linear frame field $\{\mathbf{s}, d\mathbf{r}, \mathbf{n}\}$ on $[-l, 0] \times c_1$, which we normalize to make:

$$\mathcal{V} = ds \wedge dr \wedge dn, \tag{X.18}$$

where $\{ds, dr, dn\}$ is the reciprocal coframe field.

In this frame:

$$\mathbf{B} = B^s\,\mathbf{s} + B^r\,\delta\mathbf{r} + B^n\,\mathbf{n}, \qquad \mathbf{I} = I\,\mathbf{s}, \tag{X.19}$$

so:

$$f(\delta\mathbf{r}) = IB^n \tag{X.20}$$

and:

$$f(\delta\mathbf{r})\,\#\mathbf{n} = I\,\#\mathbf{B}. \tag{X.21}$$

Hence, one ultimately finds that the total work done on c_1 by the displacement is:

$$W[-l, 0] = \int_{[-l,0]\times c_1} f(\delta\mathbf{r})\,\#\mathbf{n} = I\int_{[-l,0]\times c_1} \#\mathbf{B} = I\,\Phi_B[-l, 0], \tag{X.22}$$

in which $\Phi_B[-l, 0]$ represents the total magnetic flux through the 2-chain $[-l, 0] \times c_1$.

As long as the limit exists, we can then define the total work done bringing c_1 in from infinity as:

$$W_{\text{tot}}[c_1] = I\,\Phi_B[c_1] = I\,\lim_{l\to\infty}\Phi_B[-l, 0]\,. \tag{X.23}$$

Hence, the current couples to the total flux of the external field through the tube to infinity in the same way that charge couples to electric potential. Since the field **B** is

assumed to be conservative, the total flux will be the same for all homotopic tubes to infinity when c_1 is given as an initial loop.

For a configuration of N currents I_i, $i = 1, \ldots, N$, and only under the assumption of linear superposition, one then has:

$$W_{\text{tot}} = \tfrac{1}{2}\sum_{i=1}^{N} I_i \Phi_i[c_i], \tag{X.24}$$

In this expression, $\Phi_i[c_i]$ refers to the total magnetic flux linking the tube to infinity that has the current loop c_i as its initial loop and I_i as its current, as result of the magnetic field that is generated by the remaining $N - 1$ current loops.

We can also express the total work done in terms of the magnetic potential 1-form A ($B = \#\mathbf{B} = dA$) since:

$$\Phi_B[-l, 0] = \int_{[-l,0]\times c_1} dA = \int_{\{0\}\times c_1} A - \int_{\{-l\}\times c_1} A, \tag{X.25}$$

and:

$$\Phi_B[c_1] = \int_{\{0\}\times c_1} A = \mathcal{M}[c_1], \tag{X.26}$$

assuming that A goes to zero at infinity. Hence, the total flux linking the tube to infinity is the magnetomotive force around the loop c_1 .

In order to extend (X.24) to the continuum limit, we first point out that along any curve whose velocity vector field is $\mathbf{v}$, one has:

$$I\, A = I\, A(\mathbf{v})\, ds = A(\mathbf{I})\, ds. \tag{X.27}$$

However, if $\mathbf{I}$ is a current density that is distributed through space then we can also form a 3-form:

$$A(\mathbf{I})\, \mathcal{V} = A \wedge \#\mathbf{I} = A \wedge \#\delta\mathbf{H} = A \wedge d\#\mathbf{H}, \tag{X.28}$$

and since:

$$d(A \wedge \#\mathbf{H}) = dA \wedge \#\mathbf{H} - A \wedge d\#\mathbf{H} = B \wedge \#\mathbf{H} - A \wedge d\#\mathbf{H}, \tag{X.29}$$

when one forms the integral that corresponds to (X.24):

$$W_{\text{tot}} = \tfrac{1}{2}\int_{\text{supp}\,\mathbf{I}} A(\mathbf{I})\, \mathcal{V} = \tfrac{1}{2}\int_{\text{supp}\,\mathbf{I}} A \wedge \#\mathbf{I}, \tag{X.30}$$

an integration by parts puts this into the form:

$$W_{\text{tot}} = \tfrac{1}{2}\int_{\text{supp}\,\mathbf{I}} B \wedge \#\mathbf{H}. \tag{X.31}$$

Hence, the magnetostatic energy density takes the form:

$$w_B = \tfrac{1}{2} B \wedge \#\mathbf{H} = \tfrac{1}{2} B(\mathbf{H})\, \mathcal{V}, \tag{X.32}$$

and in component form one has:

$$B(\mathbf{H}) = B_i H^i = \tilde{\mu}^{ij} B_i B_j = \mu_{ij} H^i H^j, \tag{X.33}$$

which we can also express as:

$$B(\mathbf{H}) = \tilde{\mu}(B, B) = \mu(\mathbf{H}, \mathbf{H}). \tag{X.34}$$

One again, we see that the potential energy density takes the form of a quadratic form in either the field strengths or excitations whose components are defined by the magnetic constitutive law in effect.

c. Electromagnetic energy. Since the energy densities for both the electric field strength 1-form E and the magnetic field strength 2-form B are defined by quadratic expressions involving their respective constitutive laws, we naturally wish to examine the character of the quadratic expression $\kappa(F, F)$ when κ is the electromagnetic constitutive law for the four-dimensional spacetime manifold M, and F is the electromagnetic field strength 2-form.

We first assume that $T(M) = L(M) \oplus \Sigma(M)$ is a time-space splitting of the tangent bundle and that $\Lambda^2 = \Lambda^2_{\text{Re}} \oplus \Lambda^2_{\text{Im}}$ is the corresponding time-space splitting of the bundle of 2-forms. We can then express F as $dt \wedge E + \#\mathbf{B}$, where $dt \wedge E \in \Lambda^2_{\text{Re}}$ and $\#\mathbf{B} \in \Lambda^2_{\text{Im}}$. This allows us to write:

$$\kappa(F, F) = \kappa(dt \wedge E, dt \wedge E) + \kappa(dt \wedge E, \#\mathbf{B}) + \kappa(\#\mathbf{B}, dt \wedge E) + \kappa(\#\mathbf{B}, \#\mathbf{B}) . \tag{X.35}$$

If we revert the six-dimensional basis $\{b^i, \#\mathbf{b}_i, i = 1, 2, 3\}$ for Λ^2 so that κ can be expressed as a 6×6 block matrix:

$$\kappa^{IJ} = \left[\begin{array}{c|c} -\varepsilon^{ij} & \gamma^j_i \\ \hline \hat{\gamma}^i_j & \tilde{\mu}_{ij} \end{array}\right] \tag{X.36}$$

then:

$$\kappa(F, F) = -\varepsilon(E, E) + (\gamma + \hat{\gamma})(E, \mathbf{B}) + \tilde{\mu}(\mathbf{B}, \mathbf{B}) , \tag{X.37}$$

in which $\varepsilon(E, E)$ and $\tilde{\mu}(\mathbf{B}, \mathbf{B})$ represent the same quadratic expressions that we discussed in the static case while the middle term has the local expression:

$$(\gamma + \hat{\gamma})(E, \mathbf{B}) = (\gamma^i_j + \hat{\gamma}^j_i) E_i B^i . \tag{X.38}$$

One can similarly decompose the expression $F \wedge \#\mathfrak{h}$, with $\mathfrak{h} = \partial_t \wedge \mathbf{D} + \#^{-1} H$, into:

$$F \wedge \#\mathfrak{h} = dt \wedge E \wedge \#(\partial_t \wedge \mathbf{D}) + dt \wedge E \wedge H + \#\mathbf{B} \wedge \#(\partial_t \wedge \mathbf{D}) + \#\mathbf{B} \wedge H, \tag{X.38}$$

which then becomes:

$$F \wedge \#\mathfrak{h} = -[E(\mathbf{D}) - (\gamma + \hat{\gamma}^T)(E, \mathbf{B}) - H(\mathbf{B})]\, \mathcal{V}, \tag{X.39}$$

which equals $\kappa(F, F)\, \mathcal{V}$.

When the medium in question is not in a state of motion relative to the measurer/observer, and does not exhibit optical activity, the cross-term in the sum vanishes, and the remaining expression is:

$$F \wedge \#\mathfrak{h} = -\,[E(\mathbf{D}) - H(\mathbf{B})]\,\mathcal{V}. \tag{X.40}$$

For an isotropic medium, this takes the form:

$$F \wedge \#\mathfrak{h} = -\,[\varepsilon E^2 - (1/\mu)B^2]\,\mathcal{V}. \tag{X.41}$$

We then recognize the expressions in (X.39) as minus two times the usual electromagnetic Lagrangian density 4-form:

$$\mathcal{L}_{\text{em}} = -\tfrac{1}{2} F \wedge \#\mathfrak{h} = -\tfrac{1}{4} F_{\mu\nu} H^{\mu\nu}\, dx^0 \wedge dx^1 \wedge dx^2 \wedge dx^3. \tag{X.42}$$

Interestingly, one can also account for the usual electromagnetic Hamiltonian using the quadratic form that is defined by κ. However, one must form the "complex conjugate" of F, namely:

$$\bar{F} = dt \wedge E - \#\mathbf{B}. \tag{X.42}$$

Note that, this operation of conjugation is meaningful only for a given choice of time-space splitting. Indeed, this is consistent with the idea that energy itself is not a relativistic invariant except in the context of rest energy.

We now find that:

$$\kappa(F, \bar{F}) = -\,[\varepsilon(E, E) + \tilde{\mu}(\mathbf{B}, \mathbf{B})\,] + (\gamma - \hat{\gamma}^T)(E, \mathbf{B}), \tag{X.43}$$

and when the last term vanishes, we see that we are again dealing with minus two times the usual electromagnetic field Hamiltonian.

2. The calculus of variations in terms of vector bundles. Although we have been assuming a certain familiarity with variational field theory all along, since we are mostly concerned with fields that take the specific form of multivector fields and exterior differential forms, it is necessary to show how one can define the basic notions and carry out the basic computations of the variational calculus using such fields, in particular.

a. Extremal fields. When one is given a particular class of fields, the starting point for the calculus of variations is the definition of an action functional on the space of fields in question. When the fields are all sections $\phi\colon M \to E$ of a vector bundle over M it is somewhat convenient that the space of fields $\Gamma(E)$ is a linear space, so the action functional becomes a function on that vector space. However, one is nonetheless dealing with an infinite-dimensional vector space that is not necessarily separable, so putting a topology and a differentiable manifold structure on the space becomes problematic

enough that it is usually only heuristically useful to imagine that the action functional is a differentiable function on an infinite-dimensional manifold.

A *first-order action functional* $S[.]$ on $\Gamma(E)$ is defined by a first-order *Lagrangian density* $\mathcal{L}$: $J^1(M, E) \to \mathbb{R}$, which is at least C^2, and such a functional takes the form:

$$S[\phi] = \int_M \mathcal{L}(j^1\phi)\mathcal{V}. \tag{X.44}$$

Let us briefly recall the terminology that we introduced in Chapter VI concerning the geometry of jet manifolds in the form that it takes in the present context:

The notation $J^1(M, E)$ refers to the manifold of 1-jets of sections of the vector bundle $E \to M$, and it has local coordinate systems that take the form $(x^\mu, \phi^A, \phi^A{}_\mu)$. In the present case, the projection $\pi_{1,0} : J^1(M, E) \to E$, $j^1_x\phi \mapsto \phi$ defines a vector bundle over E in its own right. The vector spaces that comprise its fibers have the $\phi^A{}_\mu$ for coordinates, so they are nN-dimensional ($n = \dim M$, $N = \text{rank } E$).

A section s of the projection $J^1(M, E) \to M$ takes the local form $(x^\mu, s^A(x), s^A{}_\mu(x))$. It is integrable iff s is the 1-jet prolongation $j^1\phi$ of some section ϕ of $E \to M$, in which case, it takes the local form $(x^\mu, s^A(x), s^A{}_{,\mu}(x))$. Hence, s is integrable iff one locally has:

$$s^A{}_\mu(x) = s^A{}_{,\mu}(x). \tag{X.45}$$

A first-order field Lagrangian can then be expressed in the local form $\mathcal{L} = \mathcal{L}(x^\mu, \phi^A, \phi^A{}_{,\mu})$ that is so established in the physics literature.

One now sees that in order for the integral in (X.44) to converge one must restrict either the class of sections that one uses or the topology of the manifold M. Indeed, requiring the sections to vanish smoothly "at infinity" is equivalent to saying that if "infinity" is a point that one uses to compactify M then such a section can be smoothly extended to the one-point compactification of M by setting its value at infinity equal to zero.

The most-discussed one-point compactification in elementary mathematics is probably the compactification of $\mathbb{R}^2$ into a 2-sphere by stereographic projection. However, we hasten to point out that the way that one compactifies a topological space is usually determined by the type of *geometry* that one is doing, in addition to topology. One-point compactification mostly relates to conformal Euclidian geometry, in which spheres play the same fundamental role as points do in affine geometry. In projective geometry, one compactifies the affine n-plane by the addition of a "hyperplane at infinity," and in conformal Lorentzian geometry, one adds a "light-cone at infinity."

We shall tacitly assume that suitable conditions are imposed on M or the field ϕ such that the action functional always exists.

The *first variation* δS of the action functional amounts to a linear functional on the (infinite-dimensional) vector space of vector fields on $J^1(M, E)$ that can be represented in the form of an integrated Lie derivative:

$$\delta S[X] = \int_M L_X(\mathcal{L}\mathcal{V}) = \int_M [(i_X d\mathcal{L})\mathcal{V} + \mathcal{L}\, d i_X \mathcal{V}]. \tag{X.47}$$

A (substantial) *variation* of a section ϕ is a vertical vector field $\delta\phi$ on the manifold E. That is, one varies values the of the field without varying the points of M that they are associated with. If a local trivialization $U \times V$ of E over $U \subset M$ has the local coordinates (x^μ, ϕ^A) then a general vector field on E will look like:

$$X = X^\mu \frac{\partial}{\partial x^\mu} + X^A \frac{\partial}{\partial \phi^A}, \tag{X.48}$$

in which the component functions X^μ and X^A are all smooth functions on E; i.e., $X^\mu = X^\mu(x^\mu, \phi^A)$, $X^A = X^A(x^\mu, \phi^A)$.

Hence, a vertical vector field, such as $\delta\phi$, will look like:

$$\delta\phi = \delta\phi^A \frac{\partial}{\partial \phi^A}. \tag{X.49}$$

One prolongs $\delta\phi$ to a vector field $\delta^1\phi$ on $J^1(M, E)$ by differentiation. Locally, it looks like:

$$\delta^1\phi = \delta\phi^A \frac{\partial}{\partial \phi^A} + \frac{\partial(\delta\phi^A)}{\partial x^\mu} \frac{\partial}{\partial \phi^A{}_\mu}. \tag{X.50}$$

The fundamental problem of the calculus of variations is to find the *extremal sections* [2] of the first variation functional under such first prolongations of substantial variations of the fields. This basically means that one is looking for all $\phi \in \Gamma(E)$ such that $\delta S[\delta^1\phi]$ vanishes for all $\delta^1\phi$ that represent first prolongations of variations of ϕ. This means:

$$0 = \delta S[\delta^1\phi] = \int_M (i_{\delta^1\phi} d\mathcal{L})\mathcal{V} \tag{X.51}$$

for all vertical $\delta\phi$ since the second term in the integrand in (X.47) vanishes for all prolongations of vertical vector fields.

In order to perform the traditional "integration by parts," we revert to the local formulation. One has:

$$d\mathcal{L} = \frac{\partial \mathcal{L}}{\partial x^\mu} dx^\mu + \frac{\partial \mathcal{L}}{\partial \phi^A} d\phi^A + \frac{\partial \mathcal{L}}{\partial \phi^A{}_\mu} d\phi^A{}_\mu, \tag{X.52}$$

so:

$$i_{\delta^1\phi} d\mathcal{L} = \frac{\partial \mathcal{L}}{\partial \phi^A} \delta\phi^A + \frac{\partial \mathcal{L}}{\partial \phi^A{}_\mu} \delta\phi^A{}_{,\mu} = \frac{\delta \mathcal{L}}{\delta \phi^A} \delta\phi^A + \frac{\partial}{\partial x^\mu}\left(\frac{\partial \mathcal{L}}{\partial \phi^A} \delta\phi^A\right), \tag{X.53}$$

[2] We avoid the term "extremal fields" because this has a completely distinct, but established, meaning in Hamilton-Jacobi theory, namely, geodesic congruences.

in which we have introduced the *variational derivative* of $\mathcal{L}$ with respect to ϕ:

$$\frac{\delta \mathcal{L}}{\delta \phi^A} = \frac{\partial \mathcal{L}}{\partial \phi^A} - \frac{\partial}{\partial x^\mu}\frac{\partial \mathcal{L}}{\partial \phi^A{}_\mu}. \tag{X.54}$$

One can then say that:

$$\delta S[\delta^1 \phi] = \int_M \left(\frac{\delta \mathcal{L}}{\delta \phi^A} \delta \phi^A \right) \mathcal{V} + \int_{\partial M} \#\left(\mathbf{\Pi}_A \delta \phi^A \right), \tag{X.55}$$

in which we have introduced the *generalized momentum density* vector fields:

$$\mathbf{\Pi}_A = \frac{\partial \mathcal{L}}{\partial \phi^A{}_\mu}\frac{\partial}{\partial x^\mu}. \tag{X.56}$$

When one restricts oneself to either variations of ϕ that vanish on the boundary of M or variations that satisfy the transversality condition $\mathbf{\Pi}^A \delta \phi_A = 0$, the necessary and sufficient condition for an extremum is that:

$$\frac{\delta \mathcal{L}}{\delta \phi^A} = 0, \tag{X.57}$$

which represents the *Euler-Lagrange* equations.

These are generally N nonlinear partial differential equations for the components ϕ^A of the extremal fields. Hence, in order specify such fields uniquely one will generally have to pose boundary-value problems, in the elliptic case, or Cauchy problems, in the hyperbolic case.

If one introduces the *generalized force densities* f_A , along with the aforementioned generalized momentum density vector fields Π_A , on E by way of:

$$f_A = \frac{\partial \mathcal{L}}{\partial \phi^A}, \tag{X.58}$$

then the Euler-Lagrange equations can be put into the form:

$$\delta \Pi_A = f_A. \tag{X.59}$$

b. Case of differential forms. Since a k-vector field on a manifold M is a section of the vector bundle $\Lambda_k M \to M$ and a differential k-form on M is a section of the vector bundle $\Lambda^k M \to M$, one sees that all one needs to do in order to focus on the variational formulation of electromagnetism is to specialize the general results that were derived above.

One thing that needs to be addressed immediately is the fact that when considers the bundle $J^1(\Lambda^k)$ of 1-jets of exterior differential k-forms on M one sees that it is the differential of any section ϕ: $M \to \Lambda^k M$ that seems to factor most fundamentally in the

calculus of variations, whereas in the calculus of exterior differential forms, it is the completely antisymmetrized differential – viz., the exterior derivative – that plays the fundamental role. This suggests that one might wish to define a notion of "exterior 1-jet" in the predictable way – i.e., equivalence classes of k-forms that are defined in some neighborhood of a point $x \in M$ and have the same values at k as k-forms, along with the same exterior derivatives – but one eventually comes to see that this is basically unnecessary.

The real issue is differentiation with respect to fiber coordinates, and one finds that since components of k-forms are completely antisymmetrized sums a differentiation with respect to field components invariably singles out just one member of the sum. For instance, if $F_{\mu\nu} = A_{\mu,\nu} - A_{\nu,\mu}$ then:

$$\frac{\partial F_{\mu\nu}}{\partial A_{\mu,\nu}} = \delta^{\mu}_{\mu}\delta^{\nu}_{\nu} - \delta^{\nu}_{\mu}\delta^{\mu}_{\nu} = 1, \tag{X.60}$$

since $\mu \neq \nu$, by anti-symmetry.

As a consequence, one has, in particular:

$$\frac{\partial \mathcal{L}}{\partial F_{\mu\nu}} = \frac{\partial \mathcal{L}}{\partial A_{\mu\nu}}. \tag{X.61}$$

Hence, whether one differentiates with respect to $F_{\mu\nu}$ or $A_{\mu\nu}$ is irrelevant.

We shall primarily be concerned with 1-forms A as the fundamental fields, so the relevant jet bundle is $J^1(\Lambda^1) \to M$. Local coordinate charts on the total space $J^1(\Lambda^1)$ take the form $(x^{\mu}, A_{\mu}, A_{\mu\nu})$. A section s: $M \to J^1(\Lambda^1)$ of this bundle then looks like $(x^{\mu}, A_{\mu}(x^{\mu}), A_{\mu\nu}(x^{\mu}))$ and it is integrable iff $s = j^1A$, which implies that:

$$A_{\mu\nu} = A_{\mu,\nu}. \tag{X.62}$$

A Lagrangian density on $J^1(\Lambda^1)$ locally looks like $\mathcal{L}(x^{\mu}, A_{\mu}, A_{\mu\nu})$ and for the prolongation of a 1-form A it takes the form $\mathcal{L}(x^{\mu}, A_{\mu}(x), A_{\mu,\nu}(x))$. The Euler-Lagrange equations then are expressed as:

$$0 = \frac{\delta \mathcal{L}}{\delta A_{\mu}} = \frac{\partial \mathcal{L}}{\partial A_{\mu}} - \frac{\partial}{\partial x^{\nu}}\frac{\partial \mathcal{L}}{\partial A_{\mu\nu}}. \tag{X.63}$$

By introducing the notations:

$$J^{\mu} = \frac{\partial \mathcal{L}}{\partial A_{\mu}}, \qquad \mathfrak{h}^{\mu\nu} = \frac{\partial \mathcal{L}}{\partial A_{\mu\nu}} \tag{X.64}$$

the fundamental equations for the field A_{μ} become:

$$\partial_\nu \mathfrak{h}^{\mu\nu} = J^\mu. \tag{X.65}$$

From (X.61), we see that it amounts to the same thing to define $\mathcal{L}$ as a function of $F_{\mu\nu}$, instead of $A_{\mu\nu}$. Hence, if we combine the integrability condition for $F_{\mu\nu}$ with the Euler-Lagrange equations in the latter form, and the second of equations (X.64), we get the system:

$$F_{\mu\nu} = A_{\mu,\nu} - A_{\nu,\mu}, \qquad \partial_\nu \mathfrak{h}^{\mu\nu} = J^\mu, \qquad \mathfrak{h}^{\mu\nu} = \frac{\partial \mathcal{L}}{\partial F_{\mu\nu}}. \tag{X.66}$$

We can just as well express these equations in index-free form as equations in the 2-form F, the 1-form A, the bivector field $\mathfrak{h}$, and the vector field $\mathbf{J}$:

$$F = dA, \qquad \partial\mathfrak{h} = \mathbf{J}, \qquad \mathfrak{h} = \frac{\partial \mathcal{L}}{\partial F}. \tag{X.67}$$

As long as one can regard the last of these equations as essentially the constitutive law that relates F to $\mathfrak{h}$, these equations are formally identical with the pre-metric Maxwell equations. In the following sections, we shall pursue the extent to which this is a valid assumption.

3. Variational formulation of electromagnetic problems. Since the main issues in the variational formulation of the field equations for electromagnetism are the definition of a Lagrangian density and its relationship to the constitutive laws of the medium, we shall examine the specific form these matters take in the cases of static electric, static magnetic, and dynamic electromagnetic fields. We shall then discuss the variational formulation of the mechanical problem of determining the motion of charged mass distributions in external electromagnetic fields. (For some other treatments of the same basic topics, one might confer [**4**-**6**].)

a. Electrostatics. Whether one is dealing with linear or nonlinear electrostatics, one generally begins by defining a Lagrangian that is quadratic in the field strength $E \in \Lambda^1\Sigma$:

$$\mathcal{L}(x^i, E_i) = \tfrac{1}{2}\overline{\varepsilon}(E, E) = \tfrac{1}{2}\overline{\varepsilon}^{ij} E_i E_j \ . \tag{X.68}$$

For an integrable section, one has:

$$E = d\phi, \qquad (E_i = \phi_{,i}) \tag{X.69}$$

for some smooth function ϕ on Σ that is unique only up to an additive constant.

Note that even though we originally developed the formalism of calculus of variations by defining a Lagrangian density as a differentiable function $\mathcal{L}$ on $J^1(E)$, we see that as

long as we deal with gauge-invariant field Lagrangians – i.e., ones that do not depend upon ϕ explicitly – we can just as well define $\mathcal{L}$ on $\Lambda^1\Sigma$ itself.

When one derives the vector field **D** that is canonically conjugate to *E*, one obtains the electrostatic constitutive law in the form:

$$\#\mathbf{D} = \frac{\partial \mathcal{L}}{\partial E} = \#(\varepsilon^{ij} E_j \partial_i), \tag{X.70}$$

in which:

$$\varepsilon^{ij}(x^i, E_j) = \bar{\varepsilon}^{ij} + \tfrac{1}{2}\frac{\partial \bar{\varepsilon}^{kj}}{\partial E_i} E_k \,. \tag{X.71}$$

Hence, one sees that when one is concerned with nonlinear electrostatics there is a fundamental difference between the quadratic tensor field $\bar{\varepsilon}$ and the actual electrostatic constitutive tensor field ε; of course, they will coincide in the case of linear electrostatics.

The field equations that result from this choice of Lagrangian density are then:

$$E = d\phi, \qquad \delta\mathbf{D} = 0, \qquad \mathbf{D} = \varepsilon(E), \tag{X.72}$$

which can be combined into a single equation for ϕ:

$$\Delta_\varepsilon \phi = (\delta \cdot \varepsilon \cdot d)\phi = 0. \tag{X.73}$$

It is important to remember that we are really varying ϕ, not *E*, even though ϕ no longer figures explicitly in the Lagrangian density. Had we merely varied *E* directly, the only resulting field equation would be simply **D** = 0, which is a trivial outcome.

In order to include the source of the field, one must add another term to the Lagrangian density that couples the charge density ρ to the electric potential ϕ by way of

$$\mathcal{L}(x^i, \phi, \phi_i) = \tfrac{1}{2} E \wedge \#\mathbf{D} + \rho\phi\, \mathcal{V} = [\tfrac{1}{2}\bar{\varepsilon}(E,E) + \rho\phi]\, \mathcal{V} \,. \tag{X.74}$$

This makes the field equations take the form:

$$E = d\phi, \qquad \delta\mathbf{D} = \mathbf{J}, \qquad \mathbf{D} = \varepsilon(E), \tag{X.75}$$

or:

$$\Delta_\varepsilon \phi = (\delta \cdot \varepsilon \cdot d)\phi = \mathbf{J}, \tag{X.76}$$

which is of the generalized Poisson type.

b. Magnetostatics. Although the variational formulation of magnetostatics is analogous to that of electrostatics, nonetheless, there are significant differences that must be addressed. For one thing, the fundamental field *A* is a 1-form, not a 0-form, so the field strength $B = dA$ that it defines is a 2-form, not a 1-form. However, because the vector space A^2V of 2-forms over a three-dimensional vector space *V* is isomorphic to the vector space *V*, by means of Poincaré duality for a choice of volume element, it is more

conventional in classical magnetism to represent the 2-form $B = \#\mathbf{B}$ by the components B^i of the corresponding vector field **B**, and similarly, the bivector field $\mathbf{H} = \#^{-1}H$ gets represented by the components H_i of the 1-form H.

This also means that the magnetic constitutive law $\tilde{\mu} = \mu^{-1}$: $\Lambda^2\Sigma \to \Lambda_2\Sigma$ for the medium intrinsically couples 2-forms to bivector fields, although it is usually expressed by the matrix of a linear isomorphism $\tilde{\mu}: \Lambda_1\Sigma \to \Lambda^1\Sigma$ between their dual spaces. In effect, one is using:

$$\tilde{\mu}(B, B) = \tilde{\mu}(\#\mathbf{B}, \#\mathbf{B}) = (\#\tilde{\mu}\#)(\mathbf{B}, \mathbf{B}), \tag{X.77}$$

so the four-index components $\tilde{\mu}^{ijkl}$ of $\tilde{\mu}$ go to the two-index components of $\#\tilde{\mu}\#$:

$$\tilde{\mu}_{ij} = \varepsilon_{ikl}\varepsilon_{jmn}\tilde{\mu}^{klmn}. \tag{X.78}$$

From this, one defines the sourceless magnetostatic field Lagrangian density for the potential 1-form A by a quadratic form on B that then reverts to a quadratic form on **B**:

$$\mathcal{L}(x^i, B_{ij}) = \tfrac{1}{2}\bar{\mu}(\mathbf{B}, \mathbf{B}) = \tfrac{1}{2}\bar{\mu}_{ij}B^iB^j = \tfrac{1}{2}\bar{\mu}(B, B) = \tfrac{1}{4}\bar{\mu}^{ijkl}B_{ij}B_{kl}, \tag{X.79}$$

in which:

$$B^i = \tfrac{1}{2}\varepsilon^{ijk}B_{jk} = \tfrac{1}{2}\varepsilon^{ijk}(A_{j,k} - A_{k,j}) = \varepsilon^{ijk}A_{j,k} = (\nabla \times \mathbf{A})^i. \tag{X.80}$$

As before, we must emphasize that the second tensor field $\bar{\mu}$ that appears in this expression does not have to represent the magnetostatic constitutive law of the medium directly. Similarly, one sees that in the absence of sources, the Lagrangian density is independent of A – i.e., gauge-invariant – and can thus be defined as a differentiable function on $\Lambda^2\Sigma$, instead of $J^1(\Lambda^2\Sigma)$.

In order to derive the field equations for magnetostatics from this Lagrangian density, one needs only to take (X.78) into account in order to convert the partial derivative of $\mathcal{L}$ with respect to A_{ij} into a partial derivative with respect to B^i:

$$H^{ij} = \frac{\partial\mathcal{L}}{\partial A_{ij}} = \frac{\partial B_{mn}}{\partial A_{ij}}\frac{\partial B^k}{\partial B_{mn}}\frac{\partial\mathcal{L}}{\partial B^k} = \varepsilon^{ijk}\frac{\partial\mathcal{L}}{\partial B^k} = \varepsilon^{ijk}\tilde{\mu}_{kl}B^l = (\#^{-1}H)^{ij}, \tag{X.81}$$

in which:

$$H = \tilde{\mu}_{ij}B^j dx^i \tag{X.82}$$

represents the magnetic excitation 1-form that is Poincaré dual to the bivector field $\mathbf{H} = \tfrac{1}{2}H^{ij}\,\partial_i \wedge \partial_j$ and:

$$\tilde{\mu}_{ij} = \bar{\mu}_{ij} + \frac{\partial\bar{\mu}_{ij}}{\partial B^k}B^k, \tag{X.83}$$

which, as one sees, agrees with $\bar{\mu}_{ij}$ only in the case of a linear constitutive law.

The sourceless field equations for A that one derives from the given Lagrangian density take the form:

$$B = dA, \qquad \delta\mathbf{H} = 0, \qquad \mathbf{H} = \tilde{\mu}(B), \tag{X.84}$$

if one uses the 2-form B and the bivector field $\mathbf{H}$, or:

$$\mathbf{B} = \delta\#^{-1}A, \qquad dH = 0, \qquad H = \tilde{\mu}(\mathbf{B}), \tag{X.85}$$

if one uses the vector field $\mathbf{B}$ and the 1-form H.

One can also combine these equations into a single second-order system of partial differential equations for A:

$$\Delta_\mu A = (\delta \cdot \tilde{\mu} \cdot d)A = 0, \tag{X.86}$$

which is of a generalized Laplace type.

If one includes the coupling of the energy in the field $\mathbf{H}$ to the energy in the source current density $\mathbf{I}$ then one must add another term to the Lagrangian density to account for it:

$$\mathcal{L}(x^\mu, A_\mu, B_{\mu\nu}) = \tfrac{1}{2}\overline{\mu}(B,B) + A(\mathbf{I}) = \tfrac{1}{2}\overline{\mu}_{ij}B^iB^j + A_iI^i; \tag{X.87}$$

one can also express the field-source coupling term in the form:

$$A \wedge \#\mathbf{I} = A(\mathbf{I})\,\mathcal{V}. \tag{X.88}$$

The field equations now include a contribution from:

$$\frac{\partial\mathcal{L}}{\partial A} = \mathbf{I}, \tag{X.89}$$

namely:

$$B = dA, \qquad \delta\mathbf{H} = \mathbf{I}, \qquad \mathbf{H} = \tilde{\mu}(B), \tag{X.90}$$

or:

$$\mathbf{B} = \delta\mathbf{A}, \qquad dH = \#\mathbf{I}, \qquad H = \tilde{\mu}(\mathbf{B}). \tag{X.91}$$

The first set consolidates into a second-order equation for A:

$$\Delta_\mu A = \mathbf{I} \tag{X.92}$$

that has a generalized Poisson type.

c. Electromagnetism. When one puts electricity and magnetism together into a four-dimensional object, namely a 2-form F, the predictable form for a sourceless quadratic Lagrangian density is:

$$\mathcal{L}(x^\mu, F_{\mu\nu}) = \tfrac{1}{2}\overline{\kappa}(F,F) = \tfrac{1}{4}\overline{\kappa}^{\kappa\lambda\mu\nu}F_{\kappa\lambda}F_{\mu\nu}. \tag{X.93}$$

The excitation bivector field $\mathfrak{H}$ that is canonically conjugate to the 2-form F is then:

$$\mathfrak{h} = \kappa(F), \qquad (\mathfrak{h}^{\kappa\lambda} = \tfrac{1}{2}\,\kappa^{\kappa\lambda\mu\nu}F_{\mu\nu}) \tag{X.94}$$

in which:

$$\kappa^{\kappa\lambda\mu\nu} = \bar{\kappa}^{\kappa\lambda\mu\nu} + \frac{\partial \bar{\kappa}^{\kappa\lambda\mu\nu}}{\partial F_{\alpha\beta}} F_{\alpha\beta}\,. \tag{X.95}$$

As usual, this means that κ coincides with $\bar{\kappa}$ only in the case of a linear constitutive law.

The resulting field equations are then:

$$B = dA, \qquad \delta\mathfrak{h} = 0, \qquad \mathfrak{h} = \kappa(F), \tag{X.96}$$

which consolidate into a second-order equation for A:

$$\Box_\kappa A = (\delta \cdot \kappa \cdot d)A = 0 \tag{X.97}$$

that represents a generalized wave equation.

The coupling of a source current $\mathbf{J} = \rho\partial_t + \mathbf{I}$ to the field A proceeds in a manner that is analogous to the previous discussions in the electrostatic and magnetostatic cases. That is, the extra term in the Lagrangian that couples the energy of the source current to the energy of the field takes the form:

$$A \wedge \#\mathbf{J} = A(\mathbf{J})\,\mathcal{V} = \rho\phi\,\mathcal{V} + A_s \wedge \#\mathbf{I} = (\rho\phi + A_s(\mathbf{I}))\mathcal{V}, \tag{X.98}$$

which is then simply the sum of the electrostatic and magnetostatic terms.

The generalized force that appears in the field equations as a forcing term is then the charge flux 3-form:

$$\frac{\partial(A \wedge \#\mathbf{J})}{\partial A} = \#\mathbf{J}\,, \tag{X.99}$$

which makes the field equations take the form:

$$F = dA, \qquad \delta\mathfrak{h} = \mathbf{J}, \qquad \mathfrak{h} = \kappa(F), \tag{X.100}$$

or, in second-order form:

$$\Box_\kappa A = \mathbf{J}, \tag{X.101}$$

which is then a forced wave equation.

4. Motion of a charge in an electromagnetic field. The problem of determining the motion of a massive charge distribution ρ in the presence of an external electromagnetic field F brings one into the domain of mechanics rather than field theory. However, the formulation of the calculus of variations in terms of jet manifolds is sufficiently general in its application that one can adapt the methodology that was

presented above in the context of field theory to the context of mechanics. Indeed, one can treat both pointlike matter and continuously-extended matter within that formalism.

The basis for the external field approximation in this case is the assumption that the presence of ρ, which necessarily represents the source for a field of its own, does not affect the field F, except by linear superposition. Some situations in which this would break down might take the form of either nonlinear superposition, such as the combined field having a super-critical field strength in the eyes of the polarization of the medium or the cases in which the source **J** of the field F changes state, in some sense (e.g., position, shape, total charge) as a result of the presence of ρ.

First, we shall formulate the mechanical model for pointlike charged matter and then we shall discuss the issues that are associated extending the formalism to extended charged matter.

When a charge distribution is pointlike its support is along a smooth curve x: $[0, 1] \to M$. For the moment, rather than representing the distribution as a Dirac delta distribution centered on the points of x, we shall think of it as simply a real number q that is associated with γ; this also reduces the scope of the discussion to time-invariant charges. Similarly, we think of the rest mass of the point-particle as a positive real number m_0, rather than another delta distribution, and restrict the scope to the time-invariant case.

a. Variational formulation of point mechanics. Just as 1-jets of sections of vector bundles over M were the fundamental objects in the variational formulation of field theory, the 1-jets of curves in M are the fundamental objects in the variational formulation of point mechanics. However, one should note that there is certain simplification associated with point matter in the fact that the definition of the 1-jet $j^1_\tau x$ of a curve $x(\tau)$ in M at a point $x \in M$ reads the same way that the definition of the tangent vector to the curve at that point did, namely, the equivalence class of all differentiable curves though x that have the same first derivative – i.e., velocity – at that point.

We denote the manifold of all 1-jets of differentiable curves in M by $J^1(\mathbb{R}, M)$. Its source projection is $J^1(\mathbb{R}, M) \to \mathbb{R}$, $j^1_\tau x \mapsto \tau$, its target projection is $J^1(\mathbb{R}, M) \to M$, $j^1_\tau x \mapsto x$, and the projection $J^1(\mathbb{R}, M) \to \mathbb{R}\times M$, $j^1_\tau x \mapsto (\tau, x)$ plays a role that is analogous to the projection $J^1E \to E$ in the case of vector bundle. Indeed, one can regard a curve as a section of the (trivial) projection $\mathbb{R}\times M \to \mathbb{R}$, $(\tau, x) \mapsto \tau$ and then regard mechanics in general as a special case of a field theory.

Since the manifold $\mathbb{R}$ is contractible, the manifold $J^1(\mathbb{R}, M)$ is diffeomorphic to $\mathbb{R}\times T(M)$. Hence, a local coordinate chart on $J^1(\mathbb{R}, M)$ takes the form (τ, x^μ, v^μ), so a local section s: $\mathbb{R} \to J^1(\mathbb{R}, M)$ of the projection $J^1(\mathbb{R}, M) \to \mathbb{R}$, takes the local form $(\tau, x^\mu(\tau), v^\mu(\tau))$. One can also think of such a section as a differentiable curve in $T(M)$ that projects to the curve $x(\tau)$.

A section s of the latter projection is *integrable* iff it is the *1-jet prolongation* j^1x of a differentiable curve in M. In that case, it locally looks like $(\tau, x^\mu(\tau), v^\mu(\tau))$, in which the integrability condition is:

$$v^\mu(\tau) = \left.\frac{dx^\mu}{d\tau}\right|_\tau . \tag{X.102}$$

A (first-order) *Lagrangian* is a differentiable function $\mathcal{L}$ on $J^1(\mathbb{R}, M)$, which then takes the local form $\mathcal{L}(\tau, x^\mu, v^\mu)$. It defines an action functional on differentiable curves in M in the predictable way:

$$S[\gamma] = \int_\gamma \mathcal{L}(j^1\gamma)\, d\tau = \int_\gamma \mathcal{L}(\tau, x^\mu(\tau), \dot{x}^\mu(\tau))\, d\tau . \tag{X.103}$$

A *variation* δx of a curve x: $[0, 1] \to M$ is a vector field δx: $[0, 1] \to T(M)$, $\tau \mapsto \delta x(x(\tau))$ along that curve. One can think of it as the restriction to x of a vector field that is defined by differentiating a differentiable homotopy – i.e., a finite variation – of x in M. Its *prolongation* to a vector field $\delta^1 x$: $[0, 1] \to J^1(\mathbb{R}, M)$ has the local form:

$$\delta^1 x(\tau) = \delta x^\mu(\tau)\frac{\partial}{\partial x^\mu} + \left.\frac{d(\delta x^\mu)}{d\tau}\right|_\tau \frac{\partial}{\partial v^\mu} . \tag{X.104}$$

Note, in particular, that it does not have a component in the direction $\partial/\partial\tau$, which would amount to a variation of the curve parameterization.

The *first variation* of $S[.]$ is a linear functional $\delta S[X]$ on vector fields X on $J^1(\mathbb{R}, M)$ that is defined by:

$$\delta S[X] = \int_\gamma \mathrm{L}_X(\mathcal{L}\, d\tau) = \int_\gamma [(i_X d\mathcal{L})\, d\tau + \mathcal{L}\, di_X d\tau] , \tag{X.105}$$

analogously to (X.47).

Now:

$$d\mathcal{L} = \frac{\partial\mathcal{L}}{\partial\tau}d\tau + \frac{\partial\mathcal{L}}{\partial x^\mu}dx^\mu + \frac{\partial\mathcal{L}}{\partial v^\mu}dv^\mu , \tag{X.106}$$

so when $X = \delta^1 x$, one has:

$$i_{\delta^1 x} d\mathcal{L} = \frac{\partial\mathcal{L}}{\partial x^\mu}\delta x^\mu + \frac{\partial\mathcal{L}}{\partial v^\mu}\frac{d(\delta x^\mu)}{d\tau} . \tag{X.107}$$

By the usual product rule ("integration by parts") trick, this takes the form:

$$i_{\delta^1 x} d\mathcal{L} = \frac{\delta\mathcal{L}}{\delta x^\mu}\delta x^\mu + \frac{d}{d\tau}\left(\pi_\mu \delta x^\mu\right) , \tag{X.108}$$

in which:

$$\frac{\delta \mathcal{L}}{\delta x^\mu} = \frac{\partial \mathcal{L}}{\partial x^\mu} - \frac{d}{d\tau}\frac{\partial \mathcal{L}}{\partial x^\mu} \tag{X.109}$$

are the components of the variational derivative of $\mathcal{L}$ with respect to x^μ and:

$$\pi_\mu = \frac{\partial \mathcal{L}}{\partial v^\mu} \tag{X.110}$$

are the components of the *generalized momentum* 1-form that is canonically conjugate to the velocity vector field.

Hence, by an application of Stokes's theorem, the first variation of $S[.]$ in the direction δx becomes:

$$\delta S[\delta x] = \int_\gamma \left(\frac{\delta \mathcal{L}}{\delta x^\mu} \delta x^\mu \right) d\tau + \left[\pi_\mu \delta x^\mu \right]_{\tau=0}^{\tau=1}. \tag{X.111}$$

There are two basic variational problems in which the bracketed term vanishes:

1. *Fixed endpoint problems:* One considers only those differentiable curves between two fixed endpoints A and B, which then implies that $\delta x(A)$ and $\delta x(B)$ vanish.

2. *Variable endpoint-problems.* In this case, in order for the bracketed term to vanish one must restrict oneself to variations δx that satisfy the *transversality condition* that $\pi_\mu \, \delta x^\mu = 0$ at the endpoints of the curves considered.

In either case, the necessary and sufficient condition that a given curve γ be an *extremum* of the action functional – i.e., that $\delta S[\delta x]$ vanish for all allowable variations δx of γ– is that it satisfy the *Euler-Lagrange equations:*

$$\frac{\delta \mathcal{L}}{\delta x^\mu} = 0. \tag{X.112}$$

Although this condition is necessary for the extremum to be a local minimum of the action function, it is not sufficient. In order to find sufficient conditions, one must consider of the second variation of $\mathcal{L}$, which we shall not go into here.

Since the manifold $\mathbb{R}$ is one-dimensional, the Euler-Lagrange equations for point mechanics represent a system of ordinary differential equations for the curve γ. If one introduces the *generalized force* 1-form f on $J^1(\mathbb{R}, \mathrm{M})$:

$$f = \frac{\partial \mathcal{L}}{\partial x^\mu} dx^\mu \tag{X.113}$$

and sets $\pi = \pi_\mu \, dx^\mu$ then the Euler-Lagrange equations take the generalized Newtonian form:

$$f = \frac{d\pi}{d\tau}, \tag{X.114}$$

which can also be regarded as the "strong" form of the conservation law for linear momentum. (The "weak" form is Newton's first law, which says that when f vanishes π is constant along the extremal.)

b. Legendre transform. Since we have already defined equations of motion in terms of the dispersion polynomial $P[k]$ on the cotangent bundle, namely, the bicharacteristic equations, we need to relate those equations to the equations of motion for points that we just obtained for an arbitrary Lagrangian $\mathcal{L}$ on the tangent bundle. This is what the Legendre transformation accomplishes, although, as we shall see, when one does not use a quadratic dispersion polynomial, the form of the resulting Lagrangian is not quite as convenient.

Although we could have introduced the present topic above in the more general context of variational field theory, the resulting formalism seems more intuitively appealing in the context of point mechanics, in which the base manifold of the bundle in question is simply $\mathbb{R}$, so the bundle becomes the trivial one $\mathbb{R} \times M \to \mathbb{R}$, and the jet manifolds $J^1(\mathbb{R}, M)$ and $J^1(M, \mathbb{R})$ reduce to $\mathbb{R} \times T(M)$ and $T^*M \times \mathbb{R}$.

The first step in defining this transformation is to return to the association of tangent covectors with tangent vectors that one obtains from starting with a Hamiltonian function H: $T^*M \to \mathbb{R}$, which we assume to be continuously differentiable, and which was described in Chapter VIII in the section bicharacteristics. The differential dH defines a characteristic (i.e., Hamiltonian) vector field X_H by the process that was described in Chapter VIII. Any covector $k_x \in T^*M$ is then associated with a tangent vector $X_H(k_x)$ on T^*M, which then projects to a tangent vector $\mathbf{v}_x$ to M. This association of each $k_x \in T^*M$ with a corresponding $\mathbf{v}_x \in T(M)$ then defines a map $\imath_H : T^*M \to T(M)$, which we assume to be a diffeomorphism of each fiber of T^*M over $x \in M$ to the corresponding fiber T_xM. By the inverse function theorem, this implies that the differential map $d\imath_H|_{(x, k)}$ must be invertible at every point of T^*M.

Locally, the fiber map takes the form:

$$v^\mu(k_\nu) = \left.\frac{\partial H}{\partial k_\mu}\right|_{(x,k)} \tag{X.115}$$

at each x, so the local diffeomorphism constraint says that:

$$\left.\frac{\partial v^\mu}{\partial k_\nu} = \frac{\partial^2 H}{\partial k_\mu \partial k_\nu}\right|_{(x,k)} \tag{X.116}$$

must be an invertible matrix at each x.

By assumption, one can invert the relationship (X.115) to obtain $k = k(\mathbf{v})$, and one defines a Lagrangian on $T(M)$ by means of the classical Legendre transformation:

$$\mathcal{L}(x, \mathbf{v}) = k(\mathbf{v})(\mathbf{v}) - H_x(k(\mathbf{v})); \tag{X.117}$$

in local coordinate form this is:

$$\mathcal{L}(x^\mu, v^\mu) = k_\mu(\mathbf{v})\, v^\mu - H(k(\mathbf{v})). \tag{X.118}$$

The variational derivative of $\mathcal{L}$ with respect to x takes the local form:

$$\frac{\delta \mathcal{L}}{\delta x^\mu} = f_\mu - \frac{d\pi_\mu}{d\tau}, \tag{X.119}$$

in which:

$$f_\mu = \frac{\partial \mathcal{L}}{\partial x^\mu} = -\frac{\partial H}{\partial x^\mu}, \qquad \pi_\mu = k_\mu + \frac{\partial k_\nu}{\partial x^\mu}\left(v^\nu - \frac{\partial H}{\partial k_\nu}\right). \tag{X.120}$$

Hence, when one assumes the validity of the canonical equations for H, one finds that:

$$\pi_\mu = k_\mu, \qquad \frac{\delta \mathcal{L}}{\delta x^\mu} = 0, \tag{X.121}$$

and conversely; i.e., the Hamiltonian equations for H are equivalent to the Euler-Lagrange equations for $\mathcal{L}$ with the constraint that $\pi_\mu = k_\mu$.

Now, let us examine the form that the Legendre transformation takes when our Hamiltonian is defined by a homogeneous quartic polynomial:

$$P[k] = \tfrac{1}{4} P^{\kappa\lambda\mu\nu} k_\kappa k_\lambda k_\mu k_\nu. \tag{X.122}$$

The map $\imath_P : T^*M \to T(M)$, $k \mapsto \mathbf{v}(k)$, then has the local form:

$$v^\nu(k) = \frac{\partial P}{\partial k_\nu} = P^{\kappa\lambda\mu\nu} k_\kappa k_\lambda k_\mu. \tag{X.123}.$$

As we pointed out in the previous chapter, whereas this system of algebraic equations would be merely linear in the case of a quadratic $P[k]$, we see that we are presently concerned with a system of *cubic* equations, which implies that the very business of inverting them, which is unavoidable if one is to pull P over from T^*M to $T(M)$, is likely to be considerably more computationally involved than a simple matrix inversion. In particular, the inverse map i_P^{-1}, which we assume exists, must be homogeneous of degree 1/3, so it is not a polynomial map.

Locally, the invertibility of i_P means that the matrix:

$$\frac{\partial v^\nu}{\partial x^\mu} = 3P^{\kappa\lambda\mu\nu} k_\kappa k_\lambda \equiv 3\, \gamma^{\mu\nu}(x, k) \tag{X.124}$$

must be invertible for each (x, k). One sees that this matrix also represents the local components $\partial^2 P / \partial k_\mu \partial k_\nu$.

Once the invertibility assumption has been made, it is a simple matter to define a function Q: $T(M) \to \mathbb{R}$ by means of the inverse map $\imath_P^{-1}$: $T^*M \to T(M)$, $\mathbf{v} \mapsto k(\mathbf{v})$ by way of:

$$Q[\mathbf{v}] = P[k(\mathbf{v})]; \tag{X.125}$$

however, the specific nature of the resulting function is more involved than in the case where $\imath_P$ is linear on the fibers.

In particular, since the map $\imath_P$ is homogeneous of degree three, its inverse is homogeneous of degree 1/3. As the degree of P is four, this means that the degree of homogeneity of Q is 4/3:

$$Q[\lambda\mathbf{v}] = P[k(\lambda\mathbf{v})] = P[\lambda^{1/3}k(\mathbf{v})] = \lambda^{4/3}\, P[k(\mathbf{v})] = \lambda^{4/3}\, Q[\mathbf{v}]. \tag{X.126}$$

c. Variational formulation of the moving point charge problem. It is in attempting to define a Lagrangian for the motion of a charged mass point in an external electromagnetic field F that we find that reconsidering the role of a spacetime metric is also a crucial issue in mechanics, as well as electromagnetism. The key question is whether one can define the kinetic energy of motion in the absence of a metric. This is equivalent to the problem of how to associate a momentum 1-form $p(\tau)$ with the velocity vector field $\mathbf{v}(\tau)$ along a curve, since the kinetic energy will be proportional to $p(\mathbf{v})$.

Ultimately, the issue is again one of duality, in which the velocity and momentum have to be related to each other by a *mechanical constitutive law* in the same way that F and $\mathfrak{h}$ are related by an electromagnetic one and stress is related to strain by a different sort of mechanical constitutive law.

As long one deals with canonical momentum π the fact that its components are functions $\pi_\mu(\tau, x^\mu, v^\mu)$ suggests that the mechanical constitutive law is built into the form of π, or equivalently, the Lagrangian $\mathcal{L}$ that defines it. This is analogous to the way that one could define the constitutive law that coupled F to $\mathfrak{h}$ directly or indirectly by way of the Lagrangian density for F.

However, for the sake of computations in particular cases one still needs to be more specific about the form of either $\mathcal{L}$ or π and if one does not introduce a way of associating tangent vectors with covectors at some point then one gets into a vicious logical cycle by trying to avoid it. Hence, we shall assume that our manifold M is associated with an electromagnetic constitutive law, with a corresponding dispersion polynomial $P[k]$, and use the invertible map i_P: $T^*M \to T(M)$, $k \mapsto \mathbf{v}(k)$ that it defines to facilitate the definition of a Legendre transformation.

Under this transformation, the Hamiltonian $H(x, p)$ on T^*M will go to the Lagrangian:

$$\mathcal{L}(x, v) = p(\mathbf{v})(\mathbf{v}) - H(x, p(\mathbf{v})). \tag{X.127}$$

Previously, we associated frequency-wave number 1-forms with velocity vector fields. However, in order to discuss point mechanics, we must associate velocity fields along curves with energy-momentum 1-forms along those curves. Since we are dealing with the motion of a *point* particle the main problem conceptually is how to associate a frequency-wave number 1-form k with the particle when the electromagnetic field, being external, is essentially passive to the problem. We then see that, in a sense, the point particle approximation is too much of an approximation to suggest a notion of an associated wave covector field, in much the same way that point particles make the definition of angular momentum more involved than when one is dealing with extended matter.

Since the point particle concept is essentially classical – i.e., pre-quantum – we shall resolve the issue by assuming that the basic classical data of a rest energy E_0 and an energy-momentum covector field p are given and have their support on the curve of the particle. The wave covector k is then related to the energy-momentum covector field p by the de Broglie relation $p = \hbar k$, while the rest energy is related to a rest frequency ω_0 by $E_0 = \hbar\omega_0$. Furthermore, the energy-momentum covector field p satisfies the inhomogeneous characteristic equation:

$$P[p] = \hbar^4 P[k] = E_0^4 = \hbar^4 \omega_0^4; \tag{X.128}$$

hence, k satisfies:

$$P[k] = \omega_0^4. \tag{X.129}$$

In effect, the fact that the rest energy is non-vanishing has changed the nature of the dispersion law from the homogeneous one that relates to photons into the inhomogeneous one that relates to massive matter.

From the fact that the form of H is algebraically simpler than that of $\mathcal{L}$ when one is not dealing with quadratic polynomials – viz., $P[k]$ is a polynomial, while $Q[\mathbf{v}]$ is not, – one finds that it is mathematically more straightforward to deal with the Hamiltonian form of the geodesic equations, rather than their Lagrangian form.

We start with the fact that a pre-metric way of expressing kinetic energy is:

$$T = \tfrac{1}{d}\, p(\mathbf{v}), \tag{X.130}$$

in which d represents the degree of homogeneity of the function $p(x, \mathbf{v})$ in $\mathbf{v}$.

This expression can be interpreted in either the Lagrangian formalism or the Hamiltonian one.

In the Lagrangian formalism one must solve for the 1-form p as a function of the vector field $\mathbf{v}$:

$$T(x, v) = \tfrac{1}{d}\, p(\mathbf{v})(\mathbf{v}) = \tfrac{1}{d}\, p_\mu(x, v) v^\mu. \tag{X.131}$$

However, since it is $\mathbf{v} = \mathbf{v}(p)$ that is given by the elementary expression – namely, a system of homogeneous cubic equations – one sees that the inverse system will not generally admit such an elementary form.

Conversely, in the Hamiltonian formalism one must solve $\mathbf{v}$ for p:

$$T(x, p) = \tfrac{1}{d}\, p(\mathbf{v}(p)) = \tfrac{1}{d}\, p_\mu\, v^\mu(x, p). \tag{X.132}$$

Hence, since the equations that take p to $\mathbf{v}$ have the more elementary form in this case, one would regard the Hamiltonian formalism as being computationally preferable.

