Aspects of Electromagnetic Duality

D.I. Olive

University of Wales Swansea *

ABSTRACT: I briefly try to give a critical assessment of this idea, discussing what it means, the extent to which it is true, and how it relates to other ideas in mathematics and physics. Examples of related physical concepts concern solitons and integrability in space-times of two dimensions, supersymmetry, superstring theory, and so on. Relevant mathematical developments appear to be the Atiyah-Singer index theorems, homology theory, the theory of modular functions, and, maybe, Borcherds algebras.

KEYWORDS: Electromagnetic duality, homology, modular group.

Claus Montonen and I were well aware of the similarity between our 1977 proposal of a principle of S-duality in the context of a spontaneously broken $SU(2)$ gauge theory in the space-time $\mathbb{R}^{3,1}$, [1], and the much earlier (1941) result of Kramers and Wannier in the context of the Ising model in $\mathbb{R}^2$ concerning a duality transformation between high and low temperatures, [2]. We also suspected that the S transformation was just one element of a larger discrete group, such as the modular group, but were unsuccessful in developing this idea.

Now we know that the modular group is indeed relevant in $\mathbb{R}^{3,1}$ but less so in $\mathbb{R}^2$. The progress that led to the present day level of understanding came much later, in several distinct steps, including recognition of the following points:
(1) the role of supersymmetry (in the broken $SU(2)$ gauge theory), [3]
(2) the significance of the θ angle in the $SU(2)$ gauge theory, [4]
(3) the hyperkähler nature of the metric on the relative moduli space of two like BPS monopoles, [5]
(4) Sen's synthesis of this and his vision of how the modular group acts on the totality of charged single particle states, [6].

I have reviewed all this in more detail elsewhere. As the texts are readily available I shall refrain from repeating myself [7].

The S-duality principle stemmed from rather old ideas of electromagnetic duality referring to the similarity between electric and magnetic fields in $\mathbb{R}^{3,1}$. This is why I shall refer to the extended principle as extended electromagnetic duality. It seems to me to furnish what is potentially an extremely important new principle, indeed one that transcends quantum theory. However it is only applicable to certain classes of QFT, precisely which ones we do not know, even though we see that it is relevant to many physically interesting examples of QFT *e.g.* supersymmetric gauge theories, supergravity theories and superstring theories.

The reason I had to use the above qualification "potentially" is that the principle lacks any real proof. Indeed, because of its nature, it looks as if a proof would first require a concrete construction of the quantum field theory in question, an impossible task within the present level of mathematical knowledge. Nevertheless the nature of the support for the validity of the idea is rather encouraging as it stems from the application of new and powerful mathematical ideas. Indeed the eventual proof could well depend upon branches of mathematics yet to be discovered, whose creation could be triggered by the physical ideas associated with duality.

I have not yet tried to state exactly what the new principle is and that is because, in my opinion, a satisfactory formulation has yet to be found. Indeed, I have to say that much of what

I have found in the existing literature seems to me to be confusing or maybe worse. As a consequence, facilitated by the TMR network, I have made my own attempt (together with Marcos Alvarez [8,9]) to clarify the ideas by analysing a "toy" model that seems to be the simplest situation in which the principle applies and, indeed, is very nearly capable of proof.

Before trying to formulate the duality statement I shall describe this toy setup. It is rather trivial from the particle physics point of view but nevertheless rather interesting mathematically, as E Verlinde and E Witten were the first to realise independently, in 1995 [10,11]. The mathematics that turns up is quite sophisticated by the standards of ordinary theoretical physics. Seeing how it works in connection with duality opens up a number of new questions, as we shall see.

The context is somewhat like that of a supergravity theory but space-time is not dynamical, just a fixed, curved background supporting a dynamical electromagnetic field. Nor is there any supersymmetry. It is not the curvature that is important but rather that the space-time is topologically complicated with sufficient "holes" to capture frozen magnetic fluxes. This information is encoded in the homology structure of space-time. I do not want to treat this in detail other than to say that it is precisely that branch of mathematics that was originally inspired by the concepts of Faraday and Maxwell concerning the electromagnetic field.