Now, let us contrast that with the situation regarding potential energy, which has the general form:

$$U = A(\mathbf{J}) = qA(\mathbf{v}). \tag{X.133}$$

One immediately sees that this expression has a distinctly Lagrangian character, since it depends most directly upon the velocity vector field. The local functional dependencies are then:

$$U(x, v) = qA_\mu(x)\, v^\mu(x). \tag{X.134}$$

In the Hamiltonian formalism, however, one must solve $\mathbf{v}$ for p:

$$U(x, p) = qA_\mu(x)\, v^\mu(x, p). \tag{X.135}$$

Although the map that takes $\mathbf{v}$ to p is not as conveniently represented as the inverse map, we shall present the derivation of the equations of motion for a charged mass point in an external electromagnetic field in both of the aforementioned formalisms in order to see how the change from a quadratic dispersion law to a quartic one affects the equations of motion.

In Lagrangian formalism, the Lagrangian has the form $\mathcal{L} = T - U$ and the equations of motion take the form:

$$\frac{\delta T}{\delta x^\mu} = \frac{\delta U}{\delta x^\mu}. \tag{X.136}$$

By substituting for T as in (X.131) and U as in (X.134), this becomes:

$$\frac{1}{d}\left[\frac{dp_\mu}{d\tau} - \frac{\partial p_\nu}{\partial x^\mu} v^\nu + \frac{\partial^2 p_\nu}{\partial x^\kappa \partial v^\mu} v^\kappa v^\nu + \frac{\partial^2 p_\nu}{\partial v^\kappa \partial v^\mu}\frac{dv^\kappa}{d\tau} v^\nu + \frac{\partial p_\nu}{\partial v^\mu}\frac{dv^\nu}{d\tau}\right] = F_{\mu\nu} J^\nu. \tag{X.137}$$

In the quadratic case, for which $d = 2$ and $p_\mu = m_0\, g_{\mu\nu} v^\nu$, one has:

$$\frac{\partial p_\nu}{\partial x^\mu} = m_0\, g_{\kappa\nu,\mu}\, v^\kappa, \qquad \frac{\partial p_\nu}{\partial v^\mu} = m_0\, g_{\mu\nu}, \qquad \frac{\partial^2 p_\nu}{\partial x^\kappa \partial v^\mu} = m_0\, g_{\kappa\nu,\mu}, \qquad \frac{\partial^2 p_\nu}{\partial v^\kappa \partial v^\mu} = 0, \tag{X.138}$$

and one finds that the equations of motion can be put into the form:

$$\nabla_{\mathbf{v}}\, p_\mu = F_{\mu\nu} J^\nu. \tag{X.139}$$

As we pointed out, when one chooses a quartic polynomial for $P[p]$, so $p(x, v)$ is homogeneous of degree 1/3 in v, there is no simplification of the left-hand side of (X.137).

Under a Legendre transform, a Lagrangian of the form $T(x, \mathbf{v}) - U(x, \mathbf{v})$ does not generally go to a Hamiltonian of the form $T(x, p) + U(x, p)$, so in order to get some idea of how to incorporate the coupling of the charge to the external field, we first look at the result of applying a Legendre transform to the quadratic Lagrangian of that form that pertains to a Lorentzian manifold, namely:

$$\mathcal{L}(x, v) = \tfrac{1}{2} m_0 \, g_{\kappa\lambda}(x) \, v^\kappa v^\lambda - qA_\nu(x) \, v^\nu. \tag{X.140}$$

We first find that the canonical momentum takes the form:

$$p_\mu = \frac{\partial \mathcal{L}}{\partial v^\mu} = m_0 \, g_{\mu\nu} \, v^\lambda - qA_\mu. \tag{X.141}$$

Hence, the Legendre transformation produces a Hamiltonian of the form:

$$H(x, p) = \tfrac{1}{2m_0} g^{\mu\nu}(x)(p_\mu + qA_\mu)(p_\nu + qA_\nu). \tag{X.142}$$

This exhibits the "minimal electromagnetic coupling" aspect of energy-momentum, namely, that in the Hamiltonian formalism the coupling of the free charge to the external field is by way of replacing the free particle energy-momentum 1-form p with $\tilde{p}_\mu(x, p) = p + qA$. One can also think of this as the Fourier transform of the replacement of the partial derivative ∂_μ with the covariant derivative $\partial_\mu - iqA_\mu$, in which A plays the role of the $U(1)$ connection form.

One then verifies that the canonical equations for this H can be reduced to the form:

$$m_0 \, g_{\mu\nu} \nabla_{\mathbf{v}} v^\nu = F_{\mu\nu} \, v^\nu. \tag{X.143}$$

In the quartic case, a minimal coupling of A to the energy-momentum of the particle by the replacement of p_μ with $\tilde{p}_\mu(x, p)$ gives the Hamiltonian:

$$H(x, p) = \tfrac{1}{4} P^{\kappa\lambda\rho\sigma}(x) \tilde{p}_\kappa \tilde{p}_\lambda \tilde{p}_\rho \tilde{p}_\sigma. \tag{X.144}$$

The canonical equations initially take the form:

$$\frac{dx^\mu}{d\tau} = v^\nu = P^{\kappa\lambda\rho\mu} \tilde{p}_\kappa \tilde{p}_\lambda \tilde{p}_\rho, \tag{X.145a}$$

$$\frac{dp_\mu}{d\tau} = -\frac{1}{4} \frac{\partial P^{\kappa\lambda\rho\sigma}}{\partial x^\mu} \tilde{p}_\kappa \tilde{p}_\lambda \tilde{p}_\rho \tilde{p}_\sigma - qA_{\kappa,\mu} P^{\kappa\lambda\rho\sigma} \tilde{p}_\lambda \tilde{p}_\rho \tilde{p}_\sigma, \tag{X.145b}$$

although with the substitution $p = \tilde{p} - qA$ the second equation becomes:

$$\frac{d\tilde{p}_\mu}{d\tau}+\left(\frac{1}{4}P^{\kappa\lambda\rho\sigma}{}_{,\mu}\tilde{p}_\sigma - qF_{\mu\nu}P^{\kappa\lambda\rho\nu}\right)\tilde{p}_\kappa\tilde{p}_\lambda\tilde{p}_\rho = 0, \qquad \text{(X.146)}$$

and the substitution of (X.145a) in (X.145b) gives:

$$\frac{d\tilde{p}_\mu}{d\tau}+\frac{1}{4}P^{\kappa\lambda\rho\sigma}{}_{,\mu}\tilde{p}_\kappa\tilde{p}_\lambda\tilde{p}_\rho\tilde{p}_\sigma = qF_{\mu\nu}v^\nu . \qquad \text{(X.147)}$$

Although presenting the equations of motion in this form allows us to see how the introduction of an external electromagnetic field affects the free-particle equations of motion, nonetheless, if one wishes to solve for $\tilde{p}$ (if only numerically), which then allows one to derive p and **v** directly, then one must use the form (X.146).

d. Radiation damping. One finds that attempting to include the effects of the radiation reaction on the motion of a point charge in an electromagnetic field is met with the same obstacles as in attempting to account for viscous drag forces in the medium (indeed, some refer to the reaction as *radiation damping*). That is, one is no longer considering a conservative mechanical system, so the very applicability of either Lagrangian or Hamiltonian methods is no longer justified. Basically, it comes down to a question of completeness in the model for the state space of the particle-field system, since the field itself potentially contributes an infinite number of degrees of freedom, which include the radiative modes, just as accounting for the energy that gets dissipated by friction in a mechanical system would imply adding an enormous number of dimensions to the state space to account for the internal motion of the molecules that are absorbing the energy that gets lost.

When one considers an extended charge distribution instead of a pointlike one, one also adds a potential infinitude of degrees of freedom to the state space in the form of the internal degrees of freedom that pertain to the distribution, such as the various deformations, and – more to the immediate point – the wavelike modes. Indeed, the very fact that elementary charges, such as electrons, are associated with a characteristic frequency $\omega_0 = m_e c^2/\hbar$ or wave number, in the form of the Compton wave number $k_0 = \omega_0/c$ (which assumes the Lorentzian dispersion law), suggests that there is some fundamental internal motion going on, probably of the standing-wave variety. The fact that massless waves can only exist as traveling waves with a characteristic speed is undoubtedly closely relevant to that fact.

e. The motion of extended charge distributions. Although it is certainly possible to treat continuum mechanics in general, and continuous distributions of charges, more specifically (see, e.g., [**7**, **8**]), within the framework of the calculus of variations and the geometry of jet manifolds, nevertheless, since our primary focus in this book is on examining the various topics in electromagnetism that can be formulated in a pre-metric manner, we shall not go further in that direction at the moment. Suffice it to say that, rather dealing with anti-symmetric second-rank tensor fields, such as F and $\mathfrak{h}$, one is dealing with symmetric ones, such stress σ and strain e, respectively. However, the

equations of motion can still be put into a form that is analogous to the pre-metric Maxwell equations, by postulating the integrability of the strain, the conservation of energy-momentum for stress, and a mechanical constitutive law that relates stress to strain:

$$e^i{}_j = \tfrac{1}{2}(u^i{}_{,j} + u^j{}_{,i}), \qquad \sigma^j_{i\,,j} = f_i\,, \qquad \sigma^j_i = C(e^i{}_j). \tag{X.138}$$

In these equations, the displacement vector field $\mathbf{u} = u^i\partial_i$, which describes an infinitesimal deformation of a region in a medium, plays a role that is analogous to the potential 1-form in electromagnetism, except that one is symmetrizing the derivative, instead of anti-symmetrizing it. Indeed, although the symmetry of the strain tensor field is based in the symmetry of the metric tensor field in the medium (both deformed and undeformed), the assumption of symmetry of the stress tensor field can be weakened, since it is based in the absence of internal torques or moments existing in the medium. Media in which such internal torques exist are called *Cosserat media* (see, [**9**, **10**]), and constitute an intriguing possibility for modeling the structure of spacetime by analogy, since one is basically looking at objects that are embedded in the bundle of orthonormal frames over a (pseudo-) Riemannian manifold, instead of the manifold itself.

5. Fermat's principle. No discussion of the application of the calculus of variations to electromagnetism is complete without a discussion of Fermat's principle as a basis for geometrical optics [**11-15**]. This is especially true in the present context of pre-metric electromagnetism, since the action functional that Fermat defined for the light rays in an optical medium is crucially related to its electromagnetic constitutive properties, and thus represents a simplification of the problem in its full generality.

Since the space Σ of geometrical optics is two-or-three-dimensional, in order to define the variational formulation of light rays in the manner of Fermat, one must consider the 1-jets of curves in Σ; we denote this fibered manifold by $J^1(\mathbb{R}, \Sigma)$. Fermat's principle is based on the specialization of Hamilton's principle that takes the form of saying that a light ray from point A in Σ to point B is a curve from A to B that minimizes the *elapsed time* dt along the curve. (Note that since light rays cannot be parameterized by proper time, this must be understood to represent an affine parameterization.)

A key to making the transition from spacetime to space, along with the expected Legendre transformation from a Hamiltonian function to a Lagrangian function is to note the local equivalence of $J^1(\mathbb{R}, \Sigma)$ with $PT(M)$ and $J^1(\Sigma, \mathbb{R})$ with $PT^*(M)$. This is easily seen by defining local coordinate systems for each pair of manifolds in the form of (t, x^i, V^i) for the former pair and (x^i, t, n_i) for the latter. Of course, in the latter case the time coordinate t must now be regarded as a differentiable function on an open subset of Σ, not an open subset of M.

We have previously encountered the elapsed-time functional in our discussion of Huygens's principle, where we found that it comes about naturally when one treats the components k_μ of the wave covector $k = \omega dt - k_i\, dx^i$ at any point $x \in M$ of spacetime as the homogeneous coordinates of a point $[k]$ in the projectivized cotangent space $PT^*_x M$,

such that the inhomogeneous coordinates are $n_i = k_i / \omega$. When $M = \mathbb{R}\times\Sigma$, these inhomogeneous coordinates then define a spatial 1-form on Σ by way of:

$$n = n_i \, dx^i, \tag{X.139}$$

and if k were exact – i.e., $k = d\phi = \partial\phi / \partial t \, dt + \partial\phi / \partial x^i \, dx^i$ –then this would mean that:

$$n = -\frac{\partial\phi / \partial x^i}{\partial\phi / \partial t} dx^i = \frac{\partial t}{\partial x^i} dx^i = dt. \tag{X.140}$$

Hence, the integral of n along any curve segment γ will represent the elapsed time that it takes to go from one endpoint to the other.

However, in order to obtain Fermat's principle one must use this spatial path functional to define an *action* functional on curve segments. This then reverts to the problem of defining a Lagrangian function on the jets of curve segments, which would then take the form of a differentiable function $\mathcal{L}(t, x^i, V^i)$, while so far we only have a section of $T^*\Sigma \to \Sigma$, namely, n. What is missing is the input from the dispersion law $P[k]$ on T^*M, which, as we saw, becomes an inhomogeneous polynomial $[P][n]$ on PT^*M that takes the form:

$$[P][n] = P_4[n] + P_2[n] + P_0 \tag{X.141}$$

for electromagnetic waves when one factors the ω out of $P[k]$; of course, the polynomial $[P][n]$ is no longer homogeneous in its independent variables n_i .

Since the Hamiltonian $H(x, k) = 1/4P(x)[k]$ on T^*M is supposed to vanish for the physically meaningful wave covectors, as well as $[P][n]$, any multiplicative factors are superfluous, and we define the Hamiltonian on PT^*M to be:

$$H(x^i, t, n_i) = \tfrac{1}{4}[P](x)[n]. \tag{X.142}$$

In order to make the transition from a Hamiltonian on the contact manifold PT^*M to a Lagrangian on $J^1(\mathbb{R}, \Sigma)$, at least locally, we first perform a Legendre transformation on H, as it is defined on T^*M, to a Lagrangian $\mathcal{L}$ on $T(M)$:

$$\mathcal{L}(x^\mu, v^\mu) = k_\mu(\mathbf{v})v^\mu - H(x^\mu, k_\mu(\mathbf{v})). \tag{X.143}$$

Since H is homogeneous in k of degree four, $k_\mu v^\mu = 4H$ and one can say:

$$\mathcal{L}(x^\mu, v^\mu) = \tfrac{3}{4} k_\mu(\mathbf{v})v^\mu. \tag{X.144}$$

The projection of this onto $PT(M)$ is $\tfrac{3}{4}\omega v^0[1 - n_i(V)V^i]$, and since it must vanish for physically meaningful velocity vectors, the leading scalar factor can be omitted. Similarly, differentiation to obtain the equations of motion makes the unity term irrelevant and the fact that the variational derivative is set to zero also makes the sign of

the second term irrelevant. Ultimately, there is no loss in generality in defining the Lagrangian on $PT(M)$ to be:

$$\mathcal{L}(t, x^i, V^i) = \tfrac{3}{4} n_i(x, V)V^i. \tag{X.145}$$

When one forms the action functional for differentiable curves in Σ, under the assumption that $n_i(x, V)$ is independent of t, one first finds that if the curve parameter is s then $V^i = dx^i / ds$ and:

$$\mathcal{L}(t, x^i, V^i)\, ds = n_i\, dx^i, \tag{X.146}$$

which equals the elapsed-time differential in the integrable case, where $t = t(x^i)$ and $n_i = \partial t / \partial x^i$.

One notes that is it equivalent to first project the Hamiltonian $H(x^\mu, k_\mu)$ on T^*M to a Hamiltonian $H(x^i, t, n_i) = 1/4 n_i V^i(n)$ on PT^*M and then perform a Legendre transformation:

$$\mathcal{L}(t, x^i, V^i) = n_i(V^i)V^i - H(x^i, t, n_i) = \tfrac{3}{4} n_i(V^i)V^i. \tag{X.147}$$

This can be summarized in a commutative diagram:

$$\begin{array}{ccc} T^*M & \longrightarrow & T(M) \\ \downarrow & & \downarrow \\ PT^*M & \longrightarrow & PT(M) \end{array}$$

in which the horizontal arrows are the Legendre transforms and the vertical ones are the projectivizations.

Although we have succeeded in deriving the elapsed-time path functional from the dispersion law in a more general manner than one usually obtains for the quadratic polynomial that gives a Lorentzian metric on spacetime, nonetheless, one immediately finds that it practical calculations, the advantage of this expansion of scope is diminished by the disadvantage that the functions $k_\mu(\mathbf{v})$ and $n_i(\mathbf{V})$ are no longer elementary linear isomorphisms, as they would be for the quadratic case, but a set of four homogeneous functions of degree 1/3 in the former case, and a set of three functions that are not even homogeneous in the latter. Hence, these functions are easier to work with in theory than they are in practice.

The inevitable conclusion seems to be that the formulation of geometrical optics is much more natural and straightforward when one deals with cotangent objects instead of tangent ones. In other words, the wave optics of Huygens is more straightforward than the ray optics of Fermat when one is concerned with dispersion laws that are not quadratic in nature.

References

1. D. J. Saunders, *Geometry of Jet Bundles,* Cambridge University Press, Cambridge, 1989.
2. G. Sardanashvily, O. Zakharov, *Gauge Gravitation Theory,* World Scientific, Singapore, 1992.
3. J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill, New York, 1941.
4. W. Thirring, *Classical Field Theory,* Springer, Berlin, 1978.
5. F. Hehl and Y. N. Obukhov, *Foundations of Classical Electrodynamics*, Birkhäuser, Boston, 2003.
6. J. Plebanski, *Lectures on Nonlinear Electrodynamics,* NORDITA Lectures, Copenhagen, 1970.
7. A. Lichnerowicz, *Théorie relativiste de la gravitational et de l'electromagnetisme,* Masson and Co., Paris, 1955.
8. F. Gallisot, "Les formes extérieures et le mécanique des milieux continus," Ann. Inst. Fourier (Grenoble) **8** (1958), 291-335.
9. E. Cosserat and F. Cosserat, *Théorie des corps déformables,* Hermann, Paris, 1909.
10. E. Kröner, ed., *Mechanics of Generalized Continua,* Proc. of 1967 IUTAM Symposium in Freudenstadt and Stuttgart, Springer, Berlin, 1968.
11. J. L. Synge, *Geometrical Optics,* Cambridge University Press, Cambridge, 1937.
12. Born, M., Wolf, E., *Principles of Optics*, Pergamon, Oxford, 1980.
13. R. K. Luneburg, *Mathematical Theory of Optics,* The University of California Press, Berkeley, 1964.
14. M. Kline and I. W. Kay, *Electromagnetic Theory and Geometrical Optics*, Wiley-Interscience, New York, 1965.
15. V. Guillemin and S. Sternberg, *Geometric Asymptotics,* Mathematical Surveys, no. 14, Am. Math. Soc., Providence, 1977.

CHAPTER XI

Symmetry and electromagnetism

One of the most fundamental – some would say, *the* most fundamental – ways of formulating the first principles of physics is in terms of conservation laws. One can say that conservation laws can be phrased in a weak form and a strong form, where the weak form is simply an approximate version of the strong form.

The weak form of a conservation law pertains to a system that is closed in the sense of complete – i.e., having no well-defined "external" environment, or, at least no exchange of the physical quantity in question with the external environment. One must understand, though, that the "external" environment might very well be "internal" in nature, such as the unmodeled system modes due to the molecular structure of a material medium. The weak form of the conservation of the physical quantity in question states that its total value over the system is constant in time.

The strong form of a conservation law assumes that there is a well-defined distinction between the internal and external states of the system and an exchange of the physical quantity in question between the two. The strong form of the conservation law then states that the time derivative of the total value of the quantity over the internal states equals the total rate of transfer across the boundary between the internal and external systems. The weak form then represents the approximate form of the strong that one obtains by assuming that the system exterior is negligible, or, at least, the transfer of the physical quantity between the internal and external systems is negative.

In engineering mechanics, the strong form of a conservation law is called a *balance law*. One sees that a conservation law then represents a complete accounting of the *total amount* of something while a balance law represents a complete accounting of its *time rate of change.*

As an example of these two forms of a conservation law, consider Newton's first law of motion, which amounts to the weak form of the conservation of linear momentum: when the total of the external forces that act on a physical system vanishes, its total linear momentum remains constant in time. The second law of motion, by comparison, says that when the total external forces are non-vanishing the time derivative of the total linear momentum equals that resultant of the external forces, which then represents its strong form, namely the balance of linear momentum.

One of the most profound advances to the first principles of theoretical physics was made in the early Twentieth Century by Emmy Noether, who showed that the calculus of variations provides a direct way of associating transformations of the states of physical systems that preserve the action functional for the system with conserved currents – i.e., vector fields with vanishing divergence. The transformations are generally regarded as symmetries of the basic dynamical laws that govern the states of the system. For instance, symmetry under spatial translations is correlated with the conservation of linear momentum, or the vanishing of external forces. Conversely, the breaking of a symmetry of the action functional is then associated with the non-conservation of the quantity it is associated with, and implies a non-vanishing divergence of the vector field; for instance, there would be non-vanishing forces, in the case of linear momentum.

In the context of the sourceless Maxwell equations for electromagnetism, in their metric form, the problem of determining the Noether symmetries, when one formulates the field equations in Lagrangian form, was addressed by Lorentz and Poincaré, who found that the symmetries included the Poincaré group. This result was advanced in a crucial way by Bateman [**1**] and Cunningham [**2**], who found that the complete symmetry group of the action functional was the conformal Lorentz group. The extra symmetries that expansion of scope entailed were the homotheties, which are the non-zero scalar multiplications of the vectors in Minkowski space, and the inversions through the unit proper-time hyperboloid.

However, it was also becoming apparent from the work of Lie, Cartan, Vessiot, Riquier, Janet, and others[1], that not only were there systems of differential equations that could not be given a variational formulation – e.g., any non-conservative system – but there were also symmetries of systems of differential equations that were distinct from the symmetries of an action functional, should it exist. These symmetries came to play an increasingly fundamental role in modern mathematics and physics, especially in the context of nonlinear wave equations.

One of the more definitive applications of this methodology to the case of Maxwellian electromagnetism, in its metric form, was made by Harrison and Estabrook and [**4**]. As far as the author of this book is aware, to date, the only attempt to extend the Harrison-Estabrook results to pre-metric electromagnetism was his own effort [**5**]. Hence, this chapter will amount to an extended treatment of that earlier work, in one sense, while referring some of the details to the previous paper.

In the first section of this chapter, we discuss the basic terminology of Lie groups, Lie algebras, and the action of Lie groups on manifolds, for the sake of conceptual completeness. Then, we show how these notions can be applied to the calculus of variations to deduce Noether's theorem, and give several examples of physical interest. After that, we discuss the broader problem of the symmetries of systems of differential equations, which we finally specialize to the equations of pre-metric electromagnetism.

1. Transformation groups [6-9]. Since the most important role that groups play in physics seems to be in the context of groups of transformations that act on the various configuration manifolds of physics as motions, as well as acting on the fibers of bundles over them as internal symmetries – i.e., changes of gauge – we shall briefly discuss the subject of transformation groups for the sake of completeness in the presentation.

a. Lie groups. Many – but not all – of the groups of interest to physics consist of point sets that are associated with topologies and differential structures that are more interesting than merely the discrete topology and zero-dimensional manifolds. Counterexamples to this claim include the various finite groups that figure in the discussion of crystalline structures, as well as symmetries such as parity reversal, time reversal, and charge conjugation.

When the set underlying a group G is given a topology that is compatible with the group structure, in the sense that the group multiplication and inversion operations are

[1] For a historical survey of contributions to the theory of symmetries of systems of differential equations, one might read the introductory chapter of Pommaret [**3**].

both continuous maps, one calls G a *topological group*. When a topological group G is also given a differential structure – i.e., an atlas of charts with diffeomorphisms for coordinate transformations – that is compatible with the group structure, in the sense that the group operations are now required to differentiable, as well as continuous, that topological group is then called a *Lie group*.

Interestingly, the structures that the Norwegian mathematician Marius Sophus Lie was originally defining were not, in fact, Lie groups, in the sense that we just defined, but *Lie pseudogroups*, which have the more general properties that the multiplication of elements is not always defined, just as the composition of maps is defined only when the image of the first map is a subset of the domain of the second one, and furthermore there is usually more than identity. Actually, this sort of consideration becomes unavoidable when one is dealing with symmetries of systems of differential equations, which we shall touch upon later in this chapter, but for now we shall deal with the more elementary case of Lie groups.

Examples of Lie groups that are of interest to physics usually take the form of subgroups of the general linear group $GL(n; \mathbb{R})$, which consists of all invertible $n \times n$ real matrices or $GL(n; \mathbb{C})$, which consists of all invertible $n \times n$ complex matrices. In either case, the group multiplication is defined by matrix multiplication. When the scalar field is not ambiguous, we shall use the common abbreviation $GL(n)$.

As a topological space, $GL(n; \mathbb{R})$ is neither compact nor connected. Rather, it has two connected components, which correspond to the fact that a non-zero real determinant must be either positive or negative. However, this is no longer true for a non-zero complex determinant, so $GL(n; \mathbb{C})$ is non-compact, but connected.

The way that one defines subgroups of interest to physics is by imposing a condition on the matrices that often takes the form of a set of algebraic equations. Such groups are then referred to as *linear algebraic groups*. Topologically, they will all be closed subsets of $GL(n)$.

When one requires that a matrix A must satisfy the n^{th} degree polynomial equation $\det A = 1$, the subgroup that this defines is the *special linear group*, which is denoted $SL(n; \mathbb{R})$ or $SL(n, \mathbb{C})$, depending upon the field of scalars one is using. Such matrices represent linear transformations of $\mathbb{R}^n$ or $\mathbb{C}^n$ that preserve a choice of volume element. More briefly, one employs the notation $SL(n)$, when there is no risk of confusion. One sees that by demanding that the determinant of A be unity, one selects the identity component of $GL(n)$, so $SL(n)$ is a connected, but non-compact, topological space.

When the algebraic condition is the set of quadratic equations $A^T A = A A^T = I$, where T refers to matrix transposition, one defines either $O(n; \mathbb{R})$ or $O(n, \mathbb{C})$, which are the real and complex *orthogonal groups*, respectively. Such transformations preserve a choice of Euclidian scalar product on $\mathbb{R}^n$ or $\mathbb{C}^n$, resp. Topologically, the manifolds $O(n)$ have two components, since $(\det A)^2 = 1$ for orthogonal A, which makes the two components correspond to the two possibilities $\det A = \pm 1$. They are, moreover, compact in the real case and locally compact in the complex case.

When the scalar product is the Minkowski scalar product on $\mathbb{R}^4$, the orthogonal group that it defines is the *Lorentz group* $O(3, 1)$. As a topological space, $O(3, 1)$ has four components, one of which contains the identity matrix I. The other three components can be obtained by using the non-trivial elements of the finite Abelian group $\mathbb{Z}_2 \times \mathbb{Z}_2 = \{I, P, T, PT\}$, where $P(x^0, x^i) = (x^0, -x^i)$ represents spatial parity inversion, $T(x^0, x^i) = (-x^0, x^i)$ represents the inversion of time orientation, and $PT(x^0, x^i) = (-x^0, -x^i) = -(x^0, x^i)$ then represents total parity inversion. The Lorentz group is not compact, but only locally compact. Indeed, as we shall discuss in more detail in a later chapter, the identity component of $O(3, 1)$ is isomorphic to $SL(2; \mathbb{C})$, as well as $SO(3; \mathbb{C})$.

By combining the set of equations that define $O(n)$ with $\det A = 0$, one defines the subgroup $SO(n)$, which is the *special orthogonal group*. Such transformations preserve both the scalar product and the volume element. In the case of Minkowski space, the resulting group is the *special Lorentz group* $SO(3, 1)$. Since this Lie group still has two connected components as a topological space, one can restrict oneself to the component $SO_0(3, 1)$ that contains the identity matrix by further requiring that the transformations preserve the time orientation of a vector in $\mathbb{R}^4$; i.e., the sign of its time component. This subgroup is often called the *proper orthochronous Lorentz group*.

In the case of complex scalars, one can also consider complex conjugation of matrix elements, in addition to transposition, and one then defines the *Hermitian conjugate* of a matrix $A \in GL(n; \mathbb{C})$ by $A^\dagger = (A^T)^* = (A^*)^T$, where * refers to the complex conjugation operator. The subgroup of $GL(n; \mathbb{C})$ that consists of invertible complex $n \times n$ matrices that satisfy the set of quadratic equations $A^\dagger A = AA^\dagger = I$ is then called the *unitary group* and is denoted by $U(n)$. These transformations of $\mathbb{C}^n$ preserve a Hermitian inner product, which differs from the Euclidian one by complex conjugation:

$$(z^i, w^j) = \delta_{ij} z^i \overline{w}^j . \tag{XI.1}$$

One of the properties of this inner product is the fact that (z^i, z^j) is always real, since each term $z^i \overline{z}^i = \| z \|^2$ in the sum is real.

The unitary groups are all compact connected topological groups.

The determinant of any unitary matrix has unit modulus since $1 = \det(A^\dagger A) = (\det A)^* (\det A) = \| \det A \|^2$. When one combines unitarity with the requirement that an element of $U(n)$ must also preserve the volume element on $\mathbb{C}^n$ – i.e., have unit determinant – one defines the subgroup $SU(n)$, which one calls the *special unitary group*. Interestingly, although one starts out by using complex matrices, the resulting differential manifold underlying $SU(n)$ is not a complex manifold, but a real one. In particular, $SU(2)$ is diffeomorphic to a real three-dimensional sphere.

An example of a Lie group that is of interest to physics, but not initially defined as a linear algebraic group, is the conformal Lorentz group $CO(3, 1)$, which consists of all

diffeomorphisms of Minkowski space that preserve the scalar product only up to a positive conformal factor Ω^2:

$$g(A\mathbf{v}, A\mathbf{w}) = \Omega^2\, g(\mathbf{v}.\ \mathbf{w}) = g(\Omega\mathbf{v}, \Omega\mathbf{w}). \tag{XI.2}$$

Although this group contains a linear subgroup – sometimes called the *Weyl group* – that consists of Lorentz transformations multiplied by dilatations, it also includes nonlinear transformations, such as the four-dimensional translation group $\mathbb{R}^4$, which act as affine transformations (but not linear ones), and transformations that represent inversion through the unit hyperboloid $g(\mathbf{v}, \mathbf{v}) = 1$. They take any vector $\mathbf{v}$ that does not lie on the light cone and map it to $g(\mathbf{v}, \mathbf{v})^{-1}\mathbf{v}$. Hence, it will lie on the same line through the origin, although its length will be contracted or expanded to the reciprocal of its length, except for vectors on the unit hyperboloid, which remain fixed.

One can represent the conformal Lorentz group as a linear group by taking advantage of the fact that the set of all light cones at all points of $\mathbb{R}^4$ (i.e., their defining coefficients) can be made into a six-dimensional real vector space that can be given a pseudo-Euclidian structure of signature type (–, –, +, +, +, +). One then finds that $CO(3, 1)$ can be represented by the linear algebraic group $O(2, 4)$.

Because Lie groups are manifolds, among other things, they have tangent spaces at each point, and the tangent bundle $T(G)$ to any Lie group G has the important property that it is always *parallelizable;* that is, there is always a global frame field on any Lie group. This fact follows the fact that if one chooses a frame $\{\mathbf{e}_i,\ i = 1, \ldots, n\}$ in any tangent space – say, the tangent space T_eG to the identity element $e \in G$ – then one can left-translate that frame to every other point $g \in G$ in such a manner that the resulting frame field on G is a differentiable section of the bundle $GL(G) \to G$ of linear frames in $T(G)$. Such a global frame field then allows one to express $T(G)$ as the trivial bundle $G \times \mathbb{R}^n \to G$ and $GL(G) \to G$ as the trivial bundle $G \times GL(n) \to G$.

b. Lie algebras. The tangent space T_eG at the identity plays a special role in the study of Lie groups since one can left-translate (or right translate, for that matter) not only tangent frames, but tangent vectors themselves. The result of choosing a tangent vector to a point of G – say, $\mathbf{v} \in T_eG$ – and left-translating it to every other point $g \in G$ is a *left-invariant vector field* on G:

$$\mathbf{v}(g) = dL_g|_e(\mathbf{v}). \tag{XI.3}$$

Hence, the components of $\mathbf{v}$ relative to a left-invariant frame field will be constant functions of $g \in G$. This means that the vector space of left-invariant vector fields on G is linearly isomorphic to $\mathbb{R}^n$, in general, and T_eG, in particular.

Similarly, one defines *right-invariant vector fields* on G by right-translation.

Since the vector space of vector fields on any differentiable manifold always has a Lie algebra structure that is defined by the Lie bracket of vector fields, one naturally wishes to examine how the constraint of left-invariance relates to that structure. In fact, it is preserved; i.e., the Lie bracket of left-invariant vector fields is left-invariant. One calls

this resulting Lie subalgebra $\mathfrak{G}$ of $\mathfrak{X}(G)$ the Lie algebra of G; it has the same dimension as a vector space as the dimension of G as a manifold. Hence, under the association of left-invariant vector fields with tangent vectors in T_eG one can define a Lie algebra on T_eG, as well.

We use the notations $\mathfrak{gl}(n; \mathbb{R})$, $\mathfrak{gl}(n; \mathbb{C})$, $\mathfrak{sl}(n; \mathbb{R})$, $\mathfrak{sl}(n; \mathbb{C})$, $\mathfrak{so}(n; \mathbb{R})$, $\mathfrak{so}(n; \mathbb{C})$, $\mathfrak{so}(3, 1)$, $\mathfrak{su}(n)$, $\mathfrak{co}(3, 1)$ to denote the Lie algebras that are associated with the Lie groups that were defined above, in an obvious way.

One can, of course, define differentiable curves in any Lie group G, and there is a particular class of curves through the identity element e that plays an essential role in understanding the relationship of G to its Lie algebra $\mathfrak{G}$. A differentiable curve g: $\mathbb{R} \to G$, $s \mapsto g(s)$, through e – say, $g(0) = e$ – is called a *one-parameter subgroup* of G iff $g(s_1 + s_2) = g(s_1)g(s_2)$ for all $s_1, s_2 \in \mathbb{R}$. That is, the map g is a homomorphism of the group $(\mathbb{R}, +)$ with its image in G. Such a subgroup of G will be one-dimensional, and therefore Abelian. As a consequence, it can only be diffeomorphic to S^1, when it is compact, or $\mathbb{R}$, when it is not. Merely starting with a compact Lie group G is not sufficient to restrict this choice, since there are subgroups of the 2-torus $S^1 \times S^1$ that are diffeomorphic to $\mathbb{R}$, namely, ones that give "irrational flows."

When one differentiates the curve $g(s)$ at e one obtains a tangent vector in T_eG:

$$\mathfrak{g} = \left.\frac{dg}{ds}\right|_{s=0}. \tag{XI.4}$$

When $g(s)$ satisfies a set of algebraic conditions that define the subgroup of $GL(n)$ that it lies in this differentiation along curves through the identity also allows us to obtain a corresponding set of algebraic conditions on $\mathfrak{g}$ that defines the Lie subalgebra of $\mathfrak{gl}(n)$ that it lies in. For instance, if $g(s) \in SL(n)$ for all s then $\det g(s) = 1$, so, by differentiation:

$$\left.\frac{d(\det g(s))}{ds}\right|_{s=0} = \mathrm{Tr}\, \mathfrak{g} = 0. \tag{XI.5}$$

That is, the elements of the Lie algebra $\mathfrak{sl}(n)$ consist of traceless $n \times n$ matrices (regardless of the choice of scalar field).

One obtains the algebraic condition for elements of $\mathfrak{so}(n)$ by differentiating the condition $A^TA = I$, which gives:

$$\omega^T + \omega = 0, \tag{XI.6}$$

in which we represent the tangent vector to the identity by ω, this time. This means that the matrices in $\mathfrak{so}(n)$ are all anti-symmetric. This is one reason why it is unnecessary to

distinguish between $\mathfrak{o}(n)$ and $\mathfrak{so}(n)$, since all anti-symmetric matrices have trace zero. One can also recall that $SO(n)$ is the connected component of the identity in $O(n)$.

Similarly, the condition $A^\dagger A = I$ on unitary matrices differentiates to the condition:

$$\omega^\dagger + \omega = 0, \tag{XI.7}$$

on elements of $\mathfrak{su}(n)$; that is, they must be *skew-hermitian.*

A skew-hermitian matrix does not need to have a vanishing trace, since in general its trace will be imaginary. Hence, one must clearly distinguish the Lie algebra $\mathfrak{u}(n)$ from the Lie algebra $\mathfrak{su}(n)$.

One can also take each tangent vector $s\mathfrak{g}$ along the line through the origin of $\mathfrak{G}$ that is generated by $\mathfrak{g}$ and associate it with a group element $g(s)$. The map exp: $\mathfrak{G} \to G$, $\mathfrak{g} \mapsto \exp(\mathfrak{g}) = g(1)$, is called the *exponential map* for G. It does not have to be either one-to-one or onto. In the former case, when a line through the origin of $\mathfrak{G}$ maps to a circle, the map exp will "wrap around" an infinite number of times, while, in the latter case, one finds that the image of exp can only be contained in the identity component of G. Hence, it cannot be surjective on a non-connected Lie group, such as $O(3, 1)$. It is, however, surjective on the identity component, which follows from its path-connectedness.

Because of the existence of this exponential map, one can think of the elements of any Lie algebra $\mathfrak{G} = T_eG$ as being the *infinitesimal generators* of one-parameter subgroups of G. Specifically, $\mathfrak{g} \in \mathfrak{G}$ generates the one-parameter subgroup $\exp(s\mathfrak{g}) \in G$.

When $\mathfrak{g}$ is represented by a matrix, one can use the power series expansion for the function e^x to define the matrix $\exp(\mathfrak{g})$:

$$\exp(\mathfrak{g}) = \sum_{n=0}^{\infty} \frac{1}{n!} \mathfrak{g}^n , \tag{XI.8}$$

in which the product is matrix multiplication. Just as it does for real or complex numbers, this series converges for all $\mathfrak{g}$;

c. Group actions [**10**, **11**]. A Lie group G is said to be a *(differentiable) transformation group* for a differentiable manifold M iff there is a differentiable map $G \times M \to M$, $(g, x) \mapsto gx$, such that:

i) $ex = x$ for all x.

ii) $g(g'x) = (gg')x$, for all $g, g' \in G, x \in M$.

One also says that *G acts on M differentiably.*

As a consequence of the definition, for each $g \in G$, the map L_g: $M \to M$, $x \mapsto gx$ is differentiable and invertible with a differentiable inverse; hence, it is a diffeomorphism. One calls this map L_g *left-translation by g*. From conditions *i*) and *ii*) above, the map L: $G \to \mathrm{Diff}(M)$, $g \mapsto L_g$ is a group homomorphism. It does not have to be faithful, since,

for instance, if the action of G on M fixes every point of M, G will map to the identity diffeomorphism. The question of whether it is a Lie group homomorphism reverts to the question of whether Diff(M) is a Lie group. Since Diff(M) is infinite-dimensional, this is not always true, and one must be more careful about the vector space on which infinite-dimensional manifolds are modeled. One can say that when M is compact Diff(M) is a Banach Lie group; i.e., the manifold charts map to a Banach space and the coordinate transitions are Banach space diffeomorphisms. We shall not elaborate on this, however [2].

If one fixes $x \in M$ then there is a differentiable map $G \to M$, $g \mapsto gx$ whose image $G(x)$ is called the *orbit* of x under the action of G. Again, this map does not have to be one-to-one – i.e., an embedding since more than one element of G might take a give $x \in M$ to the same element. Due to the group structure on G, the question of how many elements of G take a given point x to the same point $y = gx$ reverts to the question of how many elements of G take x to itself. The set $G_x = \{g \in G | \, gx = x\}$ of all elements of G that fix a given $x \in M$ is called the *isotropy subgroup* [3] of x under the given group action. One finds that the coset space G/G_x with the quotient topology can be given a manifold structure that makes it diffeomorphic to the orbit $G(x)$ of x under the action of G. An important example of an isotropy subgroup is $G_x = G$, in which case x is a *fixed point* of the group action.

When the map $G/G_x \to M$ is a diffeomorphism, one calls the manifold M a *homogeneous space*. Since this means that any pair of points (x, y) in $M \times M$ are associated with at least one $g \in G$ such that $y = gx$, one also says that such an action is (multiply) *transitive*.

If $G_x = e$ – i.e., g is unique – then one sometimes hears the action referred to as *simply transitive*. For such a group action, the manifold M must be diffeomorphic to the manifold underlying the Lie group G.

In the multiply transitive case, one sees, by composition, that if $g \in G$ takes x to $y = gx$ then the set of all $g' \in G$ that take x to y is in one-to-one correspondence with either G_x or G_y. In particular, G_x and G_y must be isomorphic as Lie groups.

An elementary example of a homogeneous space is any affine space A^n, which has a simply transitive action of the translation group $\mathbb{R}^n$ on it. The (real or complex) n-sphere is diffeomorphic to the homogeneous space $SO(n+1)/SO(n)$, and the n-torus $T^n = S^1 \times \ldots \times S^1$ is the homogeneous space $\mathbb{R}^n / \mathbb{Z}^n$; in particular, $S^1 = \mathbb{R}/\mathbb{Z}$.

Isotropy subgroups at different points do not have to be isomorphic, although this is true for points on the same orbit. Conversely, if all isotropy subgroups are isomorphic then it does not have to be the case that there only one orbit; one refers to the set of all $x \in M$ that have the same isotropy subgroup, up to isomorphism, as the *stratum* of G_x. If one considers the example of $SO(n)$ acting on $\mathbb{R}^n - \{0\}$ then one sees that all of the isotropy subgroups are isomorphic to $SO(n-1)$, but the orbits are $n-1$-spheres of differing radii. In this example, the stratum of a given isotropy subgroup defines a manifold.

---

[2] One might confer such references as [**12**] on this matter.

[3] It is also called the *stability subgroup* and the *little subgroup*.

An important example of a group action that is perhaps the best understood case in the theory of transformation groups, while also being quite pervasive in its application to physics, is that of a *linear representation*. In that case, the manifold is a vector space V and the action $G\times V \to V$ is required to be linear, in that the left-translation operator L_g: $V \to V$ is an invertible linear map for every $g\in G$. As a consequence, the image of the group homomorphism L: $G \to \mathrm{Diff}(V)$ is going to be a subgroup of $GL(V)$. When the kernel of this homomorphism is the identity element in G one calls the representation *faithful*, since the map L then becomes an isomorphism of G with its image in $GL(V)$.

By differentiation at the identity, a representation D: $G \to GL(V)$ gives a representation $\mathfrak{D}$: $\mathfrak{G} \to \mathfrak{gl}(V)$ of the Lie algebra of G in the Lie algebra of $GL(V)$. Hence, if $a, b \in G$ then:

$$[\mathfrak{D}(a), \mathfrak{D}(b)] = \mathfrak{D}[a, b]. \tag{XI.9}$$

When G acts on two differentiable manifolds M and N and one has a differentiable map f: $M \to N$, one might wish to compare the two group actions. The actions are said to be *equivariant* iff f takes every orbit of G in M to a subset of an orbit of G in N. One also says that f *commutes* with the group actions, since equivariance is equivalent to the condition that $gf(x) = f(gx)$ for every $g \in G$ and $x\in M$. Of particular interest is the case in which $N = V$ is a vector space and the action of G on N is linear, so one is considering actions of G on a manifold M that are equivariant to a linear action on some vector space. Actually, there is an equivariant embedding theorem that says that this is *always* possible if one goes to a large enough dimension of V.

Suppose that g: $(-\varepsilon, +\varepsilon) \to G$, $s \mapsto g(s)$ is a smooth curve through the identity, and its tangent vector at the identity is $\mathfrak{g} \in \mathfrak{G}$. If G acts on M then there is a smooth curve $g(s)x$ through each point $x \in M$ such that $g(0)x = x$.

By differentiation, there is then a tangent vector to x:

$$\tilde{\mathfrak{g}}(x) = \left.\frac{d(g(s)x)}{ds}\right|_{s=0}, \tag{XI.10}$$

and the association of x with $\tilde{\mathfrak{g}}(x)$ defines a vector field on M that one calls the *fundamental vector field* that is associated with $\mathfrak{g}$. In fact, the association of $\mathfrak{g}$ with $\tilde{\mathfrak{g}}$ defines a Lie algebra homomorphism $\mathfrak{G} \to \mathfrak{X}(M)$. That is, in any case, one has:

$$\widetilde{[\mathfrak{a},\mathfrak{b}]} = [\tilde{\mathfrak{a}},\tilde{\mathfrak{b}}]. \tag{XI.11}$$

Hence, the group action defines a representation of the Lie algebra $\mathfrak{G}$ in the Lie algebra $\mathfrak{X}(M)$. If this representation is faithful then each generator of $\mathfrak{G}$ – i.e., each member e_i, $i = 1, \ldots, \dim(\mathfrak{G})$ of a basis for $\mathfrak{G}$ – is associated with a fundamental vector field $\tilde{e}_i$ on E and the elements of the set $\{\tilde{e}_i\}$ are linearly independent.

d. Groups acting on vector bundles. Suppose that a Lie group G acts smoothly on the total space to a vector bundle E from the left and commutes with the projection p: $E \to M$. To say that the action commutes with the projection is to say that for each $x \in M$ it takes every element of a given fiber E_x to elements of the same fiber E_y. Hence, there is a well-defined action of G on M by projection: $G \times M \to M$, $(g, x) \mapsto p(g\phi(x))$, where ϕ is any element of E_xM; since the action of G on E commutes with projection, the choice of ϕ is immaterial. One can then express the condition of commutativity in either of the forms $gp = pg$ or $g\phi(x) = \phi(gx)$.

There are essentially two types of actions of G on E: *vertical actions* and *motions*, depending upon whether g takes $\phi \in E_xM$ to another element $g\phi$ of the same fiber or to an element of another fiber E_yM, respectively. The vertical actions then project to the identity on M, while the motions project to non-trivial diffeomorphisms of M.

Since the fibers have a linear structure, one can also further classify vertical actions as linear or nonlinear. In the linear case, one must always have:

$$g(\alpha\phi + \beta\psi) = \alpha\, g(\phi) + \beta g(\psi) \tag{XI.12}$$

for all scalars α, β and elements ϕ, ψ of any fiber of E. This then implies that there is an action of G on the vector space $\Gamma(E)$:

$$g(\phi(x)) = (g\phi)(x). \tag{XI.13}$$

One sees that a linear action of G on a fiber E_x of a vector bundle will be the same thing as a representation of G in $GL(E_x)$. Indeed, the vector bundle E is usually defined in physical applications by first starting with a representation of G in a vector space V and a G-principal bundle $P \to M$ that defines the G-gauge structure on M and obtaining E as the *associated vector bundle* to that G-principal bundle and representation of V. This means that one first forms the manifold $P\times V$ and lets G act on it by taking (p, x) to (pg^{-1}, gx) and defining E to be the orbit space of this action; i.e., each point of E is an orbit of the action of G on $P\times V$. Although this sounds somewhat abstruse, actually, if one lets G be $GL(n)$, P be the bundle $GL(M)$ of linear frames on M, and V be $\mathbb{R}^n$ then the orbits of the action of $GL(n)$ on $GL(M)\times\mathbb{R}^n$ consist of pairs $(\mathbf{e}_i, v^i)$ that all describe the same tangent vector $v^i\, \mathbf{e}_i$; i.e., the associated vector bundle to $GL(M)$ and $\mathbb{R}^n$ is simply $T(M)$.

A typical fundamental vector field for an action of G on E will have the local form:

$$\tilde{\mathfrak{g}} = X_{\mathfrak{g}}^{\ \mu} \frac{\partial}{\partial x^\mu} + X_{\mathfrak{g}}^{\ A} \frac{\partial}{\partial \phi^A}\,. \tag{XI.14}$$

2. Symmetries of the action functional. Now let us apply some of the aforementioned concepts regarding group actions on manifolds and vector bundles to the problems of the calculus of variations.

Let $E \to M$ be a vector bundle upon which the Lie group G acts on the left, so J^1E is the manifold of all 1-jets of sections of E. Although it is possible to define the

prolongation of the action of G on E to an action of G on J^1E, for our purposes it will be sufficient to deal with the infinitesimal case, since we are only dealing with variations of fields [4].

If $\mathfrak{g} \in \mathfrak{G}$ then there is a fundamental vector field $\tilde{\mathfrak{g}}$ on E defined by the group action. One can prolong $\tilde{\mathfrak{g}}$ to a vector field $j^1\tilde{\mathfrak{g}}$ on $J^1(M, E)$ by differentiation. If $\tilde{\mathfrak{g}}$ has the local form that is given by (XI.14) then $j^1\tilde{\mathfrak{g}}$ has the local form:

$$j^1\tilde{\mathfrak{g}} = X_{\mathfrak{g}}^{\mu}\frac{\partial}{\partial x^{\mu}} + X_{\mathfrak{g}}^{A}\frac{\partial}{\partial \phi^{A}} + \frac{\partial X_{\mathfrak{g}}^{A}}{\partial x^{\mu}}\frac{\partial}{\partial \phi^{A}_{\mu}}. \qquad \text{(XI.15)}$$

Furthermore, let there be given an action functional $S[\phi]$ on E that is defined by a Lagrangian density $\mathcal{L}$ on J^1E. Its first variation functional then takes the form:

$$\delta S[X] = \int_{\operatorname{supp}\phi}[i_X d(\mathcal{L}\mathcal{V}) + d i_{X_M}(\mathcal{L}\mathcal{V})] = \int_{\operatorname{supp}\phi}[d\mathcal{L}(X) + \mathcal{L}\operatorname{div}(X_M)]\mathcal{V}, \qquad \text{(XI.16)}$$

in which X_M refers to the part of $X \in \mathfrak{X}(J^1E)$ that projects onto $T(M)$ in a non-trivial way, and we now use the notation div for $\#^{-1}\mathrm{d}\#$, so as not to be confused with the symbol for variation.

Locally it then looks like:

$$X_M = X^{\mu}\frac{\partial}{\partial x^{\mu}}. \qquad \text{(XI.17)}$$

Previously, the variations $\delta\phi$ that we considered for the sake of deriving the Euler-Lagrange equations were necessarily vertical, so this part of X would be zero in such a case.

Now, let us apply the first-variation functional $\delta S[.]$ to the vector field $j^1\tilde{\mathfrak{g}}$ as in (XI.15).

When one evaluates the integrand in (XI.16) on this vector field, one obtains, with some tedious manipulation:

$$d\mathcal{L}(X) + \mathcal{L}\operatorname{div} X_M = \frac{\delta\mathcal{L}}{\delta\phi^{A}}X_{\mathfrak{g}}^{A} + \frac{\partial}{\partial x^{\mu}}\left(\mathcal{L}X_{\mathfrak{g}}^{\mu} + \frac{\partial\mathcal{L}}{\partial\phi^{A}_{\mu}}X_{\mathfrak{g}}^{A}\right). \qquad \text{(XI.18)}$$

If the section ϕ on which $\delta S[\, j^1\tilde{\mathfrak{g}}\,]$ is being evaluated is extremal, moreover, then the first term vanishes and when the first variation functional is evaluated on $j^1\tilde{\mathfrak{g}}$ for an extremal field ϕ one ultimately obtains:

$$\delta S[\, j^1\tilde{\mathfrak{g}}\,] = \int_{\operatorname{supp}\phi}\#\operatorname{div}[\mathbf{J}(\mathfrak{g})] = \int_{\partial\operatorname{supp}\phi}\#\mathbf{J}(\mathfrak{g}), \qquad \text{(XI.19)}$$

[4] More generally, one can confer Saunders [**13**].

in which we have defined the vector field on supp ϕ:

$$\mathbf{J}(\mathfrak{g}) = \left(\mathcal{L} X_{\mathfrak{g}}^{\ \mu} + \frac{\partial \mathcal{L}}{\partial \phi^{A}_{\ \ \mu}} X_{\mathfrak{g}}^{\ A} \right) \frac{\partial}{\partial x^{\mu}} . \tag{XI.20}$$

An element $\mathfrak{g} \in \mathfrak{G}$ is said to be *an infinitesimal symmetry* of the action functional $S[.]$ iff $\delta S[\, j^1 \tilde{\mathfrak{g}}\,]$ vanishes for every extremal section ϕ. Note that this does not imply that the integrand in (XI.19) must vanish identically, since one can also add an exact n-form $d\Lambda$ such that Λ vanishes on $\partial(\text{supp}\ \phi)$ to it without changing value of the integral.

We finally arrive that one of the most pervasive theorems of the variational calculus from the standpoint of modern physics. (Confer any of the variational treatments of electromagnetism, such as [**14-17**].)

Noether's theorem:

If $\mathfrak{g} \in \mathfrak{G}$ is an infinitesimal symmetry of an action functional for an extremal field ϕ then the vector field on M:

$$\mathbf{J}(\mathfrak{g}) = \left(\mathcal{L} X_{\mathfrak{g}}^{\ \mu} + \frac{\partial \mathcal{L}}{\partial \phi^{A}_{\ \ \mu}} X_{\mathfrak{g}}^{\ A} \right) \frac{\partial}{\partial x^{\mu}} \tag{XI.21}$$

has vanishing divergence.

Such a vector field on M is called the *Noether current* that is associated with $\mathfrak{g}$. When every element of $\mathfrak{G}$ is associated with a Noether current, one has a Lie algebra homomorphism $\mathfrak{G} \to \mathfrak{X}(M)$, $\mathfrak{g} \mapsto \mathbf{J}(\mathfrak{g})$. That is, $[\mathbf{J}(\mathfrak{g}_1), \mathbf{J}(\mathfrak{g}_2)] = \mathbf{J}([\mathfrak{g}_1, \mathfrak{g}_2])$.

Now, suppose that G acts on M, as well as on E. Hence, one has a homomorphism L: $G \to \text{Diff}(M)$, $g \mapsto L_g$ of the Lie group G and a homomorphism σ: $\mathfrak{G} \to \mathfrak{X}(M)$, $\mathfrak{g} \mapsto \sigma(\mathfrak{g}) = \tilde{\mathfrak{g}}$ of the Lie algebra $\mathfrak{G}$. If we assume, moreover, that the representation is linear then we can represent the fundamental vector field $\tilde{\mathfrak{g}}$ in local form as:

$$\tilde{\mathfrak{g}} = (\sigma_a^{\mu} \mathfrak{g}^a) \partial_{\mu} , \tag{XI.22}$$

in which the components of the $n \times \dim(\mathfrak{G})$ matrices σ_a^{μ} are smooth function on $U \subset M$.

When the vector field $X_{\mathfrak{g}}$ on E is the push-forward $d\phi(\tilde{\mathfrak{g}})$ of a fundamental vector field $\tilde{\mathfrak{g}}$ on M by a section ϕ: $M \to E$, it takes the local form:

$$X_{\mathfrak{g}} = (\sigma_a^{\mu} \mathfrak{g}^a) \frac{\partial}{\partial x^{\mu}} + \left(\frac{\partial \phi^A}{\partial x^{\mu}} \sigma_a^{\mu} \mathfrak{g}^a \right) \frac{\partial}{\partial \phi^A} ; \tag{XI.23}$$

i.e.:

$$X_{\mathfrak{g}}^{\ \mu} = \sigma_a^{\mu}\mathfrak{g}^a\,, \qquad\qquad X_{\mathfrak{g}}^{\ A} = \frac{\partial \phi^A}{\partial x^{\mu}} X_{\mathfrak{g}}^{\ \mu} = \phi^A{}_{,\mu}\sigma_a^{\mu}\mathfrak{g}^a\,. \tag{XI.24}$$

More generally, the components $X_{\mathfrak{g}}^{\ A}$ also include a purely vertical contribution $\overline{X}_{\mathfrak{g}}^{\ A}$ when G acts linearly on the fibers of E as a structure – or *gauge* – group. There is then a representation D: $G \to GL(V)$, $g \mapsto D(g)$ and a corresponding representation $\mathfrak{D}$: $\mathfrak{G} \to \mathfrak{gl}(V)$, $\mathfrak{g} \mapsto \mathfrak{D}(\mathfrak{g})$. The purely vertical contribution then takes the form:

$$\overline{X}_{\mathfrak{g}}^{\ A} = \mathfrak{D}_a^A\mathfrak{g}^a = X_{\mathfrak{g}}^{\ A} + \phi^A_{,\mu}X_{\mathfrak{g}}^{\ \mu}\,, \tag{XI.25}$$

When we substitute this, along with (XI.24), into (XI.21), we see that the Noether current that is associated with $\mathfrak{g} \in \mathfrak{G}$ has the local form:

$$J^{\mu}(\mathfrak{g}) = \left[T_{\nu}^{\mu}\sigma_a^{\nu} + S_a^{\mu}\right]\mathfrak{g}^a\,, \tag{XI.26}$$

in which:

$$T_{\nu}^{\mu} = \mathcal{L}\delta_{\nu}^{\mu} - \frac{\partial \mathcal{L}}{\partial \phi^A{}_{\mu}}\frac{\partial \phi^A}{\partial x^{\nu}} = \mathcal{L}\delta_{\nu}^{\mu} - \Pi_A^{\mu}\phi^A{}_{,\nu}\,. \tag{XI.27}$$

define the components of the *canonical energy-momentum tensor* on M, and the remaining term in the bracket:

$$S_a^{\mu} = -\,\Pi_A^{\mu}\mathfrak{D}_a^A \tag{XI.28}$$

is referred to as the *spin* tensor.

Hence, we can say that the matrix components of the Noether homomorphism J: $\mathfrak{G} \to \mathfrak{X}(U)$, $\mathfrak{g} \mapsto \mathbf{J}(\mathfrak{g})$, relative to the bases for both vector spaces that have been using, are:

$$J_a^{\mu} = T_{\nu}^{\mu}\sigma_a^{\nu} + S_a^{\mu}\,. \tag{XI.29}$$

3. Examples of Noether symmetries. We shall examine some examples of the Noether representation for various groups of motions that might act on a manifold M and gauge groups that act on the fibers of a vector bundle E.

a. Translations. The first example of the Noether representation that we examine is the case of the translation group acting on M. Of course, when M is not an affine space, one must keep in mind that the very concept of translation has a largely local and approximate character on such a manifold, and that it is more natural to start with local *convections*, which are local diffeomorphisms about its points. This is equivalent to regarding extended matter distributions as physically fundamental, while regarding pointlike matter as an asymptotic approximation that implies a high degree of rigidity to the distribution.

That notwithstanding, let $\mathfrak{G} = \mathbb{R}^n$. The representation $\sigma\colon \mathbb{R}^n \to \mathfrak{X}(U)$ simply takes ε^μ to the local vector field $\varepsilon^\mu \partial_\mu$ on $U \subset M$: hence, $\sigma^\mu_\nu = \delta^\mu_\nu$. One generally does not have the translation group for a manifold acting on the fibers of a vector bundle, although it is reasonable in the case of an affine bundle. Hence, one lets $\mathfrak{D}^A_\mu = 0$.

These substitutions reduce the matrix J^μ_a of the Noether homomorphism to simply T^μ_ν. At each point $x \in U$, the vector field $J^\mu(\varepsilon^\nu) = T^\mu_\nu \varepsilon^\nu$ is then proportional to the stress that acts on a unit area tangent plane to x whose normal points in the direction specified by ε^ν. In the case of a four-dimensional spacetime manifold M, one can also interpret:

$$T^0_0 = \mathcal{L} - \Pi^0_A \phi^A{}_{,0} = -\mathcal{H} \tag{XI.30}$$

as an energy density and:

$$T^i_0 = -\Pi^i_A \phi^A{}_{,0} = -p^i \tag{XI.31}$$

as a spatial momentum flux.

This equivalence of components comes about due to the somewhat coincidental equality of the units for energy density, momentum flux, and stress (i.e., pressure), despite the fact that they describe a 3-form and two 2-forms with values in a vector space, respectively.

The fact that $\mathbf{J}(\varepsilon)$ has vanishing divergence for any ε and the fact that ε represents a constant gives the consequence that:

$$T^\mu_{\nu,\mu} = 0, \tag{XI.32}$$

which often serves as one of the fundamental sets of partial differential equations for continuum mechanics. Since the divergence on the left-hand side represents a generalized force density, one sees that the invariance of the action functional under translations implies the vanishing of all forces in spacetime. By contraposition, one can then say that presence of a force field f_ν in M "breaks" the translation invariance of the action; hence, $T^\mu_{\nu,\mu} = f_\nu$, in that case.

It is traditional at this point to note that since the component matrix T^μ_ν is not generally symmetric one can obtain a symmetric stress-energy-momentum tensor field by the Belinfante-Rosenfeld process. However, as we have defined things, it becomes clear that in a pre-metric formalism, it is meaningless to discuss the symmetry of a second-rank tensor that is of mixed type except as a component matrix when one chooses a frame, but the resulting definition of symmetry is not frame-invariant. One could only define symmetry in a frame-invariant manner by raising or lowering one index, and without a metric to define the isomorphism, the process of raising or lowering indices becomes ill-defined.

It is important to realize that the matrix T^μ_ν does not, by any means, have to be invertible. Indeed, its only non-zero component might very well be a rest energy density for T^0_0. Hence, it does not define a linear isomorphism of tangent vectors at each point, but something more like an infinitesimal deformation of tangent vectors.

Of course, one cannot avoid mentioning that the expression for T_0^0 in (XI.30) bears a less-than-coincidental resemblance to the Legendre transformation by which one associates a Hamiltonian density $\mathcal{H}$ with a Lagrangian density $\mathcal{L}$. Once again, the only way that can single out a specific component of T_ν^μ to represent an energy density is by choosing a time-space splitting of $T(M)$, which is a recurring theme in any relativistic treatment of energy, in general.

b. Rotations. Now suppose we consider only the three-dimensional spatial manifold Σ, with $G = SO(3)$, and we use the basis for $\mathfrak{so}(3)$ that is defined by $[e_i]_k^j = \varepsilon_{ijk}$, $i, j, k = 1, 2, 3$. Hence, an element $\omega \in \mathfrak{so}(3)$ can be represented in the form of the matrix $\omega^j \varepsilon_{ijk}$.

For any coordinate chart (U, x^i), one can represent the basis elements of $\mathfrak{so}(3)$ by the three fundamental vector fields:

$$L_k = \varepsilon_{ijk}\, x^i \partial_j \,, \tag{XI.33}$$

which look like:

$$L_x = y\partial_z - z\partial_y \,, \qquad L_y = z\partial_x - x\partial_z \,, \qquad L_z = x\partial_y - y\partial_x \,, \tag{XI.34}$$

explicitly.

Hence, the fundamental vector field that corresponds to ω is:

$$\begin{aligned} \tilde{\omega} &= \omega^k L_k = \varepsilon_{ijk}\, \omega^k x^i \partial_j \\ &= (\omega_y z - \omega_z y)\, \partial_x + (\omega_z x - \omega_x z)\, \partial_y + (\omega_x y - \omega_y x)\, \partial_z \,, \end{aligned} \tag{XI.35}$$

which has the same form as $\boldsymbol{\omega} \times \mathbf{r}$ from classical rotational mechanics. Hence, the fundamental vector field associated with ω represents the tangential velocity to the circle through each point of U that represents one of the orbits of the one-parameter family of rotations $g(t) = \exp(\omega t)$ that is generated by ω.

From (XI.33), we see that the matrix of the representation σ: $\mathfrak{so}(3) \to \mathfrak{X}(U)$ that we are using is:

$$\sigma_j^i = \varepsilon_{ijk}\, x^k \,. \tag{XI.36}$$

As for the action of $SO(3)$ on the fibers of E, we assume only that it is non-trivial, so the matrix $\mathfrak{D}_j^A$ is non-vanishing. By substituting (XI.36) into (XI.29), we see that the matrix J_j^i of the Noether homomorphism J: $\mathfrak{so}(3) \to \mathfrak{X}(U)$ takes the form:

$$J_j^i = T_k^i \sigma_j^k + S_j^i = T_k^i (\varepsilon_{kjl} x^l) + S_j^i = L_j^i + S_j^i \,, \tag{XI.37}$$

in which:

$$L_j^i = -\tfrac{1}{2} \varepsilon_{jkl} (T_k^i x^l - T_l^i x^k) \tag{XI.38}$$

represents the *orbital angular momentum* part of the homomorphism and:

$$S_j^i = -\,\Pi_A^i \mathfrak{D}_j^A \tag{XI.40}$$

represents the *intrinsic angular momentum* – or *spin* – part of the homomorphism.

It is important to see that the spin does not generally represent a contribution from "internal" orbital angular momenta, such as the rotation of the Earth as it orbits around the Sun, but a contribution that is more related to the nature of the action of the group in question on the fibers of E. In general, when the action is linear, so one is dealing with a representation of the Lie group in the typical fiber of E, the eigenvalues of the first-order differential operators $S_j = S^i_j \partial_i$ will be closely related to the "weight" of the representation. In particular, when ϕ is a scalar field, the weight is zero and S^i_j must vanish.

The extension from $SO(3)$ acting on Σ locally to the special Lorentz group $SO(3, 1)$ acting on M locally is straightforward. In addition to extending the coordinate indices to $\mu = 0, \ldots, 3$, etc., one must add three more generators (i.e., basis elements) K_i, $i = 1, 2, 3$ to the Lie algebra of $\mathfrak{so}(3)$ that represent the infinitesimal boosts in the spatial coordinate directions. Hence, if we index the generators of $\mathfrak{so}(3, 1)$ by $a = 1, \ldots, 6$ then we have to express the matrix of the Noether homomorphism J: $\mathfrak{so}(3, 1) \rightarrow \mathfrak{X}(U)$ in the form J^μ_a that we defined in (XI.29).

c. Homotheties. The group of *homotheties* – or *dilatations*, as they are called in continuum mechanics – is a one-dimensional subgroup of $GL(n)$ for any n that is isomorphic to the multiplicative group $(\mathbb{R}^+, \times)$ of positive real numbers, so all of its elements take the form λI, $\lambda > 0$. Hence, any non-trivial representation of it acting locally on $U \subset M$ would also have to be one-dimensional.

In effect, any vector field $\mathbf{v}$ might serve as a choice for that representation, which means that one can use the intuition gained from the geometrical representation of systems of ordinary differential equations and their singularities narrow down the choice. For one thing, as long as $\mathbf{v}$ has no zeroes, the integral curves of $\lambda\mathbf{v}$ will describe the same points in M; the only possible difference will be in the parameterization. Hence, the interesting case is when $\mathbf{v}$ has zeroes, especially isolated zeroes, such as sources and sinks.

For the representation of homotheties, the single infinitesimal generator of a dilatation centered on $x_0 \in U$, which is a zero of "source" type, is the *radius vector field* that (U, x^μ) defines:

$$\mathbf{r}(x) = x^\mu \partial_\mu . \tag{XI.41}$$

The one-parameter family of dilatations that $\mathbf{r}$ defines on U is then:

$$\exp(\lambda \mathbf{r}(x)) = e^\lambda x^\mu . \tag{XI.42}$$

Since $\mathfrak{G}$ is one-dimensional in this case, the representation matrix σ takes the form $\sigma^\mu = x^\mu$.

The action of $(\mathbb{R}^+, \times)$ on the fibers of any real vector bundle is by scalar multiplication. The fundamental vector field that it generates is a vertical vector field $\mathbf{R}$

on E – (zero section) that is sometimes called the *Liouville vector field* and assigns the radius vector $\mathbf{R}(v)$ – within the fiber – to each $v \in E$; i.e., the displacement in the fiber that takes the origin to v. For a local trivialization $U \times V$ of $E(U)$ it then takes the form:

$$\mathbf{R}(x, v) = v^A \frac{\partial}{\partial v^A}. \tag{XI.43}$$

Hence, the matrix of $\mathfrak{D}$ takes the form $\mathfrak{D}^A = v^A$, and the spin tensor becomes:

$$S^\mu = -\frac{\partial L}{\partial v^A_\mu} v^A. \tag{XI.44}$$

The Noether homomorphism J: $\mathbb{R} \to \mathfrak{X}(U)$, $\lambda \mapsto J(\lambda)$ that is associated with infinitesimal spacetime dilatations then has the components:

$$J^\mu(\lambda) = T^\mu_\nu x^\nu + S^\mu. \tag{XI.45}$$

The vanishing of its divergence implies that:

$$T^\mu_\mu = 0. \tag{XI.46}$$

Hence, when the action functional for a field theory is invariant under dilatation the trace of the energy-momentum tensor must vanish. Since this trace amounts to the rest energy density, the conservation law is basically conservation of mass-energy.

d. Gauge transformations. Now, let us consider a purely vertical action of a group G on the fibers of E, which makes G essentially a *gauge group*. More precisely, one usually has a right action of G on the "associated principal bundle" to E, namely, the bundle $GL(E)$ of all linear frames in the fibers of E. For instance, one often considers the bundles $GL(M)$ of all linear frames in $T(M)$ and $GL^*(M)$, which consists of linear frames in T^*M.

The right action then comes about, in most cases, by representing G as a subgroup of $GL(V)$, where V is the model vector space for the typical fiber of E. That is, one represents $g \in G$ by a matrix g^A_B in $GL(V)$ and its action on a frame e_A in any fiber of E gives the linear frame $e_A \tilde{g}^A_B$, where the tilde again represents the inverse of the matrx.

The action of G on the components of any vector $v = v^A e_A$ in E is then a left action that takes v^A to $g^B_A v^A$. This "contragredient" action of G on frames and components is necessary to insure that the vector v remains an invariant object in the process.

Hence, we shall think of the representation $\mathfrak{D}$: $\mathfrak{G} \to \mathfrak{gl}(V)$ as being more relevant to the action of G on $GL(E)$ than on E itself. However, due to the linear structure on each fiber one can still speak of the direct action of $\mathbb{R}^+$ by scalar multiplication and each fiber

acts on itself by translations; also, there are often various finite groups that can act naturally, such as $\mathbb{Z}_2$ acting as multiplication by ± 1.

When E is a complex vector bundle – so its fibers are complex vector spaces – one can also think of a natural action of $U(1)$ on the fibers as part of complex scalar multiplication, since the multiplicative group $\mathbb{C}^*$ of non-zero complex numbers is isomorphic to $\mathbb{R}^+ \times U(1)$ by polar representation. The subgroup $U(1)$ then consists of the points of the unit circle in $\mathbb{C}$, which are usually represented in the form $e^{i\theta}$.

If the action of G on E is the vertical action of a gauge group on the fibers then the matrix of the Noether homomorphism J: $\mathfrak{G} \to \mathfrak{X}(U)$ reduces to:

$$J_a^\mu = S_a^\mu . \tag{XI.47}$$

In the case of a one-dimensional $\mathfrak{G}$, the $\mathfrak{D}^A$ represent the components of some section of $E \to M$ and the single conserved current associated with any $\lambda \in \mathfrak{G}$ is the vector field whose local components are:

$$\lambda J^\mu = -\lambda \Pi_A^\mu \mathfrak{D}^A . \tag{XI.48}$$

One generally thinks of the conserved quantities that are described by the gauge symmetries of an action functional as "charges." When we treat the case of electromagnetic actions, in particular, we shall see that this interpretation is indeed appropriate, although in the case of gauge groups of dimension greater than one, one must be careful to note that the charge in question does not have to be electric in character. For instance, in quantum chromodynamics, which is regarded as an $SU(3)$ gauge field theory, the charges associated with the eight generators of the Lie algebra of $\mathfrak{su}(3)$ are regarded as "colors."

4. Symmetries of electromagnetic action functionals. We now apply the general methods that we just described to the case of electrostatics, magnetostatics, and electromagnetism, more generally. In each case, we start with the field Lagrangian and compute the resulting matrices in the Noether homomorphism. We particularly examine the eigenvectors and eigenvalues of the energy-momentum-stress tensor.

a. Electrostatics. The Lagrangian density for electrostatics takes the form:

$$\mathcal{L}(x, E) = \tfrac{1}{2} D^i E_i = \tfrac{1}{2} \varepsilon^{ij} E_i E_j . \tag{XI.49}$$

The matrix of the Noether homomorphism for this particular Lagrangian density is:

$$J_a^i = T_j^i \sigma_a^j + S_a^i , \tag{XI.50}$$

with:

$$T^i_j = \mathcal{L}\delta^i_j - \frac{\partial \mathcal{L}}{\partial \phi_{,i}}\phi_{,j} = \tfrac{1}{2}(D^k E_k)\delta^i_j - D^i E_j\,, \qquad S^i_a = -\,D^i\,\mathfrak{D}_a\,. \tag{XI.51}$$

One can express this construction in basis-free form as:

$$T = \mathcal{L}I - \mathbf{D}\otimes E, \qquad S = -\,D\otimes\mathfrak{D}. \tag{XI.51}$$

in which I represents the identity transformation that acts on tangent vectors.

If one thinks of T as an infinitesimal transformation of tangent vectors then its effect, when applied to a tangent vector **n** can be expressed as:

$$T(\mathbf{n}) = \mathcal{L}\mathbf{n} - E(\mathbf{n})\mathbf{D}, \tag{XI.53}$$

Since ε defines a scalar product on tangent vectors, we can express **n** as:

$$\mathbf{n} = \frac{\varepsilon(\mathbf{n},\mathbf{D})}{\varepsilon(\mathbf{D},\mathbf{D})}\mathbf{D} + \mathbf{n}_\perp = \frac{2}{\mathcal{L}}E(\mathbf{n})\mathbf{D} + \mathbf{n}_\perp, \tag{XI.54}$$

where $\mathbf{n}_\perp$ represents the part of **n** that is orthogonal to **D** under ε. One recalls that the planes that are orthogonal to the vector field **D** will be tangent to the equipotential surfaces of ϕ since they will be annihilated by the 1-form E.

This makes (XI.53) take the form:

$$T(\mathbf{n}) = \mathcal{L}\mathbf{n}_\perp - \tfrac{1}{2}E(\mathbf{n})\mathbf{D}. \tag{XI.55}$$

Note that **D**, as well as any vector field $\mathbf{n}_\perp$ that is tangent to the equipotential surfaces, is an eigenvector of T. In the former case, the corresponding eigenvalue is $-\mathcal{L}$, while in the latter case it is $+\mathcal{L}$. Indeed, from the form (XI.52) of T the eigenvectors of T are the same as the eigenvectors of $\mathbf{D}\otimes E$, while the eigenvalues of T are $\mathcal{L} - \lambda$, with $\lambda = 0$ or $E(\mathbf{D}) = 2\mathcal{L}$.

Generally, the matrix T^i_j is not symmetric, and its polarization into a symmetric and an anti-symmetric part is:

$$\overset{+}{T}{}^i_j = \tfrac{1}{2}(T^i_j + T_i{}^j) = \tfrac{1}{2}[(D^k E_k)\delta^i_j - D^i E_j - D^j E_i], \tag{XI.56a}$$

$$\overset{-}{T}{}^i_j = \tfrac{1}{2}(T^i_j - T_i{}^j) = -\tfrac{1}{2}(D^i E_j - D^j E_i). \tag{XI.56b}$$

A necessary and sufficient condition for T^i_j to be symmetric is that there be a non-zero real number α such that $D^i = \alpha E_i$, which is equivalent to saying that the dielectric is isotropic. (Sufficiency is obvious. To see necessity, contract both sides of $D^i E_j = D^j E_i$ with D^j, and note that both $D^j D^j = \delta_{ij} D^i D^j$ and $E_j D^j = E(\mathbf{D})$ are non-vanishing when **D** is.

Although the introduction of an auxiliary Euclidian metric has no physical meaning, it does serve to prove the theorem.)

The trace of T^i_j is:

$$T^i_i = \tfrac{1}{2} D^i E_i = \mathcal{L}, \tag{XI.57}$$

which is consistent with the sum of the eigenvalues, including multiplicity. Hence, the one-parameter family of transformations that T^i_i generates does not consist of volume-preserving transformations.

b. Magnetostatics. The Lagrangian for magnetostatics takes the form:

$$\mathcal{L}(x, B) = \tfrac{1}{4} H^{ij} B_{ij} = \tfrac{1}{4} \tilde{\mu}^{ijkl} B_{ij} B_{kl}\,, \tag{XI.58}$$

when expressed in terms of the 2-form B and the bivector field $\mathbf{H}$, or:

$$\mathcal{L}(x, B) = \tfrac{1}{2} H^i B_i = \tfrac{1}{2} \tilde{\mu}^{ij} B_i B_j \tag{XI.59}$$

in terms of the vector field $\mathbf{B}$ and the 1-form H.

The Noether homomorphism for this particular Lagrangian density takes the form:

$$J^i_a = T^i_j \sigma^j_a + S^i_a \tag{XI.60}$$

with:

$$T^i_j = \mathcal{L}\delta^i_j - \frac{\partial \mathcal{L}}{\partial A_{ki}} A_{k,j} = \tfrac{1}{4}(H^{kl} B_{kl})\delta^i_j - H^{ki} B_{kj}\,, \tag{XI.61a}$$

$$S^i_a = H^{ij}\, \mathfrak{D}_{ja}\,. \tag{XI.61b}$$

in terms of B and $\mathbf{H}$ or:

$$T^i_j = \mathcal{L}\delta^i_j - \varepsilon^{jkm}\varepsilon_{ikn} \frac{\partial \mathcal{L}}{\partial B^m} B^n = -\tfrac{1}{2}(B^k H_k)\delta^i_j + B^i H_j\,, \tag{XI.62a}$$

$$S^i_a = H_j \mathfrak{D}^{ij}_a \;(\mathfrak{D}^{ij}_a \equiv \varepsilon^{ijk}\mathfrak{D}_{ka}) \tag{XI.62b}$$

in terms of $\mathbf{B}$ and H.

The basis-free representation of the tensor field T is now:

$$T = -\mathcal{L}\, I + \mathbf{B} \otimes H, \tag{XI.63}$$

which has the same form as minus the result (XI.52) that we obtained above for the electrostatic case. Hence, the eigenspaces are the same as before, except that the signs of the eigenvalues are switched. Of course, the tangent planes that are orthogonal to $\mathbf{B}$ under the scalar product that is defined by $\tilde{\mu}$ are no longer tangent to equipotential surfaces, since the potentials for B are 1-forms, not 0-forms.

As before in the electrostatic case, the matrix T is not usually symmetric, and its polarization into a symmetric and an anti-symmetric part is:

$$\overset{+}{T}{}^{i}_{j} = \tfrac{1}{2}[-(B^k H_k)\delta^i_j + B^i H_j + B^j H_i], \qquad \text{(XI.64a)}$$

$$\overset{-}{T}{}^{i}_{j} = \tfrac{1}{2}(B^i H_j - B^j H_i), \qquad \text{(XI.64b)}$$

and its trace is:

$$T^i_i = -\tfrac{1}{2} B^i H_i = -\mathcal{L}, \qquad \text{(XI.65)}$$

so the one-parameter family of transformations that it generates does not consist of volume-preserving transformations in this case either. As before in the electrostatic case, symmetry is equivalent to the isotropy of the magnetic material.

c. Electromagnetism. The electromagnetic field Lagrangian has the form:

$$\mathcal{L}(x, F) = \tfrac{1}{4} H^{\mu\nu} F_{\mu\nu} = \tfrac{1}{4} \kappa^{\kappa\lambda\mu\nu} F_{\kappa\lambda} F_{\mu\nu}. \qquad \text{(XI.66)}$$

The matrix of the Noether homomorphism is then:

$$J^\mu_a = T^\mu_\nu \sigma^\nu_a + S^\mu_a \qquad \text{(XI.67)}$$

with:

$$T^\mu_\nu = \mathcal{L}\delta^\mu_\nu - \frac{\partial \mathcal{L}}{\partial A_{\kappa\mu}} A_{\kappa,\nu} = \tfrac{1}{4}(\mathfrak{h}^{\kappa\lambda} F_{\kappa\lambda})\delta^\mu_\nu - \mathfrak{h}^{\kappa\mu} F_{\kappa\nu}, \qquad \text{(XI.68a)}$$

$$S^\mu_a = -H^{\mu\nu} \mathfrak{D}_{\nu a}. \qquad \text{(XI.68b)}$$

One can use the local frame field ∂_μ and its reciprocal coframe field dx^μ to define four vector fields $\mathfrak{h}_\kappa$ ($\kappa = 0, \ldots, 3$) and four 1-forms F^κ by way of:

$$\mathfrak{h}_\kappa = \mathfrak{h}^{\kappa\mu} \partial_\mu, \qquad F^\kappa = F_{\kappa\nu}\, dx^\nu, \qquad \text{(XI.69)}$$

Of course, these quadruples of vectors and covectors do not, by any means, have to be linearly independent, and some of them might very well vanish.

We can now express the stress-energy-momentum tensor T as:

$$T = \mathcal{L} I - \mathfrak{h}_\kappa \otimes F^\kappa. \qquad \text{(XI.70)}$$

As far as eigenvectors and eigenvalues are concerned, one immediately sees that the eigenvectors of T are identical with those of $\mathfrak{h}_\kappa \otimes F^\kappa$, although the eigenvalues λ of T will differ from those λ' of $\mathfrak{h}_\kappa \otimes F^\kappa$ by the relation:

$$\lambda = \mathcal{L} - \lambda'. \qquad \text{(XI.71)}$$

Hence, one can redirect one's attention to the eigenvectors of $\mathfrak{h}_\kappa \otimes F^\kappa$.

Here, it helps to keep in mind that on a four-dimensional vector space V a non-zero 2-form F – hence, a bivector $\mathfrak{h}$ – can have a rank that equals either two or four, but nothing else. That is, in the rank-two case they can be expressed in the form:

$$F = \alpha \wedge \beta, \qquad \mathfrak{h} = \mathbf{a} \wedge \mathbf{b}, \tag{XI.72}$$

but not uniquely, and in the rank-four case, one has:

$$F = \alpha \wedge \beta + \gamma \wedge \delta, \qquad \mathfrak{h} = \mathbf{a} \wedge \mathbf{b} + \mathbf{c} \wedge \mathbf{d}. \tag{XI.73}$$

In both cases, all of the 1-forms and vectors in question are linearly independent. Hence, a rank-two 2-form defines a 2-frame $\{\mathbf{a}, \mathbf{b}\}$ in V and a 2-coframe $\{\alpha, \beta\}$ in V^*, neither of which are unique. Similarly, a rank-four 2-form defines a 4-frame $\{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$ – i.e., a basis – for V and a 4-coframe $\{\alpha, \beta, \gamma, \delta\}$ in V^*. We shall refer to their elements generically as $\mathbf{e}_a$ and θ^a, respectively, with the understanding that the range of indices is defined by the nature of the problem at hand.

In the rank-2 case, one has:

$$\mathfrak{h}_a = i_{\theta^a}\mathfrak{h} = \theta^a(\mathbf{e}_1)\, \mathbf{e}_2 - \theta^a(\mathbf{e}_2)\, \mathbf{e}_1, \tag{XI.74a}$$

$$F^a = i_{\partial_a} F = \theta^a(\mathbf{e}_1)\, \theta^2 - \theta^a(\mathbf{e}_2)\, \theta^1. \tag{XI.74b}$$

One finds that it is particularly convenient if the 2-frame and the 2-coframe are *projectively* reciprocal, in the sense that:

$$\theta^a(\mathbf{e}_b) = \begin{cases} \neq 0 & a = b, \\ = 0 & a \neq b. \end{cases} \tag{XI.75}$$

This makes:

$$\mathfrak{h}_1 = \gamma_2\, \mathbf{e}_2, \qquad \mathfrak{h}_2 = -\gamma_1\, \mathbf{e}_1, \quad F^1 = \gamma_2\, \theta^2, \qquad F^2 = -\gamma_1\, \theta^1, \tag{XI.76}$$

in which we have defined:

$$\gamma_1 = \theta^2(\mathbf{e}_2), \qquad \gamma_2 = \theta^1(\mathbf{e}_1). \tag{XI.77}$$

One now has:

$$F^1(\mathfrak{h}_1) = (\gamma_2)^2 \gamma_1, \qquad F^2(\mathfrak{h}_2) = (\gamma_1)^2 \gamma_2, \qquad F^1(\mathfrak{h}_2) = F^2(\mathfrak{h}_1) = 0. \tag{XI.78}$$

In this case, one finds that:

$$T = \mathfrak{h}_1 \otimes F^1 + \mathfrak{h}_2 \otimes F^2, \tag{XI.79}$$

and its eigenvectors are $\mathfrak{h}_1$ and $\mathfrak{h}_2$, with associated eigenvalues $F^1(\mathfrak{h}_1)$ and $F^2(\mathfrak{h}_2)$, resp.

One has a similar situation in the rank-four case, although generally one must deal with it in terms of special situations since it admits many possibilities.

Let us look at the two rank-2 cases that we already examined above, namely, electrostatics and magnetostatics. In the former case, one has:

$$F = dt \wedge E, \qquad \mathfrak{h} = \partial_t \wedge \mathbf{D}, \tag{XI.80}$$

which makes:

$$\mathfrak{h}_0 = \mathbf{D}, \qquad \mathfrak{h}_1 = -\partial_t, \qquad F^0 = E, \qquad F^1 = -E(\mathbf{D})\, dt, \tag{XI.81}$$

since one sees that the 2-frame $\{\partial_t, \mathbf{D}\}$ is projectively reciprocal to the 2-coframe $\{dt, E\}$.

This makes:

$$\begin{aligned} T &= \tfrac{1}{2} E(\mathbf{D}) I - \mathbf{D} \otimes E - E(\mathbf{D})\, \partial_t \otimes dt \\ &= -\tfrac{1}{2} E(\mathbf{D})\, \partial_t \otimes dt + \tfrac{1}{2} E(\mathbf{D})\, \partial_i \otimes dx^i - \mathbf{D} \otimes E. \end{aligned} \tag{XI.82}$$

We express this in space-time block matrix form:

$$T = \left[\begin{array}{c|c} -\mathcal{L}_E & 0 \\ \hline 0 & T^i_j(E) \end{array}\right], \tag{XI.83}$$

in which $T^i_j(E)$ the 3×3 electrostatic stress matrix (XI.49) that we derived above.

One notes that the trace of T is now zero, whereas the trace of the spatial part was found to be $\mathcal{L}_E$.

The magnetostatic case is similar, but still requires special treatment, since the role of E and $\mathbf{B}$ are actually played by $\mathbf{B}$ and H, respectively, not vice versa. One starts with:

$$F = -\#\mathbf{B} = -B^i \#\partial_i, \qquad \mathfrak{h} = -\#^{-1}H = -H_i \#^{-1} dx^i, \tag{XI.84}$$

from which it follows that:

$$F^i = \varepsilon_{ijk} B^j dx^k, \qquad \mathfrak{h}_i = \varepsilon^{ijk} H^j \partial_k, \tag{XI.85}$$

which makes:

$$\mathfrak{h}_i \otimes F^i = H(\mathbf{B})\, \partial_i \otimes dx^i - \mathbf{B} \otimes H. \tag{XI.86}$$

and:

$$\begin{aligned} T &= \tfrac{1}{2} H(\mathbf{B})\, \partial_\mu \otimes dx^\mu - H(\mathbf{B})\, \partial_i \otimes dx^i + \mathbf{B} \otimes H \\ &= \tfrac{1}{2} H(\mathbf{B})\, \partial_t \otimes dt - \tfrac{1}{2} H(\mathbf{B})\, \partial_i \otimes dx^i + \mathbf{B} \otimes H. \end{aligned} \tag{XI.87}$$

We express this in space-time block matrix form as:

$$T = \left[\begin{array}{c|c} \mathcal{L}_B & 0 \\ \hline 0 & T^i_j(B) \end{array}\right], \tag{XI.88}$$

in which $T^i_j(B)$ represents the spatial magnetostatic stress tensor (XI.61a) or (XI.62a) that we derived previously. Once again, the trace comes out to zero.

In the generic case of a rank-four 2-form, one has:

$$F = dt \wedge E - \#\mathbf{B}, \qquad \mathfrak{h} = \partial_t \wedge \mathbf{D} - \#^{-1}H. \tag{XI.89}$$

This makes:

$$F^0 = i_{\partial_t} F = E, \qquad F^i = i_{\partial_i} F = -E_i\, dt - i_{\partial_i} \#\mathbf{B} = -E_i\, dt + \#(\partial_i \wedge \mathbf{B}), \tag{XI.90a}$$

$$\mathfrak{h}_0 = i_{dt}\mathfrak{h} = \mathbf{D}, \qquad \mathfrak{h}_i = -D^i\, \partial_t - i_{dx^i} \#^{-1} H = -D^i\, \partial_t + \#^{-1}(dx^i \wedge H). \tag{XI.90b}$$

We can then compute the components of T:

$$T^0_0 = \mathcal{L}_{\text{em}} - E(\mathbf{D}), \qquad T^i_0 = -\#^{-1}(E \wedge H)^j, \tag{XI.91a}$$

$$T^0_j = \#(\mathbf{D} \wedge \mathbf{B})_j, \qquad T^i_j = \mathcal{L}_{\text{em}}\delta^i_j - D^iE_j + B^iH_j. \tag{XI.91b}$$

When the constitutive law has vanishing couplings between E and $\mathbf{B}$, so $\mathcal{L}_{\text{em}} = \frac{1}{2}[E(\mathbf{D}) - H(\mathbf{B})]$, these components take the form:

$$T^0_0 = -\tfrac{1}{2}[E(\mathbf{D}) + H(\mathbf{B})], \qquad T^i_0 = -\#^{-1}(E \wedge H)^j, \tag{XI.92a}$$

$$T^0_j = \#(\mathbf{D} \wedge \mathbf{B})_j, \qquad T^i_j = T^i_j(E) + T^i_j(\mathbf{B}). \tag{XI.92b}$$

One sees that the time-space and space-time components of the stress-energy-momentum tensor field take the form of the elementary Poynting vector $\mathbf{E} \times \mathbf{B}$. However, two points must be made concerning this: First, it is clear that when one does not identify vectors and covectors, as in elementary electromagnetism, the vector field $\#^{-1}(E \wedge H)$ is quite distinct from the covector field $\#(\mathbf{D} \wedge \mathbf{B})$, even when one identifies components using a Euclidian metric. Second, it is only in the case of propagating waves that it is physically meaningful to identify these terms with energy flux or momentum flux. Of course, in order to convert the energy flux to a momentum flux, one must divide by a propagation speed. One sees that in the general case, for which this speed is not a universal constant, it would probably make the most sense to divide by the speed of propagation in the direction of wave motion.

This brings us to the fact that one of the most important rank-2 cases, namely, the fields of electromagnetic waves, is probably better treated within the general framework of the rank-4 case that we just described. The relevant 2-form and bivector field take the form:

$$F = k \wedge A, \qquad \mathfrak{h} = \mathbf{k} \wedge \mathbf{A} = \kappa(k \wedge A), \tag{XI.93}$$

in which A is defined up to a gauge transformation of the form $A \mapsto A + \lambda k$ for some λ; similarly $\mathbf{A}$ can be replaced by $\mathbf{A} + \lambda' \mathbf{k}$ for some λ'.

In order for F to be wavelike it is necessary, but not sufficient, that one have:

$$F \wedge F = F \wedge \#\mathfrak{h} = 0. \qquad \text{(XI.94)}$$

If we put everything in time-space form, with:

$$k = \omega\, dt - k_s, \qquad \mathbf{k} = \omega \partial_t + \mathbf{k}_s\,, \qquad \text{(XI.95)}$$

then we have:

$$F = \omega\, dt \wedge A - k_s \wedge A, \qquad \mathfrak{h} = \omega \partial_t \wedge \mathbf{A} + \mathbf{k}_s \wedge \mathbf{A}\,. \qquad \text{(XI.96)}$$

This allows us to identify the E, $\mathbf{D}$, H, and $\mathbf{B}$ that are necessary in order to identify terms in (XI.92a, b):

$$E = \omega A, \qquad \#\mathbf{B} = k_s \wedge A, \quad \mathbf{D} = \omega \mathbf{A}, \qquad H = \#(\mathbf{k}_s \wedge \mathbf{A}). \qquad \text{(XI.97)}$$

One verifies directly – with an appropriate choice of gauge – that the triples $\{k_s, E, H\}$ and $\{\mathbf{k}_s, \mathbf{D}, \mathbf{B}\}$ are projectively reciprocal. That is, $k_s(\mathbf{k}_s)$, $E(\mathbf{B})$, and $H(\mathbf{B})$ are non-zero, while all of the cross terms are already zero or can be made zero by a choice of gauge for A and $\mathbf{A}$. In particular, one chooses λ and λ' such that:

$$k_s(\mathbf{A}) = A(\mathbf{k}_s) = 0. \qquad \text{(XI.98)}$$

By substitution of (XI.98) into (XI.92a, b), we obtain more specific forms for the time-space and space-time components:

$$T_0^i = \omega A(\mathbf{A})\, k^i, \qquad T_j^0 = -\,\omega A(\mathbf{A})\, k_j\,. \qquad \text{(XI.99)}$$

Hence, we see the direction of flow of the energy in the wave is defined by the wave vector or covector.

In the case of a constitutive law with no cross couplings, the energy density takes the form:

$$T_0^0 = \tfrac{1}{2}[\,\omega^2 + k_s(\mathbf{k}_s)]\, A(\mathbf{A}). \qquad \text{(XI.100)}$$

Here, one can point out that k must satisfy the dispersion relation $P[k] = 0$ that is associated with the constitutive law. Hence, in the simplest – viz., quadratic – case, where $\omega^2 = k_s(\mathbf{k}_s)$, one can express the energy density as $\omega^2 A(\mathbf{A})$. This provides another opportunity for specifying A and $\mathbf{A}$; recall that so far we have only specified the spatial parts.

5. Symmetries of systems of differential equations [18-20]. In addition to considering the symmetries of the action functional for a system of differential equations, one can also consider the symmetries of that system itself. Indeed, the latter class of symmetries is potentially richer than the former one since not every system of differential equations can be represented as the Euler-Lagrange equations for some action functional. In physics, this is especially true when one is considering non-conservative forces in mechanical systems, but even when one is still considering systems that follow from a

variational principle, one can often find symmetries beyond the ones that follow from Noether's theorem.

Loosely speaking, a transformation is said to be a *symmetry* of a system of differential equations if it takes every solution of the system to another solution of the system. Of course, in order to make this concept precise in the language of transformation groups that was discussed above one must first be more precise about how one represents the system of differential equations and its space of solutions. Of the various ways of representing such things that we discussed in Chapter VI the ones that are most naturally adapted to the problem of finding symmetries of systems of differential equations seem to be the methods of jet manifolds and exterior differential systems, so we first present the general notions in both forms and then apply them to the specific problem of the system of differential equations that is associated with pre-metric electromagnetism.

It is worth pointing out that the problem that first motivated Lie to develop his theory of differentiable transformation groups was the problem of how to adapt the methods of Galois theory, which involved associating finite groups of transformations with the solutions of systems of algebraic equations, to the problem of investigating the solutions of systems of differential equations by means of continuously infinite groups of transformations of those solutions. It is for that reason that one must recognize that the problem and methodology that he was first introducing was actually more involved than the simpler problems of Lie groups acting globally as transformations on manifolds, namely the problems of Lie pseudogroups and Lie groupoids acting locally. The obstruction to the extension of the local methods to the global ones generally involves the limits of the existence and uniqueness of the solutions to the differential equations.

a. Symmetries in terms of jets. If we recall the definitions associated with the manifold $J^1(M, N)$ of 1-jets of differentiable maps from a manifold M to a manifold N that were discussed in Chapter VI then we see that the formalism is naturally adapted to the problem of describing symmetries of systems of differential equations since both the system of equations and its solutions can be represented as submanifolds in one sense or the other.

If M is m-dimensional and N has dimension n then a system of m first-order partial differential equations for n unknown functions can be described implicitly by a submanifold in $J^1(M, N)$ of codimension one – i.e., a hypersurface – by means of some differentiable function F: $J^1(M, N) \to \mathbb{R}$. The equation then takes the local form of a level hypersurface of F:

$$F(x^i, y^a, y^a_i) = \text{const.} \tag{XI.101}$$

One generally expects that one can solve F for the derivatives y^a_i, at least in principle:

$$y^a_i = f(x^i, y^a), \tag{XI.102}$$

so this implies that one restricts F by means of the implicit function theorem to satisfy:

$$\frac{\partial F}{\partial y_i^a} \neq 0, \tag{XI.103}$$

by which we mean that not all of the derivatives vanish identically.

One might also define a differentiable function F: $J^1(M, N) \to \mathbb{R}^k$ that would define a system of k implicit differential equations.

A *solution* to the differential equation defined by such an F is then a continuously-differentiable function ϕ: $M \to N$ such that its 1-jet prolongation $j^1\phi$, which locally looks like $(x^i, y^a(x), y_a^i(x))$, satisfies the equation (XI.101):

$$F(j^1\phi) = F(x^i, y^a(x), y_a^i(x)) = \text{const.} \tag{XI.104}$$

Hence, one can regard a solution ϕ as a submanifold of M that also defines a submanifold $j^1\phi$: $M \to J^1(M, N)$ of $J^1(M, N)$.

One calls a section s: $M \to J^1(M, N)$ integrable iff $s = j^1\phi$ for some C^1 map ϕ: $M \to N$. A necessary and sufficient condition for this is that when one pulls back the contact forms:

$$\theta^a = dy^a - y_i^a dx^i \tag{XI.105}$$

by means of s the result is a set of n vanishing 1-forms:

$$0 = s^*\theta^a(x) = (y_{,i}^a - y_i^a)dx^i . \tag{XI.106}$$

That is, s is integrable iff its components satisfy the compatibility condition:

$$\frac{\partial y^a}{\partial x^i} = y_i^a . \tag{XI.107}$$

The vanishing of $\theta^a(X)$ for all $a = 1, \ldots, n$ when X is a tangent vector to $s \in J^1(M, N)$ defines n hypersurfaces in the tangent space T_sJ^1, whose intersection then defines a linear subspace C_sJ^1 of T_sJ^1 of dimension $m + n + mn - n = m(n + 1)$; i.e., of codimension n. The set of all such linear subspaces C_sJ^1 defines a vector sub-bundle $C(J^1)$ of $T(J^1)$ that we shall call the *contact bundle* to $J^1(M, N)$.

Since any vector sub-bundle of a tangent bundle can be regarded as a differential system on the manifold in question, one must naturally inquire about the integrability of the sub-bundle. From Frobenius's theorem, this is equivalent to asking whether there are 1-forms η_b^a on $J^1(M, N)$ such that:

$$-\Omega^a = d\theta^a = \eta_b^a \wedge \theta^b . \tag{XI.108}$$

By substituting (XI.105), we first get:

$$\Omega^a = dy_i^a \wedge dx^i , \tag{XI.109}$$

and (XI.108) takes the form:

$$0 = (dy_i^a - y_i^b \eta_b^a) \wedge dx^i + \eta_b^a \wedge dy^b. \qquad \text{(XI.110)}$$

Since this is not generally the case, we must conclude that in general the contact sub-bundle is not integrable.

We then call a diffeomorphism Φ: $J^1(M, N) \to J^1(M, N)$ a contact transformation iff its differential map at each point $d\Phi|_s : T_s J^1 \to T_{\Phi(s)} J^1$ takes each $C_s J^1$ to $C_{\Phi(s)} J^1$ isomorphically. Note that this is not equivalent to saying that it preserves the contact 1-forms, except up to a linear isomorphism A_b^a of $\mathbb{R}^n$ that may vary from point to point:

$$\Phi^* \theta^a|_{\Phi(s)} = A_b^a(s)\theta^b\Big|_s, \qquad \text{(some } A_b^a : J^1(M, N) \to GL(n)\text{)}. \qquad \text{(XI.111)}$$

It now becomes clear that a diffeomorphism of $J^1(M, N)$ that takes solutions of a differential equation that is defined by a level hypersurface of a C^1 function F on $J^1(M, N)$ must be, at the very least, a contact transformation, since it must take integrable sections to other integrable sections. A *symmetry* of the differential equation that is defined by F is then defined to be a contact transformation of $J^1(M, N)$ that also takes points of the level hypersurface to other points of the level hypersurface. Hence, it must also preserve F:

$$\Phi^* F = F, \qquad \text{(XI.112)}$$

which has the local form:

$$F(x^i, y^a, y_i^a) = F(\Phi^i, \Phi^a, \Phi_i^a), \qquad \text{(XI.113)}$$

in which each of the coordinate functions on the right-hand side is a function of (x^i, y^a, y_i^a), in general.

A one-parameter family of contact transformations Φ_σ, $\sigma \in (-\varepsilon, +\varepsilon)$ of F will be called *differentiable* iff for each $s \in J^1(M, N)$ the curve $\Phi_\sigma(s)$ in $J^1(M, N)$ is a differentiable function of σ. By differentiation, one can then define a vector field on $J^1(M, N)$:

$$X(s) = \frac{d\Phi_\sigma(s)}{d\sigma}\bigg|_{\sigma=0}. \qquad \text{(XI.114)}$$

However, since each $\Phi_\sigma(s)$ must preserve each θ^a only up to $A_b^a(s)$, as in (XI.108), we see that by Lie derivation the infinitesimal condition takes the form:

$$\mathrm{L}_X \theta^a = a_b^a(s)\theta^b, \qquad \text{(XI.115)}$$

in which $a_b^a : J^1(M, N) \to \mathfrak{gl}(n)$, this time.

When a vector field X on $J^1(M, N)$ satisfies this condition we shall it an *infinitesimal contact transformation.*

By an application of Cartan's formula for L_X, we see that (XI.115) also takes the form:

$$di_X\theta^a + i_X d\theta^a = a^a_b(s)\theta^b\,, \tag{XI.116}$$

and by substitution of (XI.105), this gives the following set of equations for the components of X:

$$\frac{\partial X^a}{\partial x^i} - X^a_i = y^a_j \frac{\partial X^j}{\partial x^i} - y^b_j a^a_b\,, \tag{XI.117a}$$

$$\frac{\partial X^a}{\partial y^b} - y^a_i \frac{\partial X^i}{\partial y^b} = a^a_b\,, \tag{XI.117b}$$

$$\frac{\partial X^a}{\partial y^j_b} = y^a_i \frac{\partial X^i}{\partial y^b_j}\,. \tag{XI.117c}$$

By substituting (XI.117b) into (XI.117a), the arbitrary matrix a^a_b ceases to play any role and the equations for the components of X become:

$$\frac{\partial X^a}{\partial y^b_i} = y^a_j \frac{\partial X^j}{\partial y^b_i}\,, \tag{XI.118a}$$

$$X^a_i = D_i X^a - y^a_j D_i X^j\,, \tag{XI.118b}$$

in which the operator D_i takes the form of a total derivative:

$$D_i = \frac{\partial}{\partial x^i} + y^a_i \frac{\partial}{\partial y^a}\,. \tag{XI.119}$$

Hence, from (XI.118b) we see that only the components X^i and X^a remain undetermined, except as solutions to the system of partial differential equations (XI.118a).

If X is the prolongation $j^1\xi$ of a vector field ξ on $M \times N$ then (XI.118b) gets replaced by:

$$X^a_i = X^a{}_{,i} \tag{XI.120a}$$

$$y^b_i X^a_{,b} = y^a_j D_i X^j\,. \tag{XI.120b}$$

It is illuminating to see what happens to vector fields that are annihilated by the θ^a; i.e., vector fields that are tangent to the integrable sections of $J^1(M, N) \to M$. If $i_X\theta^a = 0$ then $i_X\Omega^a = -\,a^a_b\theta^b$ (although the minus sign is unnecessary, since the matrix a^a_b is ambiguous), and upon expanding this into component equations, one gets:

$$-X^a_i dx^i + X^i dy^a_i = -a^a_b y^b_i dx^i + a^a_b dy^b\,, \tag{XI.121}$$

which implies that $a^a_b = 0$, which then makes X itself identically zero.

One concludes that any non-zero contact transformation must be transversal to any integrable section.

When we also have a differential equation defined by a C^1 function F on $J^1(M, N)$, we call the infinitesimal contact transformation X an *infinitesimal symmetry* of the differential equation that is defined by F if the local one-parameter family of contact transformations Φ_σ of $J^1(M, N)$ that it generates by integration all lie within the level hypersurface of F that defines the equation. Since each Φ_σ preserves F, by differentiating along the curves $\Phi_\sigma(s)$, we see that the Lie derivative of F with respect to X must vanish:

$$0 = \mathrm{L}_X F = XF = X^i \frac{\partial F}{\partial x^i} + X^a \frac{\partial F}{\partial y^a} + X_i^{\ a} \frac{\partial F}{\partial y_i^a} . \qquad \text{(XI.122)}$$

b. Symmetries in terms of exterior differential systems [**18-21**]. Now, let us consider a system of partial differential equations that is defined by a finite set $\{\theta^\kappa, \kappa = 1, \ldots, k\}$ of exterior differential forms on a manifold N of varying degrees as an exterior differential system:

$$\theta^\kappa = 0 \qquad \text{(all } \kappa\text{)}. \qquad \text{(XI.123)}$$

A solution to this system is a differentiable map y: $M \to N$ such that:

$$y^* \theta^\kappa = 0 \qquad \text{(all } \kappa\text{)}. \qquad \text{(XI.124)}$$

If the forms θ^κ take the local form:

$$\theta^\kappa = \frac{1}{r!} \alpha^\kappa_{a_1 \cdots a_r}(y) dy^{a_1} \wedge \cdots \wedge dy^{a_r} \qquad \text{(XI.125)}$$

then the pull-backs to M (with local coordinates x^i) take the form:

$$y^* \theta^\kappa = \frac{1}{r!} \alpha^\kappa_{a_1 \cdots a_r}(y(x)) y^{a_1}_{i_1} \cdots y^{a_r}_{i_r} dx^{i_1} \wedge \cdots \wedge dx^{i_r} . \qquad \text{(XI.126)}$$

A *(finite) symmetry* of the exterior differential system is a diffeomorphism f: $N \to N$ that takes solutions to other solutions. Note that this does not imply that it takes the θ^κ to themselves – i.e., $f^* \theta^\kappa = \theta^\kappa$ – only that $y^*(f^* \theta^\kappa) = 0$ iff $y^* \theta^\kappa = 0$. This means that there is an invertible matrix $A^\kappa_\nu \in GL(k)$ such that:

$$f^* \theta^\kappa = A^\kappa_\nu \theta^\nu . \qquad \text{(XI.127)}$$

An *infinitesimal symmetry* of the exterior differential system is then a vector field $X \in \mathfrak{X}(N)$ such that for some matrix $a^\kappa_\nu \in \mathfrak{gl}(k)$ one has:

$$\mathrm{L}_X \theta^\kappa = d i_X \theta^\kappa + i_X d\theta^\kappa = a^\kappa_\nu \theta^\nu . \qquad \text{(XI.128)}$$

Hence, the vector field X will be a solution to a system of linear first-order partial differential equations of a type that one calls *Lie equations*, since their solutions represent infinitesimal symmetries. In fact, since Lie himself was not using the global, basis-free formalism of modern Lie groups, he was more concerned with algebraic structures that were defined by solving such local systems of partial differential equations. To this day, the study of Lie pseudogroups and Lie equations still represents a class of problems that are considerably more complicated to address than those of finite-dimensional Lie groups and Lie algebras, which are certainly non-trivial, in their own right [5].

When the solutions to an exterior differential system are to take the form of sections s: $M \to E$ of a vector bundle $E \to M$ over a manifold M, the exterior differential system will be defined by a set of differential forms on E. If a local trivialization has coordinates of the form (x^i, y^a) then differential forms on E will be exterior products of factors of both the 1-forms dx^i and the 1-forms dy^a, while vector fields on E will have the local form:

$$X = X^i \frac{\partial}{\partial x^i} + X^a \frac{\partial}{\partial y^a}. \tag{XI.129}$$

c. Symmetries of the equations of pre-metric electromagnetism [**5**]. Since the equations of pre-metric electromagnetism are concerned with differential forms on the space or spacetime manifold, it seems clear that representing them as an exterior differential system on a vector bundle would be the most straightforward path to take. However, one must note that the form that we have been addressing them relates to differential operators on sections of vector bundles, not an exterior differential system. Hence, we must convert them to such a system. For the sake of specificity, we shall consider the case of spacetime.

Since the field that we are solving for is a section F: $M \to \Lambda^2 M$, we see that what we need to do first is to define the pre-metric Maxwell equations in terms of a set of differential forms on $\Lambda^2 M$ that pull down to the previous equations on M by way of a section. The way that we do this is to start with the local expression for the 2-form F as $\frac{1}{2} F_{\mu\nu}\, dx^\mu \wedge dx^\nu$ and regard the components $F_{\mu\nu}$, not as functions on M, but as coordinate functions on $\Lambda^2 M$. We then have the local expression for a 2-form F on the total space $\Lambda^2 M$ of the bundle that becomes a 2-form on M when we pull it down by way of the section F; i.e., when one replaces the components $F_{\mu\nu}$ with functions $F_{\mu\nu}(x)$ on M.

Hence, when we form the 3-form on $\Lambda^2 M$:

$$\Theta^1 = 2dF = dF_{\mu\nu} \wedge dx^\mu \wedge dx^\nu \tag{XI.130}$$

we see that the first of the Maxwell equations can be represented by the exterior differential system on $\Lambda^2 M$:

$$\Theta^1 = 0. \tag{XI.131}$$

[5] For some idea of the possible role of Lie pseudogroups and Lie groupoids in physics, see Pommaret [**3**].

We absorb the algebraic constitutive law $\mathfrak{h} = \kappa(F)$ into the differential equation for $\mathfrak{h}$ in the absence of sources, which we express in the form $d{*}F = 0$, in which ${*}F = (\kappa\cdot\#)F$; i.e., ${*} = \kappa\cdot\#$. We have not yet normalized the * diffeomorphism, so we only assume that ${*}^2 = -\lambda^2 I$.

If ${*}F$ then takes the form $\frac{1}{2}{*}F_{\mu\nu}\, dx^\mu \wedge dx^\nu$ then we see that the differential equation for ${*}F$ can be expressed by the vanishing of the 3-form on $\Lambda^2 M$:

$$\Theta^2 = 2d{*}F = d{*}F_{\mu\nu}\ dx^\lambda \wedge dx^\mu \wedge dx^\nu, \tag{XI.132}$$

that is:

$$\Theta^2 = 0. \tag{XI.133}$$

Since the coordinate functions on $\Lambda^2 M$ involve $F_{\mu\nu}$, not ${*}F_{\mu\nu}$, we see that next we must expand the differentiation:

$$d{*}F_{\mu\nu} = d(\kappa\cdot\#)F = \tfrac{1}{4} d(\kappa_{\kappa\lambda\alpha\beta}\, \varepsilon^{\alpha\beta\mu\nu})F_{\mu\nu}\,. \tag{XI.134}$$

Here, we see that the representation of the pre-metric Maxwell equations as an exterior differential system on $\Lambda^2 M$ has one natural advantage in the eyes of constitutive laws: nonlinear constitutive laws are simply the ones for which the component functions $\kappa_{\kappa\lambda\alpha\beta} = \kappa_{\kappa\lambda\alpha\beta}(x, F)$ are fully general, while linear laws have components of the form $\kappa_{\kappa\lambda\alpha\beta} = \kappa_{\kappa\lambda\alpha\beta}(x)$. This suggests that nonlinear electrodynamics is more conveniently addressed in the present form.