The electromagnetic field is assumed to be smooth and nonsingular. Since it is specified by an antisymmetric tensor it can conveniently be used to define a two-form $F = \frac{1}{2}F_{\mu\nu}dx^\mu \wedge dx^\nu$. Two of the four Maxwell equations imply that F is closed, $dF = 0$, where d is the exterior derivative. Let Σ be a two-dimensional surface embedded in space-time. Then we can define a flux of F through Σ, written simply as $\int_\Sigma F$.

If Σ is a closed two-surface, that is has no boundary, written $\partial\Sigma = 0$, and called a two-cycle, then the flux through it has remarkable invariance properties. It is unchanged either if Σ is altered to Σ' differing by the boundary of a three-surface or if F is altered by a total derivative. These are straightforward consequences of

Stokes' theorem. This means that the flux is a number associated with a (homology) class of two-cycles and a (cohomology) class of field strengths.

If Σ is itself a boundary of a three-surface, it is certainly closed. But then any flux through it must vanish. But if space-time is sufficiently complex this need not be the case and then Σ can capture a nonzero flux which is associated with the "hole" in space-time enclosed within Σ.

That the values of these fluxes cannot be arbitrary was pointed out by Dirac, in 1931, by taking account of the quantum theory that should apply to any electrically charged particles moving in space-time [12]. Suppose such a particle is spinless and carries electric charge q. Then it should have a complex scalar wave function $\phi(x)$ that can be defined all over space-time, at least in a piecemeal sense.

Now the gauge principle implies that ϕ couples to the gauge potential, A where $F = dA$. The field strength F is well defined all over space-time but A can only be obtained from it by integration. The result is unambiguous only if the integration is restricted to a topologically trivial region or neighbourhood. Likewise the wave function ϕ can also only be defined in such a neighbourhood. Space-time can be covered with such regions and when there is an overlap the two versions of A and ϕ should be capable of being related by a gauge transformation:

$$\phi(x) \to e^{\frac{iq\chi(x)}{\hbar}}, \qquad A \to A + d\chi. \qquad (1)$$

This procedure is only consistent if the fluxes satisfy certain constraints, the celebrated Dirac quantisation conditions [12]:

$$\frac{q}{2\pi\hbar}\int_\Sigma F = m(\Sigma) \qquad \in \qquad \mathbb{Z}. \qquad (2)$$

This version of the argument was due to Orlando Alvarez in 1985 [13].

Notice that electrically charged particles have entered in a surreptitious way, as has Planck's constant $\hbar$.

So far space-time has been assigned neither a metric nor even a dimension. Now we take the latter to be four and see that this implies something special, given the electromagnetic field F. For there is a natural closed four-form $F \wedge F$,

quadratic in F, defined on space-time. Its integral over space-time should be something topological (though different from the instanton number it would be if the field were non abelian instead of abelian). In fact the integral can be evaluated to be quadratic in the fluxes already introduced:

$$\left(\frac{q}{2\pi\hbar}\right)^2 \int F \wedge F = \sum_{i,j=1}^{b_2} m(\Sigma_i)Q_{ij}m(\Sigma_j), \quad (3)$$

if it is assumed that space-time is Poincaré dual, that is closed, compact and orientable. The classes of two-cycles mentioned above can be added in a perfectly natural way and, as a consequence, form a lattice, whose dimension defines the second Betti number, b_2. $\Sigma_1, \Sigma_2, \ldots \Sigma_{b_2}$ form a basis for this lattice and the number $(Q^{-1})_{ij}$ is the intersection number of Σ_i with Σ_j. This is well defined precisely because of the sum $2 + 2 = 4$, the dimension of space-time. So $(Q^{-1})_{ij}$ is an integer valued matrix. It is also symmetric, and more remarkably has determinant ± 1, because of Poincaré duality. Hence the matrix Q_{ij} occurring above also has integer entries and the expression (3) above is quantised too.

One role for expression (3) is as a possible contribution to the action. Since it is the integral of a total derivative it would leave the classical equations of motion unaffected and only affect quantum amplitudes. Indeed when multiplied by a coefficient denoted $\theta\hbar/2$ it yields what is known as the theta term.