As a further simplification, we combine the components of κ and # into the components:

$$\kappa_{\kappa\lambda}{}^{\mu\nu} = \tfrac{1}{2}\, \kappa_{\kappa\lambda\alpha\beta}\, \varepsilon^{\alpha\beta\mu\nu}, \tag{XI.135}$$

of *.

Since:

$$d\kappa_{\kappa\lambda}{}^{\mu\nu} = \kappa_{\kappa\lambda}{}^{\mu\nu}{}_{,\rho}\ dx^\rho + \tfrac{1}{2}\, \kappa_{\kappa\lambda}{}^{\mu\nu,\rho\sigma}\, dF_{\rho\sigma}\,, \tag{XI.136}$$

we get:

$$\begin{aligned} d{*}F_{\mu\nu} &= \tfrac{1}{2}(d\kappa_{\mu\nu}{}^{\kappa\lambda}\, F_{\kappa\lambda} + \kappa_{\mu\nu}{}^{\kappa\lambda}\, dF_{\kappa\lambda}) \\ &= \tfrac{1}{2}\, \kappa_{\mu\nu}{}^{\kappa\lambda}{}_{,\rho}\ F_{\kappa\lambda}\, dx^\rho + \tfrac{1}{2}(\kappa_{\mu\nu}{}^{\kappa\lambda} + \tfrac{1}{2}\, \kappa_{\mu\nu}{}^{\rho\sigma,\,\kappa\lambda}\, F_{\rho\sigma})\, dF_{\kappa\lambda}\,. \end{aligned} \tag{XI.137}$$

As usual, we introduce the notation:

$$\tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda} = \kappa_{\mu\nu}{}^{\kappa\lambda} + \tfrac{1}{2}\, \kappa_{\mu\nu}{}^{\rho\sigma,\,\kappa\lambda}\, F_{\rho\sigma}, \tag{XI.138}$$

and put (XI.137) into the form:

$$d{*}F_{\mu\nu} = \tfrac{1}{2}(\kappa_{\mu\nu}{}^{\kappa\lambda}{}_{,\rho}\ F_{\kappa\lambda}\, dx^\rho + \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda}\, dF_{\kappa\lambda}\,). \tag{XI.139}$$

This allows us to set:

$$\Theta^2 = \kappa_{\mu\nu}{}^{\kappa\lambda}{}_{,\rho}\ F_{\kappa\lambda}\ dx^\rho \wedge dx^\mu \wedge dx^\nu + \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda} dF_{\kappa\lambda} \wedge dx^\mu \wedge dx^\nu \ . \qquad \text{(XI.140)}$$

We next address the infinitesimal symmetries of the exterior differential system:

$$\Theta^a = 0, \qquad a = 1, 2, \qquad \text{(XI.141)}$$

which will then be solutions X of the system of Lie equations:

$$\mathrm{L}_X\, \Theta^a = di_X\, \Theta^a = a^a_b \Theta^b \, . \qquad \text{(XI.142)}$$

Since $\Theta^1 = dF$ and $\Theta^2 = d{*}F$ are both exact forms, this takes the form:

$$di_X dF = a^1_1 dF + a^1_2 d*F \, , \qquad di_X d{*}F = a^1_1 dF + a^1_2 d*F \, . \qquad \text{(XI.143)}$$

The possible solutions of these equations include the possibility that the underdetermined functions a^a_b are constants, in which case, the exterior derivative operator commutes with them in the right-hand sides of (XI.143), and we are left with the equations:

$$di_X dF = d(a^1_1 F + a^1_2 * F) \, , \qquad di_X d{*}F = d(a^2_1 F + a^2_2 * F) \, , \qquad \text{(XI.144)}$$

which can be solved by:

$$i_X dF = a^1_1 F + a^1_2 * F + \varepsilon_{\mu\nu} \, , \qquad i_X d{*}F = a^2_1 F + a^2_2 * F + {*}\varepsilon_{\mu\nu} \, , \qquad \text{(XI.145)}$$

in which $\varepsilon_{\mu\nu}$ and ${*}\varepsilon_{\mu\nu}$ are the components of closed 2-forms ε and ${*}\varepsilon$ on $\Lambda^2 M$.

If we expand the left-hand sides in latter equations in components we get the following algebraic equations for the components X^μ and $X_{\mu\nu}$:

$$X_{\mu\nu}\, dx^\mu \wedge dx^\nu - 2X^\mu\, dF_{\mu\nu} \wedge dx^\nu = (a^1_1 + a^1_2 *) F_{\mu\nu} dx^\mu \wedge dx^\nu \, , \qquad \text{(XI.146a)}$$

$$(3\kappa_{\mu\nu}{}^{\kappa\lambda}{}_{,\rho} F_{\kappa\rho} X^\lambda + \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda} X_{\kappa\lambda}) dx^\mu \wedge dx^\nu - 2\tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda} X^\mu dF_{\kappa\lambda} \wedge dx^\nu$$
$$= (a^2_1 + a^2_2 *) F_{\mu\nu} dx^\mu \wedge dx^\nu \, . \qquad \text{(XI.146b)}$$

From the first one we deduce:

$$X^\mu = 0, \qquad X_{\mu\nu} = (a^1_1 + a^1_2 *) F_{\mu\nu} + \varepsilon_{\mu\nu}, \qquad \text{(XI.147)}$$

and when we substitute these results in the second equation of (XI.146b), we get:

$$\tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda} X_{\kappa\lambda} = (a^1_1 + a^1_2 *) \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda} F_{\mu\nu} + \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda} \varepsilon_{\mu\nu} = (a^2_1 + a^2_2 *) F_{\mu\nu} + {*}\varepsilon_{\mu\nu} . \qquad \text{(XI.148)}$$

Hence, the only remaining issue to be resolved in order to conclude that (XI.147) represents a class of infinitesimal symmetries is whether one can find suitable constants a^a_b, and components $*\varepsilon_{\mu\nu}$ of a closed 2-form that make:

$$(a^1_1 + a^1_2 *)\tilde{\kappa} = a^2_1 + a^2_2 *, \qquad *\varepsilon = \tilde{\kappa}\varepsilon, \tag{XI.148}$$

in which we have reverted to the component-free expressions for the operators in question. Of course, the second of these equations gives us $*\varepsilon_{\mu\nu}$ directly.

In the linear case, we find that $\tilde{\kappa} = \kappa = *$, and the first equation says simply:

$$-\lambda^2 a^1_2 + a^1_1 * = a^2_1 + a^2_2 *, \tag{XI.149}$$

which admits a two-parameter family of solutions:

$$a^2_1 = -\lambda^2 a^1_2, \qquad a^1_1 = a^2_2. \tag{XI.150}$$

From (XI.147), the symmetries that we have been describing represent vertical transformations of $\Lambda^2 M$. More precisely, when the constitutive law described by κ defines a complex structure on each fiber of the real vector bundle $\Lambda^2 M$ (which is then an *almost-complex* structure on $\Lambda^2 M$), as long as the field equations are complex-linear the solution space is a complex vector space in its own right, and the symmetries of the solution space will include complex affine transformations. This follows directly from the fact that if F is a solution to $dF = d*F = 0$ then so is $\alpha F + \beta *F + \varepsilon = (\alpha + i\beta)F + \varepsilon$ for any constant real scalars α and β and any closed 2-form ε. Since any non-zero complex number can be expressed in polar form the multiplication of F by the complex scalar $\alpha + i\beta$ involves a rotation in the plane of F and $*F$ that one calls a *duality rotation.*

Of course, in the nonlinear case ($\tilde{\kappa} \neq \kappa = *$) one does not expect the space of solutions to be a vector space of any sort, in general. Hence, the question of the extent to which the pre-metric Maxwell equations admit the aforementioned class of symmetries must be dealt with in terms of specific cases of nonlinear constitutive laws.

Returning to the general case (XI.142), let us first examine the form of the general element on the right-hand side:

$$\begin{aligned} &a^a_b \Theta^b \\ &\quad = (a^a_1 \delta^{[\kappa}_\mu \delta^{\lambda]}_\nu + a^a_2 \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda}) dF_{\kappa\lambda} \wedge dx^\mu \wedge dx^\nu + a^a_2 \kappa_{\mu\nu}{}^{\kappa\lambda}{}_{,\lambda} F_{\kappa\lambda} dx^\lambda \wedge dx^\mu \wedge dx^\nu . \end{aligned} \tag{XI.151}$$

Our first clue as to how we can reduce the results of expanding the left-hand side of (XI.142) is to note that only 3-forms involving at most one factor of $dF_{\kappa\lambda}$ can appear with a non-zero coefficient. Furthermore, we see that the coefficients of $dF_{\kappa\lambda} \wedge dx^\mu \wedge dx^\nu$ take the general form of complex scalar multiplication, as long as the nonlinear-modified constitutive law $\tilde{\kappa}$ defines an almost complex structure, although, so far, we have only defined that operation on 2-forms over M, not 3-forms over $\Lambda^2 M$.

Taking $a = 1$, we find:

$$\mathrm{L}_X\Theta^1 = X_{\mu\nu,\lambda}\, dx^\lambda \wedge dx^\mu \wedge dx^\nu + \left(\tfrac{1}{2} X_{\mu\nu}{}^{,\kappa\lambda} + X^{[\kappa}_{\ ,\mu}\delta^{\lambda]}_\nu\right) dF_{\kappa\lambda} \wedge dx^\mu \wedge dx^\nu$$
$$- X^{\mu,\kappa\lambda}\, dF_{\kappa\lambda} \wedge dF_{\mu\nu} \wedge dx^\nu . \qquad \text{(XI.152)}$$

Hence, the first general conclusion we can infer about infinitesimal symmetries of the pre-metric electromagnetic field equations is the fact that:

$$\frac{\partial X^\mu}{\partial F_{\kappa\lambda}} = 0, \qquad \text{(XI.153)}$$

which says that the X^μ must be functions of only position, but not field strength. Therefore, they are the lifts of infinitesimal generators of spacetime diffeomorphisms; i.e., motions.

The other equations that follow from (XI.142), (XI.151), and (XI.152) are:

$$\frac{\partial X_{\mu\nu}}{\partial x^\lambda} = a_2^1 \kappa_{\mu\nu}{}^{\kappa\rho}{}_{,\lambda} F_{\kappa\sigma}\,, \qquad \frac{\partial X_{\mu\nu}}{\partial F_{\kappa\lambda}} + X^{[\kappa}_{\ ,\mu}\delta^{\lambda]}_\nu = a_1^1 \delta^{[\kappa}_\mu \delta^{\lambda]}_\nu + a_2^1 \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda}\,. \qquad \text{(XI.154)}$$

When one chooses the arbitrary function a_2^1 to be a function of only F, the first set of equations can be solved by:

$$X_{\mu\nu} = a_2^1 \kappa_{\mu\nu}{}^{\kappa\lambda} F_{\kappa\lambda} + \varepsilon_{\mu\nu}\,, \qquad \text{(XI.155)}$$

which is already included on the infinitesimal generators of complex affine transformations of the fibers of $\Lambda^2 M$, so it tells us nothing new. However, if one considers a homogeneous medium then the right-hand side of the first equations in (XI.154) vanishes, and this says that in such a case $X_{\mu\nu}$ must be functions $\varepsilon_{\mu\nu}(F)$ of only F; i.e., it is the infinitesimal generator of purely vertical translations. Between this and the previous observation about the character of X_μ , we see that the infinitesimal symmetries of the field equations for a homogeneous medium are the sums of infinitesimal generators of pure motions and pure vertical transformations; i.e.:

$$X(x, F) = X^\mu(x) \frac{\partial}{\partial x^\mu} + X_{\mu\nu}(F) \frac{\partial}{\partial F_{\mu\nu}}\,. \qquad \text{(XI.156)}$$

As for the second set of equations in (XI.154), it clear that it requires deeper analysis. It includes the special cases of pure motions and pure vertical transformations. The pure vertical transformations again give us complex scalar multiplication and translation in the fibers, but the pure motions give the equation:

$$X^{[\kappa}_{\ ,\mu}\delta^{\lambda]}_\nu = a_1^1 \delta^{[\kappa}_\mu \delta^{\lambda]}_\nu + a_2^1 \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda}\,. \qquad \text{(XI.157)}$$

The solutions to it include the possibility that $a_2^1 = 0$ and a_1^1 is a constant, which then gives the class of solutions:

$$X_\mu(x) = \alpha x^\mu + \varepsilon_\mu, \tag{XI.158}$$

in which the ε_μ are constants. These vector fields represent infinitesimal spacetime dilatations and translations.

When a_2^1 is non-vanishing, the motions thus defined are harder to interpret, unless $\tilde{\kappa}$ is expressible as an anti-symmetrized tensor product, such as in the Lorentzian case, where κ takes the form $g \wedge g$, so when one raises one index on the components of each g, the resulting components are simply $\delta_\mu^{[\kappa}\delta_\nu^{\lambda]}$, which gives the previous transformation.

It is in addressing (XI.142) for $a = 2$ that we find the most complexity arising, since we are differentiating the constitutive law in the process. The equations for the components of X that we obtain consist of three basic types: equations for the coefficients of $dF_{\alpha\beta} \wedge dF_{\kappa\lambda} \wedge dx^\mu$, equations for the coefficients of $dx^\lambda \wedge dx^\mu \wedge dx^\nu$, and equations for the coefficients of $dF_{\kappa\lambda} \wedge dx^\mu \wedge dx^\nu$.

The first type of coefficients must vanish, which gives the equations:

$$0 = X^\mu \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda,\alpha\beta} = \frac{\partial}{\partial F_{\alpha\beta}}\left(X^\mu \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda} \right), \tag{XI.159}$$

which says that $X^\mu \tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda}$ must be a function of only x. Although this is always true when κ describes a linear medium, it must imply that $\tilde{\kappa}$ is a function of only x in the nonlinear case. Hence, if we differentiate (XI.138) with respect to $F_{\gamma\delta}$, while suppressing the irrelevant indices, we get the following condition on κ:

$$\frac{\partial \kappa^{\delta\gamma}}{\partial F_{\alpha\beta}} F_{\alpha\beta} = -3\,\kappa^{\gamma\delta}, \tag{XI.160}$$

which says that $\kappa^{\cdot,\gamma\delta}$ is homogeneous of degree – 3 in F.

If κ does not have this property then one must have $X^\mu = 0$, which means that the general nonlinear medium will have no symmetries due to spacetime motions.

The second type of coefficients gives equations of the form:

$$3(X^\sigma \kappa_{\mu\nu}{}^{\kappa\rho}{}_{,\sigma} F_{\kappa\rho})_{,\lambda} + X_{\kappa\rho}\, \kappa_{\mu\nu}{}^{\kappa\rho}{}_{,\lambda} = a_1^2 (\kappa_{\mu\nu}{}^{\kappa\rho} F_{\kappa\rho})_{,\lambda}. \tag{XI.161}$$

Of course, for homogeneous media these equations become trivial.

In the inhomogeneous case, when we consider $X_{\kappa\rho}$ that are purely vertical – i.e., functions of only F – these equations reduce to the form:

$$3X^\sigma \kappa_{\mu\nu}{}^{\kappa\rho}{}_{,\sigma} F_{\kappa\rho} + X_{\kappa\rho}\, \kappa_{\mu\nu}{}^{\kappa\rho} = a_1^2\, \kappa_{\mu\nu}{}^{\kappa\rho} F_{\kappa\rho} + \varepsilon_{\mu\nu}(F). \tag{XI.162}$$

This includes the possibility:

$$X^\sigma \kappa_{\mu\nu}{}^{\kappa\rho}{}_{,\sigma} = \tfrac{1}{3} a_1^2\, \kappa_{\mu\nu}{}^{\kappa\rho}, \qquad X_{\kappa\rho} = \kappa^{\mu\nu}{}_{\kappa\rho}\, \varepsilon_{\mu\nu}, \tag{XI.163}$$

in which we have made use of the invertibility of κ:

Since the $\varepsilon_{\mu\nu}$ are arbitrary functions of F, the second set of equations says that $X_{\kappa\rho}$ depend on x only by way of κ:

The first set of equations is somewhat more intriguing, since it says that the components of κ must be eigenfunctions of the first-order differential operator defined by X. This related to the more general condition:

$$L_X \kappa = \lambda \kappa; \tag{XI.164}$$

which we shall return to shortly.

The third type of coefficients gives the equations:

$$(3\kappa_{\mu\nu}{}^{\kappa\rho}{}_{,\lambda}{}^{\cdot\alpha\beta} F_{\alpha\beta} + 3\kappa_{\mu\nu}{}^{\alpha\beta}{}_{,\lambda} + \tilde{\kappa}_{\lambda\nu}{}^{\alpha\beta}{}_{,\mu})X^{\lambda} + 2X^{\lambda}{}_{,\mu}\,\kappa_{\lambda\nu}{}^{\alpha\beta} + (X_{\kappa\lambda}\tilde{\kappa}_{\mu\nu}{}^{\kappa\lambda})^{\cdot\alpha\beta}$$
$$= a_1^2 \delta_{\mu}^{[\kappa}\delta_{\nu}^{\lambda]} + a_2^2 \kappa_{\mu\nu}{}^{\alpha\beta}. \tag{XI.165}$$

Once again, in a homogeneous medium, we see that the motions include dilatations and translations, as expected. The inhomogeneous case is clearly more involved, and requires further analysis.

The purely vertical transformations that satisfy these equations amount to complex scalar multiplications and complex translations of the fibers of $\Lambda^2 M$.

One can get more of an intuition for some of the symmetries of the field equations if one applies the Lie derivative operator L_X to the expressions dF and $d{*}F$, which then give $dL_X F$ and $dL_X {*}F$, respectively. The 3-forms dF and $d{*}F$ are eigenforms of the first-order differential operator L_X iff:

$$dL_X F = \lambda' dF, \qquad dL_X {*}F = \lambda'' d{*}F \tag{XI.166}$$

for some real constants λ', λ''.

These conditions become:

$$L_X F = \lambda' F + (\text{closed 2-form}),\ L_X {*}F = \lambda'' {*}F + (\text{another closed 2-form}). \tag{XI.167}$$

Hence, F and ${*}F$ are eigenforms of L_X, modulo a closed 2-form.

One sees that these conditions are equivalent iff:

$$L_X {*} = \lambda {*}. \tag{XI.168}$$

for some real function λ.

Since ${*} = \# \cdot \kappa$, this becomes:

$$L_X \# \cdot \kappa + \# \cdot L_X \kappa = \lambda \# \cdot \kappa, \tag{XI.169}$$

which then demands that:

$$L_X \# = \alpha \#, \qquad L_X \kappa = \beta \kappa; \tag{XI.170}$$

for suitable real functions α, β.

The first equation is equivalent to:

$$L_X \mathcal{V} = (\delta X)\mathcal{V} = \alpha \mathcal{V}, \tag{XI.171}$$

for some a, which is always satisfied by setting $\alpha = \delta X$. Hence, it is not necessary for one to restrict oneself to volume-preserving diffeomorphisms.

The second equation in (XI.171) is a generalization of the conformal Killing equation for the Lorentzian metric g:

$$L_X g = \Omega^2 g, \tag{XI.172}$$

as one can see by expressing κ in the form $g \wedge g$.

Since that equation gives infinitesimal generators of diffeomorphisms of Minkowski space that preserve the light cone, it appears that the corresponding group for the more general case of pre-metric electromagnetism is defined by diffeomorphisms of $\mathbb{R}^4$ that preserve the characteristic hypersurface that is defined by the dispersion law. This would include a subgroup that preserves the characteristic polynomial itself, which would then be analogous to the Lorentz group in the quadratic case. The deeper nature of these two groups then merits further study, since its role in physics is possibly as fundamental as that of the Lorentz group.

One finds that in practice the study of symmetries of the pre-metric field equations involves first choosing a specific constitutive law κ; or at least a specific class of them, such as homogeneous linear, inhomogeneous linear, homogeneous nonlinear, etc. For some progress in this direction, one might confer Delphenich [**5**].

References

1. H. Bateman, "The transformation of the electrodynamical equations," Proc. London Math. Soc. [2] **8** (1910) 223-264.
2. E. Cunningham, "The principle of relativity in electrodynamics and an extension thereof," Proc. London Math. Soc. [2] **8** (1910) 77-98.
3. J. F. Pommaret, *Lie Pseudogroups in Continuum Mechanics,* Gordon and Breach, New York, 1988.
4. B. K. Harrison and F. B. Estabrook, "Geometric Approach to Invariance Groups and Solution of Partial Differential Equations," J. Math. Phys. **12** (1971) 653-666.
5. D. H. Delphenich, Symmetries and pre-metric electromagnetism, Ann. d. Phys. (Leipzig), **14**, No. 11-12, 663-704 (2005), and gr-qc/0508035.
6. D. H. Sattinger and O. L. Weaver, *Lie Groups and Algebras, with Applications to Physics, Geometry, and Mechanics,* Springer, Berlin, 1986.
7. R. Gilmore, *Lie Groups, Lie Algebras, and Some of Their Applications,* Dover, NY, 2006.
8. C. Chevalley, *Theory of Lie Groups,* Princeton Univ. Press, Princeton, 1946.
9. J. Adams, *Lectures on Lie Groups,* U. of Chicago Press, Chicago, 1969.
10. K. Kawakubo, *The Theory of Transformation Groups,* Oxford University Press, Oxford, 1991.

11. L. Michel, "Nonlinear group action, smooth action of compact Lie groups on manifolds," in *Statistical Mechanics and Field Theory,* R. N. Sen and C. Weil, eds., Halsted Press, New York, 1972.
12. A. Pressley and G. Segal, *Loop Groups*, Clarendon Press, Oxford, 1990.
13. Saunders, D., *Geometry of Jet Bundles,* Cambridge University Press, Cambridge, 1989.
14. F. W. Hehl and Y. N. Obukhov, *Foundations of Classical Electrodynamics*, Birkhäuser, Boston, 2003.
15. J. Plebanski, *Lectures on Nonlinear Electrodynamics,* NORDITA Lectures, Copenhagen, 1970.
16. W. Thirring, *Classical Field Theory,* Springer, Berlin, 1978.
17. J. Rzewuski, *Field Theory. Volume I: Classical Theory*, Iliffe Books Ltd., London, 1964.
18. P. Olver, *Applications of Lie Groups to Differential Equations,* 2nd ed., Springer, Berlin, 1993.
19. P. Olver, *Equivalence, Invariance, and Symmetry*, Cambridge University Press, Cambridge, 1995.
20. G. W. Bluman and S. C. Anco, *Symmetry and Integration Methods for Differential Equations*, Springer, Berlin, 2002.
21. E. Cartan, *Les systèmes différentielle extérieurs et leur applications géométriques*, Hermann, Paris, 1971.
22. Y. Choquet-Bruhat, *Géométrie différentielle et systèmes différentieles exterieure*, Dunod, Paris, 1964.
23. W. Slebodzinski, *Exterior Differential Forms and Applications,* Polish Scientific Publishers, Warsaw, 1970.
24. R. L. Bryant, S. S. Chern, R. B. Gardner, H. L. Goldschmidt, P. A. Griffiths, *Exterior Differential Systems,* Springer, Berlin, 1991.

CHAPTER XII

PROJECTIVE GEOMETRY AND ELECTROMAGNETISM

One of the lasting contributions that general relativity made to the foundations of physics was the idea that such an everyday natural phenomenon as gravitation could nonetheless ultimately be a manifestation of something as abstract as the geometric structure of the spacetime manifold itself. Up until Einstein succeeded in showing the link between the distribution of matter in that manifold and its curvature, the study of non-Euclidian geometries had been of interest primarily to pure mathematicians, some of whom, such as William Kingdon Clifford and Henri Poincaré, had already speculated on the possibility of such a link.

A predictable consequence of the revelation that spacetime geometry was at the root of gravitation was that theoretical physics devoted more attention to the "geometrization" of all of its other fundamental principles. Mechanics drifted further into the geometry of contact and symplectic manifolds, gauge field theory focused more on the description of gauge structures for field theories in terms of connections on principal bundles, and the conjecture emerged that perhaps one could generalize spacetime geometry in some way that would encompass the other fundamental forces (or fundamental *interactions*, to the quantum field theorists).

In the original form of the field unification problem, Einstein conjectured that there was such a broadening of the scope of spacetime geometry that would subsume both gravitation and electromagnetism. He made many attempts along these lines, along with Mayer, Cartan, Oskar Klein, Kaluza, Jordan, Schrödinger, Veblen, Vranceanu, and others [1]. However, the most common results of these theories were either elegant mathematical theories that nevertheless admitted unphysical solutions or consistent unifications of gravitation and electromagnetism that said nothing new about either; one might call such a theory a "concatenation" of the two field theories.

The scope of the unification problem eventually expanded to include all of the fundamental interactions, presumably in a gauge theory for a suitably large gauge group, but the new obstacle emerged that the mathematical formalism of quantum field theories was mostly oriented towards describing the interaction of matter in the scattering approximation, not the specific details of the time evolution of the system during the actual time interval in which the interaction takes place. Hence, since general relativity does not spend much time addressing the scattering of gravitating bodies, any more than most quantum field theories address the Cauchy problem for field solutions, simply trying to reconcile the mathematical methodologies become a complex problem, in its own right [2].

---

[1] A good historical discussion of some of these attempts can be found in the book by Vizgin [**1**], and a more mathematically detailed discussion of two them – viz., the Kaluza-Klein and Einstein-Schrödinger theories – can be found in Part II of Lichnerowicz [**2**].

[2] An illuminating insight into the difference between the general relativistic approach to fundamental interactions and the quantum field theoretic approach can be gleaned from reading the Feynman lectures on

We have seen how the theory of electromagnetism can be formulated in the absence of a background spacetime metric, since the Lorentzian structure eventually "emerges" as a consequence of the dispersion law that follows from the constitutive law, by way of the field equations. This suggests that perhaps the unification problem that Einstein posed, although reasonable by analogy to the unification of electricity and magnetism into electromagnetism, might still be the wrong problem to pose. Indeed, we already have a definitive link between the two theories by way of the fact that the Lorentzian structure that accounts for gravitation consists of *light* cones, not *gravity* cones; i.e., cones that pertain to the propagation of electromagnetic waves, not merely the propagation of gravitational waves.

The question then arises of how one might still geometrize pre-metric electromagnetism in the absence of a spacetime metric. In order to address this question, one might consider the geometrical hierarchy that was proposed by Felix Klein, who once decreed that "projective geometry is all geometry." From that level of generality, one could then reduce one's scope to affine geometry or metric geometry by considering the hyperplane at infinity or introducing a metric on the projective space, respectively. One could also introduce the metric on an affine space, which is more akin to the approach of Riemannian geometry and general relativity. The resulting hierarchy can then be schematically described as in Fig. 13.

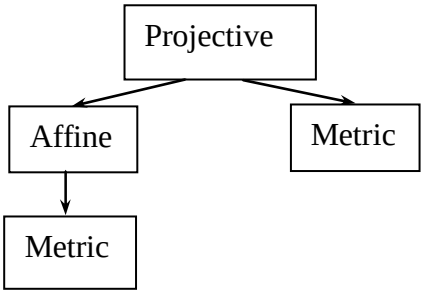

Figure 13. Klein's hierarchy of geometries.

What we shall attempt to do in this chapter is establish that projective geometry plays a role at the fundamental level of physics in both mechanics and field theories. We begin by summarizing some of the relevant notions of projective analytic geometry [**4-7**], and then show how some of these notions play a role in mechanics. We then go on to discuss the use of the algebra of exterior forms and multivector fields to represent projective geometric situations by way of the Plücker-Klein embedding, and then apply that to the exterior forms and multivector fields that are of interest in electromagnetism.

It is our hope that by establishing the role of projective geometry in physics at the fundamental level, this might bring about a shift of emphasis in physics from the Riemannian geometry defined by a metric on the tangent bundle to the Kleinian geometry of subspaces in the tangent spaces, as they are represented by the exterior algebras over the tangent and cotangent bundles. The fundamental physical concept that supplants the

---

gravitation [**3**], in which he attempts to extend the methodology of quantum field theory to imply the Einstein equations by regarding the graviton as an elementary particle like any other, while avoiding the necessity of differential geometry, which he apparently regarded as "fancy-schmancy mathematics."

distribution of massive matter is then the electromagnetic constitutive law, which might itself be determined by the distribution of charged matter, by way of polarization. It is entirely reasonable, considering the relative magnitudes of the coupling constants, that gravity is then simply the shadow cast by electromagnetism.

1. Elementary projective geometry. Something else that Klein once said, which has received considerably more attention by the mathematics community at large, was the suggestion that he made in his Erlangen Programme that geometries could be best characterized by the invariants of group of transformations that were distinctive to the class of problems at hand. For instance, the fundamental concept in affine geometry is parallelism, so the relevant group is the group of transformations of affine space that preserve the parallelism of lines, which is then the group of translations. In metric geometry, the key concept is that of the distance between points of the space, which implies that the relevant group is the group of isometries of the space; in conformal geometry, distance is replaced by angle, and the isometry group is then enlarged to the conformal isometry group. Note that in order to make sense of an angle one requires the presence of *two* lines, which then suggests that one is implicitly considering the points of a projective space; i.e., and angle between two lines through a point in an affine space is equivalent to a distance between two points in the corresponding projective space.

As Hilbert and Cohn-Vossen envisioned the situation [**8**], the key concept in projective geometry was that of *incidence*, which refers to the way that subspaces of projective space intersect. Hence, the relevant group that pertains to projective geometry should be the group of transformations of a projective space that preserve the incidence of subspaces. We shall then begin by clarifying the concepts of projective subspaces and incidence.

a. Projective subspaces and incidence. For the sake of (physically useful) generality, we first assume our field $\mathbb{K}$ of scalars is either $\mathbb{R}$ or $\mathbb{C}$, since many of the definitions and results of projective geometry work the same in either case. In most cases, the differences do not appear until one starts looking at solving algebraic equations or making use of the conjugation operator.

Although a projective space, such as $\mathbb{K}\mathrm{P}^n$, is defined by a process of projectivization, i.e., projecting from a vector space, such as $\mathbb{K}^{n+1}$, to the set of all lines through its origin, for the purposes of projective geometry it is better to regard the elements of that set as points of a projective space, not lines in a vector space. Similarly, under the projection $\mathbb{K}^{n+1} - \{0\} \rightarrow \mathbb{K}\mathrm{P}^n$, a 2-plane through the origin projects to a line in a projective space, and, more generally, a k+1-plane through the origin of $\mathbb{K}^{n+1}$ projects to a k-plane in $\mathbb{K}\mathrm{P}^n$. We shall then call a subset of $\mathbb{K}\mathrm{P}^n$ that is the image of a a k+1-plane through the origin of $\mathbb{K}^{n+1}$ under this projection a *k-dimensional (projective) subspace* of $\mathbb{K}\mathrm{P}^n$.

As subsets, projective subspaces can be partially ordered by inclusion; i.e., one can speak of a subspace as being a subspace of some other subspace or not. However, this is not precisely the relation of incidence, which, unlike a partial ordering, is symmetric. In particular, two subspaces are *incident* iff one of them is a subspace of the other one [3]. For instance, a point and a line are incident iff the point in question is one of the points of the line, while two lines are incident iff they coincide. One then sees that it is possible for two subspaces to be neither incident nor disjoint, such as when a line intersects a plane without being contained in it.

One of the binary operations that pertain to the partial ordering of subsets that carries over to the partial ordering of subspaces is that of intersection. However, one refers to the intersection of two subspaces S_1 and S_2 as their *meet* and denotes it by $S_1 \wedge S_2$. Under the defining projection of $\mathbb{K}^{n+1} - \{0\}$ onto $\mathbb{K}\mathrm{P}^n$, the meet of two projective subspaces will be the image of the intersection of the linear subspaces that projected onto them. One also sees that $-1 \leq \dim(S_1 \wedge S_2) \leq \min\{\dim(S_1), \dim(S_2)\}$; by convention, $\dim\{\varnothing\} = -1$.

For example, the intersection of a line and a plane in $\mathbb{K}\mathrm{P}^3$ can be either a point or a line, in which case, the line is incident on the plane. The intersection of two planes can be a line or a plane since they are projections of 3-planes in $\mathbb{K}^4$, which can intersect in a 2-plane or a 3-plane. The only pairs of projective subspaces of $\mathbb{K}\mathrm{P}^3$ that can intersect vacuously are a point and some subspaces of any dimension and some pairs of lines.

One sees why the concept of parallelism is not as useful in the context of projective geometry by considering the intersection of two lines in $\mathbb{K}\mathrm{P}^2$. Since they both represent the projections of 2-planes through the origin of $\mathbb{K}^3$, and all such 2-planes must intersect in a linear subspace of dimension one or two the projection of their intersection will be either a point or a line in $\mathbb{K}\mathrm{P}^2$, but a non-vacuous subset, in either event. In other words, any two lines in $\mathbb{K}\mathrm{P}^2$ must intersect, so no lines can be parallel to each other. However, as we shall discuss below, if one regards a projective space as an affine space plus a hyperplane "at infinity" then one is saying that all of the parallel lines in the affine space will intersect at points in the hyperplane at infinity. Hence, one sees how projective spaces can be regarded as the completions of affine spaces, in one sense. This makes the claim of Klein about the level of generality that projective geometry represents seem more reasonable.

By contrast, one does not usually consider the union of two subspaces, as much as their *join*, which is denoted by $S_1 \vee S_2$. This is then the smallest subspace (with respect to inclusion) that contains both subspaces. Under the defining projection, the join of two projective subspaces will be the projection of the join of the linear subspaces that projected onto them, in the sense of the smallest linear subspace that contained both of them. This join will then be generated by all finite linear combinations of vectors from both spaces. Hence, $\max\{\dim(S_1), \dim(S_2)\} \leq \dim(S_1 \vee S_2) \leq \dim(S_1) + \dim(S_2)$. In fact:

[3] More elaborate axioms for incidence than the ones give here can be found in Van der Waerden [**9**].

$$\dim(S_1 \vee S_2) = \dim(S_1) + \dim(S_2) - \dim(S_1 \wedge S_2). \qquad \text{(XII.1)}$$

For example, the join of two distinct points is a line and the join of a line and a point not on that line is a plane. When two linear subspaces V_1 and V_2 of $\mathbb{K}^{n+1}$ are transversal, in the sense that $V_1 \wedge V_2 = 0$ and $V_1 \vee V_2 = \mathbb{K}^{n+1}$ the corresponding condition on projective subspaces S_1 and S_2 of $\mathbb{K}\mathrm{P}^n$ is that $S_1 \wedge S_2 = \varnothing$ and $S_1 \vee S_2 = \mathbb{K}\mathrm{P}^n$. That is, they are disjoint and their join is the entire space.

Although one cannot speak of linear combinations of elements in $\mathbb{K}\mathrm{P}^n$, and consequently, bases for subspaces, one can speak of "projective frames." A set $\{p_0, \ldots, p_k\}$ of k+1 elements $p_i \in \mathbb{K}\mathrm{P}^n$ is said to be a *projective frame* [4] for the k-dimensional subspace S_k iff S_k is the join $p_0 \vee \ldots \vee p_k$ of all its elements (one can define that notion by recursion since the join operation is associative) and no proper subset of that set will span S_k. Hence, two distinct points frame a line, three non-collinear points frame a plane, and so on. The pre-images of the points in a projective frame then define a linear frame for the pre-image of S_k .

When given the meet and the join operations, the partially ordered set of projective subspaces of $\mathbb{K}\mathrm{P}^n$ becomes a *lattice* (see, e.g., Birkhoff [**10**]). It also has a greatest element and a least element in the form of $\mathbb{K}\mathrm{P}^n$ and $\varnothing$, respectively. Hence, every subspace is a subspace of $\mathbb{K}\mathrm{P}^n$ and $\varnothing$ is a subspace of every subspace.

This lattice is, moreover, a *complemented* lattice in the sense that for every subspace S there is a subspace S' such that $S \vee S' = \mathbb{K}\mathrm{P}^n$. However, this complement is by no means unique, which would be the case for an *orthocomplemented* lattice. An example of such a lattice is given by the linear subspaces of an orthogonal space, since one can uniquely specify an orthogonal complement to any subspace in that event.

In one sense, we can regard projective geometry as being primarily concerned with transformations of $\mathbb{K}\mathrm{P}^n$ that preserve its lattice of projective subspaces, in the sense that meets go to meets and joins go to joins. Consequently, such a transformation will also preserve incidence.

b. Duality. One can just as well define the projectivization of $\mathbb{K}^{n*} = (\mathbb{K}^n)^*$ by looking at all of the lines through its origin. The resulting projective space will then be denoted by $\mathbb{K}\mathrm{P}^{n*}$, although it will be projectively equivalent to $\mathbb{K}\mathrm{P}^n$, in a sense that we shall define shortly. We shall refer to $\mathbb{K}\mathrm{P}^{n*}$ as the *dual* of the projective space $\mathbb{K}\mathrm{P}^n$.

[4] The classical term for a projective frame was a "reference simplex," or "reference tetrahedron," in the three-dimensional case. We shall not use that terminology since the concept of a frame is closer to the group-theoretic methods of geometry.

The lattice of projective subspaces of $\mathbb{K}\mathrm{P}^{n*}$ will be in one-to-one correspondence with that of $\mathbb{K}\mathrm{P}^n$, except that a k-dimensional projective subspace of $\mathbb{K}\mathrm{P}^{n*}$ will define an $n–k$-dimensional projective subspace of $\mathbb{K}\mathrm{P}^n$, by annihilation. That is, all of the elements of the linear subspace V_{k+1} of $\mathbb{K}^{n+1}$ that is the pre-image of some k-dimensional projective subspace S_k of $\mathbb{K}\mathrm{P}^n$ will be annihilated by all of the elements of some unique subspace V^*_{n-k} of $(\mathbb{K}^{n+1})^*$ that is the pre-image of a unique $n–k$-dimensional subspace S^*_{n-k} of $\mathbb{K}\mathrm{P}^{n*}$. As we shall see, the fact that any k-dimensional subspace in $\mathbb{K}\mathrm{P}^n$ is associated with a unique $n–k$-dimensional subspace in $\mathbb{K}\mathrm{P}^{n*}$ is at the root of the Poincaré duality between k-vectors and $n–k$-forms that we have been using all along.

The one-to-one correspondence between subspaces of $\mathbb{K}\mathrm{P}^n$ and subspaces of $\mathbb{K}\mathrm{P}^{n*}$ does not, however, preserve the operations of meet and join. Rather, it inverts them, just as subset complementation inverts the operations of intersection and union. That is, the dual of the meet of two subspaces in $\mathbb{K}\mathrm{P}^n$ is the join of the duals of the subspaces, and conversely.

One must be cautioned at this point not to assume that the one-to-one correspondence between *subspaces* in $\mathbb{K}\mathrm{P}^n$ and subspaces of its dual implies that there is an actual one-to-one correspondence between the *points* of these spaces. Indeed, under the present duality, a point in one space is associated with a *hyperplane* in the other. Such an association of points would be called a "correlation," which we shall discuss below.

c. Hyperplane at infinity. As we pointed out in chapter II, the projective spaces $\mathbb{K}\mathrm{P}^n$ cannot be covered by a single coordinate chart, since they are all compact, while $\mathbb{K}^n$ is not. However, it is in examining the homogeneous coordinate charts for $\mathbb{K}\mathrm{P}^n$ that we gain a deeper insight into the relationship between affine and projective spaces.

Let $(x^0, x^1, \ldots, x^n)$ be one such chart that projects onto the corresponding inhomogeneous chart $(X^1, \ldots, X^n)$ by the prescription $X^i = x^i/x^0$. For each $x^0 \neq 0$, this map is a diffeomorphism of the affine subspace of $\mathbb{K}^{n+1}$ that is parameterized by setting x^0 equal to a constant and allowing the remaining coordinates to vary arbitrarily in the vector space $\mathbb{K}^n$.

However, as x^0 goes to 0 the inhomogeneous coordinates all become indefinitely large. Hence, none of the points of $\mathbb{K}^{n+1}$ that take the form $(0, x^1, \ldots, x^n)$ project to inhomogeneous coordinates. One refers to the points of this hyperplane as the *hyperplane at infinity.* Nevertheless, they still describe points of $\mathbb{K}\mathrm{P}^n$, even though they

do not define inhomogeneous coordinates. Indeed, one can regard $\mathbb{K}\mathrm{P}^n$ as consisting of the affine space $\mathbb{K}^n$ completed by the addition of the hyperplane at infinity, which also renders the resulting space compact.

For instance, if one looks at $\mathbb{K}\mathrm{P}^1$, which consists of lines through the origin of the plane, in terms of the homogeneous coordinates (x^0, x^1) and represents the inhomogeneous coordinates by $(1, X)$ with $X = x^1/x^0$ then it becomes clear that the point of $\mathbb{K}\mathrm{P}^1$ that gets left out by the omission of $x^0 = 0$ is the vertical line in the plane. The inhomogeneous coordinate X is, in fact, the tangent of the angle between the line through the point (x^0, x^1) and the x^0 axis. Hence, to completely describe all of the lines through the origin, one must add the "point at infinity" that is described by the vertical line. We illustrate this situation in Fig. 14. One should also notice that since all of the lines through the origin in this case are described by either homogeneous coordinates with $x^0 > 0$ or $x^0 < 0$ individually, the projection of $\mathbb{K}^2 - (0, x^1)$ onto $\mathbb{K}\mathrm{P}^1$ is two-to-one and thus represents a double covering map. This fact is at the root of the spin representations of the orthogonal groups.

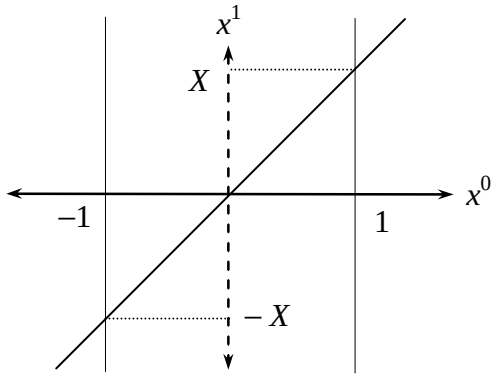

Figure 14. The projective line as an affine line plus a point at infinity.

d. Projective transformations. We shall call a map f: $\mathbb{K}\mathrm{P}^n \to \mathbb{K}\mathrm{P}^m$ of an n-dimensional projective space to an m-dimensional one a *projective transformation* iff it maps the lattice of projective subspaces of $\mathbb{K}\mathrm{P}^n$ into the lattice of projective subspaces of $\mathbb{K}\mathrm{P}^m$ consistently. That is, if $S \subset S'$ in $\mathbb{K}\mathrm{P}^n$ then $f(S) \subset f(S')$, and for any subspaces S_1, S_2 one has $f(S_1 \wedge S_2) \subset f(S_1) \wedge f(S_2)$, $f(S_1 \vee S_2) \subset f(S_1) \vee f(S_2)$.

Under the projectivization $\mathbb{K}^{n+1} - \{0\} \to \mathbb{K}\mathrm{P}^n$, linear subspaces of $\mathbb{K}^{n+1}$ go to projective subspaces of $\mathbb{K}\mathrm{P}^n$, and one finds that a projective transformation of $\mathbb{K}\mathrm{P}^n$ to $\mathbb{K}\mathrm{P}^m$ is covered by a map of $\mathbb{K}^{n+1}$ to $\mathbb{K}^{m+1}$ that takes linear subspaces of $\mathbb{K}^{n+1}$ to linear subspaces of $\mathbb{K}^{m+1}$ in a manner that preserves the lattice of linear subspaces.

This way of characterizing projective transformations subsumes the classical concept of a *collineation*, which takes lines in $\mathbb{K}^{n+1}$ to other lines in $\mathbb{K}^{n+1}$, hence, points of $\mathbb{K}\mathrm{P}^n$ to other points. However, since a projective transformation must, among other things, preserve all joins, and any k-dimensional projective subspace can be expressed as the join of $k+1$ points, one sees that it is sufficient to specify the effect on points.

Of course, in order to preserve the dimension of subspaces, one must be dealing with invertible projective transformations. We call such invertible projective transformations *projective equivalences.*

When one considers how a projective transformation must take projective frames to projective frames, hence, it must be covered by a map that takes linear frames to linear frames, one sees that projective transformations must be covered by linear transformations. Of course, they are not unique, but are only defined up to scalar multiplication. The usual way of representing these linear transformations is:

$$\rho y^a = A^a_i x^i, \quad \rho \neq 0. \tag{XII.2}$$

One then sees that a projective transformation of $\mathbb{K}\mathrm{P}^n$ to $\mathbb{K}\mathrm{P}^m$ corresponds to a line through the origin in the vector space $L(n+1, m+1)$ of linear maps from $\mathbb{K}^{n+1}$ to $\mathbb{K}^{m+1}$. That is, it defines a point in the projective space that is defined by that vector space.

In the invertible case, one must have $n = m$ to begin with, and the image of a projective k-frame under the map must be a projective k-frame for all $0 < k \leq n$. As for the linear transformation that covers the projective transformation, the determinant of the matrix A^i_j must be non-vanishing (for any linear frame on $\mathbb{K}^{n+1}$). Hence, one is looking at the intersection of a line through the origin in the $\mathfrak{gl}(n+1; \mathbb{K})$ with the elements of the group $GL(n+1; \mathbb{K})$. One finds that this set $PGL(n; \mathbb{K})$ of line intersections is a group under the multiplication of matrices since invertible matrices that are defined up to a multiplicative scalar constant will multiply to produce another invertible matrix that is also defined up to a multiplicative scalar constant. This group is referred to as the *projective linear group* in dimension n.

One sees that each of these equivalence classes contains an element of $SL(n+1; \mathbb{K})$ that is unique, up to sign, since:

$$\det(-A) = (-1)^{n+1} \det A. \tag{XII.3}$$

Hence, when $n+1$ is even – i.e., when n is odd – a change of sign will make no change in the determinant. One can then say that:

$$PGL(n; \mathbb{K}) \cong \begin{cases} SL(n+1; \mathbb{K}), & n \text{ even}, \\ SL(n+1; \mathbb{K})/\mathbb{Z}_2, & n \text{ odd}. \end{cases}$$

These invertible projective transformations are called, variously, *homographies, collineations,* and *projectivities* in the literature. They are ultimately generated by finite products of elementary transformations called *perspectivities*, which show us how projective geometry relates to the older study of perspective in visual perception. A typical example of how a perspectivity, relative to a point *P*, takes a projective 3-frame *ABC* for a projective plane to another projective 3-frame *A′ B′ C′* is illustrated in Fig. 15. One can think of the three-dimensional space in which the connecting lines *PAA′, PBB′, PCC′* exist as the space of homogeneous coordinates.

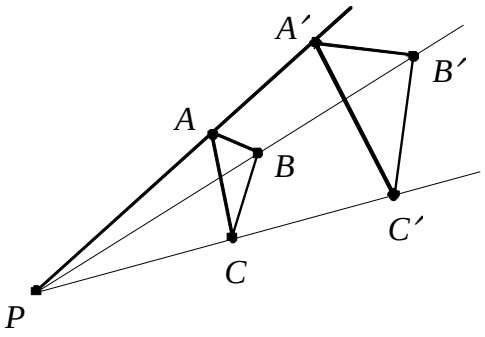

Figure 15. An example of a perspectivity in the projective plane.

It is important to see how the linear transformations of homogeneous coordinates relate to transformations of inhomogeneous coordinates because one finds that generally they will not be linear, but rational transformations. In particular, let $(x^0, x^1, \ldots, x^n)$ be homogeneous coordinates on $\mathbb{K}\mathrm{P}^n$ with corresponding inhomogeneous coordinates $(X^1, \ldots, X^n)$, and similarly one has $(y^0, y^1, \ldots, y^m)$ and $(Y^1, \ldots, Y^m)$ for $\mathbb{K}\mathrm{P}^m$. For convenience, we assume that x^0 and y^0 are non-zero, so $X^i = x^i/x^0$ and $Y^a = y^a/y^0$.

If a linear map A: $\mathbb{K}^{n+1} \to \mathbb{K}^{m+1}$ is represented by a matrix A^α_μ then the transformation of X^i into Y^a that is induced by A is:

$$Y^a = y^i/y^0 = \frac{A^a_0 x^0 + A^a_i x^i}{A^0_0 x^0 + A^0_i x^i} = \frac{A^a_0 + A^a_i X^i}{A^0_0 + A^0_i X^i}. \qquad \text{(XII.4)}$$

In the invertible case, for which $m = n$, these transformations, which are sometimes called *fractional bilinear* transformations, then fall into four elementary categories:

1) *Homotheties:* In this case, the only non-zero submatrices of A^μ_ν are A^0_0 and A^i_j, which equals δ^i_j, so X^i goes to $A^0_0 X^i$.

2) *Translations:* For these, the non-zero matrices are $A^0_0 = 1$, $A^i_j = \delta^i_j$, and at least one component of A^i_0. The transform of X^i is then $X^i + A^i_0$.

3) *Inversions:* Now, $A^0_0 = A^i_0 = 0$, $A^i_j = \delta^i_j$, and at least one A^0_j is non-vanishing. These transformations take X^i to $X^i / A^0_j X^j$, as long as X^i does not live in the hyperplane $A^0_j X^j = 0$. One sees that points of the affine hyperplane $A^0_j X^j = 1$ will be fixed by these transformations. Hence, the inversion in question is an inversion *through* the affine hyperplane $A^0_j X^j = 1$.

4) *Linear transformations:* $A^0_0 = 1$ and A^i_j are the only non-zero submatrices.

In general, one can express A^{μ}_{ν} in block matrix form as:

$$A^{\mu}_{\nu} = \left[\begin{array}{c|c} A^0_0 & A^0_j \\ \hline A^i_0 & A^i_j \end{array}\right]. \qquad \text{(XII.5)}$$

The direct sum decomposition of $\mathbb{K}^{n+1}$ that this corresponds to is $\mathbb{K} \oplus \mathbb{K}^n$, in which the first summand consists of all (x^0, 0, …, 0) while the second summand represents all (0, x^1, …, x^n), which is the hyperplane at infinity. One then sees that the affine subgroup $A(n; \mathrm{K})$ of $SL(n+1; \mathbb{K})$ consists of those elements that take the hyperplane at infinity to itself, which will then have $A^0_j = 0$; the stray factor of A^0_0 can be eliminated by the constraint that the determinant of A^{μ}_{ν} is unity. The affine group in n dimensions can also be represented by the transformations that take the affine hyperplane $x^0 = 1$ to itself, which implies that A^0_0 must be set to unity.

e. Correlations. As pointed out above, the duality that associates each k-plane in $\mathbb{K}\mathrm{P}^n$ with a unique $n–k$-plane in $\mathbb{K}\mathrm{P}^{n*}$ is not sufficiently fine-grained to also associate a point of $\mathbb{K}\mathrm{P}^n$ with a point of $\mathbb{K}\mathrm{P}^{n*}$. Such an invertible map $[C]$: $\mathbb{K}\mathrm{P}^n \to \mathbb{K}\mathrm{P}^{n*}$, is called a *correlation*. Under projectivization, it will be covered by an invertible linear map C: $\mathbb{K}^{n+1} \to (\mathbb{K}^{n+1})^*$, which is unique up to a non-zero scalar multiple.

Even though the points of $\mathbb{K}\mathrm{P}^{n*}$ no longer represent linear functionals, as the elements of $(\mathbb{K}^{n+1})^*$ do, nonetheless one can still speak of the evaluation of an element $[\alpha] \in \mathbb{K}\mathrm{P}^{n*}$ on an element $[\mathbf{v}] \in \mathbb{K}\mathrm{P}^n$. The trick is to understand that the "field of projective scalars" associated with $\mathbb{K}$ consists of 0 and "$\neq 0$." That is, the only issue in the eyes of projective geometry is whether the number $\alpha(\mathbf{v})$ is zero or not. Since any choices of representive non-zero vectors and covectors $\mathbf{v}$ and α for the points $[\mathbf{v}]$ and $[\alpha]$ will differ by non-zero scalar multiples, so will all possible values of the number $\alpha(\mathbf{v}) \in \mathbb{K}$. Hence, $\alpha(\mathbf{v})$ will either be zero for all representatives or non-zero for all representatives and the evaluation $[a][\mathbf{v}]$ will either be zero or non-zero unambiguously.

The geometric interpretation for the vanishing or non-vanishing of $[\alpha][\mathbf{v}]$ is then based in incidence: $[\alpha][\mathbf{v}] = 0$ iff the point $[\mathbf{v}]$ is incident on the hyperplane $[\alpha]$.

One can then define the evaluation $[C][\mathbf{v}][\mathbf{w}]$ of $[C][\mathbf{v}] \in \mathbb{K}\mathrm{P}^{n*}$ on the point $[\mathbf{w}] \in \mathbb{K}\mathrm{P}^n$, which allows one to define the "projectively bilinear" functional on $\mathbb{K}\mathrm{P}^n$:

$$[C]([\mathbf{v}], [\mathbf{w}]) = [C][\mathbf{v}][\mathbf{w}], \qquad \text{(XII.6)}$$

whose value will either be zero or not. Geometrically, one sees that $[C]([\mathbf{v}], [\mathbf{w}])$ will vanish iff the line $[\mathbf{w}]$ is incident on the hyperplane $[C][\mathbf{v}]$.

This functional on $\mathbb{K}\mathrm{P}^n$ will be covered by a bilinear functional on $\mathbb{K}^{n+1}$:

$$C(\mathbf{v}, \mathbf{w}) = C(\mathbf{v})(\mathbf{w}). \tag{XII.7}$$

Since the linear map C: $\mathbb{K}^{n+1} \rightarrow (\mathbb{K}^{n+1})^*$ is defined only up to a nonzero scalar multiple, so is the bilinear functional that it defines. Hence, we are now dealing with a point in the projective space that is obtained from the vector space $(\mathbb{K}^{n+1})^* \otimes (\mathbb{K}^{n+1})^*$.

One can then consider the symmetry of the bilinear form $[C]$. When the functional $[C]$ is symmetric one calls the map $[C]$ itself a *polarity*. Hence, $[\mathbf{v}]$ is incident on $[C][\mathbf{w}]$ iff $[\mathbf{w}]$ is incident on $[C][\mathbf{v}]$. Furthermore, if a point $[\mathbf{v}] \in \mathbb{K}\mathrm{P}^n$ is regarded as the *pole* of a polarity then the point $[C][\mathbf{v}] \in \mathbb{K}\mathrm{P}^{n^*}$ is then regarded as its *polar*, in classical terminology.

The bilinear map C that projects to $[C]$ can be either symmetric or anti-symmetric. Either are called *involutions* since the expression $[C]([\mathbf{v}], [\mathbf{w}])$ will not have a sign, and will then be sensitive to the symmetry of C only by means of the absolute value; i.e., $[C]([\mathbf{v}], [\mathbf{w}])$ is symmetric iff $|C(\mathbf{v}, \mathbf{w})| = |C(\mathbf{w}, \mathbf{v})|$ for any representative non-zero vectors $\mathbf{v}, \mathbf{w} \in \mathbb{K}^{n+1}$. When C is anti-symmetric, the map $[C]$ is called a *null correlation*.

Of particular interest is the case where $[C][\mathbf{v}][\mathbf{v}]$ vanishes, since this means that $[\mathbf{v}]$ is incident on the hyperplane $[C][\mathbf{v}]$; such a point will be called *isotropic*, since it is the projection of a non-zero vector $\mathbf{v} \in \mathbb{K}^{n+1}$ such that $C(\mathbf{v}, \mathbf{v}) = 0$. Clearly, when C is anti-symmetric this will always be the case. The set of all isotropic vectors in $\mathbb{K}^{n+1}$ is therefore a quadric hypersurface. If all $\mathbf{v}$ are isotropic then C must be anti-symmetric since:

$$0 = C(\mathbf{v} - \mathbf{w}, \mathbf{v} - \mathbf{w}) = -[C(\mathbf{w}, \mathbf{v}) + C(\mathbf{v}, \mathbf{w})] \tag{X.8}$$

when all vectors are isotropic.

f. Homogeneous functions. The usage of the words "homogeneous" and "inhomogeneous" in the context of coordinates for projective space is actually quite consistent with their usage in the context of functions. This is due to the fact that any homogeneous function on $\mathbb{K}^{n+1}$ is associated with a unique inhomogeneous function on $\mathbb{K}\mathrm{P}^n$ and conversely, since if $f(\mathbf{x})$ is homogeneous of degree r on $\mathbb{K}^{n+1}$ then, by definition:

$$f(\lambda\mathbf{x}) = \lambda^r f(\mathbf{x}) \tag{X.9}$$

for all scalars $\lambda \in \mathbb{K}$. This implies that $f(\mathbf{x}) = 0$ iff $f(\lambda\mathbf{x}) = 0$ for all λ. Hence, f has the same value as a projective scalar (viz., 0 or $\neq 0$) for all representatives $\mathbf{x}$ of the point $[\mathbf{x}] \in \mathbb{K}\mathrm{P}^n$. This means that the level hypersurface $f(\mathbf{x}) = 0$ in $\mathbb{K}^{n+1}$ projects to a level hypersurface $[f][\mathbf{x}] = 0$ in $\mathbb{K}\mathrm{P}^n$.

The case in which $r = 0$ defines a class of homogeneous functions that one calls *ray functions*, since they are constant on the lines generated by the non-zero $\mathbf{x}$ in $\mathbb{K}^{n+1}$. They therefore define functions on $\mathbb{K}\mathrm{P}^n$ by the association $[f][\mathbf{x}] = f(\mathbf{x})$ for any $\mathbf{x}$ that generates $[\mathbf{x}]$.

The way that one defines the inhomogeneous function $[f]$ on $\mathbb{K}\mathrm{P}^n$, when it is described by inhomogeneous coordinates $X^i = x^i/x^0$ is:

$$[f](X^1, \ldots, X^n) = f(x^0, x^i) = (x^0)^r f(1, X^i), \tag{X.10}$$

when one has chosen a non-zero value of x^0; of course, if f is a ray function then this choice is immaterial.

Conversely, when one is given $[f]$ one can define the homogeneous function f by means of:

$$f(x^0, x^i) = [f](x^i/x^0), \tag{X.11}$$

as long as x^0 is non-vanishing.

One finds that a polynomial function of degree d on $\mathbb{K}^{n+1}$ is homogeneous iff it is the sum of monomials of degree d. In particular, quadratic forms are homogeneous of degree two.

Of particular interest in classical geometry are the quadratic functions. If $\mathbb{K}$ represents the space of inhomogeneous coordinates X for $\mathbb{K}\mathrm{P}^1$ then an inhomogeneous quadratic polynomial:

$$[P][X] = aX^2 + bX + c \tag{X.12}$$

corresponds to the homogeneous polynomial on $\mathbb{K}^2$:

$$P[x, y] = ax^2 + bxy + cy^2 \tag{X.13}$$

when one omits the factor $1/y^2$; as long as the only issue is the vanishing of the polynomials, this is permissible. If the level hypersurface $[P] = 0$ consists of two points X_1 and X_2 in $\mathbb{K}\mathrm{P}^1$ – viz., the roots of the polynomial – then the level hypersurface $P = 0$ consists of two lines in $\mathbb{K}^2$ that are generated by any points (x_1, y_1) and (x_2, y_2) such that $X_i = y_i/ x_i$, $i = 1, 2$.

Going to the next dimension, an inhomogeneous polynomial on $\mathbb{K}\mathrm{P}^2$:

$$[P][X, Y] = aX^2 + bXY + cY^2 + dX + eY + f \tag{X.14}$$

becomes, when one sets $X = x/z$ and $Y = y/z$:

$$P[x, y, z] = ax^2 + bxy + cy^2 + dxz + eyz + fz^2; \tag{X.15}$$

again, we have eliminated the irrelevant multiplicative factor $1/z^2$.

We now see that the quadratic polynomials in two variables, which describe conic sections, also correspond to quadratic forms on $\mathbb{K}^3$. Indeed, that is the space in which the cone is defined, and a section of the cone is defined by a choice of intersecting affine plane in $\mathbb{K}^3$. The fact that one is dealing with a cone is due to the fact that the homogeneity of the function $[P]$ implies that the level hypersurface $[P] = 0$ contains the lines through each of its points and the origin. When $\mathbb{K} = \mathbb{R}$, one sees that the quadratic form $[P]$ must be hyperbolic in order for $[P] = 0$ to consist of anything but the origin.

We shall discuss the case $\mathbb{K} = \mathbb{R}$, $n = 3$ in the section since it is directly relevant to special relativity.

2. Projective geometry and mechanics. In order to justify the importance of a purely mathematical concept or technique in physics, one must show its point of application. Furthermore, that point of application can be either specialized or fundamental in character, where a specialized application of a mathematical technique is generally of no interest outside of some particular problem. For instance, a trick that makes an integral more manageable would fall into this category.

The purpose of this chapter is to show that projective geometry can be applied to the mathematical modeling of physical phenomena in many contexts, including at the most fundamental level of mechanics and field theories. In effect, in this section we will be going beyond the scope of pre-metric electromagnetism into "pre-metric mechanics" to show that the issues in physics that touch upon the methods and concepts of projective geometry are found at the very root level of physics itself, namely, the process of measurement that serves as the empirical, phenomenological basis for the theories that emerge eventually.

a. The geometry of measurement. The simplest level at which relativistic physics and quantum physics overlap seems to be at the level of physical measurements and observations. Indeed, one can think of an observation as a special kind of measurement, namely, a *passive* measurement. By this, we mean that the measurer/observer is not interacting with the source of the information that is being measured, only the information that is being received. The obvious example of such measurements would be the ones that are made by astronomers concerning the photons that are emitted by distant stars.

By contrast, an *active* measurement involves the measurer emitting information that is intended to interact with the system in question so that the measurement will involve the way that the information that comes back from the system has been altered by the state of

the system. This is the spirit of elementary particle physics, which depends upon collisions of test particles with target particles to deduce information about the structure of the state space – i.e., the field space – in which the particle/fields live. It also accounts for most of the measurements of geophysics that are used to construct models for the Earth's interior, although the mere act of recording an unprovoked earthquake would constitute a passive measurement.

Of course, in the eyes of quantum mechanics, or rather, the statistical interpretation of wave mechanics (see, e.g., Dirac [**11**]), the concept of a measurement includes both types. Hence, although it is absurd to say that observing a distant star through a telescope has changed the state of the star since the event being observed is outside of your causal future, nonetheless, it is correct to say that the state of the star changed as a result of the emission of the photons that are being observed, however slightly. Indeed, the necessity of passing from classical to quantum physics seems to be most unavoidable when one can no longer justify the existence of "test" particles, in the sense of particles whose interaction with the system in question will not change its state appreciably. For instance, a collision with a microwave photon from a traffic policemen's radar gun will not change the state (e.g., position and momentum) of an oncoming motor vehicle appreciably, but it might very well change the state of an atomic electron.

In order to see how this all relates to projective geometry, one must start with the observation of Max Born that all measurements are carried out in the rest space of the measuring device. One must then take a closer look at the very nature of a rest space, along with the process of measurement itself, which is where projective geometry becomes applicable, because, in a sense, it is the geometry of perception. The recurring theme in what follows is that geometrically a rest space is not an affine space, but a projective space, so the affine spaces of conventional physics represent either the hyperplane at infinity in the projective space or the vector space that it projects onto it in its definition; i.e., the space of homogeneous coordinates.

First, let us examine the geometry of vision, which amounts to the statement that the effect of a converging lens on light rays is analogous to the effect of projecting from homogeneous to inhomogeneous coordinates. We refer to Fig. 16 in our explanation.

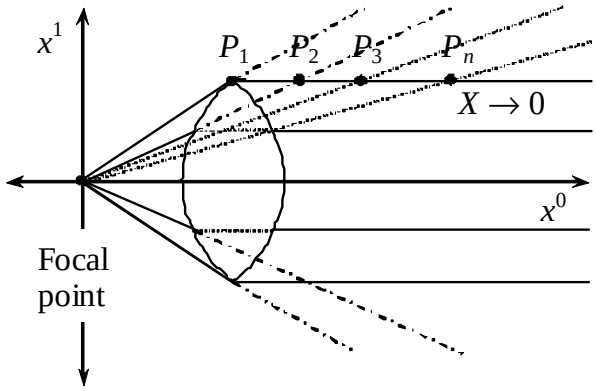

Figure 16. The effect of a converging lens on parallel lines.

The key geometrical attribute of a converging lens is the fact that it bends parallel lines into lines that all intersect in a single point, which is called the *focal point*. One can see that, for all practical purposes, the parallel lines "live" in the affine space $\mathbb{R}^2$, while the lines through the focal point "live" in the projective line $\mathbb{R}\mathrm{P}^1$. Furthermore, the projection of points in $\mathbb{R}^2$ to points in $\mathbb{R}\mathrm{P}^1$ that is implied by the process of refraction is precisely the projection of homogeneous coordinates for $\mathbb{R}\mathrm{P}^1$ onto inhomogeneous coordinates. For instance, if one considers the sequence of points P_1, P_2,

… along the top incoming light ray in the diagram then one sees that their homogeneous coordinates (x^0, x^1) will all have the same value of x^1 and increasing values of x^0, while the corresponding sequence of inhomogeneous coordinates $X^1, ^2 \ldots$ with $X = x^1/x^0$ will be converging to zero. Indeed, this same fact would be true for any other light ray that is parallel to the one considered. In effect, they all converge to a "point at infinity" – the *vanishing point*, as they say in the language of perspective – in the same way that the refracted rays intersect at the focal point. Indeed, in terms of homogeneous coordinates the x^1 axis becomes the point at infinity relative to the projection of $\mathbb{R}^2 - \{0\}$ onto $\mathbb{R}\mathrm{P}^1$ and the focal point is the ideal point $\{0\}$ that is not projected. Hence, the way that the scene to the right of the lens appears to the observer at the focal point is like an inversion through the vertical line through the center of the lens that maps the focal point to the vanishing point.

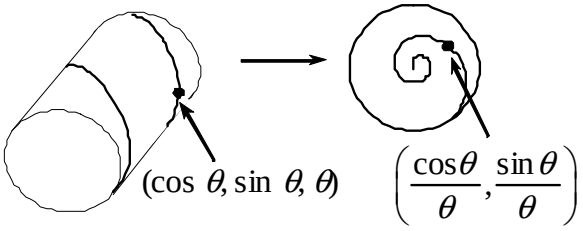

Figure 17. The projection of a helix from homogeneous to inhomogeneous coordinates.

As an example of the analogous situation in a three-dimensional affine space projecting onto a two-dimensional projective space, consider the way that a cylindrical helix about the *z*-axis in $\mathbb{R}^3$, which we rename the θ axis, appears when viewed along that axis, namely, as a spiral. We illustrate this in Fig. 17.

b. Special relativity. Certainly no one would deny the applicability of projective geometric concepts to the problem of measurement/observation in the purely optical sense that we just described. However, the somewhat surprising fact is that one can also apply precisely the same considerations to the problem of projecting events in a four-dimensional spacetime manifold M into the three-dimensional rest space Σ of a measurer/observer.

The key to understanding this is in understanding the difference between the proper time parameter τ of a worldline $x(\tau)$ in M and the time coordinate $t = x^0/c$ associated with a coordinate chart $(x^0, \ldots, x^3)$ on an open subset U in M. One must understand that proper time is valid only in the rest space of a measurer/observer. Hence, the parameter τ refers to the rest space of the point whose worldline is described by $x(\tau)$. One must then think of the time coordinate t as simply being the proper time parameter for another measurer/observer whose rest space involves the coordinates of the chart as measurements.

Since:

$$\frac{dt}{d\tau} = \left(1 - \frac{v^2}{c^2}\right)^{1/2} \tag{XII.16}$$

one sees that the parameter τ and the coordinate t will differ iff the measurer/observer that is described by the coordinate chart is not comoving with the object that is described by the worldline $x(\tau)$, which would make the relative speed v vanish.

As a result of this, if:

$$\mathbf{v}(\tau) = v^{\mu}(\tau)\frac{\partial}{\partial x^{\mu}} \qquad [g(\mathbf{v}, \mathbf{v}) = c^2] \tag{XII.17}$$

is the velocity vector field along $x(\tau)$ then if one wishes to describe how it will appear in the rest space of the coordinates, one cannot simply project $(v^0, \ldots, v^3)$ onto its spatial components (v^1, v^2, v^3), which would be appropriate if the tangent spaces to M looked like $\mathbb{R}\oplus\Sigma$, but one must also change from the proper-time parameter τ to the time coordinate t as a parameter. (In (XII.17), we have assumed a Lorentzian metric g so the constraint on **v** that we make corresponds to the constraint that $x(\tau)$ is parameterized by proper time.)

By the chain rule for differentiation, this means that the spatial components of **v**, as seen by the measurer/observer that defines the chart will be:

$$V^i(t) = \frac{dx^i}{dt} = \frac{d\tau}{dt}\frac{dx^i}{d\tau} = \frac{v^i}{v^0}. \tag{XII.18}$$

Hence, the projection of $(v^0, \ldots, v^3)$ onto the rest space described by the chart gives the inhomogeneous coordinates (V^1, V^2, V^3), instead of the homogeneous ones (v^1, v^2, v^3).

We then conclude that, at least as far as velocity is concerned, rest spaces are projective spaces, not affine ones. The affine space that usually gets employed to describe the projection of velocity four-vectors onto three-vectors amounts to the plane at infinity in $\mathbb{R}P^3$, which corresponds to the hyperplane $v^0 = 0$ in the space of homogeneous coordinates.

One can give this situation a coordinate-free description by defining the *projectivized tangent bundle* $PT(M)$ to M to be set of all lines through the origins of the tangent spaces in $T(M)$. Hence, there will be a map $T(M) - Z(M) \to PT(M)$, $\mathbf{v} \mapsto [\mathbf{v}]$ that takes each non-zero tangent vector to the line through the origin in its tangent space that it generates; we are, of course, using the notation $Z(M)$ to refer to the zero section of $T(M)$.

Since we have established the result above only for velocity vectors, it is illuminating to examine the form that it takes for other four-dimensional measurements. In particular, one must examine the effect of projecting the coordinates $(x^0, \ldots, x^3)$ themselves onto the corresponding inhomogeneous coordinates $X^i = x^i/x^0$, under the assumption that $x^0 = ct$ is non-null. One sees that the only thing that has changed is to rescale the distance measurements x^i into dimensionless units, such as light-seconds. This amounts to using light rays as one's universal yardsticks for the measurement of distance. Hence, geometry again emerges from the propagation of electromagnetic waves.

We naturally need to examine the effect of a coordinate transformation $\overline{x}^{\mu} = \overline{x}^{\mu}(x^{\nu})$ on the inhomogeneous coordinates V^i of the velocity four-vector **v**. As one knows, from the chain rule for differentiation the homogeneous coordinates transform as:

$$\overline{v}^{\mu} = \frac{d\overline{x}^{\mu}}{d\tau} = \frac{\partial \overline{x}^{\mu}}{\partial x^{\nu}}\frac{dx^{\nu}}{d\tau} = A^{\mu}_{\nu}v^{\nu}, \tag{XII.19}$$

in which we have abbreviated the matrix $\partial\overline{x}^{\mu}/\partial x^{\nu}$ of the differential of the coordinate transformation by A^{μ}_{ν}.

We then see that the resulting effect on the inhomogeneous coordinates is to take $V^i = v^i/v^0$ to:

$$\overline{V}^i = \frac{A^i_0 + A^i_j V^j}{A^0_0 + A^0_j V^j}. \tag{XII.19}$$

We recover the conventional Galilean transformations of velocity 3-vectors by setting A^i_0 equal to the components V^i_r of the relative velocity of the measurer/observers described by the two charts, and getting $A^i_j = \delta^i_j$, $A^0_0 = 1$, and $A^0_j = 0$.

Of course, there is a relative speed limit in spacetime that originates in the fact that if two measurer/observers were moving with a relative speed greater than that of light there would be no way for them to communicate information anymore. In conventional relativity, this means that one must erect light cones in the tangent spaces of M and restrict oneself to relative velocities vectors that lie inside them. Of course, as we have seen, when the dispersion law for electromagnetic waves is quartic, not quadratic, the light cones become more involved as algebraic manifolds.

In the quadratic case, one introduces the light cone:

$$c^2(v^0)^2 - g_{ij}\, v^i\, v^j = 0 \tag{XII.20}$$

in the space of homogeneous coordinates for velocity four-vectors. One can also write this in the form:

$$v^0 = \frac{1}{c}(g_{ij} v^i v^j)^{1/2}, \tag{XII.21}$$

which shows that in the Galilean limit as c grows infinite the coordinate v^0 goes to 0. Hence, in the Galilean limit the light cone for a finite c converges to the hyperplane at infinity for $\mathbb{R}\mathrm{P}^3$. Perhaps for this reason, the light cone is referred to as the *absolute quadric* in projective geometry. In a sense, it represents a deformation of the hyperplane at infinity into a quadric hypersurface "at infinity."

Now, let us absorb the factor of c into the units of the homogeneous coordinate x^0 so that the components $g_{\mu\nu}$ are dimensionless. Under projection, the homogeneous quadratic polynomial $P[x^{\mu}] = g_{\mu\nu}\, x^{\mu}\, x^{\nu}$ in $\mathbb{R}^4$ goes to the inhomogeneous polynomial in $\mathbb{R}\mathrm{P}^3$:

$$[P][X^i] = g_{00} + 2g_{0i}\, X^i + g_{ij}\, X^i\, X^j. \tag{XII.22}$$

One finds that there is always a frame (indeed, an infinitude of frames) in $\mathbb{R}^4$ that makes g_{0i} disappear, since this is the case for any orthonormal frame relative to g, and therefore any frame that is obtained from it by an invertible spatial transformation, which does not have to be orthogonal. Furthermore, we rescale the components of the $[P][X]$ by

factoring out the g_{00} and absorbing it, too, into the coordinates X^i. We now have $[P][X]$ in the form:

$$[P][X^i] = 1 + G_{ij}\, X^i\, X^j \qquad (G_{ij} = g_{ij}/\, g_{00}). \qquad \text{(XII.22)}$$

Since our transformation from $P[x]$ to $[P][X]$ has involved only non-zero scalar multiplications, we see that the hypersurface $P[x] = 0$ in $\mathbb{R}^4$ corresponds to the hypersurface $[P][X] = 0$ in $\mathbb{R}\mathrm{P}^3$. In particular, the light cone:

$$P[x] = g_{00}(x^0)^2 - g_{ij}\, x^i\, x^j = 0 \qquad \text{(XII.23)}$$

in $\mathbb{R}^4$ corresponds to the unit sphere in $\mathbb{R}\mathrm{P}^3$:

$$G_{ij}\, X^i\, X^j = 1. \qquad \text{(XII.22)}$$

Hence, we can just as well regard the transition from introducing a Euclidian metric on $\mathbb{R}^3$ to introducing a Minkowski metric on $\mathbb{R}^4$ as being like replacing the affine space $\mathbb{R}^3$ with the projective space $\mathbb{R}\mathrm{P}^3$ and introducing the Euclidian metric on $\mathbb{R}\mathrm{P}^3$ directly.

It is also essential to understand what happens to Lorentz transformations of $\mathbb{R}^4$ when they are projected to fractional bilinear transformations of $\mathbb{R}\mathrm{P}^3$. In fact, one can just as well extend to the Weyl group $\mathbb{R}^*\times O(3, 1)$, by including the non-zero scalar factor ρ; as we mentioned before, this is the linear subgroup of the conformal Lorentz group. We think of the elements of this group as being invertible linear transformations of $\mathbb{R}^4$ that preserve the light cone, in the sense that if $\overline{x}^\mu = A^\mu_\nu x^\nu$ are the transforms of the x^μ then $P[\overline{x}^\mu] = 0$ iff $P[x] = 0$. Hence, the corresponding transformations of $\mathbb{R}\mathrm{P}^3$ are invertible fractional bilinear transformations that preserve the unit sphere defined by G.

It is important to note that this does not mean that they preserve the metric G itself. It means that $G(AX, AX) = 1$ iff $G(X, X) = 1$. Indeed, only the $O(3)$ subgroup ($A^0_0 = 0$, $A^0_j = A^i_0 = 0$, $A^i_j \in O(3)$) will preserve the metric G for all $X \in \mathbb{R}\mathrm{P}^3$.

As for the Lorentz boosts, it is illuminating to see what they become when represented as collineations of $\mathbb{R}\mathrm{P}^3$. For instance, a boost along the x^1 axis takes the following form in homogeneous coordinates ($x^0 = ct$):

$$\overline{x}^0 = \gamma(x^0 - v/c\, x^1), \qquad \overline{x}^1 = \gamma(-\, v/c\, x^0 + x^1), \quad \overline{x}^i = x^i \ \ (i = 1, 2), \qquad \text{(XII.23)}$$

in which we have introduced the Fitzgerald-Lorentz contraction factor:

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1/2} . \qquad \text{(XII.24)}$$

The matrix of this as a linear transformation of $\mathbb{R}^4$ is then:

$$A^{\mu}_{\nu} = \left[\begin{array}{c|ccc} \gamma/c & -\gamma v/c & 0 & 0 \\ \hline -\gamma v/c & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] . \qquad \text{(XII.25)}$$

In terms of projective transformations, such a boost is then going to involve all four types of transformations, namely, homotheties, inversions, spatial translations, and linear transformations.

The corresponding fractional bilinear transformation of $\mathbb{R}\mathrm{P}^3$ is:

$$\bar{X}^1 = \frac{\bar{x}^1}{\bar{x}^0} = \frac{-v + cX^1}{1 - (v/c)X^1} , \qquad \bar{X}^i = \frac{\bar{x}^i}{\bar{x}^0} = \frac{\gamma^{-1} X^i}{1 - (v/c)X^1} \qquad (i = 2, 3). \qquad \text{(XII.26)}$$

Note, in particular, that the coordinates X^2 and X^3 are altered, even though the coordinates x^2 and x^3 were not. Moreover, although the factor γ has dropped out of $\bar{X}^1$, it appears in the other two coordinates.

These transformations admit an immediate physical interpretation when the $X^i = V^i/c$ represent the components of a spatial velocity vector in dimensionless units. They then take the form:

$$\bar{V}^1 = \frac{v + V}{1 + vV/c^2} , \qquad \bar{V}^i = \frac{(c\gamma)^{-1} V^i}{1 + vV/c^2} . \qquad \text{(XII.27)}$$

The first expression is, of course, nothing but the usual relativistic rule for the addition of spatial velocities, which we now see to be projective-geometric in its origin [5].

c. The SL(2; $\mathbb{C}$) *representation of the Lorentz group.* One of the fundamental refinements that quantum wave mechanics, both relativistic and non-relativistic, introduced into special relativity was the fact that the empirically-verified existence of electron spin suggested that there was something more fundamental about the representation of the Lorentz group as the group *SL*(2; $\mathbb{C}$), since the two-to-one homomorphism of that group with the identity component of the Lorentz group seemed

[5] Many of these associations between special relativity and projective geometry were made in the early years of relativity when projective geometry was more mainstream in mathematics than it seems to be nowadays. One might confer the treatment it is given in Kommerell [**5**], for instance. More recently, it has been discussed by Gschwind [**12**] and, of course, the author of the present work [**13**].

to account for the existence of up and down spin states in electron wavefunctions. Since we now see that the groups $SL(n+1; \mathbb{R})$ describe the groups of collineations of the projective spaces $\mathbb{R}P^n$, we should now examine the fact that $SL(2; \mathbb{C})$ has a similar significance in terms of the complex projective space $\mathbb{C}P^1$.

The complex projective space $\mathbb{C}P^1$ is, of course, defined by the set of all *complex* lines through the origin of $\mathbb{C}^2$. That is, if the complex vector $\mathbf{v} \in \mathbb{C}^2$ is represented by the pair (z_1, z_2) then the corresponding point $[\mathbf{v}] \in \mathbb{C}P^1$ that it defines is the set of all complex scalar multiples $\lambda(z_1, z_2) = (\lambda z_1, \lambda z_2)$, $\lambda \in \mathbb{C}$.

Now, a complex line can also be regarded as a real plane, just as $\mathbb{C}^2$ can be regarded as the real vector space $\mathbb{R}^4 = \mathbb{R}^2 \oplus i\mathbb{R}^2$. (Here, we use the symbol i more as an identifier than anything else). However, not every real plane in $\mathbb{R}^4$ represents a complex line, since the complex line spanned by any $(z_1, z_2) = (x_1, x_2) + i(y_1, y_2)$ is the orbit of the action of all scalar multiples by scalars of the form $\lambda = \alpha + i\beta$:

$$\lambda(z_1, z_2) = (\alpha x_1 - \beta y_1, \alpha x_2 - \beta y_2) + i(\alpha y_1 + \beta x_1, \alpha y_2 + \beta x_2). \qquad \text{(XII.28)}$$

For instance, the real plane spanned by all points of the form $(x, y) + i(0, 0)$ is not a complex line in $\mathbb{R}^4$. Hence, there is a difference between the real manifold $\mathbb{C}P^1$ and the Grassmanian manifold V_2^4 of all planes through the origin in $\mathbb{R}^4$.

In fact, one easily sees that the real dimension of $\mathbb{C}P^1$ is two, since the complex dimension is one. One simply regards an element (z_1, z_2) of $\mathbb{C}^2 - \{0\}$ as the homogeneous coordinates of a complex line through the origin while the single inhomogeneous coordinate is then either:

$$Z = z_2 / z_1 \qquad \text{(XII.29)}$$

when z_1 is non-zero or the reciprocal when it is zero.

In real form, (XII.29) looks like:

$$X + iY = \frac{x_1 x_2 + y_1 y_2}{|z_1|^2} + i\frac{x_1 y_2 - y_1 x_2}{|z_1|^2} = \left\|\frac{z_2}{z_1}\right\|(\cos\psi + i\sin\psi), \qquad \text{(XII.30)}$$

in which ψ is the angle between the real lines through (x_1, x_2) and (y_1, y_2).

Since the only point that is being omitted by the set of all inhomogeneous coordinates corresponds to $z_1 = 0$, and the completion of the complex line by a point is the Riemann

sphere, one sees how the two-dimensional real manifold that describes $\mathbb{C}P^1$ is actually the 2-sphere.

Physically, one can think of this 2-sphere as the celestial sphere; i.e., a sphere in "space" at infinity. Under the projection of $\mathbb{C}^2 - \{0\}$ onto $\mathbb{C}P^1$, one also sees that each point of the celestial sphere gets associated with a complex line – i.e., a real plane – as a fiber in $\mathbb{C}^2 - \{0\}$. In fact, this complex line bundle is sometimes referred to as the *canonical line bundle* on $\mathbb{C}P^1$.

As we observed above, the 2-sphere also shows up in special relativity as the projection of the light cone in Minkowski space onto $\mathbb{R}P^3$, although in that case, the resulting radius was finite (viz., unity). However, since the points of the hyperplane at infinity ($x^0 = 0$) are in one-to-one (perspective) correspondence with the points of the hyperplane at $x^0 = 1$, one can see how the celestial sphere is essentially the same as the light sphere.

d. Time-space splittings. The concept of a time-space splitting $[\mathbf{t}](M) \oplus \Sigma(M)$ of the tangent bundle $T(M)$ to a manifold M or an analogous splitting $[\tau](M) \oplus \Sigma^*(M)$ of T^*M is deeply rooted in projective-geometric notions. For the sake of simplicity, we confine ourselves to a (finite-dimensional) vector space V and its dual V^*, in general, rather than the tangent spaces to a particular manifold; however, the transition from vector spaces to vector bundles is entirely straightforward.

If one is given a direct sum decomposition of V into $[\mathbf{t}] \oplus \Sigma$, where $[\mathbf{t}]$ is the line generated by a non-zero vector $\mathbf{t} \in V$ and Σ is a complementary hyperplane, then since any hyperplane, such as Σ, in V defines a unique element $[\tau] = \#^{-1}\Sigma \in PV^*$, we see that in order to complete $[\tau]$ to a direct sum decomposition $[\tau] \oplus \Sigma^*$ of V^*, we need only to specify the complementary hyperplane Σ^*. However, $\Sigma^* = \#[\mathbf{v}]$ is dual to a unique element $[\mathbf{v}]$ in PV, which we then choose to be $[\mathbf{t}]$. Hence, one can either uniquely characterize the splitting $[\mathbf{t}] \oplus \Sigma$ by means of the pair $([\mathbf{t}], [\tau]) \in PV \times PV^*$ or the pair $(\Sigma, \Sigma^*) \in PV^* \times PV$.

The transversality requirements on the aforementioned two splittings of V and V^* can be written in terms of the bilinear pairings of vectors and covectors as:

$$\tau(\mathbf{t}) \neq 0, \qquad \tau(\mathbf{v}) = \nu(\mathbf{t}) = 0, \qquad \mathbf{v} \in \Sigma, \nu \in \Sigma^*. \tag{XII.31}$$

Since the conditions only pertain to zero, it is immaterial which representative vector one uses from the subspaces involved. However, in practice, one might choose τ and $\mathbf{t}$ to make $\tau(\mathbf{t}) = 1$.

The physical process by which a splitting comes about generally begins with the selection of $\mathbf{t}$ to be a velocity vector field on the spacetime manifold M, whose congruence of integral curves represents the motion of a particular measurer/observer. Hence, for practical reasons, one does not generally expect the support of $\mathbf{t}$ to be all of M,

but only some proper subset, such as the diffeomorphic image of a *world tube* of the form $\mathbb{R}\times O$, where O is a three-dimensional "object" such as a 3-chain or compact 3-manifold.

In the absence of a Lorentzian structure, one cannot choose the complementary spatial sub-bundle $\Sigma(M)$ of $T(M)$ uniquely, but in the Lorentzian case, one can choose it to be the orthocomplement of $[\mathbf{t}]$. Since $\Sigma(M)$ is dual to a line field $[\tau]$ in T^*M – i.e., a section of $PT^*M \to M$ – we can also choose the line field $[\tau]$ arbitrarily, except for the transversality conditions in (XII.31), when one defines Σ to be $\#^{-1}[\tau]$ and Σ^* to be $\#[\mathbf{t}]$; hence, only the first condition is not automatic. We shall then refer to the pair $([\mathbf{t}], [\tau])$ as a (pre-metric) choice of measurer/observer.

A choice of measurer/observer $([\mathbf{t}], [\tau])$ defines a splitting $\Lambda_k V = \Lambda_k^{\mathrm{Re}} \oplus \Lambda_k^{\mathrm{Im}}$ ($\Lambda^k V = \Lambda^k_{\mathrm{Re}} \oplus \Lambda^k_{\mathrm{Im}}$, resp.) into time-space k-vectors (k-forms, resp.) and purely spatial ones. In the former case, one can factor out $\mathbf{t}$ (τ, resp.) as an exterior factor, while in the latter case, one has $\Lambda_k^{\mathrm{Im}} = \Lambda_k\Sigma$ ($\Lambda^k_{\mathrm{Im}} = \Lambda^k\Sigma$, resp.) (see Delphenich [**14**] for more discussion of this subject). The reason that there are no time-time-space, etc. summands in $\Lambda_k V$ or its dual is because subspaces $[\mathbf{t}]$ or $[\tau]$ are one-dimensional, so only one exterior factor of $\mathbf{t}$ or τ can appear without driving the resulting expression to zero identically.

By now, we are getting some inkling that whenever our perception of space is based in local – i.e., tangential – objects, we should really be thinking of the space we perceive as a projective space, not an affine one. Indeed, to enlarge the scope of Max Born's observation, one should think of all measurements as local measurements that are carried out in the rest space of the measuring device, and the appropriate representation for the rest space of the measurer/observer at a point of spacetime is the projectivization of the tangent space at that point. The actual splitting of the manifold itself is a deeper matter that involves subtle question of integrability.

e. Wave motion. If we regard the wave covector field $k = \omega\, dt - k_i\, dx^i$ as the fundamental kinematical object in wave mechanics that is analogous to the velocity vector field $\mathbf{v}$ in continuum mechanics then we find that wave motion is dual to point motion in the sense of the word "duality" that projective geometry recognizes. That is, a non-zero tangent vector $\mathbf{v}$ defines a point $[\mathbf{v}]$ in the projectivized tangent bundle $PT(M)$, while a non-zero cotangent vector k defines a point $[k]$ in the projectivized cotangent bundle PT^*M. Hence, whereas a non-zero tangent vector generates a tangent line, a non-zero cotangent vector generates a tangent hyperplane.

We have already observed in Chapter VIII that the effect of projecting the four-dimensional components k_μ of the wave covector k as if they were homogeneous coordinates projecting onto inhomogeneous coordinates is to produce components that have the dimensions of 1/ (phase) velocity:

$$K_i = \frac{k_i}{\omega} = \frac{1}{v_p^i}, \tag{XII.32}$$

although the dimensionless numbers cK_i are not necessarily the principal indices of refraction n_i that one obtains from Fresnel analysis.

Note that if one defines the energy-momentum 1-form p as $\hbar k$ then this has no effect on the corresponding inhomogeneous coordinates.

If we look at the vector field $\mathbf{v} = (\partial P[k(x)] / \partial k_\mu)\, \partial/\partial x^\mu$ then we see that, although we have previously identified the components as belonging to a four-velocity vector field, nonetheless its components actually have the dimensions of time (period) and distance (wavelength). However, the inhomogeneous coordinates that it defines are the components of group velocity, up to sign:

$$V^i = \frac{\partial P / \partial k_i}{\partial P / \partial \omega} = -v_g^i . \tag{XII.33}$$

Hence, we see that whereas the components K_i describe a point in PT^*M the components V^i describe a point in $PT(M)$.

Since the characteristic polynomial $P[k]$ is homogeneous of degree four in k, under the projection of the k_μ onto the K_i , it will define a inhomogeneous polynomial $[P][K]$ of degree four on $\mathbb{R}\mathrm{P}^3$. The characteristic hypersurface that $P[k]$ defines by its vanishing will then produce a corresponding hypersurface in $\mathbb{R}\mathrm{P}^3$. Of course, the geometry of this situation will be more complicated to describe than the projection of a cone onto a sphere.

Furthermore, by homogeneity, we have:

$$P[k] = \tfrac{1}{4}\frac{\partial P}{\partial k}(k) = \tfrac{1}{4}k(\mathbf{v}) \tag{XII.34}$$

so we see that the effect of the dispersion law $P[k] = 0$ is to make $k(\mathbf{v})$ vanish, as well. That is, the vectors $\mathbf{v}(x)$ are incident on the tangent hyperplanes that are annihilated by k. However, we shall still regard $\mathbf{v}$ as the dual vector field to k, even though we do not have the transversality that would make $k(\mathbf{v})$ non-vanishing, as we assumed for $\tau(\mathbf{t})$, above. This "isotropy" condition is based in the physical reality that there is no rest space for electromagnetic wave motion.

We recall the important property of a homogeneous function $P[k]$ – of any degree r – that since Euler's formula takes the form:

$$k_\mu v^\mu = rP[k], \tag{XII.35}$$

the hypersurface $P[k] = 0$ in $\mathbb{R}^{4*}$ corresponds to the hypersurface:

$$k_\mu v^\mu = 0 \tag{XII.36}$$

in $\mathbb{R}^{4*} \times \mathbb{R}^4$ and the hypersurface:

$$K_i V^i = 1 \tag{XII.37}$$

in $\mathbb{R}\mathrm{P}^{3*} \times \mathbb{R}\mathrm{P}^3$ (Here, the components km are of the form $(\omega, -k_i)$).

We previously encountered this situation in the context of the Fresnel normal surface and ray surface in the context of geometrical optics.

3. The Plücker-Klein embedding. Although projective geometry begins by dealing with lines through the origin in a vector space – say, $\mathbb{K}^{n+1}$ – nonetheless, as we saw above, it is also concerned with 2-planes, 3-planes, and so on up to hyperplanes in $\mathbb{K}^{n+1}$. In the case of dimension four, one is then concerned with points, lines, and planes in $\mathbb{K}P^3$. Furthermore, by duality, they will also be represented by planes, lines, and points in $\mathbb{K}P^{3*}$, respectively.

a. Plücker-Klein embedding. The connection between projective geometry and electromagnetism comes about once one discovers that one can represent any k-plane through the origin in $\mathbb{K}^n$ by either a decomposable k-vector in $\Lambda_k(\mathbb{K}^n)$ or a decomposable n–k-form in $\Lambda^{n-k}(\mathbb{K}^n)$; both are unique up to a non-zero scalar factor. Hence, the representation of a k-plane by a point in the projective spaces $P\Lambda_k(\mathbb{K}^n)$ or $P\Lambda^{n-k}(\mathbb{K}^n)$ is unique. Furthermore, the operations of exterior and interior multiplication are closely related to the meet and join operations, although not isomorphically.

Suppose V_k is a k–plane through the origin in $\mathbb{K}^n$. Let $\mathbf{e}_i$, $i = 1, \ldots, k$ be a k-frame that spans it. Since the members of any frame are linearly independent, the k-vector $\mathbf{e}_1 \wedge \ldots \wedge \mathbf{e}_k$ is non-zero. Now, observe the effect of changing to another frame $\mathbf{f}_i$ for V_k that is related to the first one by the formula $\mathbf{f}_i = A_i^j \mathbf{e}_j$. When one forms the new k-vector $\mathbf{f}_1 \wedge \ldots \wedge \mathbf{f}_k$, one finds:

$$\mathbf{f}_1 \wedge \ldots \wedge \mathbf{f}_k = (\det A)\, \mathbf{e}_1 \wedge \ldots \wedge \mathbf{e}_k . \qquad \text{(XII.38)}$$

That is, the two k-vectors differ only by a non-zero scalar multiple.

We define the *Grassmanian manifold* $V_{k,n}(\mathbb{K})$ to be set of all k-planes through the origin in $\mathbb{K}^n$. (Previously, we encountered this concept in the more specific context of 2-planes in $\mathbb{R}^4$.) When $\mathbb{K} = \mathbb{C}$ ($\mathbb{R}$, resp.) it can be given the topology of a complex (real, resp.) differentiable manifold of dimension equal to kn. Indeed, the most direct way of representing the coordinates of chart is by choosing a basis $\mathbf{e}_a$, $a = 1, \ldots, n$ for $\mathbb{K}^n$, such as the canonical basis, and noting that since the members of any k-frame, such as $\mathbf{f}_i$, can be represented as linear combinations of the basis elements $\mathbf{e}_a$ as $\mathbf{f}_i = A_i^a \mathbf{e}_a$, the most direct way of associating the k-plane spanned by the $\mathbf{f}_i$ as a set of coordinates is to associate it with the matrix A_i^a.

The map $V_{k,n+1}(\mathbb{K}) \to \mathrm{P}\Lambda_k(\mathbb{K}^n)$, $\mathrm{span}\{\mathbf{f}_i\} \mapsto [\mathbf{f}_1 \wedge \ldots \wedge \mathbf{f}_k]$ is not only one-to-one, it is, in fact an embedding that one calls the *Plücker-Klein embedding* [6]. The aforementioned image of this embedding consists of all decomposable k-vectors. If one defines a basis $\mathbf{e}_i$ for $\mathbb{K}^n$ and expresses the decomposable k-vector $\mathbf{f}_1 \wedge \ldots \wedge \mathbf{f}_k$ in terms of the basis $\mathbf{e}_i$:

$$\mathbf{f}_1 \wedge \ldots \wedge \mathbf{f}_k = \frac{1}{k!} V^{i_1 \cdots i_k} \mathbf{e}_{i_1} \wedge \cdots \wedge \mathbf{e}_{i_k} \tag{XII.39}$$

then the components of the $V^{i_1 \cdots i_k}$ can be regarded as the homogeneous coordinates of the corresponding point in $\mathrm{P}\Lambda_k(\mathbb{K}^n)$; they are usually called the *Plücker coordinates* of the k-plane.

Dually, one can also represent k-planes through the origin of $\mathbb{K}^{n*}$, which collectively define a manifold that we denote by $V^*_{n-k,n}(\mathbb{K})$, by k-forms, up to a non-zero scalar multiple. Let V^*_k be such a k-plane and let θ^i, $i = 1, \ldots, k$ be a k-frame that spans it. The k-form $\theta^1 \wedge \ldots \theta^k$ is then unique up to a non-zero scalar factor, as before, and we also have an embedding of $V^*_{n-k,n}(\mathbb{K})$ in $\mathrm{P}\Lambda^k(\mathbb{K}^n)$ that takes $\mathrm{span}\{\theta^i\}$ to $[\theta^1 \wedge \ldots \theta^k]$.

By duality, since one has both the diffeomorphism of $V_{k,n}(\mathbb{K})$ with $V^*_{n-k,n}(\mathbb{K})$ and the projective equivalence of $\mathrm{P}\Lambda_k(\mathbb{K}^n)$ with $\mathrm{P}\Lambda^{n-k}(\mathbb{K}^n)$, one can either represent a k-plane through the origin of $\mathbb{K}^n$ as a projective equivalence class of either k-vectors or n–k-forms over the vector space $\mathbb{K}^n$.

b. Line geometry. In the case of 2-planes in $\mathbb{K}^n$, one sees that they will be represented by either decomposable 2-vectors on $\mathbb{K}^n$ or decomposable n–2-forms. Since a 2-plane in $\mathbb{K}^n$ projects to a line in $\mathbb{K}\mathrm{P}^{n-1}$, the study of 2-planes by their representation as 2-vectors is sometimes referred to as *line geometry* [**18**].

One can further characterize the decomposable 2-vectors by the fact that their *rank* equals two. The rank of a k-vector $\mathbf{A}$ on $\mathbb{K}^n$ can be defined, equivalently, as the minimum number of linearly independent vectors in that it takes $\mathbb{K}^n$ to express $\mathbf{A}$ as a linear combination of k-fold exterior products of vectors or as the minimum integer $r \geq 0$ such that $\mathbf{A} \wedge \ldots \wedge \mathbf{A} = 0$.

In the case of 2-vectors, the rank must be an even integer, and the set of all 2-vectors $\mathbf{A}$ of rank 2 is then characterized by the quadratic equation:

[6] In addition to the mathematical references on projective geometry one might also confer [**16**, **17**] for applications to physics.

$$\mathbf{A} \wedge \mathbf{A} = 0. \qquad \text{(XII.40)}$$

Hence, the set of all decomposable 2-vectors in the vector space $\Lambda_2(\mathbb{K}^n)$ is a quadric hypersurface that one calls the *Klein quadric.*

The exterior product is actually closely related to the question of incidence of projective subspaces, although, unfortunately, it does not give an isomorphic representation of the lattice of projective subspaces in the sense of representing the meet and join operations precisely. What one can say is that if a k-plane V_k in $\mathbb{K}^n$ is represented by a decomposable k-vector $\mathbf{A}$ and an m-plane V_m is represented by a decomposable m-vector $\mathbf{B}$ then $\mathbf{A} \wedge \mathbf{B} = 0$ iff the meet (i.e., intersection) of V_k and V_m has a dimension that is greater than zero. For instance, when 2-planes are represented by 2-vectors the vanishing of their exterior product is equivalent to the statement that the 2-planes intersect in a line. Hence, when regarded as projective subspaces in $\mathbb{K}\mathrm{P}^{n-1}$, one is describing two lines that intersect at a point.

Since the (exterior) polynomial that (XII.40) defines is homogeneous of degree two, it also defines a quadric hypersurface in the associated projective space $\mathrm{P}\Lambda_2(\mathbb{K}^n)$.

4. Projective geometry and electromagnetism. As we have seen from the outset, the mathematics of electromagnetism, in its most general formulation, involves bivector fields and 2-forms. In the previous section, we pointed out that the exterior algebra over any vector space has a close relationship with the lattice of linear subspaces of that vectors space, as well as the lattice of projective subspaces of its associated projective space. Therefore, in this section we shall apply the one result to the other to establish the extent to which projective geometry relates to pre-metric electromagnetism as metric geometry relates to gravitation.

When $n = 4$ and the field of scalars is $\mathbb{R}$, the basic vector space is $\mathbb{R}^4$, and its associated projective space is $\mathbb{R}\mathrm{P}^3$. The only linear subspaces of $\mathbb{R}^4$ that one needs to consider are lines, planes, and hyperplanes. These, in turn, correspond to points, lines, and planes in $\mathbb{R}\mathrm{P}^3$. We have already seen that representation of lines and hyperplanes is natural to $\Lambda_1(\mathbb{R}^4)$ and $\Lambda^1(\mathbb{R}^4)$, respectively, so we see that the only remaining k-vectors and k-forms that produce projective subspaces are bivectors and 2-forms, which represent 2-planes in $\mathbb{R}^4$ or lines in $\mathbb{R}\mathrm{P}^3$, in the decomposable case.

Hence, in order to be sure that the projective geometry – in fact, *line* geometry – of spacetime is rooted in the physics of electromagnetism, one must first gain an intuition for the physical nature of electromagnetic fields that can be described by a decomposable 2-form F and a decomposable bivector $\mathfrak{h}$. What one finds is that they generally represent the "elementary" fields, while the electromagnetic fields that are described by 2-forms and bivector fields of rank four are more elaborate linear superpositions of elementary

fields. For instance, any static electric field is described by a 2-form $F = dt \wedge E$ and a bivector field $\mathfrak{h} = \partial_t \wedge \mathbf{D}$, any static magnetic field is represented by a 2-form $F = \#_s\mathbf{B}$ and a bivector field $\mathfrak{h} = \#_s^{-1} H$, and an electromagnetic wave field has $F = k \wedge u$, $\mathfrak{h} = \mathbf{k} \wedge \mathbf{u}$.

The best way to distinguish between static electric, static magnetic, and wavelike 2-forms is given by the constitutive law κ: $\Lambda^2 M \to \Lambda_2 M$, $F \mapsto \mathfrak{h} = \kappa(F)$, at least in the linear case. In that case, one can define a bilinear pairing on $\Lambda^2 M$ by means of:

$$(F, G) \equiv \kappa(F)(G) = \tfrac{1}{4}\kappa^{\kappa\lambda\mu\nu} F_{\kappa\lambda} G_{\mu\nu}. \qquad \text{(XII.41)}$$

In general, this pairing is non-degenerate, but not symmetric, so in order to obtain a symmetric, non-degenerate, bilinear pairing – i.e., a scalar product – one must either restrict κ to its symmetric part or assume that the constitutive law has the property of symmetry to begin with. We shall call a κ with the property:

$$(F, G) = (G, F) \qquad \text{(XII.42)}$$

self-adjoint.

If one wishes to consider only the quadratic form (F, F) then it is unnecessary to make such a restriction, since only the symmetric part of κ will be involved; i.e., the skewon part plays no role.

Furthermore, as we pointed out in the discussion of constitutive laws, there is another non-degenerate bilinear pairing on $\Lambda^2 M$ that is defined by any volume element $\varepsilon \in \Lambda_4 M$, namely:

$$\langle F, G\rangle \equiv \#(F)(G) = \tfrac{1}{4}\varepsilon^{\kappa\lambda\mu\nu} F_{\kappa\lambda} G_{\mu\nu}. \qquad \text{(XII.43)}$$

As we have seen, the following are equivalent:

1. $\langle F, F\rangle = 0$,
2. F is decomposable,
3. F has rank two,
4. F lies on the Klein quadric.

This pairing is always symmetric, and therefore defines a scalar product on $\Lambda^2 M$. Hence, in order to properly separate the effects of κ from those of #, one must use only the principal part of κ in the definition (XII.41). We shall assume that this is the case from now on.

One sees that scalar products can just as well be defined on $\Lambda_2 M$ using the inverse isomorphisms to # and κ:

$$\langle \mathbf{A}, \mathbf{B}\rangle = \#^{-1}(\mathbf{A})(\mathbf{B}) = \tfrac{1}{4}\varepsilon_{\kappa\lambda\mu\nu} A^{\kappa\lambda} B^{\mu\nu}. \qquad \text{(XII.44a)}$$

$$(\mathbf{A}, \mathbf{B}) = \kappa^{-1}(\mathbf{A})(\mathbf{B}) = \tfrac{1}{4}\kappa_{\kappa\lambda\mu\nu} A^{\kappa\lambda} B^{\mu\nu}. \qquad \text{(XII.44b)}$$

One can also express the scalar products $\langle F, G\rangle$ and (F, G) as the Poincaré duals of the 4-forms $F \wedge G$ and $F \wedge \#\kappa(G)$:

$$\langle F, G\rangle\varepsilon = F \wedge G, \qquad (F, G)\varepsilon = F \wedge \#\kappa(G). \qquad \text{(XII.45)}$$

One can now distinguish three distinct type of 2-forms F, according to the quadratic form (F,F):

1. Electric: $(F,F) > 0$,
2. Isotropic: $(F,F) = 0$,
3. Magnetic: $(F,F) < 0$.

The reason that we did not call the isotropic 2-forms "wavelike" is because it is conceivable that an isotropic 2-form might represent a superposition of static electric and magnetic fields.

Now suppose we have a measurer/observer $(\mathbf{t}, \tau)$ that defines a time-space splitting of $T(M)$ and T^*M, with corresponding splittings $\Lambda_2 = \Lambda_2^{\mathrm{Re}} \oplus \Lambda_2^{\mathrm{Im}}$, $\Lambda_2 = \Lambda^2_{\mathrm{Re}} \oplus \Lambda^2_{\mathrm{Im}}$. We see that when $F = \tau \wedge E + \#_s\mathbf{B}$, $\mathfrak{h} = \mathbf{t} \wedge \mathbf{D} + \#_s^{-1} H$ the quadratic forms take the form [7]:

$$\langle F, F\rangle = 2\mathbf{V}(\tau \wedge E, \#_s\mathbf{B}) = -2E_iB^i, \tag{XII.47a}$$

$$\langle \mathfrak{h}, \mathfrak{h}\rangle = 2V(\mathbf{t} \wedge \mathbf{D}, \#_s^{-1} H) = -2H_iD^i, \tag{XII.47b}$$

$$\begin{aligned}(F, F) &= \kappa(\tau \wedge E, dt \wedge E) + 2\kappa(\tau \wedge E, \#_s\mathbf{B}) + \kappa(\#_s\mathbf{B}, \#_s\mathbf{B})\\ &= \varepsilon(E, E) + 2\gamma(E, \mathbf{B}) - \mu^{-1}(\mathbf{B}, \mathbf{B}),\\ &= \varepsilon^{ij}E_iE_j + 2\gamma^i_jE_iB^j - \tilde{\mu}_{ij}B^iB^j,\end{aligned} \tag{XII.48c}$$

$$\begin{aligned}(\mathfrak{h}, \mathfrak{h}) &= \kappa^{-1}(\mathbf{t} \wedge \mathbf{D}, \partial_t \wedge \mathbf{D}) + 2\kappa^{-1}(\mathbf{t} \wedge \mathbf{D}, \#_s^{-1} H) - \kappa^{-1}(\#_s^{-1} H, \#_s^{-1} H)\\ &= \varepsilon^{-1}(\mathbf{D}, \mathbf{D}) + 2\gamma^{-1}(\mathbf{D}, H) - \mu(H, H),\\ &= \tilde{\varepsilon}_{ij}D^iD^j + 2\gamma^i_jH_iD^j - \mu^{ij}H_iH_j.\end{aligned} \tag{XII.48d}$$

In the isotropic case, the last two take the form:

$$(F, F) = \varepsilon_0 E^2 - 1/\mu_0 B^2, \qquad (\mathfrak{h}, \mathfrak{h}) = 1/\varepsilon_0 D^2 - \mu_0 H^2, \tag{XII.49}$$

which is the form that gets the most attention from theoreticians.

One sees that the quadratic forms (F, F) and $\langle F, F\rangle$ are proportional to the Lorentz-invariant field invariants $\mathcal{F}$ and $\mathcal{G}$, resp., that were introduced by Mie [**19**] and currently play such a crucial role in phenomenological Lagrangians, such as the Born-Infeld and Heisenberg-Euler ones.

In the rank-two case, we see that the 2-form F and the bivector field $\mathfrak{h}$ define 2-planes, which we denote by $[F]$ and $[\mathfrak{h}]$, in the tangent spaces of the spacetime manifold. The main issue in the eyes of projective geometry is how they intersect. Naively, they can intersect transversally ($[F] \wedge [\mathfrak{h}] = 0$), in a line ($[F] \wedge [\mathfrak{h}] = [\mathbf{k}]$), or they can be

[7] We now represent our volume elements by $V \in \Lambda^4$ and $\mathbf{V} \in \Lambda_4$ to avoid confusion with the electric part of κ;

concurrent ($[F] = [\mathfrak{h}]$). However, the last possibility is not realized, which follows from the fact that $\mathfrak{h} = \kappa(F)$. Hence, the ultimate question is whether or not:

$$F \wedge \#\mathfrak{h} = \kappa(F, F)\, V \tag{XII.50}$$

vanishes; i.e., whether the scalar product (F, F) vanishes. Its vanishing is, in turn, equivalent to the possibility that the intersection is a line.

We then see that the homogeneous quadratic equation on $\Lambda^2 M$:

$$(F, F) = 0 \tag{XII.51}$$

defines not only a second quartic hypersurface in each fiber, in addition to the Klein quadric, but also a hypersurface in each projectivized tangent space $PT_x(M)$, namely, the set of all points $[\mathbf{k}]$ that are defined by the intersections of the lines $[F]$ and $[\mathfrak{h}]$ in each projective tangent space.

Now, let us try to get a better physical intuition for the nature of the 2-planes $[F]$ and $[\mathfrak{h}]$ by examining the form that they take in various elementary cases.

In the electrostatic case, they are of the form $[\tau \wedge E]$ and $[\mathbf{t} \wedge \mathbf{D}]$ and they intersect transversally. Hence, if $\mathbf{D} = D_x\, \partial_x$, while $E = E_x\, dx$ then the plane of $[\tau \wedge E]$ is the yz-plane, while the plane of $[\mathbf{t} \wedge \mathbf{D}]$ is the tx-plane. When one intersects them with the spatial subspaces $\Sigma(M)$ of $T(M)$, for the relevant choice of measurer/observer, the plane of $[\mathbf{t} \wedge \mathbf{D}]$ becomes the line generated by $\mathbf{D}$, when it is non-zero, and the plane $[\tau \wedge E]$ is tangent to the equipotential surfaces for $E = d\phi$.

In the magnetostatic case, the situation is reversed. If $F = \#_s\mathbf{B}$, $\mathfrak{h} = \#_s^{-1} H$, where $\mathbf{B} = B_x\, \partial_x$ and $H = H_x\, dx$ then $\#_s\mathbf{B} = B_x\, dy \wedge dz$, $\#_s^{-1} H = H_x\, \partial_y \wedge \partial_z$, and the plane of $[\#_s\mathbf{B}]$ is the tx-plane while the plane of $[\#_s^{-1} H]$ is the yz-plane. The spatial intersections are then the line generated by $\mathbf{B}$, when it is non-zero, and the plane that is annihilated by the 1-form H, when it is non-zero.

If $\mathbf{B} = \delta_s\mathbf{A} = \#_s^{-1} d_s \#_s \mathbf{A}$, where $\mathbf{A} \in \Lambda^2(\Sigma)$ is the magnetic potential bivector field then $\mathbf{A}$ spans a plane in each tangent space when it is non-zero, but the resulting rank-two sub-bundle of $\Sigma(M)$ does not have to be integrable, as a differential system, since the issue at stake, from Frobenius, is whether the 3-form:

$$A \wedge d_s A = A \wedge B \tag{XII.52}$$

vanishes, in which we have set $A = \#_s\mathbf{A}$, $B = \#_s\mathbf{B}$, which are then a 1-form and a 2-form, respectively. We see that this 3-form is of Chern-Simons type, if one regards A as a $\mathfrak{u}(1)$ connection form.