To specify the action fully it is necessary to introduce a Minkowski metric on space-time (thereby limiting its possible topology). With this it is possible to define a Hodge star operation mapping p-forms onto $4 - p$-forms. This therefore maps 2-forms onto themselves and it does so in a way which is unaltered if the metric $g_{\mu\nu}(x)$ is rescaled:

$$g_{\mu\nu}(x) \rightarrow \lambda(x)g_{\mu\nu}(x), \quad (4)$$

i.e. a Weyl transformation. In this notation, the conventional Maxwell term in the action reads as $\frac{1}{2}\int F \wedge *F$, making altogether the free action

$$\frac{\hbar}{4\pi}\left(\frac{q}{\hbar}\right)^2 \int F \wedge \left(\frac{\theta}{2\pi} + \frac{2\pi\hbar}{q^2}*\right) F.$$

This can be simplified by introducing the "geometric" version of the field strength, $f = qF/\hbar$, rescaled to have dimensionless fluxes according to (2). Then the quantum phase factor

$$e^{i\frac{W_{MINKOWSKI}}{\hbar}} = e^{\frac{i}{4\pi}\int f\wedge\hat{\tau} f}, \quad (5)$$

where

$$\hat{\tau} = \frac{\theta}{2\pi} + \frac{2\pi\hbar}{q^2}*. \quad (6)$$

The metric has Minkowski signature, $** = -1$, and so $*$ acts like an imaginary unit. Accordingly we can define a complex variable τ by (6) with $*$ replaced by i. Notice this variable is intrinsically dimensionless and that its imaginary part is essentially the inverse of the fine structure constant, and hence positive.

Evidently the transformation $\theta \rightarrow \theta + 2\pi$ corresponds to $\tau \rightarrow \tau + 1$ and is a symmetry of the quantum phase (5), providing the matrix Q in (3) has even integer entries on the diagonal. If not, Q is said to be odd, and the phase (5) has the lesser symmetry $\theta \rightarrow \theta + 4\pi$ or $\tau \rightarrow \tau + 2$.

We are now ready to start discussing the electromagnetic duality transformations.

The Euler-Lagrange equations resulting from varying the gauge potential in the action $W_{MINKOWSKI}$ in (5) is simply $d(\hat{\tau} f) = 0$. Since we also have $df = 0$ the following transformations provide symmetries of the equations of motion, transporting solutions into solutions:

$$\hat{\tau}f \rightarrow A\hat{\tau} f + B f \quad (7a)$$

$$f \rightarrow C\hat{\tau} f + D f, \quad (7b)$$

where the coefficients A, B, C and D are simply real numbers [14,15]. The field strength can be eliminated between these two equations yielding a fractional linear transformation of the complex dimensionless coupling τ:

$$\tau \rightarrow \frac{A\tau + B}{C\tau + D}. \quad (8a)$$

Before discussing what this can mean, let us note that it implies that τ_2, the imaginary part of τ and the inverse of the fine structure constant, transforms as

$$\tau_2 \rightarrow \frac{(AD - BC)\tau_2}{|C\tau + D|^2}. \quad (8b)$$

Since τ_2 is intrinsically positive the determinant $AD - BC$ should be positive, and in fact unity, for reasons to be seen below. Let us accept this and look at the subgroup of the $SL(2, \mathbb{R})$ group of transformations leaving τ unchanged. Obviously the fine structure constant is unchanged, so, by (8b), $C\tau + D$ is a phase. This means that there exists an angle ϕ for which $C\tau_1 + D + iC\tau_2 = e^{i\phi} = \cos\phi + i\sin\phi$, and we have an $SO(2)$ subgroup. Hence inserting in (7b),

$$f \to f\cos\phi + *f\sin\phi.$$

If the Minkowski metric is flat we recognise this as describing electromagnetic duality rotations through ϕ between the electric and magnetic fields. Rotation through $\pi/2$ supplies the original electromagnetic duality transformation.

It is very familiar that these electromagnetic duality rotations that do not affect the coupling τ leave the energy momentum tensor invariant. We now consider the effect of the more general transformations (7) that do change τ. Since the energy momentum tensor $\theta^{\mu\nu}$ is proportional to the response of the action to a variation of the metric $g^{\mu\nu}$, it cannot depend on the angle θ and hence τ_1, since expression (3) is topological. In fact

$$\theta_{\mu\nu}/\hbar = \frac{\tau_2}{4\pi}\left(f_{\mu\lambda}f_{\sigma\nu} + *f_{\mu\lambda}*f_{\sigma\nu}\right)g^{\lambda\sigma}$$

in terms of the geometric field strengths, f. The effect of the transformations (7) is to produce an overall factor $AD - BC$. Since energy is intrinsically positive, this factor again has to be positive. The demand that the energy-momentum tensor be invariant requires the factor to be unity so that the transformations form the group $SL(2, \mathbb{R})$. Notice that to achieve this when τ is not fixed, the transformations have to be applied to the geometric field strengths f rather than the original physical ones F.

Evidently, if τ is altered, the transformations (7) provide not a symmetry of a given theory but rather relationships between similar theories at different values of the coupling. But there is something fishy for it is the dimensionless coupling that is altered and this involves Planck's constant according to (6). So it is not a classical effect that is at stake but rather a quantum effect. We should therefore be be taking account of

quantum mechanics, for example the Dirac quantisation conditions (2).

Indeed we have already met the transformation T sending τ to $\tau + 1$ which is sometimes a symmetry of the quantum phase (5), when account is taken of of (2) and (3). It corresponds to the $SL(2, \mathbb{R})$ matrix $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

The S transformation of Montonen and myself corresponded to the matrix $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ sending τ to $-1/\tau$ and very dramatically interchanging strong and weak coupling regimes. In the supersymmetric $SU(2)$ gauge theory it was realised by interchanging a magnetically charged soliton, the 't Hooft-Polyakov monopole [16,17], with a gauge particle. This is the relation between duality, solitons and integrability, all themes of our TMR network.

Now S and T generate an infinite subgroup of $SL(2, \mathbb{R})$, denoted $SL(2, \mathbb{Z})$, and essentially the modular group. How can the action of S be realised in the toy model I have beeen describing? Certainly not as a symmetry of the action or even the quantum phase (5), but maybe it is possible in terms of some physically observable quantities such as the energy, H.

Unfortunately this cannot be constructed explicitly but the partition function, $Z(\tau) = Tr(e^{-H})$, can. According to Feynman's ideas this emphatically Minkowski metric quantity can be reexpressed as a Feynman path integral involving an action which is Euclidean rather than Minkowski because of the imaginary time implied [18]. Furthermore, the trace implies periodicity in this time and the effective four-dimensional spacetime in the action therefore factorises into a circle S^1 times a 3-manifold. It is also possible to consider an extended partition function (which is no longer real) and involves any (Poincaré dual) four-manifold with a Euclidean metric. It could even have non-vanishing Euler number and consequently lack any Minkowski metric and, indeed, any obvious physical interpretation.

The effect of the Dirac quantisation conditions (2) is an unexpected simplification of the evaluation of the path integral. The field configurations integrated over fall into topological sectors labelled by the flux quanta m. By a

theorem due to Hodge, there is just one solution to Maxwell's equations in each sector. This makes it easy to apply the semi-classical approximation which is essentially exact. The result is a sum over the integers, m, of an exponential of a quadratic in m. This is a generalised sort of theta function associated with the unimodular matrix Q (and the metric). For example, if four-dimensional space-time is the complex projective space $CP(2)$, the Betti number b_2 is just one and the result is simply proportional to the familiar theta function associated with the unimodular lattice consisting of integers, $\mathbb{Z}$, namely $\sum_{n \in \mathbb{Z}} e^{i\pi\tau n^2}$. This transforms nicely under S because of the Poisson summation formula and so do all the other theta functions because the matrix Q is unimodular by virtue of Poincaré duality.

When the matrix Q is even, the previous argument suffices to show that the partition function is invariant under T and hence transforms nicely under the whole modular group acting on the coupling τ. Thus quantum effects do break the $SL(2, \mathbb{R})$ group action (8a) on the coupling τ to the discrete subgroup given by restricting to integer entries. Division by the invariant subgroup consisting of plus or minus the unit matrix yields the modular group.