Of course, A is defined only up to the addition of a closed 1-form α, so $A \wedge B$ is defined only up to the addition of a 3-form $\alpha \wedge B$. Hence, the question of whether the rank-two sub-bundle defined by $\mathbf{A}$ is integrable into magnetostatic equipotential surfaces is equivalent to the question of whether a suitable choice of gauge A will make $A \wedge B$

vanish. All uniform magnetic fields have this property, as one verifies in the example $A = -B_0 y\, dx + B_0 x\, dy$, $B = B_0\, dx \wedge dy$, where B_0 is a non-zero constant. An example of a non-integrable B is given by starting with a potential 1-form A of the form $A_x(y, z)\, dx + A_y(x, z)\, dy$, which then makes $A \wedge dA = (A_x A_{y,z} - A_y A_{x,z})\, dx \wedge dy \wedge dz$, which does not have to vanish.

When $F = k \wedge u$ and $\mathfrak{h} = \mathbf{k} \wedge \mathbf{u}$ are isotropic, one has:

$$0 = F \wedge \#\mathfrak{h} = k(\mathbf{k})u(\mathbf{u}) - k(\mathbf{u})u(\mathbf{k}) = -k(\mathbf{u})u(\mathbf{k}); \qquad \text{(XII.53)}$$

the vanishing of the first term follows from the dispersion law for k. Hence, either $k(\mathbf{u})$ vanishes or $u(\mathbf{k})$ vanishes.

As a consequence of (XII.53), the planes $[k \wedge u]$ and $[\mathbf{k} \wedge \mathbf{u}]$ intersect in a line $[\mathbf{l}]$, where one can express the vector $\mathbf{l}$ as $\alpha\mathbf{k} + \beta\mathbf{u}$ for appropriate scalars α, β. Since $\mathbf{l}$ is incident on the plane $[k \wedge u]$, one must have:

$$0 = i_{\mathbf{l}}(k \wedge u) = \beta k(\mathbf{u})\, u - [\alpha u(\mathbf{k}) + \beta u(\mathbf{u})]k \qquad \text{(XII.54)}$$

after one includes the dispersion law for $k(\mathbf{k})$. This gives the conditions:

$$0 = \beta k(\mathbf{u}) = \alpha u(\mathbf{k}) + \beta u(\mathbf{u}). \qquad \text{(XII.55)}$$

Now, since the vector field $\mathbf{u}$ is defined only up to the addition of a scalar multiple of $\mathbf{k}$, one can choose it to be any vector in $[\mathbf{k} \wedge \mathbf{u}]$ except $\mathbf{k}$. Similarly, the 1-form u can be replaced with any 1-form in the plane spanned by k and u, except k itself. However, these choices of gauge are not independent, since F and $\mathfrak{h}$ are connected by the constitutive law κ. We then choose $\mathbf{u}$ and u to make:

$$0 = k(\mathbf{u}) = u(\mathbf{k}), \qquad u(\mathbf{u}) \neq 1, \qquad \text{(XII.56)}$$

which then forces β to vanish. Hence, with this choice of "gauge" for $\mathbf{u}$ and u, one must have $[\mathbf{l}] = [\mathbf{k}]$. That is, the line of intersection of the planes $[\mathbf{k} \wedge \mathbf{u}]$ and $[k \wedge u]$ is the direction of motion for the electromagnetic wave fronts, when the field in question is wavelike.

Since the planes $[k \wedge u]$ and $[\mathbf{k} \wedge \mathbf{u}]$ are not identical, but only intersect in the line $[\mathbf{k}]$, their join is three-dimensional. In order to find a third linearly-independent vector field $\mathbf{v}$ and covector field v, we consider the bivector field $\#^{-1}F$, which is also decomposable, and defines a plane that intersects $[\mathfrak{h}]$ in a line since:

$$\#^{-1}F \wedge \mathfrak{h} = F(\mathfrak{h})\mathbf{V} = 0. \qquad \text{(XII.57)}$$

Since the line of intersection is also the line $[\mathbf{k}]$, one can express $\#^{-1}F$ and $\#\mathfrak{h}$ in the form:

$$\#^{-1}F = \mathbf{k} \wedge \mathbf{v}, \quad \#\mathfrak{h} = k \wedge v \qquad \text{(XII.58)}$$

for a suitable vector field **v** and covector field *v*.

The vector field **v** must be non-collinear with both **k** and **u**, and must span the three-dimensional space [**k** ∧ **u**] ∨ [**k** ∧ **v**]. Hence, **k** ∧ **u**, **k** ∧ **v**, and **u** ∧ **v** must all be non-vanishing, as well as **k** ∧ **u** ∧ **v**; one derives analogous statements for *k*, *u*, and *v*.

In order to see how the 3-frame {**k**, **u**, **v**} relates to the 3-coframe {*k*, *u*, *v*}, we note that the vanishing of the expressions $F \wedge F$, $\mathfrak{h} \wedge \mathfrak{h}$, $F \wedge \mathfrak{h}$, and $\#^{-1}F \wedge \mathfrak{h}$ implies that:

$$0 = k(\mathbf{v})\, u(\mathbf{k}) = k(\mathbf{u})\, v(\mathbf{k}) = k(\mathbf{u})\, u(\mathbf{k}) = k(\mathbf{v})\, v(\mathbf{k}), \tag{XII.59}$$

which have the solution:

$$0 = k(\mathbf{v}) = u(\mathbf{k}) = k(\mathbf{u}) = v(\mathbf{k}). \tag{XII.60}$$

Hence, although the 3-frame {**k**, **u**, **v**} and the 3-coframe{*k*, *u*, *v*} are not projectively reciprocal, since *k*(**k**) vanishes, due to the dispersion law, they are dual, in the three-dimensional sense that the plane [**k** ∧ **u**] is the intersection of the hyperplane [*v*] with the three-dimensional space that is spanned by {**k**, **u**, **v**}, with analogous statements for [**k** ∧ **v**] and [**u** ∧ **v**]. We illustrate this situation in Fig. 18.

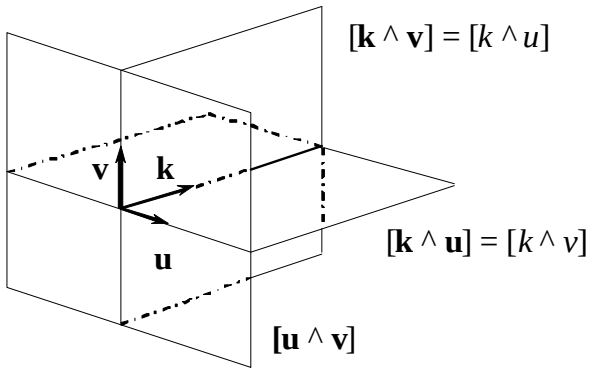

Figure. 18. The planes defined by an isotropic *F*.

Note that so far we have characterized [**u** ∧ **v**] as only the intersection of the hyperplane [*k*] with the three-dimensional space that is spanned by {**k**, **u**, **v**}, but not in terms of the Poincaré dual of a 2-form. This is where we must notice a geometric subtlety: whereas the planes [**k** ∧ **u**] and [**k** ∧ **v**] are uniquely defined by *F* and $\mathfrak{h}$, nevertheless, the choice of **u** and **v**, as well as the plane that they spanned, was an arbitrary gauge choice, except that neither could be collinear with **k**. Hence, the Poincaré dual to the plane [**u** ∧ **v**] also becomes open to a choice of gauge, in the form of a choice of measurer/observer, as defined by the pair (**t**, τ), where we choose $\tau(\mathbf{t}) = 1$. The pair (**t**, τ) allows us to decompose *k* into $\omega\tau - k_s$ and **k** into $\omega' \mathbf{t} + \mathbf{k}_s$, ($k(\mathbf{k}) = 0$) which then makes:

$$F = \omega\tau \wedge u - k_s \wedge u = \omega\tau \wedge u + \#(\omega' \mathbf{t} \wedge \mathbf{v}), \tag{XII.61a}$$

$$\mathfrak{h} = \omega' \mathbf{t} \wedge \mathbf{u} + \mathbf{k}_s \wedge \mathbf{u} = \omega' \mathbf{t} \wedge \mathbf{u} + \#^{-1}(\omega\tau \wedge v). \tag{XII.61b}$$

The "electric" parts of the *F* and $\mathfrak{h}$ are then proportional to *u* and **u**, respectively:

$$E = i_{\mathbf{t}} F = \omega u, \qquad \mathbf{D} = i_{\tau} \mathfrak{h} = \omega' \mathbf{u}, \tag{XII.62}$$

and we can consistently define **v** and *v* by their "magnetic" parts:

$$\mathbf{B} = i_{\tau}\#^{-1}F = -\omega'\mathbf{v}, \qquad H = i_{\mathbf{t}}\#\mathfrak{h} = \omega\nu. \tag{XII.63}$$

That is, **u** is in the direction of **D**, **v** is in the direction of **B**, *u* is in the direction of *E*, and *v* is in the direction of *H*, relative to this measurer/observer.

Since **u**, **v**, *u*, and *v* follow naturally when one chooses (**t**, τ), we see that the choice of a gauge for an isotropic electromagnetic field *F* is closely related to a choice of measurer/observer. We also see that the bivector **u** ^ **v** is proportional to the Poynting bivector field **D** ^ **B**, while the 2-form *u* ^ *v* is proportional to the Poynting 2-form *E* ^ *H*.

By contrast, the 2-planes [*k* ^ *u*] and [**k** ^ **u**] are defined independently of this choice of gauge or measurer/observer. One calls either the *polarization plane* for the electromagnetic wave in question.

References

1. V. P. Vizgin, *Unified Field Theories*, Birkhäuser, Boston, 1994.
2. A. Lichnerowicz, *Théorie relativiste de la gravitational et de l'electromagnetisme*, Masson and Co., Paris, 1955.
3. R. Feynman, *Feynman Lectures on Gravitation*, Westview Press, Boulder, 2003.
4. F. Klein, *Nicht-Euklidische Geometrie*, Chelsea, NY, 1927.
5. K. Kommerel, *Vorlesungen über Analytische Geometrie des Raumes*, K. F. Koehler Verlag, Leipzig, 1940.
6. J. G. Semple and G. T. Kneebone, *Algebraic Projective Geometry*, Clarendon Press, Oxford, 1952.
7. J. Verdina, *Projective Geometry and Point Transformations*, Allyn and Bacon, Boston, 1971.
8. D. Hilbert and S. Cohn-Vossen, *Geometry and the Imagination*, Chelsea, NY, 1999.
9. B. L. Van der Waerden, *Algebraische Geometrie*, Dover, New York, 1945.
10. G. Birkhoff, *Lattice Theory*, Amer. Math. Soc., Providence, 1940.
11. P. A. M. Dirac, *The Principles of Quantum Mechanics*, Oxford University Press, Oxford, 1982.
12. P. Gschwind, *Projektive Microphysik*, Goetheanum, Dornach, 2004.
13. D. H. Delphenich, "Projective Geometry and Special Relativity," Ann. Phys. (Leipzig) **15** (2006), 216-246.
14. D. H. Delphenich, "Complex geometry and pre-metric electromagnetism," LANL Archive gr-qc/0412048.
15. V. I. Arnol'd, "Contact Geometry and Wave Propagation," L'Enseignement Mathématique, 1989.
18. R. Penrose and W. Rindler, *Spinors and Spacetime, v. 1: two-spinor calculus and relativistic fields*, Cambridge University Press, Cambridge, 1984.
19. R. S. Ward and R. O. Wells, *Twistor Geometry and Field Theory*, Cambridge University Press, Cambridge, 1990.
18. V. Hlavaty, *Differential Line Geometry*, Noordhoff, Groenigen, 1953.
19. G. Mie, "Grundlagen einer Theorie der Materie," Ann. Phys. (Leipzig), **37** (1912), 511-534; *ibid.*, **39** (1912), 1-40; **40** (1913), 1-66.

CHAPTER XIII

COMPLEX RELATIVITY AND LINE GEOMETRY

Since the primary focus of this work has been to investigate the mathematical and physical considerations that strictly precede the introduction of a spacetime metric, the subject of the present chapter may seem out of place. However, since a recurring theme all along has been that is it is necessary to shift one's geometric intuition from the tangent bundle to the bundle of 2-forms on spacetime, it is important to see how the geometry that follows from the definition of a Lorentzian structure can be represented in terms of things that are more closely associated with the bundle of 2-forms.

The key to making the transition from $T(M)$ to $\Lambda^2(M)$ is in the isomorphism of the identity component of the Lorentz group with the Lie group $SO(3; \mathbb{C})$. It has the effect of saying that there is a one-to-one correspondence between oriented, time-oriented Lorentzian frames on Minkowski space and complex orthogonal frames in $\mathbb{C}^3$ that have unit volume. As long as one gives the real vector bundle $\Lambda^2(M)$ an almost-complex structure, which turns its fibers into three-dimensional complex vector spaces, any choice of complex frame in the fiber will define a complex-linear isomorphism of the fiber with $\mathbb{C}^3$. If one introduces a complex orthogonal structure on the fiber, which follows naturally from the introduction of a complex structure, then any complex orthogonal frame in a fiber defines an isometry of the fiber with $\mathbb{C}^3$, when it is given the complex Euclidian inner product. If one further introduces a complex volume element on the bundle $\Lambda^2(M)$ – which is not to be confused with a real volume element on $T(M)$ – then one can represent the Lie group $SO(3; \mathbb{C})$ by its action on complex orthogonal frames in $\Lambda^2(M)$ that have unit volume.

One can associate a principal fiber bundle with any vector bundle, and its elements are the linear frames in the fibers of that vector bundle. When the vector bundle is the tangent bundle $T(M)$ to a manifold, the associated bundle is the bundle $GL(M)$ of linear frames. One can reduce this bundle to various sub-bundles whose structure groups G are subgroups of $GL(n)$, and are called "G-structures," in general. For instance, unit-volume frames correspond to $G = SL(n)$, orthonormal frames, to $G = O(p, q)$, if the signature type of the metric is (p, q), and a global frame field would correspond to a complete reduction to $G = \{e\}$.

If one starts with the associated principal bundle to $\Lambda^2(M)$, which we denote by $GL(\Lambda^2)$, and whose structure group is $GL(6; \mathbb{R})$, then one can carry out similar reductions of this bundle by examining the frames that correspond to the various subgroups of $GL(6; \mathbb{R})$. These subgroups include $GL(3; \mathbb{C})$ and $SO(3; \mathbb{C})$, which correspond to the complex linear frames in the fibers of $\Lambda^2(M)$ and the complex orthogonal ones with unit complex

volume, respectively. It is in this latter reduction that we find the representation of Lorentzian relativity as something that also lives in the structure of the bundle $\Lambda^2(M)$.

In section 1 we first discuss the elementary concepts that we shall be applying to the context of vector bundles that first appear in complex linear algebra. Then, in section 2 we show how this relates to the representations of the Lorentz group by complex orthogonal transformations with unit determinant.

Since the literature of general relativity only occasionally presents the differential geometry that it uses in the language of connection 1-forms on the bundle of linear frames over spacetime, in section 3 we provide a section of this chapter that serves that purpose. The transition to expressing the same concepts in terms of the bundle of complex linear frames in $\Lambda^2(M)$ then becomes more natural.

Finally, in section 4 we bring the various elementary pieces together into the representation of general relativity in terms of the geometry of $\Lambda^2(M)$. Along the way, we point out that when this bundle has been given an almost-complex structure it is actually unnecessary to further complexify it, as is commonly done in complex relativity, and upon closer inspection, one sees that one is using only half of the resulting six-complex-dimensional vector space.

Once one has recast the geometry of gravitation as a sub-geometry of the geometry of electromagnetism, it is only natural to ponder the question of whether there is as much ultimate physical significance to better understanding the geometry of electromagnetism as their seems to be in the geometry of gravitation. Since pre-metric electromagnetism seems to be most relevant to the realm of strong electromagnetic fields, this also suggests possible inroads into the realm of quantum electrodynamics, which has always been obscured from geometry by its phenomenological formalism that deals with interactions and scattering amplitudes more that it deals with electromagnetic fields or spacetime structures directly.

1. Complex structures on real vector spaces [1, 2]. Although sometimes it is simpler to merely deal with complex scalars from the outset and consider complex vector spaces as being modeled on $\mathbb{C}^n$, nevertheless, it is sometimes more illuminating to regard the field $\mathbb{C}$ of complex numbers as an algebra over the real vector space $\mathbb{R}^2$ and then treat complex vector spaces as real vector spaces that have been given a "complex structure," when suitably defined. In order to motivate the concept of a complex structure on a real vector space, we first show how one exhibits $\mathbb{C}$ as an algebra over $\mathbb{R}^2$.

Recall that a $\mathbb{K}$-*algebra* over a vector space V (for some specified field of scalars $\mathbb{K}$) is a bilinear map $V \times V \to V$, $(a, b) \mapsto ab$; that is, it is a binary operation that respects the linear structure on each factor. Hence, since V already has an Abelian group structure defined by vector addition, and bilinearity amounts to the statement that the algebra multiplication (left and right) distributes over addition:

$$a(b + c) = ab + ac, \qquad (a + b)c = ac + bc, \tag{XIII.1}$$

we see that any algebra can also be regarded as a ring in the eyes of abstract algebra.

As a ring, an algebra does not need to be multiplicatively commutative, or even associative. Lie algebras are the primary example of a non-associative algebra, as far as physics is concerned, but the Cayley algebra over $\mathbb{R}^8$ has this property, as well. Similarly, an algebra does not need to have a unity – i.e., a multiplicative identity – and elements of an algebra do not need to possess multiplicative inverses. In the event that they do, they are called *units*, and if all non-zero elements of an algebra are units then one calls it a *division algebra*. In the extreme case where the multiplicative structure defines a group structure on the non-zero elements of the algebra, one calls it a *field*.

When the field is $\mathbb{R}$ and the vector space is $\mathbb{R}^2$, we can define a bilinear multiplication by imitating the effect of complex multiplication. Since:

$$(x + iy)(u + iv) = (xu - yv) + i(xv + yu) \qquad \text{(XIII.2)}$$

we define the product of two elements $(x, y), (u, v) \in \mathbb{R}^2$ to be:

$$(x, y)(u, v) = (xu - yv, xv + yu). \qquad \text{(XIII.3)}$$

As an $\mathbb{R}$-algebra, the complex numbers represent a commutative field. The multiplicative identity is $1 = (1, 0)$, and if $z = a + ib \neq 0$ then its inverse is $z^{-1} = 1/z = \overline{z} / \| z \|^2$, in which we have introduced the complex conjugate $\overline{z}$ of z and its modulus $\| z \|$ in the usual fashion. In its real representation, this says:

$$(a, b)^{-1} = \frac{1}{a^2 + b^2}(a, -b)\,. \qquad \text{(XIII.4)}$$

Hence, if we regard $\mathbb{C}$ as a *real* vector space (by restricting the scalars to real numbers) then the map $\mathbb{C} \to \mathbb{R}^2$, $x + iy \mapsto (x, y)$ is not only a linear isomorphism of real vector spaces, which we describe by saying that the map is an $\mathbb{R}$-linear isomorphism, but, from (XIII.2), it also an isomorphism of $\mathbb{R}$-algebras.

Since the algebra product is bilinear when one fixes one of its factors – say the left factor a – the remaining map L_a: $V \to V$, $b \mapsto ab$ is $\mathbb{K}$-linear. One calls this map *left multiplication* by a; one can define *right multiplication* by a analogously. In the event that the algebra product is commutative, it is unnecessary to distinguish between left and right multiplication. From (XIII.3), if $a = x + iy$ then the matrix of left (or right) multiplication by a, relative to the canonical basis on $\mathbb{R}^2$, is:

$$[L_a] = \begin{bmatrix} x & -y \\ y & x \end{bmatrix} = xI + yJ, \tag{XIII.5}$$

in which I is the 2×2 identity matrix and:

$$J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{XIII.6}$$

represents multiplication by the imaginary i.

One sees that:

$$\det L_a = x^2 + y^2 = \| a \|^2. \tag{XIII.7}$$

Hence, as long as $a \neq 0$, L_a will be invertible. We pause to note how the Euclidian quadratic form on $\mathbb{R}^2$ has been replaced by a determinant on a matrix algebra.

Since the product $L_a L_b$ of the matrices associated with $a, b \in \mathbb{C}$ is L_{ab}, we have a homomorphism L_a: $\mathbb{C}^* \to GL(2; \mathbb{R})$ of the group $(\mathbb{C}^*, \times)$ of non-zero complex numbers under multiplication in the group of invertible real 2×2 matrices under matrix multiplication. As the only element in $\mathbb{C}^*$ that goes to the identity element I is 1, the homomorphism is injective and we can call L_a a faithful representation of the group $(\mathbb{C}^*, \times)$ in $GL(2; \mathbb{R})$. Since not all matrices in $GL(2; \mathbb{R})$ can be represented in the form (XIII.5), the representation is not an isomorphism, except onto its image, which is a two-dimensional Abelian Lie subgroup of the four-dimensional non-Abelian Lie group $GL(2; \mathbb{R})$.

Now, one can generalize (XIII.5) to define the action of the *real* algebra $\mathbb{C}$ on $V \times V$ for any $\mathbb{K}$-vector space [1] V by defining L_a: $V \times V \to V \times V$ to have a matrix of the form:

$$[L_a] = \left[\begin{array}{c|c} xI & -yI \\ \hline yI & xI \end{array}\right] = xI + yJ, \tag{XIII.8}$$

in which I now represents the $2n \times 2n$ identity matrix and:

$$J = \left[\begin{array}{c|c} 0 & -I \\ \hline I & 0 \end{array}\right] \tag{XIII.9}$$

[1] We are tacitly assuming that $\mathbb{R}$ is a sub-field of $\mathbb{K}$ and the injection is by way of $a \mapsto a1$, where $1 \in \mathbb{K}$ is the unity element; as long as the only other field of interest to us is $\mathbb{K} = \mathbb{C}$, this is no problem.

has replaced the multiplication by i. In particular, J: $V\times V \to V\times V$ is a linear isomorphism such that:

$$J^2 = -I. \tag{XIII.10}$$

We shall call the vector space $V\times V$, together with a J that has this property a *complexification* of V. One can define such a structure for any $V\times V$ by making $J(v, w) = (-w, v)$.

By contrast, when J: $V \to V$ is an $\mathbb{R}$-linear isomorphism such that (XIII.10) is valid one says that J defines a *complex structure* on V. However, whereas any real vector space can be complexified, not every real vector space admits a complex structure. In particular, in order such a J will exist iff dim V is even. Clearly, $V\times V$ will always admit a complex structure

When a real vector space V has a complex structure J one can define complex scalar multiplication on V by:

$$(a + ib)v = av + bJv. \tag{XIII.11}$$

If V has dimension $2n$ as a real vector space then a basis $\{\mathbf{e}_i, i = 1, \ldots, 2n\}$ for V is called a *complex basis* iff it is compatible with J:

$$\mathbf{e}_{i+n} = J\mathbf{e}_i, \qquad i = 1, \ldots, n. \tag{XIII.12}$$

Hence, if $\mathbf{v} = v^i\, \mathbf{e}_i + v^{i+n}\, \mathbf{e}_{i+n}$ with real components v^i, v^{i+n} we can also say that:

$$\mathbf{v} = (v^i + iv^{i+n})\mathbf{e}_i \tag{XIII.13}$$

in which the components $v^i + iv^{i+n}$ are now complex numbers.

Hence, whereas the set $\{\mathbf{e}_i, i = 1, \ldots, 2n\}$defines an $\mathbb{R}$-linear isomorphism $V \to \mathbb{R}^{2n}$, as long as this set satisfies (XIII.12), the set $\{\mathbf{e}_i, i = 1, \ldots, n\}$ defines a $\mathbb{C}$-linear isomorphism $V \to \mathbb{C}^n$. This justifies our terminology in calling the set a complex basis on V.

Another way of defining the *complexification* of a real vector space is by defining the complex scalar multiplication on elements of $V^{\mathbb{C}} = V\otimes\mathbb{C}$, which one regards as the *real* vector space of $\mathbb{R}$-linear maps of V into the *real* vector space $\mathbb{C} = \mathbb{R}^2$; strictly speaking, we should probably use $V^*\otimes\mathbb{C}$ for this purpose, but it seems to be convention that one always sees the present notation.

One then defines complex scalar multiplication on the elements of $V^{\mathbb{C}}$ by:

$$\begin{aligned}(a + ib)\, v\otimes z &= v\otimes(a + ib)z = v\otimes az + v\otimes ibz \\ &= \alpha v\otimes z + \beta v\otimes iz.\end{aligned} \tag{XIII.14}$$

We can then see that the complex structure on $V\otimes\mathbb{C}$ also comes from:

$$J(v\otimes z) = v\otimes iz\ . \tag{XIII.15}$$

If V has real dimension n then $V^{\mathbb{C}}$ has complex dimension n and real dimension $2n$.

One can define an $\mathbb{R}$-linear isomorphism of $V\otimes\mathbb{C}$ with $V\times V$ that takes $v\otimes(a + ib)$ to (av, bv). Since, from (XIII.16), $J(v\otimes(a + ib)) = v\otimes(-b + ia)$ one can then define $J(v, w) = (-w, v)$, as before, and see that the two ways of complexifying a real vector space are equivalent.

Some elementary examples of complexification are given by $\mathbb{R}^{\mathbb{C}} = \mathbb{C}$ and $(\mathbb{R}^n)^{\mathbb{C}} = \mathbb{C}^n$. A somewhat more confusing example is given by $\mathbb{C}^{\mathbb{C}} = \mathbb{C}^2$, which only makes sense if one regards $\mathbb{C}$ as a *real* vector space of real dimension 2 and $\mathbb{C}^2$ as a complex vector space of complex dimension two. The element $(x + iy)\otimes(u + iv)$ in $\mathbb{C}^{\mathbb{C}}$ goes to $(x + iy, u + iv)$ in $\mathbb{C}^2$ or (x, y, u, v) in $\mathbb{R}^4$. Note that the complex vector space of $\mathbb{C}$-linear maps from $\mathbb{C}$ to itself has complex dimension one, which means real dimension two; hence, the possible confusion.

Elements of $\mathbb{C}^{\mathbb{C}}$ can also be represented by 2×2 real matrices, and, in fact, the representation is an $\mathbb{R}$-linear isomorphism that takes the basis elements $1\otimes1$, $i\otimes1$, $1\otimes i$, $i\otimes i$ to the basis elements:

$$\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}, \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}, \begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix}, \begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix},$$

resp.

2. Complex orthogonal representations of the Lorentz group [3]. When it comes to alternative ways of representing the Lorentz group, most of the attention to date has been focused on the two-to-one homomorphism from $SL(2;\mathbb{C})$ to $SO_0(3, 1)$, which, you will recall, denotes the identity component of the Lorentz; viz., the proper orthochronous Lorentz group. Most of the justification for the popularity of that representation comes from its use in the representation of the relativistic quantum-mechanical wave equation, namely, the Dirac equation.

In the previous chapter, we discussed the projective-geometric significance of the group $SL(2;\mathbb{C})$ as the group of projective transformations of $\mathbb{C}P^1$, which is diffeomorphic to the two-sphere as a real manifold. In this section, we shall discuss another isomorphic

representation of $SO_0(3, 1)$ that is quite naturally suggested by the representation of 2-planes in $\mathbb{R}^4$ by decomposable bivectors and 2-forms.

a. Isomorphism of $SO_0(3, 1)$ *and* $SO(3; \mathbb{C})$. In addition to the isomorphism $SO_0(3, 1)$ and $SL(2; \mathbb{C})$ there is also an isomorphism of Lie groups between $SO_0(3, 1)$, $SL(2; \mathbb{C})$, and $SO(3; \mathbb{C})$. The latter group is the subgroup of complex invertible matrices in $GL(3; \mathbb{C})$ that not only preserve a volume element but also the complex Euclidian scalar product:

$$<\alpha, \beta> = \delta_{ij}\, \alpha^i\, \beta^j, \qquad (\alpha, \beta \in \mathbb{C}^3,\ \alpha^i, \beta^j \in \mathbb{C}). \tag{XIII.16}$$

Hence, the only difference between the geometry of $\mathbb{C}^3$ with this metric and $\mathbb{R}^3$ with the same metric is in the field to which the scalars and vector components belong. In particular, a matrix $A \in SO(3; C)$ satisfies:

$$\det A = 1, \qquad AA^{\mathrm{T}} = A^{\mathrm{T}}A = I. \tag{XIII.17}$$

One sees that the Lie algebra $\mathfrak{so}(3; \mathbb{C})$ of $SO(3; \mathbb{C})$ then consists of complex 3×3 matrices ω that satisfy:

$$\mathrm{Tr}\ \omega = 0, \qquad \omega^{\mathrm{T}} + \omega = 0; \tag{XIII.18}$$

although the scalars are complex, the first requirement follows automatically from the anti-symmetry that is implied by the second one. A matrix $\omega \in \mathfrak{so}(3; \mathbb{C})$ is then the infinitesimal generator of a one-parameter subgroup of complex rotations in $\mathbb{C}^3$; namely, $e^{t\omega}$.

It follows immediately that $\mathfrak{so}(3; \mathbb{C}) = \mathfrak{so}(3; \mathbb{R})\otimes\mathbb{C}$; i.e., $\mathfrak{so}(3; \mathbb{C})$ is the complexification of $\mathfrak{so}(3; \mathbb{R})$. Hence, if I_i, $i = 1, 2, 3$ is the basis for $\mathfrak{so}(3; \mathbb{R})$ that consists of the elementary anti-symmetric matrices $[I_i]_{jk} = \varepsilon_{ijk}$:

$$I_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad I_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \qquad I_3 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{XIII.19}$$

then these matrices also define a basis for $\mathfrak{so}(3; \mathbb{C})$. Note that these matrices also give the adjoint representation of the Lie algebras $\mathfrak{so}(3; \mathbb{R})$ and $\mathfrak{so}(3; \mathbb{C})$.

Since we defined $\mathfrak{so}(3; \mathbb{C})$ as a matrix Lie algebra – i.e., a sub-algebra of $\mathfrak{gl}(3; \mathbb{C})$ – we observe that the isomorphism of $\mathfrak{so}(3; \mathbb{R})$ with $\mathbb{R}^3$, when it is given the vector cross product, also complexifies to an isomorphism of $\mathfrak{so}(3; \mathbb{C})$ with $\mathbb{C}^3$, when it is given the same cross product. This is really somewhat amusing since one of the selling points of using the exterior product in place of the cross product is the fact that the exterior product generalizes to vector spaces of any dimension, while the cross product is only defined in the vector spaces of dimension three. However, we now see that the transition from non-relativistic physics to relativistic physics does not have to define the transition from a real three-dimensional Euclidian space to a real four-dimensional Minkowski space, but simply to a complex three-dimensional Euclidian space. In that sense, relativistic physics is the complexification of non-relativistic physics.

We then see that if $\mathbf{z} = z^i\, \mathbf{e}_i \in (\mathbb{C}^3, \times)$ then its representation by a complex 3×3 matrix in $\mathfrak{so}(3; \mathbb{C})$ is:

$$\mathrm{ad}(\mathbf{z}) = z^i\, \mathrm{ad}(\mathbf{e}_i) = z^i\, I_i \tag{XIII.20}$$

since:

$$\mathrm{ad}(\mathbf{z})(\mathbf{v}) = [\mathbf{z}, \mathbf{v}] = (\varepsilon_{ijk}\, z^j\, v^k)\, \mathbf{e}_i = \mathbf{z} \times \mathbf{v}. \tag{XIII.21}$$

This makes the components of the 3×3 complex matrix ad(**z**) equal to:

$$[\mathrm{ad}(\mathbf{z})]_{ij} = \varepsilon_{ijk}\, z^k. \tag{XIII.22}$$

The isomorphism of $SO_0(3, 1)$ with $SO(3; \mathbb{C})$ is easiest to see by first defining the isomorphism of the corresponding Lie algebras, which implies that $\mathfrak{so}(3; \mathbb{C})$ must be regarded as a real Lie algebra, not a complex one, and exponentiating them. The isomorphism of Lie algebras is then simplest to describe in terms of real bases for both.

We use the basis $\{J_i, K_i, i = 1, 2, 3\}$ for $\mathfrak{so}(3, 1)$, where the J_i are of the form:

$$J_i = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & I_i \end{array}\right], \quad i = 1, 2, 3, \tag{XIII.23}$$

and the K_i are the elementary infinitesimal boosts:

$$K_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad K_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad K_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}. \tag{XIII.24}$$

By direct computation, the structure constants of this Lie algebra are obtained from:

$$[J_i, J_j] = \varepsilon_{ijk} J_k\,, \qquad [J_i, K_j] = \varepsilon_{ijk} K_k\,, \qquad [K_i, K_j] = -\,\varepsilon_{ijk}\, J_k\,. \qquad \text{(XIII.25)}$$

Notice that the last of these commutation relations implies that the three-dimensional vector subspace of $\mathfrak{so}(3, 1)$ that is spanned by the infinitesimal boosts is not a Lie subalgebra. It indirectly leads to Thomas precession, since it says, in effect, that the composition of boosts along different axes will produce a rotation in addition to the combined boost. However, the first set does imply that the Lie algebra $\mathfrak{so}(3; \mathbb{R})$ is included in $\mathfrak{so}(3, 1)$ by the replacement of the I_i with the J_i.

By complexification, the I_i also define a complex basis for the complex Lie algebra $\mathfrak{so}(3; \mathbb{C})$. The commutation rules for this basis are then identical to those of so(3; $\mathbb{R}$):

$$[I_i, I_j] = \varepsilon_{ijk} I_k\,, \qquad \text{(XIII.26)}$$

The simplest way to expand this into a real basis is by way of $\{I_i, iI_i, i = 1, 2, 3\}$. From the $\mathbb{C}$-bilinearity of the Lie bracket, the commutation rules for this real basis are then:

$$[I_i, I_j] = \varepsilon_{ijk} I_k\,, \qquad [I_i, iI_j] = i\varepsilon_{ijk} I_k\,, \qquad [iI_i, iI_k] = -\,\varepsilon_{ijk} I_k\,, \qquad \text{(XIII.26)}$$

which are formally the same as (XIII.25). Hence, the $\mathbb{R}$-isomorphism of $\mathfrak{so}(3, 1)$ with $\mathfrak{so}(3; \mathbb{C})$ is then effected by the association of the J_i with the I_i and the K_i with the iI_i.

This amounts to regarding a boost as essentially a rotation through an imaginary angle. Of course, this is really the opposite of what one should think, as one sees in the elementary case of the isomorphism of $\mathbb{R} \oplus \mathfrak{so}(2; \mathbb{R})$ with $\mathbb{C}$ (= $\mathbb{R}^2$), when one represents the $\mathbb{R}$ summand by matrices of the form αK with:

$$K = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \qquad \text{(XIII.27)}$$

The isomorphism simply takes $\alpha K + \beta J$ to $\alpha + i\beta$. Under exponentiation, since the Lie algebra is Abelian, one has:

$$\exp(\alpha K + \beta J) = \exp(\alpha K)\exp(\beta J) = \begin{bmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{bmatrix} \begin{bmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{bmatrix} \qquad \text{(XIII.28)}$$

and:

$$e^{\alpha + i\beta} = e^{\alpha}\, e^{i\beta} = e^{\alpha}(\cos\beta + i \sin\beta). \qquad \text{(XIII.29)}$$

Hence, it is more precise to say that rotations are the imaginary entities.

The isomorphism of the Lie group $SO_0(3, 1)$ with $SO(3; \mathbb{C})$ is then obtained by exponentiation of the isomorphism of Lie algebras. The actual association of a 4×4 real matrix with a complex 3×3 matrix is more involved at the group level than it was at the algebra level, though. Although one first exponentiates the basis matrices to obtain generators of the group, the remaining elements of the group are obtained by taking matrix products, not matrix sums.

b. The representation of $SO(3, 1)$ *in* $\Lambda_2(\mathbb{R}^4)$ *or* $\Lambda^2(\mathbb{R}^4)$. Both $\Lambda_2(\mathbb{R}^4)$ and $\Lambda^2(\mathbb{R}^4)$ are real vector spaces that are isomorphic to $\mathbb{R}^6$ by a choice of frame or coframe, resp. Hence, the set of all real linear frames for either vector space is described by the elements of the Lie group $GL(6; \mathbb{R})$.

However, this set of frames ignores the fact that both vector spaces have an additional structure that is defined by the fact that both of them are exterior products of four-dimensional real vector spaces: $\Lambda_2(\mathbb{R}^4) = \mathbb{R}^4 \wedge \mathbb{R}^4$, $\Lambda^2(\mathbb{R}^4) = \mathbb{R}^{4*} \wedge \mathbb{R}^{4*}$. Hence, we can identify a subgroup of $GL(6; \mathbb{R})$ that amounts to the image of $GL(4; \mathbb{R})$ under the anti-symmetrized tensor product representation.

We describe this representation by starting with a 4-frame $\{\mathbf{e}_\mu\,,\ \mu = 0, \ldots, 3\}$ on $\mathbb{R}^4$ and associating it with the 6-frame $\{\mathbf{E}_I,\ I = 1, \ldots, 6\}$ on $\Lambda_2(\mathbb{R}^4)$ that we defined previously:

$$\mathbf{E}_i = \mathbf{e}_0 \wedge \mathbf{e}_i\,, \qquad \mathbf{E}_{i+3} = \tfrac{1}{2}\varepsilon_{ijk}\,\mathbf{e}_j \wedge \mathbf{e}_k\,. \tag{XIII.30}$$

When $\mathbf{e}_\mu$ is subjected to a linear frame change to $\overline{\mathbf{e}}_\mu = A^\nu_\mu \mathbf{e}_\nu$, the resulting transformation of $\mathbf{E}_I$ is obtained from expanding:

$$\overline{\mathbf{E}}_i = \overline{\mathbf{e}}_0 \wedge \mathbf{e}_i = A^\mu_0 A^\nu_i \mathbf{e}_\mu \wedge \mathbf{e}_\nu\,, \qquad \overline{\mathbf{E}}_{i+3} = \tfrac{1}{2}\varepsilon_{ijk}\,\overline{\mathbf{e}}_j \wedge \overline{\mathbf{e}}_k = \tfrac{1}{2}\varepsilon_{ijk} A^\mu_j A^\nu_k \mathbf{e}_\mu \wedge \mathbf{e}_\nu \tag{XIII.31}$$

and identifying submatrices.

If D: $GL(4; \mathbb{R}) \to GL(6; \mathbb{R})$, $A \mapsto D(A)$ is the resulting representation and the matrices of $GL(6; \mathbb{R})$ are represented in the block matrix form:

$$D(A) = \left[\begin{array}{c|c} D_{tt}(A)^i_j & D_{ts}(A)^i_j \\ \hline D_{st}(A)^i_j & D_{ss}(A)^i_j \end{array}\right] \tag{XIII.32}$$

then the submatrices take the form:

$$D_{tt}(A)^i_j = A^0_0 A^i_j - A^i_0 A^0_j\,, \qquad D_{ts}(A)^i_j = \varepsilon_{jkl} A^0_k A^i_l\,, \tag{XIII.33a}$$

$$D_{st}(A)^i_j = \varepsilon_{ikl} A^k_0 A^l_j, \qquad D_{ss}(A)^i_j = \tfrac{1}{2}\varepsilon_{ikl}\varepsilon_{jmn} A^k_m A^l_n. \qquad \text{(XIII.33b)}$$

It is instructive to examine the form that these matrices take for various subgroups of $GL(4;\mathbb{R})$.

For the homotheties, one will have $A^0_0 = \lambda \neq 0, A^0_j = A^i_0 = 0, A^i_j = \delta^i_j$, and the corresponding matrix for $D(A)$ will take the form:

$$\left[\begin{array}{c|c} \lambda\delta^i_j & 0 \\ \hline 0 & \delta^i_j \end{array}\right].$$

For the $GL(3;\mathbb{R})$ subgroup that is represented by $A^0_0 = 1, A^0_j = A^i_0 = 0, A^i_j \in GL(3;\mathbb{R})$, one has a matrix of the form:

$$\left[\begin{array}{c|c} A^i_j & 0 \\ \hline 0 & A^i_j \end{array}\right].$$

Naturally, any subgroup of $GL(3;\mathbb{R})$, such as $SO(3;\mathbb{R})$, will be represented in this form, as well.

For the translations of $\mathbb{R}^3$, with $A^0_0 = 1, A^0_j = 0, A^i_0 = a^i, A^i_j = \delta^i_j$ the matrix looks like:

$$\left[\begin{array}{c|c} \delta^i_j & 0 \\ \hline -\varepsilon_{ijk}a^k & \delta^i_j \end{array}\right].$$

For the inversions of $\mathbb{R}^3$ that are described by $A^0_0 = 1, A^0_j = b^i, A^i_0 = 0, A^i_j = \delta^i_j$, it looks like:

$$\left[\begin{array}{c|c} \delta^i_j & -\varepsilon_{ijk}b^k \\ \hline 0 & \delta^i_j \end{array}\right].$$

Of particular interest in relativistic electromagnetism is the representation of a boost along the x axis. If the 4×4 matrix takes the form:

$$A = \begin{bmatrix} \cosh\zeta & -\sinh\zeta & 0 & 0 \\ -\sinh\zeta & \cosh\zeta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \gamma & -\gamma v/c & 0 & 0 \\ -\gamma v/c & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad \text{(XIII.34)}$$

in which $\gamma = (1 - v^2/c^2)^{-1/2}$ is the Fitzgerald-Lorentz contraction factor, then its image under the exterior product representation is:

$$D(A) = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \cosh\zeta & 0 & 0 & 0 & \sinh\zeta \\ 0 & 0 & \cosh\zeta & 0 & -\sinh\zeta & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\sinh\zeta & 0 & \cosh\zeta & 0 \\ 0 & \sinh\zeta & 0 & 0 & 0 & \cosh\zeta \end{array}\right]. \qquad \text{(XIII.35)}$$

In order to transform the components b^I of a bivector $\mathbf{b} = b^I \mathbf{E}_I$, one must use the inverse matrix to $D(A)$, which is:

$$D(A^{-1}) = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \cosh\zeta & 0 & 0 & 0 & -\sinh\zeta \\ 0 & 0 & \cosh\zeta & 0 & \sinh\zeta & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \sinh\zeta & 0 & \cosh\zeta & 0 \\ 0 & -\sinh\zeta & 0 & 0 & 0 & \cosh\zeta \end{array}\right]. \qquad \text{(XIII.36)}$$

When one applies this matrix to $\mathfrak{h} = [D^i, H^i]^{\mathrm{T}}$ the new components are:

$$\bar{D}^1 = D^1, \qquad \bar{D}^i = \gamma[\mathbf{D} + 1/c(\mathbf{v} \times \mathbf{H})]^i, \qquad i = 2, 3, \qquad \text{(XIII.37a)}$$
$$\bar{H}^1 = H^1, \qquad \bar{H}^i = \gamma[-\mathbf{D} + 1/c(\mathbf{v} \times \mathbf{H})]^i, \qquad i = 2, 3, \qquad \text{(XIII.37b)}$$

which agrees with the usual formulae, as one might find in Jackson [**4**].

The representation of $GL(4; \mathbb{R})$ in $\Lambda^2(\mathbb{R}^4)$ is contragredient to the one in $\Lambda_2(\mathbb{R}^4)$ that we just described. That is, if $\mathbf{e}_\mu$ transforms by way of A^μ_ν then if θ^μ is the reciprocal coframe to $\mathbf{e}_\mu$, in order to preserve the relation $\theta^\mu(\mathbf{e}_\nu) = \delta^\mu_\nu$ the coframe θ^μ must transform by means of the inverse matrix $\tilde{A}^\mu_\nu$. Thus, $\theta^\mu \wedge \theta^\nu$ must transform as:

$$\bar{\theta}^\mu \wedge \bar{\theta}^\nu = \tilde{A}^\mu_\kappa \tilde{A}^\nu_\lambda \, \theta^\kappa \wedge \theta^\lambda . \qquad \text{(XIII.38)}$$

Since we are defining our 6-coframe for $\Lambda^2(\mathbb{R}^4)$ by means of:

$$E^i = \theta^0 \wedge \theta^i, \qquad E^{i+3} = \tfrac{1}{2} \varepsilon_{ijk} \, \theta^j \wedge \theta^k, \qquad \text{(XIII.39)}$$

this serves to define the representation, which is then D': $GL(4;\ \mathbb{R}) \rightarrow GL(6,\ \mathbb{R})$, $A \mapsto D(A^{-1})$. Hence, the matrix of $D'(A) = D(A^{-1})$ is similar to (XIII.32), except that the components of A are replaced by the components of A^{-1}. The transformation of the components of a 2-form $b = b_I E^I$ is then by way of $D'(A^{-1}) = D(A)$, whose matrix is then given by (XIII.32) directly.

The transformation of $F = [E_i, B_i]$ is then entirely analogous to (XIII.37a, b), except that one right-multiplies by $D(A)$, instead of left-multiplying by $D(A^{-1})$. However, the end result is a similar set of equations to (XIII.37a, b), with the symbols for the components of $\mathfrak{h}$ replaced with those of F.

c. The representation of SO(3; $\mathbb{C}$) *in* $\Lambda_2(\mathbb{R}^4)$ *or* $\Lambda^2(\mathbb{R}^4)$. Another structure that one can impose on the six-dimensional real vector space $\Lambda_2(\mathbb{R}^4)$ is a complex structure, since its dimension is even. This time, we denote the complex structure by an $\mathbb{R}$-linear isomorphism $*$: $\Lambda_2(\mathbb{R}^4) \rightarrow \Lambda_2(\mathbb{R}^4)$, $\mathbf{b} \mapsto *\mathbf{b}$, such that:

$$*^2 = -I. \tag{XIII.40}$$

We can then define complex scalar multiplication by the usual means:

$$(\alpha + i\beta)\mathbf{b} = \alpha\mathbf{b} + \beta *\mathbf{b}. \tag{XIII.41}$$

An $\mathbb{R}$-linear frame$\{\mathbf{E}_I\ ,\ I = 1\ ,\ \ldots.,\ 6\}$ is *complex* iff:

$$\mathbf{E}_{i+3} = *\mathbf{E}_i = i\mathbf{E}_i\ , \qquad i = 1,\ 2,\ 3. \tag{XIII.42}$$

This terminology is then consistent with (XIII.41) and one can regard $\{\mathbf{E}_i\ ,\ i = 1\ ,\ 2,\ 3\}$ as a complex 3-frame on the complex vector space $\Lambda_2(\mathbb{R}^4)$, so any $\mathbf{b} \in \Lambda_2(\mathbb{R}^4)$ can be expressed in the form:

$$\mathbf{b} = b^i\ \mathbf{E}_i = \alpha^i\ \mathbf{E}_i + \beta^i\ *\mathbf{E}_i\ , \qquad i = 1,\ 2,\ 3, \tag{XIII.43}$$

in which the components $b^i = \alpha^i + i\beta^i$ are generally complex numbers.

Not all $\mathbb{R}$-linear frames on $\Lambda_2(\mathbb{R}^4)$ will have the property (XIII.42) that makes them complex. This is simple enough to show if one considers the transformation from one complex 3-frame $\{\mathbf{E}_i\ ,\ i = 1\ ,\ 2,\ 3\}$ on $\Lambda_2(\mathbb{R}^4)$ to another one $\{\bar{\mathbf{E}}_i\ ,\ i = 1\ ,\ 2,\ 3\}$. Since the former set defines a complex frame, one must have a 3×3 complex matrix A^i_j of components that makes $\bar{\mathbf{E}}_i = A^j_i \mathbf{E}_j$. Since $\bar{\mathbf{E}}_i$ is another complex 3-frame, the matrix A^i_j will be invertible. Hence, the matrix A^i_j is an element of the Lie group $GL(3;\ \mathbb{C})$. One can

then see that the set of real 6-frames on $\Lambda_2(\mathbb{R}^4)$ is in one-to-one correspondence with the elements of $GL(6; \mathbb{R})$, while the set of complex 3-frames is one-to-one correspondence with the elements of $GL(3; \mathbb{C})$. As the real dimension of $GL(6; \mathbb{R})$ is 36 and that of $GL(3; \mathbb{C})$ is 18 (complex dimension = 9), it is clear that there are "more" real frames than complex ones.

In order for an $\mathbb{R}$-linear map A: $V \to V$ on a real vector space V that has been given a complex structure *: $V \to V$ to behave like a $\mathbb{C}$-linear map, it must commute with * in the same way that $A(i\mathbf{v}) = iA\mathbf{v}$ when $\mathbf{v}$ belongs to a complex vector space; i.e.:

$$A* = *A. \tag{XIII.44}$$

In fact, this is not only necessary, but sufficient for a real 6×6 matrix $A \in GL(6; \mathbb{R})$ to be associated with a complex 3×3 matrix $C \in GL(3; \mathbb{C})$. The association of the one with the other is quite straightforward: If $C^i_j = \alpha^i_j + i\beta^i_j$ is the matrix of C then the corresponding A has a matrix that is given in block form by:

$$A^I_J = \left[\begin{array}{c|c} \alpha^i_j & -\beta^i_j \\ \hline \beta^i_j & \alpha^i_j \end{array}\right] = \left[\begin{array}{c|c} \alpha^i_j & 0 \\ \hline 0 & \alpha^i_j \end{array}\right] + * \left[\begin{array}{c|c} \beta^i_j & 0 \\ \hline 0 & \beta^i_j \end{array}\right]. \tag{XIII.45}$$

with respect to a complex basis.

If one goes back to (XIII.5) then one sees that the way that one constructs a real 6×6 matrix that represents a complex 3×3 matrix is closely analogous to the way that one makes a real 2×2 matrix out of a complex number. In particular, the identity map and * take the block matrix form:

$$I = \left[\begin{array}{c|c} \delta^i_j & 0 \\ \hline 0 & \delta^i_j \end{array}\right], \qquad * = \left[\begin{array}{c|c} 0 & -\delta^i_j \\ \hline \delta^i_j & 0 \end{array}\right]. \tag{XIII.46}$$

If $\{\mathbf{E}_I, I = 1, \ldots, 6\}$ is a complex frame for $\Lambda_2(\mathbb{R}^4)$ then one can also decompose $\Lambda_2(\mathbb{R}^4)$ into a direct sum $\Lambda_2^{\mathrm{Re}} \oplus \Lambda_2^{\mathrm{Im}}$ of 3-dimensional subspaces that are spanned by the frames $\{\mathbf{E}_i, i = 1, 2, 3\}$ and $\{*\mathbf{E}_i, i = 1, 2, 3\}$, respectively, since they then have the property:

$$*\Lambda_2^{\mathrm{Re}} = \Lambda_2^{\mathrm{Im}}, \qquad *\Lambda_2^{\mathrm{Im}} = \Lambda_2^{\mathrm{Re}}.$$

More precisely, if $\mathbf{b}_{\mathrm{R}} \in \Lambda_2^{\mathrm{Re}}$ and $\mathbf{b}_{\mathrm{I}} \in \Lambda_2^{\mathrm{Im}}$ then $*\mathbf{b}_{\mathrm{R}} \in \Lambda_2^{\mathrm{Im}}$ and $*\mathbf{b}_{\mathrm{I}} \in \Lambda_2^{\mathrm{Re}}$.

This situation seems to justify our terminology of "Re" and "Im" since these subspaces correspond to real and imaginary subspaces under the complex structure.

However, it is important to understand that when one is given a complex structure * on a vector space V the decomposition of V into real and imaginary subspaces is not generally canonical, as it is when $V = \mathbb{R}^n$, so one must choose a real+imaginary splitting in the same way that one might choose a frame.

If $[\mathbf{t}] \oplus \Sigma$ is a time+space splitting of $\mathbb{R}^4$ then one has an induced splitting [2] of $\Lambda_2(\mathbb{R}^4)$ into $[\mathbf{t}] \wedge \Sigma \oplus \Lambda_2(\Sigma)$, in which the subspace $[\mathbf{t}] \wedge \Sigma$ is spanned by all elements of the form $\mathbf{t} \wedge \mathbf{v}$, where $\mathbf{v} \in \Sigma$ and the subspace $\Lambda_2(\Sigma)$ is spanned by all elements of the form $\mathbf{v} \wedge \mathbf{w}$ with $\mathbf{v}, \mathbf{w} \in \Sigma$. Whether the splitting is an actual real+imaginary decomposition will depend upon the choice of time+space splitting, since not all of them will induce a splitting with the property $*([\mathbf{t}] \wedge \Sigma) = \Lambda_2(\Sigma)$, $*(\Lambda_2(\Sigma)) = [\mathbf{t}] \wedge \Sigma$.

Previously, we decomposed an electromagnetic bivector field into an electric part and a magnetic part. Now, we can now see that this latter decomposition is closely related to imposing a complex structure on $\Lambda_2(\mathbb{R}^4)$. It is amusing to regard electric bivectors as real and magnetic ones as imaginary, since one knows that elementary magnetic fields are, in a sense, "fictitious" fields that appear as a result of the choice of rest space in much the same way that Coriolis and centripetal forces appear.

Conversely, it is not true that any real+imaginary splitting of $\Lambda_2(\mathbb{R}^4)$ will imply a corresponding time+space splitting of $\mathbb{R}^4$. The main issue is an algebraic one: in order for a three-dimensional subspace of $\Lambda_2(\mathbb{R}^4)$ to take the form $[\mathbf{t}] \wedge \Sigma$ all of the elements must have a common exterior factor that is a scalar multiple of some $\mathbf{t} \in \mathbb{R}^4$, which is not always the case. Moreover, it is not the case that if $\Lambda_2(\mathbb{R}^4)$ is decomposed into a pair of three-dimensional summands then at least one of then must be of the form $[\mathbf{t}] \wedge \Sigma$ for some $[\mathbf{t}] \in \mathbb{R}^4$.

As we have seen before, if $\mathbb{R}^4$ has been given a volume element $V \in \Lambda^4$ then one can define a scalar product on $\Lambda_2(\mathbb{R}^4)$ by way of $\langle \mathbf{A}, \mathbf{B} \rangle = V(\mathbf{A} \wedge \mathbf{B})$. Furthermore, if one also has a complex structure * then one can define a second scalar product by means of $(\mathbf{A}, \mathbf{B}) = \langle \mathbf{A}, *\mathbf{B} \rangle = V(\mathbf{A} \wedge *\mathbf{B})$, as long as * is self-adjoint; i.e., $\mathbf{A} \wedge *\mathbf{B} = *\mathbf{A} \wedge \mathbf{B}$ for all $\mathbf{A}, \mathbf{B} \in \Lambda_2(\mathbb{R}^4)$.

One can define a complex orthogonal structure on $\Lambda_2(\mathbb{R}^4)$ by means of:

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathbb{C}} = (\mathbf{A}, \mathbf{B}) + i\langle \mathbf{A}, \mathbf{B} \rangle. \qquad \text{(XIII.47)}$$

[2] The details of this statement are discussed in greater rigor in Delphenich [**1**].

Actually, one could also define such a structure by means of $\langle \mathbf{A}, \mathbf{B}\rangle + i(\mathbf{A}, \mathbf{B})$, but it is clear that this is equal to the complex conjugate of $-i\langle \mathbf{A}, \mathbf{B}\rangle_{\mathbb{C}}$, which is not fundamentally different. This amounts to the observation that complex orthogonal quadratic forms all have the Euclidian signature type.

One can then further reduce the set of complex 3-frames on $\Lambda_2(\mathbb{R}^4)$ to the set of complex orthogonal 3-frames, which then satisfy:

$$\langle \mathbf{E}_i, \mathbf{E}_j\rangle_{\mathbb{C}} = \delta_{ij} \,. \tag{XIII.48}$$

A 3×3 complex matrix A^i_j that relates a complex orthogonal 3-frame $\bar{\mathbf{E}}_i$ to another one $\mathbf{E}_i$ by way of $\bar{\mathbf{E}}_i = A^j_i \mathbf{E}_j$ will then be an element of $O(3; \mathbb{C})$. Hence, $\langle A\mathbf{a}, A\mathbf{b}\rangle_{\mathbb{C}} = \langle \mathbf{a}, \mathbf{b}\rangle_{\mathbb{C}}$ for any elements $\mathbf{a}, \mathbf{b} \in \Lambda_2(\mathbb{R}^4)$ and, as a result, $A^{-1} = A^{\mathrm{T}}$.

In order reduce this set of complex orthogonal frames further to a subset of complex orthogonal 3-frames with unit-volume one must define a volume element on the complex vector space $\Lambda_2(\mathbb{R}^4)$. Here, we must be careful not to confuse this with a volume element on $\mathbb{R}^4$, since a volume element on the three-dimensional complex vector space $\Lambda_2(\mathbb{R}^4)$ will be a non-zero complex 3-form on $\Lambda_2(\mathbb{R}^4)$ itself, not a non-zero 4-form on $\mathbb{R}^4$. Hence, if $\mathbf{E}_i$ is a complex 3-frame on $\Lambda_2(\mathbb{R}^4)$ and E^i is its reciprocal 3-frame on $\Lambda_2(\mathbb{R}^4)$ then one can define such a volume element by means of:

$$\mathcal{V} = E^1 \perp E^2 \perp E^3 = \frac{1}{3!}\, \varepsilon_{ijk}\, E^i \perp E^j \perp E^k, \tag{XIII.49}$$

in which the symbol $\perp$ is used to denote the exterior product in the exterior algebra over the complex vector space $\Lambda_2(\mathbb{R}^4)$ in order to minimize the confusion with the exterior algebra over $\mathbb{R}^4$. Since the complex dimension of the vector space $\Lambda_2(\mathbb{R}^4)$ is three, its exterior algebra will consist of complex scalars, vectors, bivectors, and trivectors. The volume of the parallelepiped spanned by a complex linear 3-frame $\mathbf{F}_i = A^j_i \mathbf{E}_j$ is then:

$$\mathcal{V}(\mathbf{F}_1 , \mathbf{F}_2 , \mathbf{F}_3) = \det A\, \mathcal{V}(\mathbf{E}_1 , \mathbf{E}_2 , \mathbf{E}_3) = \det A. \tag{XIII.50}$$

Hence, the transformation preserves the volume of the 3-frame iff $\det A = 1$.

This allows us to associate complex unit-volume 3-frames with elements of $SL(3; \mathbb{C})$ and complex unit-volume orthogonal 3-frames with elements of $SO(3; \mathbb{C})$. Hence, we have finally arrived at a representation of the proper orthochronous Lorentz group $SO_0(3;$

1) as an isomorphic subgroup of the group that is responsible for the transformations of frames in the vector space of bivectors on $\mathbb{R}^4$.

Since we made no mention of defining a Lorentzian structure on $\mathbb{R}^4$ that would induce the complex orthogonal structure on $\Lambda_2(\mathbb{R}^4)$ the question arises whether the complex orthogonal structure on $\Lambda_2(\mathbb{R}^4)$ will induce a Lorentzian structure on $\mathbb{R}^4$.

We have already seen that a complex orthogonal structure on $\Lambda_2(\mathbb{R}^4)$ will allow us to construct a light cone out of the lines of intersection of all isotropic decomposable bivectors and their duals under *, which then determines a conformal class of Lorentzian metrics. We now see that since $SO(3; \mathbb{C})$ is isomorphic to $SO_0(3, 1)$ every unit-volume complex orthogonal 3-frame on $\Lambda_2(\mathbb{R}^4)$ is associated with a unique time-oriented unit-volume Lorentzian frame on $\mathbb{R}^4$, as long as one has agreed on the way that 4-frames in $\mathbb{R}^4$ relate to complex 3-frames in $\Lambda_2(\mathbb{R}^4)$; the one that we have been using all along would suffice, as long as the 6-frame defined by $\mathbf{e}_\mu \wedge \mathbf{e}_\nu$ is complex under *. Hence, the isomorphism of $SO(3; \mathbb{C})$ and $SO_0(3, 1)$, together with one 4-frame $\mathbf{e}_\mu$ on $\mathbb{R}^4$ whose corresponding 6-frame $\mathbf{E}_I$ on $\Lambda_2(\mathbb{R}^4)$ is complex orthogonal will define the set of all time-oriented unit-volume Lorentzian frames on $\mathbb{R}^4$, which is equivalent to a choice of Lorentzian structure.

Of course, analogous constructions to the foregoing ones can be made for the vector space $\Lambda^2(\mathbb{R}^4)$ that is dual to $\Lambda_2(\mathbb{R}^4)$. One must use the transpose of * to define a complex structure on $\Lambda^2(\mathbb{R}^4)$ in order to make $*A(\mathbf{A}) = A(*\mathbf{A})$ for every $A \in \Lambda^2(\mathbb{R}^4)$ and $\mathbf{A} \in \Lambda_2(\mathbb{R}^4)$.

d. Hermitian structures on $\Lambda_2(\mathbb{R}^4)$ *and* $\Lambda^2(\mathbb{R}^4)$. Since we are starting with such a large group of frame transformations on $\Lambda_2(\mathbb{R}^4)$ – namely, $GL(6; \mathbb{R})$ – we have potentially more subgroups to consider than when we start with $GL(4; \mathbb{R})$, which pertains to frames in $\mathbb{R}^4$ itself. This is especially true when one puts a complex structure on either of the vector spaces $\Lambda_2(\mathbb{R}^4)$ or $\mathbb{R}^4$, which allows one to consider complex frames and the groups $GL(3; \mathbb{C})$ and $GL(2; \mathbb{C})$, respectively. Of course, the physical significance of such a complex structure is more established in the case of $\Lambda_2(\mathbb{R}^4)$ than it seems to be in the case of $\mathbb{R}^4$.

One of the intriguing consequences of introducing a complex structure * on $\Lambda_2(\mathbb{R}^4)$ is that in addition to the complex orthogonal structure that such an isomorphism defines one can also define a Hermitian structure, as long as one also has a choice of real+imaginary decomposition of $\Lambda_2(\mathbb{R}^4)$, as well. Furthermore, the physical significance of the construction is actually quite fundamental to electromagnetism.

The reason that one needs to choose a real+imaginary decomposition $\Lambda_2^{\mathrm{Re}} \oplus \Lambda_2^{\mathrm{Im}}$ of $\Lambda_2(\mathbb{R}^4)$ is because otherwise there is no unambiguous way to define the operator of complex conjugation. However, when one is given such a decomposition the definition is immediate. If $\mathbf{a} = \mathbf{a}_R + *\mathbf{a}_I$ with $\mathbf{a}_R, \mathbf{a}_I \in \Lambda_2^{\mathrm{Re}}$ then its complex conjugate relative to this splitting is:

$$\overline{\mathbf{a}} = \mathbf{a}_R - *\mathbf{a}_I , \tag{XIII.51}$$

We now examine what happens to the complex Euclidian structure when we include the effect of conjugation. We define:

$$(\mathbf{a}, \mathbf{b})_{\mathbb{C}} = <\mathbf{a}, \overline{\mathbf{b}}>_{\mathbb{C}} = (\mathbf{a}, \overline{\mathbf{b}}) + i <\mathbf{a}, \overline{\mathbf{b}}> . \tag{XIII.52}$$

Suppose $\mathbf{b}_i$, i = 1, 2, 3 is a complex orthonormal frame for $\Lambda_2(\mathbb{R}^4)$, so we assume:

$$<\mathbf{b}_i, \mathbf{b}_j>_{\mathbb{C}} = \delta_{ij}, \tag{XIII.53}$$

and define the bilinear form associated with the complex Euclidian structure to be:

$$\delta = \delta_{ij}\, b^i \otimes b^j, \tag{XIII.54}$$

where b^i, i = 1, 2, 3 is the reciprocal coframe to $\mathbf{b}_i$. As long as we understand the notation $\overline{b}^i$ to mean the composition of the $\mathbb{C}$-linear functional b^i with the complex conjugation operation on $\mathbb{C}$ then we should have:

$$(\mathbf{b}_i, \mathbf{b}_j)_{\mathbb{C}} = <\mathbf{b}_i, \overline{\mathbf{b}}_j>_{\mathbb{C}} = <\mathbf{b}_i, \mathbf{b}_j>_{\mathbb{C}} = \delta_{ij} , \tag{XIII.55}$$

since $\mathbf{b}_i$ is "real."

This allows us to set the bilinear form h that is associated with the new inner product equal to:

$$h = \delta_{ij} b^i \otimes \overline{b}^j , \tag{XIII.56}$$

in this frame and:

$$h = h_{ij} \beta^i \otimes \overline{\beta}^j , \tag{XIII.57}$$

for a general frame.

Hence, when $\mathbf{A} = A^i \mathbf{b}_i$ and $\mathbf{B} = B^j \mathbf{b}_j$ the value of the inner product is:

$$(\mathbf{A}, \mathbf{B})_{\mathbb{C}} = h(\mathbf{A}, \mathbf{B}) = \delta_{ij} A^i \overline{B}^j . \qquad \text{(XIII.58)}$$

This shows that what we have defined is indeed a Hermitian structure on the complex vector space $\Lambda_2(\mathbb{R}^4)$, and the $\mathbb{C}$-linear isomorphism of $\Lambda_2(\mathbb{R}^4)$ with $\mathbb{C}^3$ that is defined by a Hermitian frame becomes a unitary isomorphism when one gives $\mathbb{C}^3$ the usual Hermitian structure $\delta_{ij}\theta^i \otimes \overline{\theta}^j$ with θ^i being the canonical coframe.

Due to the assumption of the self-adjointness of *, when one computes $(\mathbf{a}, \mathbf{a})_{\mathbb{C}}$, one obtains:

$$(\mathbf{a}, \mathbf{a})_{\mathbb{C}} = [(\mathbf{a}_R, \mathbf{a}_R) + (\mathbf{a}_I, \mathbf{a}_I)] + i[\langle \mathbf{a}_R, \mathbf{a}_R \rangle + \langle \mathbf{a}_I, \mathbf{a}_I \rangle] = (\mathbf{a}_R, \mathbf{a}_R) + (\mathbf{a}_I, \mathbf{a}_I). \qquad \text{(XIII.59)}$$

The vanishing of the imaginary part follows from the fact that $\mathbf{a}_R$ and $\mathbf{a}_I$ belong to a three-dimensional (real) vector space, while $\langle \mathbf{a}_R, \mathbf{a}_R \rangle$ and $\langle \mathbf{a}_I, \mathbf{a}_I \rangle$ are both defined by 4-vectors, which must vanish identically, since $\mathbf{a}_R$, $\mathbf{a}_I$ belong to three-dimensional vector spaces.

If we apply this to the case of the electromagnetic excitation bivector $\mathfrak{h} = \mathbf{t} \wedge \mathbf{D} + *(\mathbf{t} \wedge \mathbf{H})$ then we see that:

$$\begin{aligned}(\mathfrak{h}, \mathfrak{h})_{\mathbb{C}} &= (\mathbf{D}, \mathbf{D}) + 2(\mathbf{D}, \mathbf{H}) + (\mathbf{H}, \mathbf{H}) \\ &= \varepsilon^{-1}(\mathbf{D}, \mathbf{D}) + 2\gamma^{-1}(\mathbf{D}, \mathbf{H}) + \mu(\mathbf{H}, \mathbf{H}). \qquad \text{(XIII.60)}\end{aligned}$$

In the case where the electromagnetic couplings (= γ) vanish this becomes:

$$(\mathfrak{h}, \mathfrak{h})_{\mathbb{C}} = \varepsilon^{-1}(\mathbf{D}, \mathbf{D}) + \mu\,(\mathbf{H}, \mathbf{H}) = E(\mathbf{D}) + H(\mathbf{B}), \qquad \text{(XIII.61)}$$

and in the isotropic case:

$$(\mathfrak{h}, \mathfrak{h})_{\mathbb{C}} = 1/\varepsilon D^2 + \mu H^2 = \varepsilon E^2 + 1/\mu B^2. \qquad \text{(XIII.62)}$$

Hence, we see that the Hermitian structure that we defined is intimately related to the energy density of the electromagnetic field. The fact that it can only be defined after one has made a choice of real+imaginary decomposition is entirely consistent with the fact that one cannot speak of energy in relativistic mechanics until one has made a choice of time+space splitting of spacetime.

As is well-known, quadratic expressions such as (XIII.62) are associated with the Hamiltonian of the simple harmonic oscillator in mechanics, and electromagnetic wave motion is envisioned to involve a continuous distribution of simple harmonic oscillators throughout space. Hence, the generalization defined by (XIII.60) shows that one can generalize the oscillators accordingly to three-dimensional anisotropic harmonic oscillators.

What is truly intriguing about the appearance of a Hermitian structure on the (complex) three-dimensional vector spaces of interest to electromagnetism is that this naturally allows one to reduce the group $GL(3;\mathbb{C})$ of complex linear frame transformations to the group $U(3)$ of unitary frame transformations, and further to $SU(3)$, if one requires that they preserve the volume element, as well. Ordinarily, the group $SU(3)$ does not begin to figure in physics until one is dealing with strong interactions, but now we see that it is clearly relevant to the energetics of electromagnetism, as well. Since the strong interaction was introduced into physics in order to account for the stability of most nuclei in the face of the mutual electrostatic repulsion of their constituent protons, this makes one wonder if there is some natural transition from the theory of electromagnetism into the theory of the strong interaction within the pre-metric formalism. After all, we have already seen that it affords a natural transition from electromagnetism into gravitation by way of the electromagnetic dispersion law for the medium.

For more musings along these lines, the reader is invited to peruse the author's paper [**5**]

3. General relativity in terms of Lorentzian frames. In order to facilitate the transition from conventional Lorentzian general relativity, which deals with Lorentzian frames in the tangent bundle, to the form that things take in the context of complex orthogonal frames in the bundle of 2-forms, we briefly summarize the formulation of general relativity using the formalism of connection 1-forms on the bundle of oriented, time-oriented, Lorentzian frames on spacetime. The approach to differential geometry that involves defining connection 1-forms on principal bundles is quite commonplace in gauge field theories, but apparently still not completely accepted in general relativity and gravitation theory [3].

a. Reductions of the bundle of linear frames. In order to define an oriented, time-oriented, Lorentzian frame in a tangent space T_xM to a point x in real four-dimensional spacetime manifold M, one will clearly need three things: a volume element $V \in \Lambda^4$, a time+space splitting of $T(M)$ into $[\mathbf{t}]\oplus\Sigma$, and a Lorentzian metric g.

The presence of a volume element allows one to reduce from the bundle $GL(M) \to M$ of linear frames in $T(M)$ to the bundle $GL^+(M) \to M$ of oriented frames to the bundle $SL(M) \to M$ of unit-volume frames. Of course, one must assume that M – or rather, $T(M)$ – is orientable to begin with, which we will.

One can also regard a volume element as a real-valued function V: $GL(M) \to \mathbb{R}$, $\mathbf{e}_\mu \mapsto V(\mathbf{e}_\mu) = V(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, where we are abusing the notation by using the same symbol for the function and the non-zero 4-form. It is, moreover, required to be equivariant under the right action of $GL(4;\mathbb{R})$ on $GL(M)$ and its action on $\mathbb{R}$ by way of the determinant. That is, when the frame $\mathbf{e}_\mu$ is changed to $\mathbf{e}_\nu A^\nu_\mu$ the numerical value of $V(\mathbf{e}_0,$

[3] Some of the references on general relativity that use this formalism to a greater or lesser degree are [**6-12**].

$\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$) is changed to $\det(A)V(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. One can then reconstruct the 4-form at $x \in M$ by means of:

$$V_x = \theta^0 \wedge \theta^1 \wedge \theta^2 \wedge \theta^2 = \frac{1}{4!}\,\varepsilon_{\kappa\lambda\mu\nu}\,\theta^\kappa \wedge \theta^\lambda \wedge \theta^\mu \wedge \theta^\nu. \qquad \text{(XIII.63)}$$

The actual reduction of $GL(M)$ to $GL^+(M)$ then follows from considering the subset of $GL(M)$ that consists of linear frames for which $V(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ is positive. Since each fiber of $GL(M)$ consists of two components – namely, $V(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) < 0$ and $V(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) > 0$ – one can see that the homogeneous space $\mathcal{O}_x = GL_xM\,/GL_x^+$ consists of two points, up to homotopy equivalence. We can then regard the bundle $\mathcal{O}(M) \to M$ as the *orientation bundle* over M, so an orientation on M is a global section of this bundle. Hence, the manifold $\mathcal{O}(M)$ is diffeomorphic to the orientable covering manifold of M.

The reduction from $GL^+(M)$ to $SL(M)$ follows from looking at the level hypersurface of unity for the function V. That is, one chooses a set of oriented frames at each point that one wishes to call *unit-volume* frames. Since the homogeneous space $GL_x^+/\,SL_x$ is homotopically equivalent to $\mathbb{R}$ (by way of the determinant), which is contractible, this reduction is never obstructed by homotopy.

The time+space splitting allows one to further reduce from $SL(M)$ to the bundle $SL_0(M) \to M$ of time-oriented unit-volume frames. Again, one must assume that the line bundle $[\mathbf{t}](M) \to M$ is orientable, which amounts to the existence of a *non-zero* vector field $\mathbf{t} \in \mathfrak{X}(M)$ that generates the line $[\mathbf{t}](x)$ at each $x \in M$. If M is compact then it is necessary that its Euler-Poincaré characteristic vanish in order for this to be possible. However, requiring that a given line field admit a non-zero section is a stronger condition than requiring that the tangent bundle admit a non-zero section, so it is not a sufficient condition in this case.

The equivariant maps on $SL(M)$ and its dual $SL^*(M)$ – which is then the bundle of unit-volume coframes on M – that define the time+space splitting of $T(M)$ are then the maps $\mathbf{t}$: $SL(M) \to \mathbb{R}^4$, $\mathbf{e}_\mu \mapsto \mathbf{t}(\mathbf{e}_\mu) = t^\mu$ and τ: $SL^*(M) \to \mathbb{R}^{4*}$, $\theta^\mu \mapsto \tau(\theta^\mu) = \tau_\mu$, which are equivariant under the right action of $SL(4; \mathbb{R})$ on the bundles and its left action on $\mathbb{R}^4$ and its dual in a manner that is consistent with regarding $\mathbf{t}$ as a vector field and τ as a covector field. As usual, we assume that $\tau(\mathbf{t}) \neq 0$.

The Lorentzian metric g then allows one to reduce $SL_0(M)$ to the bundle $SO_0(3, 1)(M)$ M of oriented, time-oriented, Lorentzian frames in $T(M)$. When M is compact the introduction of a Lorentzian metric is also obstructed by the non-vanishing of the Euler-Poincaré characteristic, but for a non-compact M, such a metric always exists [4].

One can regard the metric g as also defining a map g: $GL(M) \to \mathrm{Lor}(4)$, $\mathbf{e}_\mu \mapsto g(\mathbf{e}_\mu) = g_{\mu\nu}(x)$, which takes any linear frame $\mathbf{e}_\mu$ in T_xM to the matrix $g_{\mu\nu}(x)$ of components of g with respect to this frame. The manifold $\mathrm{Lor}(4) = GL(4; \mathbb{R})/O(3, 1)$ is the homogeneous

[4] See the paper of Markus [**13**] on the subject of the obstructions to the existence of Lorentzian metrics. Some general remarks were also made in Steenrod [**14**].

space of all invertible 4×4 real matrices that can serve as the component matrices of Lorentzian metrics. That is, in addition to their invertibility they must also be symmetric and congruent to the matrix $\eta_{\mu\nu}$ = diag[+1, −1, −1, −1]. Furthermore, one requires that the map g, as just defined, must be equivariant under the right action of $GL(4;\mathbb{R})$ on $GL(M)$ and the left action on Lor(4). This is really just the usual tensorial transformation law of the metric components, since it says that when the frame $\mathbf{e}_\mu$ is changed to $\mathbf{e}_\nu A^\nu_\mu$ the component matrix $g_{\mu\nu}$ goes to $A^\kappa_\mu A^\lambda_\nu g_{\kappa\lambda}$.

If θ^μ is reciprocal coframe to $\mathbf{e}_\mu$ then one can reconstruct the metric $g(x)$ at the point x M by means of:

$$g(x) = g_{\mu\nu}(x)\, \theta^\mu \otimes \theta^\nu . \tag{XIII.64}$$

Since g is equivariant under frame changes, this construction will be unambiguous at each point of M even when no global frame field on M exists.

The actual reduction from $GL(M)$ to $O(3, 1)(M)$ then comes about by defining $O(3, 1)(M)$ to be the level hypersurface of $\eta_{\mu\nu}$ under the map g. Since the map g is not unique, neither is this reduction. Another way of seeing this is to observe that any linear frame $\mathbf{e}_\mu$ in T_xM can be regarded as orthonormal for some Lorentzian metric; i.e.:

$$g(\mathbf{e}_\mu, \mathbf{e}_\nu) = \eta_{\mu\nu} . \tag{XIII.65}$$

In fact, one simply *defines* that metric by means of:

$$g = \eta_{\mu\nu}\, \theta^\mu \otimes \theta^\nu . \tag{XIII.66}$$

The frame $\mathbf{e}_\mu$ defines an orbit in the manifold GL_xM under the action of $O(3, 1)$, but not all linear frames in GL_xM will belong to a common orbit. For instance, frames that differ by a non-trivial dilatation will not be on the same orbit. Hence, one can see that a choice of Lorentzian metric on M is also equivalent to a global section of the fiber bundle $\mathrm{Lor}(M) \to M$ whose fibers are diffeomorphic to the orbit spaces we just described. The fact that a global section does not have to exist follows from the fact this orbit space is not contractible; in fact, it is homotopically equivalent to $\mathbb{R}\mathrm{P}^3$. More precisely, Lor_xM is homotopically equivalent to PT_xM. Hence, the existence of a Lorentzian metric is homotopically equivalent to the existence of a global line field.

b. Canonical 1-forms on frame bundles. The bundle $\pi\colon GL(M) \to M$ has a *canonical 1-form* θ^μ with values in $\mathbb{R}^4$ that is defined on it. If $x \in M$ and $\mathbf{e} \in GL_xM$ then for any tangent vector $\mathbf{v} \in T_{\mathbf{e}}GL_x(M)$ the numbers $\theta^\mu(\mathbf{v})$ are the components v^μ of the tangent vector $\pi_*\mathbf{v}$ in T_xM:

$$\theta^\mu(\mathbf{v}) = \theta^\mu(\pi_*\mathbf{v}) = v^\mu \qquad (\text{so } \pi_*\mathbf{v} = v^\mu \mathbf{e}_\mu). \tag{XIII.67}$$

Again, we are abusing notation on the grounds that the 1-forms θ^μ on GL_xM are associated with the reciprocal coframe field in T_xM.

The canonical 1-form θ^μ on $GL(M)$ has a predictable variance under the action of $GL(4;\mathbb{R})$ on $GL(M)$ to the right, if one looks at the parenthetical comment in (XIII.66). That is, when the linear frame $\mathbf{e}_\mu$ goes to $\mathbf{e}_\nu A^\nu_\mu$, since the components v^μ of the tangent vector $\pi_*\mathbf{v}$ must go to $\tilde{A}^\mu_\nu v^\nu$ (tilde = matrix inverse) and $v^\mu = \theta^\mu(\mathbf{v})$, one must have that θ^μ goes to $\tilde{A}^\mu_\nu \theta^\nu$ in order to be consistent.

Although this probably sounds devilishly esoteric and confusing to many relativity theorists, who, to this day, seem to prefer the local component formulation of differential geometry to the exclusion of its modern formulation, it is really quite elementary when one goes to local components, since when one has a local frame field $\mathbf{e}$: $U \to GL(U)$, $x \mapsto \mathbf{e}_\mu(x)$ the 1-forms θ^μ on $GL(U)$ pull down to the reciprocal coframe field $\theta^\mu(x)$ to $\mathbf{e}_\mu(x)$.

The canonical 1-form θ^μ on $GL(M)$ also defines a canonical 1-form on all of its reductions by restriction. The only thing that changes is the subgroup of $GL(4;\mathbb{R})$ that one uses to make the frame changes.

Since so many important geometric constructions can be associated with reductions of the bundle $GL(M) \to M$ that are defined by reductions of the structure group to a subgroup $G \subset GL(4;\mathbb{R})$, differential geometry long since evolved a general notion for this process. When G is a Lie subgroup of $GL(n)$ a reduction of the bundle $GL(M) \to M$ of linear frames on an n-dimensional differentiable manifold M to a bundle $G(M) \to M$ is called a *G-structure*. In addition to being fundamental in the purely geometric context (cf., e.g., [**15-17**]), this notion is also quite fundamental in physics, as well (cf., [**18**]).

In general, it relates to a particular form of the general process of the spontaneous breaking of gauge symmetries in gauge field theories, which has been emerging over the years as a natural process that is more general than simply its application to elementary particle physics would suggest. Indeed, its application in condensed matter physics is just as important, since many of the phenomena of that branch of physics are more intuitively tractable than those of the sub-visible microcosmos.

c. Connections on frame bundles. There are many ways of introducing a connection on a fiber bundle, depending upon which ultimate class of problems one intends upon addressing, and they almost all relate to each other at varying levels of generality. When the bundle in question is a frame bundle – i.e., a G-structure $G(M) \to M$ – over a manifold M, some of them are more natural than others. Since we mostly wish to define a connection 1-form ω on $G(M)$ that will allow us to define the parallel translation of frames along curves in M, at least locally, we shall start with that as the definition.

If G is a Lie subgroup of $GL(4;\mathbb{R})$ and $\mathfrak{g}$ is its Lie algebra then a $\mathfrak{g}$-*connection* on a G-structure $G(M) \to M$ is a 1-form ω: $T(G(M)) \to \mathfrak{g}$, $\mathbf{v} \mapsto \omega^\mu_\nu(\mathbf{v})$ such that its restriction to the vertical [5] sub-bundle $V(G(M))$ is a linear isomorphism of each $V_\mathbf{e}$ with the vector

[5] Recall that a tangent vector to any fiber bundle $\pi: B \to M$ is *vertical* iff it projects to zero under the differential map $d\pi$. Such a tangent vector will also be tangent to some fiber of B. In the case of $GL(M)$ vertical tangent vector fields generate one-parameter families of linear frame transformations with the fibers of $GL(M)$.

space on which $\mathfrak{g}$ is defined, and when the frame $\mathbf{e}_\mu$ is changed to $\mathbf{e}_\nu A^\nu_\mu$ the components of the matrix ω^μ_ν go to $\tilde{A}^\mu_\kappa \omega^\kappa_\lambda A^\lambda_\nu$.

This last requirement eventually leads to the usual transformation law for a connection matrix when one chooses a local G-frame field $\mathbf{e}$: $U \to G(U)$ and pulls the 1-form ω on $G(U)$ down to a $\mathfrak{g}$-valued 1-form on U that we also denote by ω^μ_ν. In such a case, a change of local frame field involves a smooth transition function A: $U \to G$, $x \mapsto A^\mu_\nu(x)$. If one considers how the differential map $D\mathbf{e}$ to the frame field $\mathbf{e}$ relates to the differential map $D\overline{\mathbf{e}}$ to the frame field $\overline{\mathbf{e}} = \mathbf{e}A$ then one finds that:

$$D\overline{\mathbf{e}} = D(\mathbf{e}A) = (D\mathbf{e})A + \mathbf{e}\otimes dA = [D\mathbf{e} + \mathbf{e}\otimes(dAA^{-1})]A. \qquad \text{(XIII.68)}$$

That is, the transformation of the differential is not *covariant* – i.e., effected by means of A alone. Hence, one must *replace* the differential operator D with the *covariant differential* operator:

$$\nabla\mathbf{e}_\mu = D\mathbf{e}_\mu + \mathbf{e}_\nu \otimes \omega^\nu_\mu. \qquad \text{(XIII.69)}$$

One now has:

$$\nabla\overline{\mathbf{e}}_\mu = D\overline{\mathbf{e}}_\mu + \overline{\mathbf{e}}_\nu \otimes \overline{\omega}^\nu_\mu = [D\mathbf{e}_\nu + \mathbf{e}_\kappa \otimes (dA^\kappa_\lambda \tilde{A}^\lambda_\nu + A^\kappa_\lambda \overline{\omega}^\lambda_\rho \tilde{A}^\rho_\nu)]A^\nu_\mu. \qquad \text{(XIII.70)}$$

Hence, as long as:

$$\omega^\nu_\mu = A^\nu_\kappa \overline{\omega}^\kappa_\lambda \tilde{A}^\lambda_\mu + dA^\nu_\kappa \tilde{A}^\kappa_\mu \qquad \text{(XIII.71)}$$

one will have

$$\nabla\overline{\mathbf{e}}_\mu = \nabla\mathbf{e}_\nu\, A^\nu_\mu; \qquad \text{(XIII.72)}$$

i.e., the covariant differential will indeed be covariant.

Thus, the local version of the Ad^{-1}-equivariance of the 1-form ω on $G(M)$ will take the form:

$$\omega = A\overline{\omega}A^{-1} + dAA^{-1}. \qquad \text{(XIII.73)}$$

The matrix of local 1-forms ω^μ_ν on U can be related to the coframe θ^μ that is reciprocal to the chosen frame field $\mathbf{e}_\mu$ by means of:

$$\omega^\mu_\nu = \Gamma^\mu_{\kappa\nu}\theta^\kappa \qquad \text{(XIII.74)}$$

for a unique set of smooth functions $\Gamma^\mu_{\kappa\nu}$ on U that are analogous to the Riemann-Christoffel symbols of the Levi-Civita connection, which we shall discuss shortly.

The introduction of the covariant derivative allows us to introduce the notation of parallel translation along a curve $\gamma(t)$ in U. The local frame field $\mathbf{e}_\mu$ on U is *parallel along* the curve $\gamma(t)$ iff the covariant derivative of $\mathbf{e}_\mu$ in the direction of $\mathbf{v} = d\gamma/dt$ vanishes; i.e.:

$$0 = \nabla_{\mathbf{v}}\mathbf{e}_\mu = i_{\mathbf{v}}(\nabla\mathbf{e}_\mu) = v^\nu \frac{\partial \mathbf{e}_\mu}{\partial x^\nu} + \Gamma^\nu_{\kappa\mu} v^\kappa \mathbf{e}_\nu \,. \tag{XIII.75}$$

If **w** is a vector field on U then one can say that if $\mathbf{e}_\mu$ is parallel along $\gamma(\tau)$ then **w** is parallel along $\gamma(\tau)$ iff its components w^μ relative to $\mathbf{e}_\mu$ are constant. More generally, we define the covariant derivative of **w** along **v** to be:

$$\nabla_{\mathbf{v}}\mathbf{w} = v^\mu \nabla_{\mathbf{e}_\mu}(w^\nu \mathbf{e}_\nu) = v^\mu(\mathbf{e}_\mu w^\nu + w^\nu \nabla_{\mathbf{e}_\mu}\mathbf{e}_\nu) = v^\mu(\mathbf{e}_\mu w^\nu + \Gamma^\nu_{\mu\kappa} w^\kappa)\mathbf{e}_\nu \tag{XIII.76}$$

for an arbitrary G-frame field $\mathbf{e}_\mu$ on U. In these computations, we have set:

$$\nabla_{\mathbf{e}_\mu}\mathbf{e}_\nu = \Gamma^\kappa_{\mu\nu}\mathbf{e}_\kappa. \tag{XIII.77}$$

This means that the components of $\nabla_{\mathbf{v}}\mathbf{w}$ with respect to this local frame field are:

$$(\nabla_{\mathbf{v}}\mathbf{w})^\mu = v^\nu \mathbf{e}_\nu w^\mu + \Gamma^\mu_{\kappa\nu} v^\kappa w^\nu \,. \tag{XIII.78}$$

The vector field **w** is then parallel along the curve $\gamma(\tau)$ iff $\nabla_{\mathbf{v}}\mathbf{w}$ vanishes.

If the local frame field in question is the natural frame field ∂_μ for a local coordinate system (U, x^μ), with reciprocal coframe field dx^μ then (XIII.78) will take the form:

$$(\nabla_{\mathbf{v}}\mathbf{w})^\mu = v^\nu \frac{\partial w^\mu}{\partial x^\nu} + \Gamma^\mu_{\kappa\nu} v^\kappa w^\nu = \frac{dw^\mu}{d\tau} + \Gamma^\mu_{\kappa\nu} v^\kappa w^\nu \,. \tag{XIII.79}$$

Of particular interest are *geodesics*, which are curves along which the velocity vector $\mathbf{v}(t)$ is itself parallel-translated. One will then have the vanishing of the *proper acceleration* of the curve:

$$\mathbf{a}(\tau) = \nabla_{\mathbf{v}}\mathbf{v}, \tag{XIII.80}$$

which has the component form:

$$0 = v^\nu \mathbf{e}_\nu v^\mu + \Gamma^\mu_{\kappa\nu} v^\kappa v^\nu \tag{XIII.81}$$

with respect to an arbitrary local frame field and:

$$0 = \frac{dv^\mu}{d\tau} + \Gamma^\mu_{\kappa\nu} v^\kappa v^\nu \,. \tag{XIII.82}$$

with respect to a natural one.

The map that takes any G-frame $\mathbf{e}_\mu$ to its reciprocal G-coframe θ^μ defines a canonical isomorphism of the bundle $G(M) \to M$ of G-frames on M with the bundle $G^*(M) \to M$ of G-coframes on M. It too has a canonical 1-form with values in $\mathbb{R}^4$ that we also denote by θ^μ, because for a local coframe field $\theta\colon U \to G^*(M)$, $x \mapsto \theta^\mu_x$ the canonical 1-form pulls down to the coframe field θ^μ itself.

This isomorphism also allows one to map the connection ω on $G(M)$ over to a corresponding $\mathfrak{g}$-connection on $G^*(M)$. The covariant derivative of a local frame field θ^μ on $U \subset M$ along a curve $\gamma(\tau)$ in U whose velocity vector field is $\mathbf{v}(\tau)$ is defined to be:

$$\nabla_{\mathbf{v}}\theta^\mu = i_{\mathbf{v}} D\theta^\mu - \omega^\mu_\nu(\mathbf{v})\theta^\nu ; \tag{XIII.83}$$

Hence, this coframe field is parallel along the curve in question iff this covariant derivative vanishes.

The covariant derivative of a covector field $\alpha = \alpha_\mu \theta^\mu$ is then defined to have components with respect to θ^μ that are equal to:

$$(\nabla_{\mathbf{v}}\alpha)_\mu = \mathbf{v}\alpha_\mu - \omega^\nu_\mu(\mathbf{v})\alpha_\nu \tag{XIII.84}$$

with respect to an arbitrary coframe field and:

$$(\nabla_{\mathbf{v}}\alpha)_\mu = \frac{d\alpha_\mu}{d\tau} - \Gamma^\nu_{\kappa\mu} v^\kappa \alpha_\mu \tag{XIII.85}$$

with respect to a natural one.

One can define a stronger condition on vector fields, frame fields, and the like by eliminating the requirement that the field in question be parallel only along a specified curve and generalizing it to the requirement that the field itself be parallel; i.e., parallel along all curves. If one wishes that a local frame field $\mathbf{e}_\mu$ on $U \subset M$ be *parallel* then it is necessary and sufficient that its covariant differential:

$$\nabla \mathbf{e}_\mu = D\mathbf{e}_\mu + \omega^\nu_\mu \otimes \mathbf{e}_\nu \tag{XIII.86}$$

must vanish at all points of U. That is:

$$D\mathbf{e}_\mu = - \omega^\nu_\mu \otimes \mathbf{e}_\nu . \tag{XIII.87}$$

This condition amounts to a system of partial differential equations for the members of the local frame field while the condition that it be parallel along a specified curve gave a system of ordinary differential equations for the restriction of those members to the curve in question. Hence, the question of the integrability of the system of differential equations is more involved in the present case. In fact, one finds that a connection cannot admit parallel local frame fields unless its curvature vanishes; we shall discuss curvature in the next subsection, but, for now, we simply characterize it as an obstruction to the integrability of the equation (XIII.87).

One can also say that a vector field $\mathbf{v} = v^\mu \mathbf{e}_\mu$ on U is parallel on U if its covariant differential, which we define by:

$$\nabla \mathbf{v} = dv^\mu \otimes \mathbf{e}_\mu + v^\mu \nabla \mathbf{e}_\mu = (dv^\mu + \omega^\mu_\nu v^\nu) \otimes \mathbf{e}_\mu + v^\mu D\mathbf{e}_\mu , \tag{XIII.88}$$

vanishes at all points of *U*.

In the case of a natural frame, for which $D\mathbf{e}_\mu = 0$, we find that the covariant differential of **v** takes the form:

$$\nabla \mathbf{v} = \left(\frac{\partial v^\mu}{\partial x^\nu} + \Gamma^\mu_{\kappa\nu} v^\kappa \right) dx^\nu \otimes \partial_\mu . \tag{XIII.89}$$

To clarify what we mean by the expression $D\mathbf{e}_\mu$, let ∂_μ be a natural frame field on *U*, so we can represent $\mathbf{e}_\mu$ as $e^\nu_\mu(x)\partial_\nu$ for some unique invertible matrix of smooth functions $e^\nu_\mu(x)$. We then define:

$$D\mathbf{e}_\mu = de^\nu_\mu \otimes \partial_\nu = e^\nu_{\mu,\kappa} dx^\kappa \otimes \partial_\nu . \tag{XIII.90}$$

Hence, the vanishing of $D\mathbf{e}_\mu$ means that $\mathbf{e}_\mu$ differs from a natural frame field only by a constant transition function on *U*. In other words, the local frame field $\mathbf{e}_\mu$ could be called *integrable*, or, in physics terminology, *holonomic*; the non-integrable local frame fields are then *anholonomic.*

One can similarly define the covariant differential of a local coframe field θ^μ:

$$\nabla \theta^\mu = D\theta^\mu - \omega^\mu_\nu \otimes \theta^\nu , \tag{XIII.91}$$

and a covector field $\alpha = \alpha_\mu \theta^\mu$:

$$\nabla \alpha = d\alpha_\mu \otimes \theta^\mu + \alpha_\mu \nabla \theta^\mu = (d\alpha_\mu - \omega^\nu_\mu \alpha_\nu) \otimes \theta^\mu + \alpha_\mu D\theta^\mu . \tag{XIII.92}$$

In a natural frame ($D\theta^\mu = 0$), we then have:

$$\nabla \alpha = \left(\frac{\partial \alpha_\mu}{\partial x^\nu} - \Gamma^\kappa_{\mu\nu} \alpha_\kappa \right) dx^\mu \otimes dx^\nu . \tag{XIII.93}$$

As above, if dx^μ is a natural coframe field on *U* and $\theta^\mu = \theta^\mu_\nu(x)\, dx^\nu$ then we are defining:

$$D\theta^\mu = d\theta^\mu_\nu \otimes dx^\nu = \theta^\mu_{\nu,\kappa}\, dx^\kappa \otimes dx^\nu . \tag{XIII.94}$$

There is then an analogous notion of integrability and anholonomity that is associated with local coframe fields.

The corresponding notion of parallelism in the above cases is also the vanishing of the various covariant differentials.

The relationship between the covariant differential and the covariant derivative along a curve – or, more generally, a vector field **v** on *U* – is simply that of the relationship of a differential to a directional derivative; i.e.:

$$\nabla_{\mathbf{v}} = i_{\mathbf{v}}\nabla. \tag{XIII.95}$$

Since every possible tensor field on U can be expressed as a finite linear combination of tensor products of frame fields and coframe fields on U, the covariant differential operator can be extended to the full tensor algebra over $T(U)$ and T^*M by requiring that it behave like a derivation. That is, if $T = a \otimes b$ then:

$$\nabla T = \nabla a \otimes b + a \otimes \nabla b. \tag{XIII.96}$$

d. Torsion and curvature. In addition to the covariant form of the differential operator on tensor fields there is also a covariant form of the exterior derivative operator on differential forms, at least when the forms take their values in vector spaces.

For this situation, it becomes more convenient to represent a tensor field of rank $r+s$ on a manifold M as a smooth map:

$$T\colon G(M) \to \mathbb{R}^{n^*} \otimes \ldots \otimes \mathbb{R}^{n^*} \otimes \mathbb{R}^n \otimes \ldots \otimes \mathbb{R}^n,\ \mathbf{e} \mapsto T(\mathbf{e}) = T^{\mu_1 \cdots \mu_r}_{\nu_1 \cdots \nu_s}(x)$$

that is equivariant under the right action of G on frames and the tensor product representation of G in $GL(\mathbb{R}^{n^*} \otimes \ldots \otimes \mathbb{R}^{n^*} \otimes \mathbb{R}^n \otimes \ldots \otimes \mathbb{R}^n)$. That is, T associates a G-frame $\mathbf{e}_\mu$ at $x \in M$ with the components $T^{\mu_1 \cdots \mu_r}_{\nu_1 \cdots \nu_s}(x)$ of a tensor field on M with respect to that frame in a manner that obeys the usual rules of transformation for the components of a tensor field:

The tensor field on M is then defined globally, even in the absence of a global frame field, by the construction rule:

$$T(x) = T^{\mu_1 \cdots \mu_r}_{\nu_1 \cdots \nu_s}(x)\ \mathbf{e}_{\mu_1} \otimes \ldots \mathbf{e}_{\mu_r} \otimes \theta^{\nu_1} \otimes \ldots \otimes \theta^{\nu_s}. \tag{XIII.97}$$

As an example, suppose $G = GL(n; \mathbb{R})$. One can then represent a metric tensor field g on M as a smooth function $g\colon GL(M) \to \mathbb{R}^n \otimes \mathbb{R}^n$, $\mathbf{e} \mapsto g(\mathbf{e}) = g_{\mu\nu}(x)$ and reconstruct the tensor field g by way of:

$$g(x) = g_{\mu\nu}(x)\ \theta^\mu \otimes \theta^\nu. \tag{XIII.98}$$

Although a smooth map is essentially a 0-form, this definition of a tensor field generalizes immediately to the case of k-forms on M with values in vector spaces, and, in particular, the vector spaces that are expressible as the tensor product of other vector spaces. For instance, we can regard the canonical 1-form θ^μ on $G(M)$ as a 1-form on $G(M)$ with values in $\mathbb{R}^n$ and a connection form ω is a 1-form on $G(M)$ with values in $\mathfrak{g}$.

The exterior derivative operator that acts on elements of $\Lambda^*(G(M))$ can be extended to an *exterior covariant derivative* operator on $\Lambda^*(G(M)) \otimes V$, which is how we generically represent equivariant differential forms on $G(M)$ with values in a vector space V on which G acts linearly. Since the definition will depend upon the choice of action, rather than

define the general case, we simply define the way that it works for the cases of immediate interest to us.

In the case of the canonical 1-form θ^μ, as we observed above, $V = \mathbb{R}^n$, and when G is represented as a subgroup of $GL(n; \mathbb{R})$ the action of G on $\mathbb{R}^n$ is simply the defining representation; viz., the multiplication of a matrix and a vector or covector. We then define the exterior covariant derivative of the $\mathbb{R}^n$-valued 1-form θ^μ on $G(M)$ to be the $\mathbb{R}^n$-valued *torsion* 2-form on $G(M)$:

$$\Theta^\mu = \nabla^\wedge \theta^\mu = d\theta^\mu + \omega^\mu_\nu \wedge \theta^\nu . \tag{XIII.99}$$

We clarify that the second term on the right-hand side of this definition takes any pair of vector field **v**, **w** on $G(M)$ to the vector in $\mathbb{R}^n$:

$$\omega^\mu_\nu \wedge \theta^\nu (\mathbf{v}, \mathbf{w}) = \tfrac{1}{2}[\omega^\mu_\nu(\mathbf{v})\theta^\nu(\mathbf{w}) - \omega^\mu_\nu(\mathbf{w})\theta^\nu(\mathbf{v})] = \tfrac{1}{2}(\Gamma^\mu_{\kappa\nu} - \Gamma^\mu_{\nu\kappa})v^\kappa w^\nu . \tag{XIII.100}$$

When one defines a local G-frame field **e**: $U \to G(M)$, one can pull all of the differential forms in (XIII.99) down to forms on U that we denote by the same symbols. In particular, θ^μ represents the reciprocal local coframe field to $\mathbf{e}_\mu$ on U.

The first term on the right-hand side of (XIII.99) is the *anholonomity* of the local coframe field θ^μ. Since all k-forms on U can be represented in terms of θ^μ we have:

$$d\theta^\mu = -\tfrac{1}{2}\, c^\mu_{\kappa\lambda}\ \theta^\kappa \wedge \theta^\lambda, \tag{XIII.101}$$

for a set of functions $c^\mu_{\kappa\lambda}$ on U that we call the *structure functions* of θ^μ; the reason for the minus sign is rooted in the fact that one also has:

$$[\mathbf{e}_\kappa, \mathbf{e}_\lambda] = c^\mu_{\kappa\lambda}\, \mathbf{e}_\mu . \tag{XIII.102}$$

One sees that the anholonomity amounts to an obstruction to the integrability of the local system of differential equations $\theta^\mu = dx^\mu$ on U, since its vanishing is a necessary and sufficient condition for such functions x^μ to exist.

If $\theta^\mu = h^\mu_\nu dx^\nu$ for some natural coframe field then we can represent the $c^\mu_{\kappa\lambda}$ by:

$$c^\mu_{\kappa\lambda} = h^\mu_{\kappa,\lambda} - h^\mu_{\lambda,\kappa} . \tag{XIII.103}$$

If we represent ω^μ_ν as $\Gamma^\mu_{\kappa\nu}\theta^\kappa$ then the torsion 2-form takes the form:

$$\Theta^\mu = \tfrac{1}{2} S^\mu_{\kappa\lambda}\, \theta^\kappa \wedge \theta^\lambda, \tag{XIII.104}$$

with:

$$S^\mu_{\kappa\lambda} = -c^\mu_{\kappa\lambda} + \Gamma^\mu_{\kappa\lambda} - \Gamma^\mu_{\lambda\kappa} . \tag{XIII.105}$$

In a holonomic local frame field, the $c^{\mu}_{\kappa\lambda}$ vanish and the torsion 2-form then represents the "anti-symmetric part" of the connection. In such a frame field, the vanishing of torsion is equivalent to the symmetry of the connection components $\Gamma^{\mu}_{\kappa\nu}$ in their lower indices.

Although (XIII.104) and (XIII.105) were defined on $G(M)$, they take the same form when pulled down to $U \subset M$ by means of a local G-frame field.

Another way of looking at torsion is that when it vanishes the equation:

$$d\theta^{\kappa} = -\omega^{\mu}_{\nu} \wedge \theta^{\nu} \tag{XIII.106}$$

says that the exterior differential system $\theta^{\kappa} = 0$ on $G(M)$ is completely integrable. As the 1-form θ^{κ} vanishes precisely when evaluated on vertical tangent vectors one sees that connections with vanishing torsion will always exist since the integral submanifolds to this exterior differential system will be the fibers of G, which always exist. Hence, non-vanishing torsion is only weakly an obstruction to the integrability of this exterior differential system.

Since the action of G on the Lie algebra $\mathfrak{g}$ is the adjoint action, we define the exterior covariant derivative of the $\mathfrak{g}$-valued connection 1-form ω to be the $\mathfrak{g}$-valued *curvature* 2-form:

$$\Omega^{\mu}_{\nu} = \nabla\hat{}\,\omega^{\mu}_{\nu} = d\omega^{\mu}_{\nu} + \omega^{\mu}_{\kappa} \wedge \omega^{\kappa}_{\nu}. \tag{XIII.107}$$

In the right-hand side of this expression, we clarify that the term $\omega^{\mu}_{\kappa} \wedge \omega^{\kappa}_{\nu}$, when evaluated on two vector fields **v**, **w** on $G(M)$, will have the value:

$$\omega^{\mu}_{\kappa} \wedge \omega^{\kappa}_{\nu}(\mathbf{v}, \mathbf{w}) = \tfrac{1}{2}[\omega^{\mu}_{\kappa}(\mathbf{v})\omega^{\kappa}_{\nu}(\mathbf{w}) - \omega^{\mu}_{\kappa}(\mathbf{w})\omega^{\kappa}_{\nu}(\mathbf{v})] = -\tfrac{1}{2}[\omega(\mathbf{v}), \omega(\mathbf{w})]. \tag{XIII.108}$$

When ω^{μ}_{ν} is expressed as $\Gamma^{\mu}_{\kappa\nu}\theta^{\kappa}$, the curvature 2-form can be expressed as:

$$\Omega^{\mu}_{\nu} = \tfrac{1}{2} R^{\mu}_{\kappa\lambda\nu}\, \theta^{\kappa} \wedge \theta^{\lambda}, \tag{XIII.109}$$

in which:

$$R^{\mu}_{\kappa\lambda\nu} = \Gamma^{\mu}_{\lambda\nu,\kappa} - \Gamma^{\mu}_{\kappa\nu,\lambda} + \Gamma^{\mu}_{\kappa\sigma}\Gamma^{\sigma}_{\lambda\nu} - \Gamma^{\mu}_{\lambda\sigma}\Gamma^{\sigma}_{\kappa\nu} \tag{XIII.110}$$

gives the components of the curvature tensor field in its conventional local form.

One can regard the curvature 2-form as an obstruction to the complete integrability of the exterior differential system $\omega^{\mu}_{\nu} = 0$ since its vanishing would make:

$$d\omega^{\mu}_{\nu} = -\omega^{\mu}_{\kappa} \wedge \omega^{\kappa}_{\nu}. \tag{XIII.111}$$

The 1-forms ω^{μ}_{ν} vanish precisely when they are evaluated on tangent vectors to $G(M)$ that are, by definition, *horizontal*. if one denotes the vector space of all horizontal tangent vectors at $\mathbf{e} \in G(M)$ by $H_{\mathbf{e}}(G(M))$ and the resulting horizontal sub-bundle of $T(G(M))$ by $H(G(M))$ then one will have a Whitney sum decomposition:

$$T(G(M)) = H(G(M)) \oplus V(G(M)).$$

However, unlike the vertical sub-bundle $V(G(M))$, which is always integrable, the horizontal sub-bundle $H(G(M))$ does not always have to be integrable. Since its fibers have dimension n, and, in fact, the 1-form θ^μ defines a linear isomorphism of each horizontal tangent space with $\mathbb{R}^n$, the integral submanifolds of the sub-bundle $H(G(M))$ would have to represent diffeomorphic copies of M in $G(M)$ that are transverse to the fibers; i.e., global sections of the fibration $G(M) \to M$. This possibility is obstructed by topology, though, so the existence of a $\mathfrak{g}$-connection with vanishing curvature is necessary for the triviality of $G(M)$; viz., its parallelizability.

Often (see, e.g., [**7, 9, 10, 12, 15, 16**]), equations (XIII.99) and (XIII.106) are collectively referred to as the *Cartan structure equations* for the connection ω. As we have presented them, they are simply the definitions of the torsion and curvature of the connection, but if one starts with a more general definition of a $\mathfrak{g}$-connection, such as a complementary horizontal sub-bundle to the vertical sub-bundle of $T(G(M))$, then one can derive the equations as a consequence.

If one is given an $\mathbb{R}^n$-valued 2-form Θ^μ on $G(M)$ and a $\mathfrak{g}$-valued 2-form Ω^μ_ν that have the same transformation properties as the torsion and curvature 2-forms then the integrability conditions for the systems of partial differential equations for a connection 1-form ω^μ_ν that the structure equations represent are given by computing the square of the exterior covariant derivative operator, which we express generically as:

$$\nabla^\wedge\nabla^\wedge\alpha = d(d\alpha + \omega^\wedge \alpha) + \omega^\wedge (d\alpha + \omega^\wedge \alpha) = \Omega \wedge \alpha. \qquad \text{(XIII.112)}$$

When this formula is applied to the canonical 1-form and the connection 1-form, one derives the *Bianchi identities:*

$$\nabla^\wedge\Theta^\mu = \Omega^\mu_\nu \wedge \theta^\nu, \qquad \nabla^\wedge\Omega^\mu_\nu = 0. \qquad \text{(XIII.113)}$$

One can associate the curvature 2-form Ω^μ_ν on $G(M)$ with a 1-form $\mathcal{R}_\mu$ that takes its values in $\mathbb{R}^{n*}$, which we shall the *Ricci curvature* tensor, even though that tensor usually gets defined in the context of metric connections. It takes the G-frame $\mathbf{e}_\mu \in G(M)$ to the 1-form:

$$\mathcal{R}_\mu = i_{\mathbf{e}_\nu}\Omega^\nu_\mu = R_{\mu\nu}\theta^\nu, \qquad \text{(XIII.114)}$$

in which:

$$R_{\mu\nu} = R^\kappa{}_{\mu\kappa\nu}. \qquad \text{(XIII.115)}$$

In the general case of a $\mathfrak{g}$-connection, these components do not have to be symmetric, though.

An aspect of the definition of a $\mathfrak{g}$-connection on $G(M)$ that is important to the structure of the spacetime manifold is the fact that even though the manifold M does not

have to be parallelizable, nonetheless, the manifold $GL(M)$ is always parallelizable. This is because the 1-form θ^μ defines a linear isomorphism of each horizontal subspace on $GL(M)$ with $\mathbb{R}^n$, while the connection 1-form ω^μ_ν defines a linear isomorphism of each vertical subspace with $\mathfrak{gl}(n; \mathbb{R})$. Collectively, the set $\{\theta^\mu, \omega^\mu_\nu\}$ defines a linear isomorphism of each tangent space to $GL(M)$ with the vector space $\mathbb{R}^n \oplus \mathfrak{gl}(n; \mathbb{R})$. If one also chooses a basis for $\mathfrak{gl}(n; \mathbb{R})$ then one can think of the set $\{\theta^\mu, \omega^\mu_\nu\}$ as defining a linear isomorphism of each $T_{\mathbf{e}}G(M)$ with $\mathbb{R}^{n(n+1)}$; i.e., a global coframe field on $GL(M)$. By restriction, the same is true for any reduction of $GL(M)$, as well, with a suitable adjustment to the dimension.

The set $\{\theta^\mu, \omega^\mu_\nu\}$ can be thought of as a 1-form $\tilde{\theta}^a$, $a = 1, \ldots, n + \dim \mathfrak{g}$ on $G(M)$ with values in the vector space $\mathbb{R}^{n+\dim \mathfrak{g}}$. Hence, one can represent $\tilde{\theta}^a$ in block matrix form as:

$$\tilde{\theta}^a = \left[\begin{array}{c} \theta^\mu \\ \hline \omega^\mu_\nu \end{array}\right], \tag{XIII.116}$$

although it would be more precise to denote the elements of the matrix ω^μ_ν as a singly-indexed column vector by choosing a basis E_α, $\alpha = 1, \ldots, \dim \mathfrak{g}$, so that ω gets expressed in the form $\omega^\alpha E_\alpha$.

We define the connection on $GL(G(M)) = G(M) \times \mathbb{R}^{n+\dim \mathfrak{g}}$ that makes the global frame field $\tilde{\theta}^a$ parallel; one calls such a connection either a *teleparallelism connection*, in general, or a *Cartan connection*, in the present case.

It is no longer necessary to specify both the torsion and curvature of such a connection 1-form, since both the torsion and the curvature 2-forms of ω^μ_ν get absorbed into the torsion 2-form $\tilde{\Theta}^a$:

$$\tilde{\Theta}^a = d\tilde{\theta}^a = \left[\begin{array}{c} \Theta^\mu \\ \hline \Omega^\mu_\nu \end{array}\right], \tag{XIII.117}$$

while the curvature vanishes identically, due to parallelizability.

In physics, teleparallelism usually refers to making the assumption that the spacetime manifold M is parallelizable and working with the geometry of such manifolds in the hopes of unifying the theories of gravitation and electromagnetism, since the manifold $GL(4; \mathbb{R})$ has the same dimension – viz., sixteen – as the sum of the dimensions of the manifold of Lorentzian structures Lor(4), namely, ten, and the vector space $\Lambda^2(\mathbb{R}^4)$, which has dimension six. Interestingly, when Einstein, Mayer, and others [**19**] were doing this work in the late 1920's, they were using purely local expressions for frame fields and made no mention of the global topological obstructions to parallelizability.

Indeed, the first definitive work on topology and teleparallelism, which was published by Stiefel [**20**], did not appear until 1935. Hence, one can only wonder whether a more topologically thorough analysis of the situation might extend the physical theory accordingly, especially since the obstruction to parallelizability involves non-vanishing curvature, which is bound to contribute to the physics when it could not in the parallelizable case.

e. Metric connections. Now that we have defined the covariant differential of tensor fields in general, we can apply it to the various fundamental tensor fields that relate to the reduction of the bundle of linear frames to its various G-structures.

The first reduction that we encountered was from $GL(M)$ to $GL^+(M)$, but that did not involve a fundamental tensor field, only a choice of orientation. It was in the next reduction to $SL(M)$ that we had to introduce a tensor field in order to define this reduction, in the form of a volume element V on $T(M)$. Hence, a connection ω on $GL(M)$ that is compatible with this reduction must preserve the volume of a linear frame under parallel translation along a curve. In order for this to happen it is necessary and sufficient that ω must be Ad^{-1}-equivariant under the action of $SL(n;\mathbb{R})$ on $SL(M)$, which is equivalent to saying that ω must take its values in the Lie algebra $\mathfrak{sl}(n;\mathbb{R})$.

Now, we can draw upon a general result from the geometry of G-structures [**15-17**] that when a reduction from $GL(M)$ to a G-structure $G(M)$ is effected by means of a fundamental tensor field τ: $GL(M) \to V$ the necessary and sufficient condition for a $\mathfrak{gl}(n)$-connection 1-form ω on $GL(M)$ to be reducible to a $\mathfrak{g}$-connection on $G(M)$ is that τ be covariantly constant with respect to the connection ω; i.e.:

$$0 = \nabla^{\wedge}\tau = d\tau + \omega^{\wedge}\tau . \qquad \text{(XIII.118)}$$

In the case of V, we regard it, not as a 4-form on $GL(M)$, but as a 0-form with values in $\Lambda^4(\mathbb{R}^4)$. This makes dV vanish by dimensionality, and the remaining condition on ω is derived from:

$$\begin{aligned} 0 = \omega^{\wedge}\tau &= \frac{1}{4!}\varepsilon_{\kappa\lambda\mu\nu}(\omega^{\kappa}_{\alpha}\theta^{\alpha}\wedge\theta^{\lambda}\wedge\theta^{\mu}\wedge\theta^{\nu} + \theta^{\mu}\wedge\omega^{\lambda}_{\alpha}\theta^{\alpha}\wedge\theta^{\mu}\wedge\theta^{\nu} \\ &\qquad + \theta^{\kappa}\wedge\theta^{\lambda}\wedge\omega^{\mu}_{\alpha}\theta^{\alpha}\wedge\theta^{\nu} + \theta^{\kappa}\wedge\theta^{\lambda}\wedge\theta^{\mu}\wedge\omega^{\nu}_{\alpha}\theta^{\alpha}) \\ &= 4\,\mathrm{Tr}(\omega)\,V. \end{aligned} \qquad \text{(XIII.119)}$$

The last step follows from the fact that:

$$\frac{1}{4!}\varepsilon_{\kappa\lambda\mu\nu}\omega^{\kappa}_{\alpha}\theta^{\alpha}\wedge\theta^{\lambda}\wedge\theta^{\mu}\wedge\theta^{\nu} = \omega^{\alpha}_{\alpha}\frac{1}{4!}\varepsilon_{\kappa\lambda\mu\nu}\theta^{\kappa}\wedge\theta^{\lambda}\wedge\theta^{\mu}\wedge\theta^{\nu}, \text{ etc.} \qquad \text{(XIII.120)}$$

Hence, since V is, by assumption, non-vanishing, one must have:

$$0 = \mathrm{Tr}\ \omega = \omega^{\mu}_{\mu}. \qquad \text{(XIII.121)}$$

This is, of course, the condition for the matrix ω to have membership in $\mathfrak{sl}(n; \mathbb{R})$.

In the case of spacetime, the next reduction is from $SL(M)$ to $SO(3, 1)(M)$, which is associated with the fundamental tensor field g: $GL(M) \rightarrow SL(4)/SO(3,1)$; i.e., the Lorentzian metric on $T(M)$. From (XIII.118), the necessary and sufficient condition that an $\mathfrak{sl}(4; \mathbb{R})$-connection be reducible to an $\mathfrak{so}(3;1)$-connection is then:

$$0 = \nabla g_{\mu\nu} = dg_{\mu\nu} + \omega^{\kappa}_{\mu} g_{\kappa\nu} + \omega^{\kappa}_{\nu} g_{\mu\kappa}. \qquad \text{(XIII.122)}$$

One often finds the 1-form $Q_{\mu\nu} = \nabla g_{\mu\nu}$ on $GL(M)$ with values in $\mathrm{Lor}(4) = SL(4)/SO(3,1)$ referred to as the *non-metricity* tensor of ω (cf., e.g., [**11**]).

When g is restricted to the bundle $SO(3,1)(M)$ of Lorentzian frames, $g_{\mu\nu} = \eta_{\mu\nu}$, and the compatibility condition for ω to be a metric connection is:

$$0 = \omega^{\kappa}_{\mu}\eta_{\kappa\nu} + \omega^{\kappa}_{\nu}\eta_{\mu\kappa} = \omega_{\mu\nu} + \omega_{\nu\mu}\ . \qquad \text{(XIII.123)}$$

Again, this is simply the condition that ω take its values in $\mathfrak{so}(3, 1)$.

If we wish to further reduce to $SO_0(3, 1)(M)$ then we need a global timelike vector field $\mathbf{t}$. However, this is not a fundamental tensor field for the reduction from $SO(3,1)(M)$ to $SO_0(3,1)(M)$ since the homogeneous space $SO(3,1)/SO_0(3, 1)$ consists of two points. We note that when one has such a vector field $\mathbf{t}$, it does represent the fundamental tensor field for the reduction from $SO(3,1)(M)$ to $SO(3)(M)$ since the homogeneous space $SO(3,1)/SO(3)$ is diffeomorphic to $\mathbb{R}^3$. Hence, we can regard $\mathbf{t}$ as an $SO(3,1)$-equivariant map t: $SO(3,1)(M) \rightarrow \mathbb{R}^3$, $\mathbf{e}_{\mu} \mapsto t(\mathbf{e}_{\mu}) = t^i$, where the t^i represent the spatial components of $\mathbf{t}$ relative to $\mathbf{e}_{\mu}$. Hence, the reduction from $SO(3,1)(M)$ to $SO(3)(M)$ is defined by $t^i = 0$, which means all oriented Lorentzian frames with $\mathbf{e}_0 = \mathbf{t}$. This means that one is reducing to the rest space of a measurer/observer that is defined by $\mathbf{t}$ and its g-orthogonal complementary hyperspace, which has the 1-form $\tau = i_{\mathbf{t}} g = t_{\mu}\theta^{\mu}$.

The compatibility condition for this latter reduction of ω is that $\mathbf{t}$ be a parallel vector field for ω:

$$0 = \nabla \mathbf{t} = \nabla t^{\mu} \otimes \mathbf{e}_{\mu}. \qquad \text{(XIII.124)}$$

As a consequence, its flow must be geodesic.

Even though all $\mathfrak{gl}(n)$-connection 1-forms take their values in a vector space – namely, $\mathfrak{gl}(n)$ – nonetheless, the set $\Gamma(GL(M))$ of all $\mathfrak{gl}(n)$-connections on $GL(M)$ is not a vector space, but only an affine space. This is because for any two $\mathfrak{gl}(n)$-connection 1-forms ω and ω' the sum of the connections does not have to be a connection. However, their *difference 1-form*:

$$\tau = \omega' - \omega, \qquad \text{(XIII.125)}$$

is tensorial. Hence, the vector space on which the affine space $\Gamma(GL(M))$ is modeled is the space of equivariant 1-forms on $GL(M)$ that take their values in $\mathfrak{gl}(n)$, which is, as one sees, an infinite-dimensional vector space.

One of the big differences between affine spaces and vector spaces is that in an affine space there is usually no distinguished point that could serve as a canonical choice of origin. In the case of connections, one finds that the set of connections with vanishing torsion can sometimes define such a distinguished point in $\Gamma(GL(M))$. Although the set of torsionless connections on $GL(M)$ and $SL(M)$ is of dimension higher than zero, one finds that the space of $\mathfrak{so}(3, 1)$-connections does have a distinguished element, which one calls the *Levi-Civita* connection. By definition, it will be the unique connection ω that has both vanishing torsion and metricity:

$$0 = \Theta^{\mu}, \qquad 0 = Q_{\mu\nu}\,. \tag{XIII.126}$$

The local component equations for these conditions are then:

$$\Gamma^{\kappa}_{\mu\nu} = \Gamma^{\kappa}_{\nu\mu}\,, \qquad g_{\mu\nu,\,\kappa} = -\,\Gamma_{\mu\nu\kappa} - \Gamma_{\kappa\nu\mu}\,. \tag{XIII.127}$$

in a holonomic local frame field.

The components of the Levi-Civita connection – viz., the *Riemann-Christoffel* symbols – are then:

$$\Gamma^{\kappa}_{\mu\nu} = \tfrac{1}{2} g^{\kappa\alpha}(\, g_{\mu\alpha,\,\nu} + g_{\alpha\nu,\,\mu} - g_{\mu\nu,\,\alpha}\,). \tag{XIII.128}$$

The Riemann curvature tensor $\Omega^{\mu}_{\nu} = \nabla\omega^{\mu}_{\nu}$ that is obtained from the Levi-Civita connection differs from the general curvature 2-form for a linear connection mostly by the fact that it takes its values in $\mathfrak{so}(3, 1)$.

However, the Ricci curvature 1-form $\mathcal{R}_{\mu} = R_{\mu\nu}\theta^{\nu}$ that we defined above is now symmetric:

$$R_{\mu\nu} = R_{\nu\mu}\,, \tag{XIII.129}$$

and one can also associate Ω^{μ}_{ν} with a *scalar curvature:*

$$R(\mathbf{e}_{\mu}) = g^{\mu\nu} i_{\mathbf{e}_{\mu}} \mathcal{R}_{\nu} = g^{\mu\nu} R_{\mu\nu}\,. \tag{XIII.130}$$

From equivariance, this function on $G(M)$ is associated with a corresponding function $R(x)$ on M.

When Einstein was first looking for a way of coupling the spacetime curvature to the symmetric energy-momentum tensor $\tau_{\mu} = T_{\mu\nu}dx^{\nu}$, he had to take into account that, whereas the conservation of energy-momentum would dictate that the covariant divergence of $\tau^{\mu} = g^{\mu\nu}\,\tau_{\nu}$ should vanish:

$$0 = \nabla_{\mu}\,\tau^{\mu} = T^{\mu}{}_{\nu;\,\mu}\,dx^{\nu} \qquad (T^{\mu}{}_{\nu} = g^{\mu\kappa}T_{\kappa\nu}) \tag{XIII.131}$$

nonetheless, from the Bianchi identities, one should have:

$$\nabla_\mu \mathcal{R}^\mu = R^\mu{}_{\nu;\mu}\ dx^\nu = \tfrac{1}{2} R_{,\nu}\, dx^\nu \qquad (R^\mu{}_\nu = g^{\mu\kappa} R_{\kappa\nu}) \qquad \text{(XIII.132)}$$

Hence, the divergenceless part of the Ricci curvature tensor is called the *Einstein tensor*, which we can represent as an equivariant 1-form $\mathcal{E}^\mu$ on $GL(M)$ with values in $\mathbb{R}^4$:

$$\mathcal{E}^\mu(\mathbf{e}) = (R^\mu{}_\nu - \tfrac{1}{2} R\, \delta^\mu_\nu)\ \theta^\nu . \qquad \text{(XIII.133)}$$

The Einstein field equations for gravitation then express the idea that $\mathcal{E}^\mu$ should be proportional to τ^μ:

$$\mathcal{E}^\mu = 8\pi\kappa\tau^\mu, \qquad \text{(XIII.134)}$$

or:

$$R^\mu{}_\nu - \tfrac{1}{2} R\, \delta^\mu_\nu = 8\pi\kappa T^\mu{}_\nu, \qquad \text{(XIII.135)}$$

which also implies:

$$R = -\,8\pi\kappa T^\mu{}_\mu\ . \qquad \text{(XIII.136)}$$

Hence, the scalar curvature is directly proportional to minus the trace of the energy-momentum tensor. This is particularly relevant to the case of an energy distribution that is purely due to electromagnetic radiation, since the Faraday energy-momentum tensor then has that property. In one of the early stages of the Big Bang, the matter in the universe was assumed to be purely radiation-dominated.

This implies that (XIII.135) can also be written in the form:

$$R^\mu{}_\nu = 8\pi\kappa(T^\mu{}_\nu - \tfrac{1}{2} T^\mu{}_\mu\, \delta^\mu_\nu). \qquad \text{(XIII.137)}$$

Hence, the sourceless field equations for g are simply the vanishing of the Ricci curvature tensor:

$$R^\mu{}_\nu = 0. \qquad \text{(XIII.138)}$$

That this is consistent with the equations:

$$R^\mu{}_\nu = -\tfrac{1}{2} R\, \delta^\mu_\nu \qquad \text{(XIII.139)}$$

then follows from the vanishing of scalar curvature with the vanishing of $T^\mu{}_\nu$, as one sees from (XIII.136).

In differential geometry, the condition for a manifold with a metric g to be an *Einstein space* is usually weakened slightly to the condition that $R_{\mu\nu}$ be proportional to $g_{\mu\nu}$.

One can give the definition of the Einstein tensor $\mathcal{E}_\mu$ and the scalar curvature R a somewhat more concise and elegant formulation by means of the *curvature 3-form* $\mathcal{E}_\mu$ on $GL(M)$, which takes a frame $\mathbf{e}_\mu$ to:

$$\mathcal{E}_\mu(\mathbf{e}) = -\,\theta^\nu \wedge {}^*\Omega_{\mu\nu} = -\,\varepsilon_{\mu\nu\kappa\lambda}\, \theta^\nu \wedge \Omega^{\kappa\lambda}\ , \qquad \text{(XIII.140)}$$

in which we have defined:

$$\Omega_{\mu\nu} = g_{\mu\nu}\,\Omega^{\kappa}_{\nu}\,, \qquad \Omega^{\kappa\lambda} = g^{\kappa\alpha}\Omega^{\lambda}_{\alpha} = \tfrac{1}{4}R^{\kappa\lambda}{}_{\alpha\beta}\,\theta^{\alpha}\wedge\theta^{\beta}. \qquad \text{(XIII.141)}$$

Substituting (XIII.141) into (XIII.140) gives:

$$\begin{aligned}\mathcal{E}_{\mu}(\mathbf{e}) &= -\tfrac{1}{4}\,\varepsilon_{\mu\nu\kappa\lambda}\,R^{\kappa\lambda}{}_{\alpha\beta}\,\theta^{\nu}\wedge\theta^{\alpha}\wedge\theta^{\beta} = -\tfrac{1}{4}\,\varepsilon_{\mu\nu\kappa\lambda}\,\varepsilon^{\nu\alpha\beta\rho}\,R^{\kappa\lambda}{}_{\alpha\beta}\,\#\mathbf{e}_{\rho}\\ &= -\tfrac{1}{4}\,\delta^{\alpha\beta\rho}_{\kappa\lambda\mu}\,R^{\kappa\lambda}{}_{\alpha\beta}\,\#\mathbf{e}_{\rho} = (R^{\nu}_{\mu} - \tfrac{1}{2}R\delta^{\nu}_{\mu})\,\#\mathbf{e}_{\nu} = \#(E^{\nu}_{\mu}\mathbf{e}_{\nu})\end{aligned} \qquad \text{(XIII.142)}$$

In these computations, we have made use of the fact that the identity map on 3-forms has the components:

$$\delta^{\alpha\beta\rho}_{\kappa\lambda\mu} = \delta^{\alpha}_{\kappa}\delta^{\beta}_{\lambda}\delta^{\rho}_{\mu} + \delta^{\beta}_{\kappa}\delta^{\rho}_{\lambda}\delta^{\alpha}_{\mu} + \delta^{\rho}_{\kappa}\delta^{\alpha}_{\lambda}\delta^{\beta}_{\mu}\,. \qquad \text{(XIII.143)}$$

Hence, the curvature 3-form $\mathcal{E}_{\mu}(\mathbf{e})$, which takes its values in $\mathbb{R}^4$, is Poincaré dual to a set of four vector fields:

$$\mathbf{E}_{\mu} = E^{\nu}_{\mu}\mathbf{e}_{\nu} \qquad \text{(XIII.144)}$$

that describe the Einstein tensor for this choice of frame. They are not generally linearly independent, though, so they do not usually define a 4-frame.

If we form the 4-form on $GL(M)$:

$$\theta^{\mu}\wedge\mathcal{E}_{\mu}(\mathbf{e}) = -\,\theta^{\mu}\wedge\theta^{\nu}\wedge{*}\Omega_{\mu\nu} \qquad \text{(XIII.145)}$$

then we find that since θ^{μ} is reciprocal to $\mathbf{e}_{\mu}$ (XIII.142) gives:

$$\theta^{\mu}\wedge\theta^{\nu}\wedge{*}\Omega_{\mu\nu} = -\,(R^{\mu}_{\mu} - \tfrac{1}{2}R\delta^{\mu}_{\mu})\,V = R\,V, \qquad \text{(XIII.146)}$$

which is precisely the *Einstein-Hilbert* Lagrangian for the gravitational field that is described by the Lorentzian metric g.

Thus, we see that much of the geometric information that pertains to the physics of gravity is concisely summarized in the forms Ω^{μ}_{ν}, $\theta^{\mu}\wedge{*}\Omega_{\mu\nu}$, and $\theta^{\mu}\wedge\theta^{\nu}\wedge{*}\Omega_{\mu\nu}$.

4. General relativity in terms of complex orthogonal frames [1, 21, 22]. Now that we have discussed the isomorphism of $SO_0(3, 1)$ with $SO(3; \mathbb{C})$ and the corresponding association of oriented, time-oriented, Lorenztian frames in $T(M)$ with oriented, complex, orthogonal frames in $\Lambda^2(M)$, and presented general relativity in terms of connection 1-forms on the bundle $SO_0(3,1)(M) \to M$, it will be relatively straightforward to recast it in terms of the bundle $SO(3; \mathbb{C})(\Lambda) \to M$.

a. The isomorphism of $SO_0(3,1)(M)$ with $SO(3;\mathbb{C})(\Lambda)$. Since the Lie groups $SO_0(3,1)$ and $SO(3;\mathbb{C})$ are diffeomorphic as manifolds, the principal bundles $SO_0(3,1)(M)$ and $SO(3;\mathbb{C})(\Lambda)$ have diffeomorphic fibers; in fact, the bundles themselves are also isomorphic. The diffeomorphism of the fibers at a given $x \in M$ is not canonical, but simply takes an oriented, time-oriented, Lorentzian 4-frame $\mathbf{e}_\mu$ to an oriented, complex, orthogonal 3-frame Z_i. When both objects are local frame fields on $U \subset M$, one has:

$$Z_i = \tfrac{1}{2} Z_i^{\mu\nu}(x)\mathbf{e}_\mu \wedge \mathbf{e}_\nu . \tag{XIII.147}$$

For instance, the choice that we have been making all along has:

$$Z_i^{0j} = \delta_i^j , \qquad Z_i^{jk} = \varepsilon^{ijk}. \tag{XIII.148}$$

There is a corresponding isomorphism of the bundles of coframes, $SO_0(3,1)^*(M)$ and $SO(3;\mathbb{C})^*(\Lambda)$, that takes a coframe θ^μ to a coframe Z^i with:

$$Z^i = \tfrac{1}{2} Z^i_{\mu\nu}(x)\theta^\mu \wedge \theta^\nu \tag{XIII.149}$$

locally.

Since the bundle $SO_0(3,1)(M)$ is isomorphic to the bundle $SO_0(3,1)^*(M)$ by the association of any frame with its reciprocal coframe, and similarly for the bundles $SO(3;\mathbb{C})(\Lambda)$ and $SO(3;\mathbb{C})^*(\Lambda)$, one has the isomorphism of all four bundles.

b. Canonical 1-form on $SO(3;\mathbb{C})(\Lambda)$. Just as the frame bundle $GL(M)$ has a canonical real-valued 1-form θ^μ defined on it, the complex frame bundle $GL_{\mathbb{C}}(\Lambda^2)$ has a canonical complex-valued 1-form Z^i defined on it. One finds that the origin of either canonical form is based in the idea that one can regard a frame E_i, $i = 1, \ldots, n$ in a vector space V over a field of scalars $\mathbb{K}$ as a linear isomorphism $E_i : \mathbb{K}^n \to V$, $(v^1, \ldots, v^n) \mapsto v^i E_i$, while a coframe is a linear isomorphism $E^i : V \to \mathbb{K}^n$, $\mathbf{v} \mapsto v^i = E^i(\mathbf{v})$. Hence, a frame and a coframe are reciprocal iff these isomorphisms are inverse to each other.

If $E \to M$ is a vector bundle over M whose fiber is isomorphic to $\mathbb{K}^n$ then any frame E_i in E_x is associated with a set of set of n 1-forms E^i in $\Lambda^1_x(M)$. The canonical 1-form E^i on the bundle $\pi: GL(E) \to M$ of E-frames is obtained by pulling back the E^i along π:

$$E^i(\mathbf{v}) = \pi^* E^i(\mathbf{v}) = E^i(\pi_*\mathbf{v}). \tag{XIII.150}$$

Again, the reason that we are abusing notation by denoting both the canonical 1-form and the coframe with the same symbol is because when one chooses a local E-frame field over $U \subset M$ the canonical 1-form pulls down to the local reciprocal coframe field.

In the case where the elements of E are complex 3-frames in $\Lambda_2 M$, the canonical complex-valued 1-form Z^i on $GL_{\mathbb{C}}(\Lambda_2)$ pulls down to a set of three local real *2-forms* Z^i on $U \subset M$ by way of a local section Z_i: $U \to GL_{\mathbb{C}}(\Lambda_2)$. The 2-forms Z^i can then be expressed with respect to a local coframe field on U by expressions of the form (XIII.149). One way of seeing how the 1-form on $GL_{\mathbb{C}}(\Lambda_2)$ turns into a 2-form on U is to note that the vectors of the fibers of $\Lambda_2 M$ are actually *bivectors* to begin with. Hence, linear functionals on such vectors are 2-forms.

The canonical 1-form Z^i on $GL_{\mathbb{C}}(\Lambda_2)$ restricts to a canonical 1-form on any G-reduction of $GL_{\mathbb{C}}(\Lambda_2)$ where G is a Lie subgroup of $GL(3; \mathbb{C})$. Similarly, it induces a corresponding canonical 1-form on the coframe bundle $GL^*(3; \mathbb{C})$ by the isomorphism of those two bundles.

c. Connections on $GL_{\mathbb{C}}(M)$ and its reductions. The reason for introducing a connection on the bundle $GL_{\mathbb{C}}(M) \to M$ is analogous to the reason for introducing one on $GL(M) \to M$. That is, if $\mathbf{F}(x(\tau)) = F^i Z_i$ is a complex 3-vector along a differentiable curve $x(\tau)$ in $U \subset M$ one wishes to give some precise sense to the condition that this complex 3-vector be "parallel" along the curve. By differentiating with respect to τ, we get, with $\mathbf{v}(\tau) = dx/d\tau$:

$$\frac{d\mathbf{F}}{d\tau} = \frac{dF^i}{d\tau} Z_i + F^i (i_{\mathbf{v}} dZ_i) = \left(\frac{dF^i}{d\tau} + F^j \varpi^i_j(\mathbf{v}) \right) Z_i , \qquad \text{(XIII.151)}$$

in which:

$$\varpi^i_j = dZ_j \otimes_{\mathbb{C}} Z^i \qquad \text{(XIII.152)}$$

represents the generalized complex "angular velocity" of the 3-frame Z^i along the curve in question; here, we use the notation $\otimes_{\mathbb{C}}$ to indicate that we are concerned with the tensor product over the complex vector space that Z^i lives in, not the real vector space of 2-forms that it is associated with. One can see that it represents a 1-form on U with values in the Lie algebra $\mathfrak{gl}(3; \mathbb{C})$. Although one could propose to say that $\mathbf{F}$ is constant along $x(\tau)$ if this derivative vanishes, the condition would not be true for all choices of local frame field Z^i.

In order to make the vanishing of the derivative covariant under all changes of local frame fields, including ones for which the transition function is non-constant, one must replace the ordinary derivative with a covariant derivative that is defined by making a

choice of $\mathfrak{gl}(3;\mathbb{C})$-connection 1-form σ^i_j on the bundle $GL_{\mathbb{C}}(M) \to M$. Hence, this will be a 1-form on $GL_{\mathbb{C}}(M)$ with values in $\mathfrak{gl}(3;\mathbb{C})$ that is Ad^{-1}-equivariant under the right action of $GL(3;\mathbb{C})$ on complex 3-frames:

$$\sigma^i_j\big|_{Z_k A} = A(\sigma|_{Z_i})A^{-1} \tag{XIII.153}$$

and defines a linear isomorphism of each vertical tangent space on $GL_{\mathbb{C}}(M)$ with $\mathfrak{gl}(3;\mathbb{C})$ in such a way that if $a \in \mathfrak{gl}(3;\mathbb{C})$ and $\tilde{a}$ is the fundamental vector field on $GL_{\mathbb{C}}(M)$ that is associated with a then one has:

$$\sigma(\tilde{a}) = a \tag{XIII.154}$$

under this linear isomorphism.

If Z_i: $U \to GL_{\mathbb{C}}U$ is a local complex 3-frame field over U and Z^i is its reciprocal complex coframe field then σ can be expressed as a 1-form on U with values in $\mathfrak{gl}(3;\mathbb{C})$ by way of:

$$\sigma^i_j = \sigma^i_{\mu j}\theta^\mu, \tag{XIII.155}$$

in which the $\sigma^i_{\mu j}$ are smooth functions on U. Note that in general the choice of Z_i does not imply an unambiguous choice of θ^μ, although this will be true in the reduced case of oriented, complex orthonormal frames and oriented, time-oriented, Lorentzian frames.

The 1-forms σ^i_j transform to another coframe field $\bar{Z}^i = A^i_j Z^j$ by way of:

$$\bar{\sigma}^i_j = \tilde{A}^i_k \sigma^k_m A^m_j + \tilde{A}^i_k dA^k_j. \tag{XIII.156}$$

One can reduce the bundle $GL_{\mathbb{C}}(\Lambda_2)$ to the bundle $SL_{\mathbb{C}}(\Lambda_2)$ by defining a complex volume element $\mathcal{V}$ on $\Lambda_2 M$, which can also be defined as the 3-form on $GL_{\mathbb{C}}(\Lambda_2)$:

$$\mathcal{V} = \frac{1}{3!}\varepsilon_{ijk} Z^i \perp Z^j \perp Z^k, \tag{XIII.157}$$

in which Z^i is the canonical 2-form; again, the notation $\perp$ is used to distinguish the exterior algebra over the complex vector space in which one finds Z^i, not the real vector space of 2-forms that the Z^i are associated with.

A $\mathfrak{gl}(3;\mathbb{C})$-connection σ on $GL_{\mathbb{C}}(\Lambda_2)$ is reducible to an $\mathfrak{sl}(3;\mathbb{C})$-connection iff the exterior covariant derivative of $\mathcal{V}$ using σ vanishes:

$$\nabla\mathcal{V} = d\mathcal{V} + \sigma\mathcal{V} = \mathrm{Tr}(\sigma)\mathcal{V}. \qquad \text{(XIII.158)}$$

Hence, the connection is reducible iff the trace of the matrix σ^i_j vanishes identically.

One can define a complex orthogonal structure γ on $\Lambda_2 M$ by means of an equivariant 0-form γ_{ij} on $GL_{\mathbb{C}}(\Lambda_2)$ with values in the homogeneous space $GL(3;\mathbb{C})/O(3;\mathbb{C})$:

$$\gamma(Z_i) = \gamma_{ij}\, Z^i \otimes_{\mathbb{C}} Z^j. \qquad \text{(XIII.159)}$$

The reduction is then achieved by restricting to all complex orthonormal 3-forms, which will then have $\gamma_{ij} = \delta_{ij}$.

The $\mathfrak{gl}(3;\mathbb{C})$-connection σ is reducible to an $\mathfrak{o}(3;\mathbb{C})$-connection iff the tensor field γ is covariantly constant:

$$0 = \nabla\gamma_{ij} = d\gamma_{ij} - \sigma^k_i\gamma_{kj} - \sigma^k_j\gamma_{ki}\,. \qquad \text{(XIII.160)}$$

For a complex orthonormal frame, this says that:

$$0 = \sigma_{ij} + \sigma_{ji}\,; \qquad \text{(XIII.161)}$$

i.e., σ_{ij} is anti-symmetric.

Of course, in order for σ to be an $\mathfrak{so}(3;\mathbb{C})$-connection, the matrix σ must satisfy both conditions that it have vanishing trace and satisfy (XIII.160).

c. Curvature and torsion. Now let G be a Lie subgroup of $GL(3;\mathbb{C})$ and let $\mathfrak{g}$ be its Lie algebra. Suppose furthermore that σ is a $\mathfrak{g}$-connection on $GL_{\mathbb{C}}(\Lambda^2)$.

Since Z^i can be interpreted as either a 1-form on $GL_{\mathbb{C}}(\Lambda^2)$ with values in $\mathbb{C}^3$ or a local 2-form on some $U \subset M$, we need to clarify what it means to define the torsion of a $\mathfrak{g}$-connection σ on $GL_{\mathbb{C}}(\Lambda^2)$ to be the "exterior covariant derivative" of the canonical 1-form:

$$\Psi^i = \nabla^{\wedge} Z^i = dZ^i + \sigma^i_j \wedge Z^j\,. \qquad \text{(XIII.162)}$$

In particular, one needs to explain what its two terms represent.

It seems simplest to use the local interpretation, since it is inevitable that physics will require the necessity of going to local expressions. Hence, we shall regard Z^i as a 2-form on U with values in $\mathbb{C}^3$ that is given by (XIII.149), so Ψ^i will be a 3-form with values in $\mathbb{C}^3$, along with both of the terms in its definition. This makes:

$$dZ^i = \tfrac{1}{2}[dZ^i_{\mu\nu} - Z^i_{\kappa\nu}c^{\kappa}_{\lambda\mu}\theta^{\lambda}] \wedge \theta^{\mu} \wedge \theta^{\nu}\,, \qquad \text{(XIII.163)}$$

in which we have substituted the appropriate expression for $d\theta^\mu$, and:

$$\sigma^i_j \wedge Z^j = \tfrac{1}{2}\sigma^i_{[\kappa j} Z^j_{\mu\nu]}\, \theta^\kappa \wedge \theta^\mu \wedge \theta^\nu\,, \qquad \text{(XIII.164)}$$

in which the [.] notation implies that we have completely anti-symmetrized the components $\kappa\mu\nu$:

$$\sigma^i_{[\kappa j} Z^j_{\mu\nu]} = \tfrac{1}{3}(\sigma^i_{\kappa j} Z^j_{\mu\nu} + \sigma^i_{\mu j} Z^j_{\nu\kappa} + \sigma^i_{\nu j} Z^j_{\mu\kappa})\,. \qquad \text{(XIII.165)}$$

Combining expressions, we see that the local form of the torsion 3-form is:

$$\Psi^i = \tfrac{1}{2}[\mathbf{e}_\kappa Z^i_{\mu\nu} + \sigma^i_{[\kappa j} Z^j_{\mu\nu]} - Z^i_{\lambda\nu} c^\lambda_{\kappa\mu}]\theta^\kappa \wedge \theta^\mu \wedge \theta^\nu\,; \qquad \text{(XIII.166)}$$

that is, its components with respect to this local coframe field are:

$$\Psi^i{}_{\kappa\mu\nu} = \mathbf{e}_\kappa Z^i_{\mu\nu} + \sigma^i_{[\kappa j} Z^j_{\mu\nu]} - Z^i_{\lambda\nu} c^\lambda_{\kappa\mu}\,. \qquad \text{(XIII.167)}$$

Hence, the vanishing of torsion becomes an algebraic condition on σ.

Two simplifying special cases present themselves: local coframe fields in which the $Z^i_{\mu\nu}$ are constant, which we call *canonical* coframe fields, and anholonomic coframe fields, for which the structure functions $c^\kappa_{\mu\nu}$ vanish. In a holonomic canonical coframe field the vanishing of torsion takes the form of the algebraic identity:

$$0 = \sigma^i_{[\kappa j} Z^j_{\mu\nu]}\,. \qquad \text{(XIII.168)}$$

In the case of curvature, since only σ is involved, if we express it locally as a 1-form on U with values in $\mathfrak{g}$ then the definition of its exterior covariant derivative is more conventional. It is a 2-form on U with values in $\mathfrak{g}$:

$$\Sigma^i_j = \nabla\hat{}\sigma^i_j = d\sigma^i_j + \sigma^i_k \wedge \sigma^k_j\,, \qquad \text{(XIII.169)}$$

which has the components:

$$\Sigma^i_{j\mu\nu} = \mathbf{e}_\mu \sigma^i_{\nu j} - \mathbf{e}_\nu \sigma^i_{\mu j} + \sigma^i_{\mu k}\sigma^k_{\nu j} - \sigma^i_{\nu k}\sigma^k_{\mu j}\,. \qquad \text{(XIII.170)}$$

The Bianchi identities work the same way as before:

$$\nabla\hat{}\Psi^i = \nabla\hat{}\nabla\hat{}Z^i = \Sigma^i_j \wedge Z^j, \qquad \text{(XIII.171a)}$$

$$\nabla\hat{}\Sigma^i_j = \nabla\hat{}\nabla\hat{}\sigma^i_j = 0. \qquad \text{(XIII.171b)}$$

Of course, the expression $\nabla\hat{}\Psi^i$ is locally a 4-form, this time.

d. Penrose-Debever decomposition of the Riemann curvature tensor. The curvature 2-form Σ^i_j, at least when it is considered locally, can be regarded as a linear map Σ_x :

$(\Lambda_2)_x M \to \mathfrak{so}(3;\mathbb{C})$, $\mathbf{a} \wedge \mathbf{b} \mapsto [\Sigma_x]^i_j(\mathbf{a},\mathbf{b})$, from one real vector space of dimension six to another; hence, $\Sigma_x \in \Lambda^2_x M \otimes_{\mathbb{R}} \mathfrak{so}(3;\mathbb{C})$. Furthermore, both real vector spaces are assumed to be given complex structures, which are defined by * in the case of $\Lambda^2_x M$ and i in the case of $\mathfrak{so}(3;\mathbb{C})$.

i. The space of curvature tensors. The basis for the Penrose-Debever decomposition of the Riemann curvature tensor is that since $\Lambda^2_x M$ and $\mathfrak{so}(3;\mathbb{C})$ have the same dimension as vector spaces – whether real or complex – one can identify their elements by choosing a basis for each. Of course, this is somewhat naïve, as far as the algebraic structures on the two vector spaces are concerned, and comes down to the fact that one can associate any element $\omega^\mu_\nu \in \mathfrak{so}(p, q)$ in a matrix representation of an orthogonal Lie algebra with an anti-symmetric matrix $\omega_{\mu\nu} = g_{\mu\kappa}\omega^\kappa_\nu$, regardless of the signature type (p, q) of the metric $g_{\mu\nu}$.

In particular, since the Riemann curvature tensor Ω^μ_ν takes its values in $\mathfrak{so}(3, 1)$ in the case of general relativity, one can associate its value $\Omega^\mu_\nu(\mathbf{v}_x, \mathbf{w}_x)$ when applied to tangent vectors $\mathbf{v}_x$, $\mathbf{w}_x$ at a point $x \in M$ with a 2-form:

$$\Omega(\mathbf{v}_x, \mathbf{w}_x) = \tfrac{1}{2}\Omega_{\mu\nu}(\mathbf{v}_x, \mathbf{w}_x)\, dx^\mu \wedge dx^\nu, \qquad \text{(XIII.172)}$$

by setting:

$$\Omega_{\mu\nu}(\mathbf{v}_x, \mathbf{w}_x) = g_{\mu\nu}(x)\,\Omega^\mu_\nu(\mathbf{v}_x, \mathbf{w}_x). \qquad \text{(XIII.173)}$$

Since Ω^μ_ν is a 2-form to begin with, this means that we can regard Ω as defining linear maps from each vector space $\Lambda_{2,x}M$ to the corresponding vector space $\Lambda^2_x M$; i.e., Ω is a section of $\Lambda^2 M \otimes \Lambda^2 M \to M$ that one can locally represent as:

$$\Omega = \tfrac{1}{4} R_{\kappa\lambda\mu\nu}(dx^\kappa \wedge dx^\lambda) \otimes (dx^\mu \wedge dx^\nu). \qquad \text{(XIII.174)}$$

In fact, from the symmetry of $R_{\kappa\lambda\mu\nu}$ in the first and last index pairs, one can write this as:

$$\Omega = \tfrac{1}{4} R_{\kappa\lambda\mu\nu}(dx^\kappa \wedge dx^\lambda)(dx^\mu \wedge dx^\nu), \qquad \text{(XIII.175)}$$

with the symmetrized tensor product implied by the absence of a multiplication symbol.

If one goes the route of the other isomorphic copy of both $\Lambda^2_x M$ and $\mathfrak{so}(3;\mathbb{C})$ – namely, $\mathbb{C}^3$ – by way of the isomorphisms Z^i, then one can represent the Riemann curvature tensor (or really its value at any point $x \in M$) by means of an element Σ of $\mathbb{C}^3 \odot \mathbb{C}^3$:

$$\Sigma = \Sigma_{ij}\, Z^i\, Z^j, \qquad \Sigma_{ij} = \gamma_{ik}\Sigma^k_j\,. \tag{XIII.176}$$

Here, we can pause to point out that even though the Euclidian metric on $\mathbb{C}^3$ that is defined by γ_{ij} can be obtained by first defining a Lorentzian structure $g_{\mu\nu}$ on $T(M)$, actually, that definition is not *necessary*. It is *sufficient* to define a linear electromagnetic constitutive tensor $\kappa \in \Lambda^2 M \otimes \Lambda^2 M$ and then obtain the Euclidian scalar product by means of the association:

$$\gamma_{ij} = \tfrac{1}{4}\,\kappa_{\kappa\lambda\mu\nu}\, Z_i^{\kappa\lambda} Z_j^{\mu\nu}\,. \tag{XIII.177}$$

Note that under symmetrization the skewon part of κ will not contribute to γ_{ij}.

Therefore, just as the observation that the Lorentzian metric contributed to the Maxwell equations only by way of the isomorphism * represented the starting point for pre-metric electromagnetism, so does this latter association represent the starting point for "pre-metric gravitation." Of course, what this would really mean is not that there is no metric on anything involved, but only that the more fundamental metric is a complex Euclidian metric on the bundle of 2-forms over spacetime, not a Lorentzian metric on the bundle of tangent vectors. Moreover, the fundamental character of the metric γ_{ij} is then purely electromagnetic in origin.

ii. The double dual operator. When one is given two real vector spaces V and W that have complex structures * and $\overline{*}$ defined on them, a natural problem to investigate is the way that a given real-linear map, such as Σ_x, affects the two complex structures.

Indeed, the basic question is whether an $\mathbb{R}$-linear map A: $V \to W$ between two real vector spaces V and W that have been given complex structures * and $\overline{*}$ also defines a $\mathbb{C}$-linear map. The necessary and sufficient condition for this is that A must commute with both the isomorphisms * and $\overline{*}$:

$$A* = \overline{*}A. \tag{XIII.178}$$

We can then define an $\mathbb{R}$-linear map $\overline{(\cdot)}$: $\mathrm{Hom}_{\mathbb{R}}(V, W) \to \mathrm{Hom}_{\mathbb{R}}(V, W)$, $A \mapsto \overline{A}$, where:

$$\overline{A} = \overline{*}A*, \tag{XIII.179}$$

which we refer to as the *double dual* operator.

Since $*^2 = -I$ and $\overline{*}^2 = -I$, one sees that this map $\overline{(\cdot)}$ is then an involutory isomorphism on $\mathrm{Hom}_{\mathbb{R}}(V, W)$:

$$\overline{\overline{A}} = A. \tag{XIII.180}$$

This also says that the eigenvalues of the map $\overline{(\cdot)}$ are ± 1. Indeed, the real vector space $\mathrm{Hom}_{\mathbb{R}}(V, W)$ splits into a direct sum $\mathrm{Hom}_+ \oplus \mathrm{Hom}_-$ of two subspaces of equal dimension that correspond to the positive and negative eigenvalues, respectively. In fact, the map that takes A to $A_+ + A_-$ is just polarization using the involution $\overline{(\cdot)}$:

$$A_+ = \tfrac{1}{2}(A + \overline{A}), \qquad A_- = \tfrac{1}{2}(A - \overline{A}). \tag{XIII.181}$$

It is traditional to call the elements of the subspaces Hom_+ and Hom_- *self-dual* and *anti-self-dual*, resp., but we can see from (XIII.178) that the elements of Hom_- represent $\mathbb{C}$-linear maps, while the elements of Hom_+ are $\mathbb{C}$-antilinear. One must note that the signs behave in the opposite manner to one's intuition regarding $\mathbb{C}$-linearity.

The way this double dual map is defined in most of the literature of complex relativity is to consider the space of curvature tensors as sections of the vector bundle $\Lambda^2 M \otimes \mathfrak{so}(3; \mathbb{C}) \to M$ and give the fibers of the factors $\Lambda^2 M$ and $\mathfrak{so}(3; \mathbb{C})$ the complex structures defined by * and i, respectively.

One then defines:

$$\overline{Z}^i = i{*}Z^i, \qquad (\text{so } {*}Z^i = -i\overline{Z}^i) \tag{XIII.182}$$

which can be interpreted as meaning that * takes the 2-form Z^i to its dual 2-form and i takes its value $Z^i(\mathbf{b})$ in $\mathfrak{so}(3; \mathbb{C})$, when evaluated on a bivector $\mathbf{b}$ to the complex 3-vector $iZ^i(\mathbf{b})$. When $\overline{Z}^i = \pm Z^i$ one has ${*}Z^i = \mp iZi$, which makes the self-dual (anti-self-dual, resp.) 2-forms take the form of eigenvectors of * with eigenvalue $-i$ ($+i$, resp.).

iii. The Einstein equations. In order to account for the symmetry of $R_{\kappa\lambda\mu\nu}$ in its first last index pairs, we see that it becomes more convenient to represent Riemann curvature tensors as elements of either $\Lambda^2 M \odot \Lambda^2 M$ or $\mathbb{C}^3 \odot \mathbb{C}^3$, which we then refer to as the *real* and *complex* representations, respectively.

In the complex representation, the decomposition into self-dual and anti-self-dual parts allows one to express Σ in the form:

$$\Sigma = C'_{ij} Z^i Z^j + iE_{ij} Z^i \overline{Z}^j. \tag{XII.183}$$

This means that the complex matrix C'_{ij} is symmetric and E_{ij} is Hermitian.

Furthermore, since we already have a symmetric second rank tensor on $\mathbb{C}^3$ that is defined by γ, which we express in the form:

$$\gamma = \gamma_{ij} Z^i Z^j, \tag{XII.184}$$

we can project Σ onto that one-dimensional subspace of $\mathbb{C}^3 \odot \mathbb{C}^3$ by taking its trace:

$$\mathrm{Tr}(\Sigma) = \Sigma^i_i = \gamma^{jk}\Sigma_{ki} = \mathrm{Tr}(C'). \tag{XII.185}$$

The traceless part of C'_{ij} is then denoted by:

$$C_{ij} = C'_{ij} - \tfrac{1}{3}\mathrm{Tr}(\Sigma)\,\gamma_{ij}\,. \tag{XII.186}$$

Therefore, we have the complex representation of the *Penrose-Debever decomposition* of Σ into:

$$\Sigma = C + E + \tfrac{1}{2}(\mathrm{Tr}\,\Sigma)\gamma. \tag{XII.187}$$

This can be given a corresponding real form as a decomposition of $R_{\kappa\lambda\mu\nu}$ by means of the isomorphisms defined by Z^i. It can be shown (cf., e.g., [**21, 22**]) that:

$$\mathrm{Tr}\,\Sigma = \tfrac{1}{4}R\,, \tag{XII.188}$$

where R is the scalar curvature of Ω, and E_{ij} corresponds to:

$$E_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{4}Rg_{\mu\nu}, \tag{XII.189}$$

which is the traceless part of the Ricci curvature tensor.

Hence, although this differs from the Einstein tensor by a term equal to $-\tfrac{1}{4}Rg_{\mu\nu}$, nevertheless, since the vacuum Einstein equations implied that R had to vanish, we see that the vacuum Einstein equations are equivalent to the equations:

$$E_{ij} = 0. \tag{XII.190}$$

The remaining part of Σ that is defined by C corresponds to the *Weyl curvature tensor* of the connection ω, which is also called the *conformal curvature tensor*. Its vanishing implies that the metric g is conformal to the Minkowskian one:

$$g_{\mu\nu} = \alpha^2\eta_{\mu\nu}\,, \tag{XII.191}$$

with α^2 as the conformal factor.

One should observe that so far we have obtained the complex form of only the vacuum Einstein equations. Hence, we need to extend the analysis give to the complex form of the Einstein equations with sources. This step then depends upon first finding the complex form of the energy-momentum tensor.

As pointed out by Krasnov [**23**], in the Plebanski [**24**] formulation of complex relativity one can define the complex form of the energy-momentum tensor T^μ_ν by first splitting it into its trace $T = T^\mu_\mu$ and its traceless part $\tilde{T}^\mu_\nu = T^\mu_\nu - \tfrac{1}{4}T\delta^\mu_\nu$ and then defining:

$$\tilde{T}^i_j = \tfrac{1}{4}\tilde{T}^\kappa_\mu Z^i_{\nu\kappa}\bar{Z}^{j\mu\nu}, \qquad \bar{Z}^{i\mu\nu} = \gamma^{ij}\bar{Z}^{\mu\nu}_j . \tag{XIII.192}$$

Note that since the T^μ_ν form of the energy-momentum tensor is pre-metric in its definition, and involves a mechanical constitutive law as the mechanism for associating tangent objects with cotangent ones, the metric is not introduced into the definitions until one uses γ^{ij}, which, as we pointed out above, can be defined by the electromagnetic constitutive law alone.

The Einstein equations, with the source term included then take the form:

$$E^{ij} = -2\pi G\tilde{T}^{ij}, \qquad \mathrm{Tr}(\Sigma) = -\tfrac{1}{4}\Lambda - 2\pi GT, \tag{XIII.193}$$

if one includes the cosmological constant Λ.

In order to make the comparison more immediate, one should use following real form of the Einstein equations:

$$R^\mu_\nu - \tfrac{1}{4}R\delta^\mu_\nu = 8\pi G\tilde{T}^\mu_\nu, \qquad R = -\Lambda - 8\pi GT, \tag{XIII.194}$$

in place of the usual formulation in terms of the divergenceless part of the Ricci tensor.

5. Discussion. In its conventional formulation (e.g., [**21, 22**]), complex relativity generally involves a slight redundancy in its basic definitions. In particular, although one assumes an almost-complex structure on Λ^2M, in the form of *, one also complexifies this bundle to $\Lambda^2M\otimes\mathbb{C}$ in order to treat the mathematical expression $*F = \pm iF$ as an eigenvalue equation that allows one to define a decomposition of $\Lambda^2M\otimes\mathbb{C}$ into a direct sum of self-dual and anti-self-dual 2-forms, which then correspond to the positive and negative imaginary eigenvalues, respectively.

However, since * defines an almost-complex structure on Λ^2M it would seem unnecessary to complexify it again. One need only regard the expression $*F = \pm iF$ as giving two possible *definitions* of how the imaginary unit i acts on the fibers of Λ^2M, and thus complex scalars, more generally.

The concept of self-duality reasserts itself in the context of the Debever-Penrose decomposition of the curvature 2-form. However, at that point one can observe that an $\mathbb{R}$-linear map from Λ_2M to either $\mathfrak{so}(3;\mathbb{C})$ or $\mathbb{C}^3$ can commute with the complex structures on both or not. Hence, we can define a (double) duality operator on $\Lambda^2M\otimes\mathfrak{so}(3;\mathbb{C})$ – or $\Lambda^2M\otimes\Lambda^2M$, for that matter – that allows one to define the aforementioned decomposition, just the same. One then finds that the anti-self-dual $\mathbb{R}$-linear maps are precisely the $\mathbb{C}$-linear maps.

Since the theme of this book all along has been that the electromagnetic structure of the spacetime manifold, as encoded in the constitutive map κ, implies the Lorentzian geometry as a consequence of the dispersion law that follows from κ by way of the field

equations, we first observe that even the primary emphasis in the relativistic theory of gravitation can be shifted from the Lorentzian metric $g_{\mu\nu}$ on tangent vectors to the complex Euclidian metric γ_{ij} in 2-forms. Since this metric can be defined in terms of only κ, one sees that in a sense one can also define a sort of "pre-metric gravitation" as well as pre-metric electromagnetism. Of course, as we have observed a number of times, there is a difference between the algebraic sort of metric structure that comes from κ directly and the differential sort that comes from defining field equations that involve κ and passing to the symbol of the differential operator $\Box_\kappa$ that defines them.

A deep question then arises whether perhaps the field κ, which is like a generalization of the metric tensor, might be itself subject to field equations. One sees that in order to make physical sense of this coupling one would have to go into deeper detail about the nature of the medium in question and how electric polarizability and magnetization can come about in it. If one is dealing with macroscopic media, such as optical ones, this resolution of the macroscopic picture to a microscopic one is generally more straightforward – e.g., crystal lattices, electron orbitals – than when one is addressing the electromagnetic vacuum state of quantum electrodynamics. In that case, one must mostly rely upon effective models, such as the Heisenberg-Euler model.

Of course, if one is attempting to account for the gravitational fields of astronomical bodies then one must realize that the energy, momentum, and stress that couple to the spacetime metric in the form of κ must be essentially affecting the state of electric and magnetic polarization of the vacuum of space, at least in the immediate neighborhood of elementary massive matter, in order to produce a spacetime metric that is not merely obtained from a linear, isotropic, homogeneous electromagnetic constitutive law that is based on the constants ε_0 and μ_0. Of course, this is precisely what one gets from the Heisenberg-Euler dispersion law, which attributes the perturbation of the Minkowski metric in a region of spacetime to the Faraday stress-energy-momentum tensor of an electromagnetic field that permeates it. Although the electric and magnetic field strengths are generally close to the critical values only in the small neighborhoods of elementary charge distributions, nevertheless, one can still think of those distributions as the ultimate sources of gravitational fields, as well. Perhaps the gravitational field of a star or planet is then best viewed as the macroscopic effect of a large number of microscopic perturbations to the metric at the elementary level due to vacuum polarization.

The main objective of this chapter was therefore to embed the usual Lorentzian geometry of general relativity in the framework of the geometry that pertains to the bundle of 2-forms instead of the tangent bundle. At this point, one can only speculate on what form the field equations of κ might take, although undoubtedly the answer to that problem will probably follow from a better understanding of the role that connections on the bundle of frames in $\Lambda^2 M$ and its reductions and their curvatures take in physics.

References

1. D. H. Delphenich, "Complex geometry and pre-metric electromagnetism," arXiv gr-qc/0412048.
2. H. K. Nickerson, D. C. Spencer, and N. E., Steenrod, *Advanced Calculus*, Van Nostrand, Princeton, 1959.

3. D. H. Delphenich, "A more direct representation for complex relativity," Ann. Phys. (Leipzig) **16**, No. 9 (2007), 615-639.
4. J. D. Jackson, *Classical Electrodynamics*, 2nd ed., Wiley, New York, 1976.
5. D. H. Delphenich, "Hermitian structures defined by linear electromagnetic constitutive laws," arXiv:0710.5156.
6. W. Thirring, *Classical Field Theory*, Springer, Berlin, 1978.
7. T. Frankel, *The Geometry of Physics: an introduction*, Cambridge University Press, Cambridge, 1997.
8. A. Lichnerowicz, *Théorie relativiste de la gravitational et de l'electromagnetisme*, Masson and Co., Paris, 1955.
9. A. Lichnerowicz, *Global theory of connections and holonomy*, Noordhoff, Leyden, 1976.
10. A. Trautmann, "On the structure of the Einstein-Cartan equations," Symp. Math. **12** (1973), 139-162.
11. F.W. Hehl, J. D. McCrea, E.W. Mielke, Y. Ne'eman, "Metric-Affine Theories of Gravitation," Phys. Rep., (1995).
12. T. Eguchi, P. Gilkey, and A. Hanson, "Gravitation, Gauge Theories, and Differential Geometry," Phys Rep. **66** (1980), 213-393.
13. L. Markus, "Line Elements Fields and Lorentz Structures on Differentiable Manifolds," Ann. Math., 62 (1955), 411-417.
14. N. Steenrod, *The Topology of Fiber Bundles*, Princeton Univ. Press, Princeton, 1951.
15. S. Sternberg, *Lectures on Differential Geometry 2nd ed.*, Chelsea, New York, 1983.
16. A. Fujimoto, *Theory of G-structures*, Publications of the Study Group of Geometry, 1972.
17. D. Barnard, "Sur la géométrie différentielle des G-structures," Ann. Inst. Fourier, Grenoble **10** (1960), 151-270.
18. D. H. Delphenich, "Spacetime G-structures I: topological defects," arXiv:gr-qc/0304016; "Spacetime G-structures II: geometry of the ground states," arXiv:gr-qc/0401089.
19. A. Einstein, Sitzungber. Preuss. Akad. Wiss. (1928), 217-221; (1929), 2-7; 156-159; (1930), 18-23; and W. Mayer, 110-120, 410-412.
20. E. Stiefel, "Richtungsfelder und Fernparallelismus in *n*-dimensionalen Mannigfaltigkeiten," Comm. Math. Helv., **8** (1936), 3-51.
21. W. Israel, *Differential forms in general relativity*, Lecture notes, Dublin Institute for Advanced Studies, 1970.
22. R. Debever, "Le rayonnement gravitational," Cahiers de Physique **8** (1964), 303-349.
23. K. Krasnov, "Plebanski formulation of general relativity: a practical introduction," arXiv:0904.0423.
24. J. Plebanski, "On the separation of Einsteinian substructures," J. Math. Phys. **18** (1977), 2511.