When the matrix Q is odd only T^2 works and not T. But there is then another difficulty, associated with possible spin 1/2 particles that will compensate. The difficulty is that real spinor wave functions are now forbidden on such space-times because of what mathematicians know as a Stiefel-Whitney obstruction. Complex (and thus electrically charged) spinor wave functions are possible only if some if the integral flux units in (2) become half-integral. To accommodate this, a second partition function is needed in addition to the one for integral fluxes associated with scalar wave functions and spinless particles. The pair (together with a third) then support the action of the whole modular group. Details of this are explained in the two papers by M Alvarez and myself [8,9], but the main message is an interplay between topology, the Atiyah-Singer index theorem and the properties of unimodular lattices and associated theta functions.

Many questions remain, even in this limited context. The most impressive demonstration appears to be that of the modular transformation properties of the extended partition functions associated with an arbitary (Poincaré dual) four-manifold. These partition functions require the assignment of a Euclidean metric (which is always possible) but their transformation properties are independent of this choice as they are characterised by weights which are linear combinations of topological invariants that have the property of being local, that is integrals of local quantities over the four manifold (the Euler number and the Hirzebruch signature) [11]. The trouble is that electromagnetic duality is a feature of space-time with a Minkowski signature and these space-times cannot be assigned such a metric unless the Euler number vanishes.

The only case with an obvious direct physical interpretation is when the four-manifold factorises as a circle times a three-manifold. Then the partition function is genuinely of the form $Tr(e^{-H})$ and is indeed exactly modular invariant as the weights vanish.

It is desirable to find other quantities which can be identified physically and exhibit good behaviour under the modular group. One possibility concerns a sort of correlation function involving loop operators [10].

The $SL(2, \mathbb{R})$ transformations (7) were first considered in the contexts of supergravity theories where there is an important difference from our context. In supergravities the quantity τ is not a coupling but a function of additional scalar fields in the theory. Thus (8a) is just another field transformation and what is being considered is a continuous group of classical transformations forming symmetries of the energy momentum-tensor in a single theory. This interpretation is conceptually quite different from that discussed in the concept of the toy model.

In supergravity theories there naturally occur several copies of the Maxwell field, N, say. Obviously these can be rotated into each other and the full set of linear transformations extending (7) forms the group $Sp(2N, \mathbb{R})$ [15]. Of course $Sp(2, \mathbb{R})$ and $SL(2, \mathbb{R})$ are isomorphic. It is natural to expect that quantum effects would break this to the discrete subgroup $Sp(2N, \mathbb{Z})$.

The ideas can be extended to space-times

of higher even dimension, $2n$ say, equipped with closed mid-forms, *i.e.* n-forms F. Because $F^1 \wedge F^2 = (-1)^n F_2 \wedge F_1$ and $** = (-1)^{n+1}$, it matters whether n is even or odd. The relevant transformation groups turn out to be $Sp(2N, \mathbb{R})$ and $O(N, N, \mathbb{R})$ respectively [19]. Presumably, when the $(n-1)$-form potential couples to the world-volume of a $(n-2)$-brane [20], these continuous groups break to the discrete subgroups $Sp(2N, \mathbb{Z})$ and $O(N, N, \mathbb{Z})$ respectively. However it is not clear how to generalise the analysis described above for $n = 2$. This is because it is not really known how to formulate the wave-function of a brane and distinguish the scalar and spinor cases. Thus one is not sure which fluxes are integral or half-integral and how this relates to suitably generalised Stiefel-Whitney classes. Notice that if $n = 1$ and $N = 1$ the relevant group is $O(1, 1, \mathbb{Z})$. When this is divided by the invariant subgroup $\pm I$ the result is a Z_2 generated by the S transformation, agreeing with the old result of Kramers and Wannier [2].

Superstring theories display an even richer structure which is still being explored. Nevertheless one sees an expanding role for generalisations of the modular group and associated functions. That is one reason why the theory of those very large algebras called Borcherds algebras may be important. They associate algebraic structures generalising ones originating in string theory with new sorts of modular group.

Acknowledgments

References

[1] C. Montonen and D.I. Olive: "Magnetic monopoles as gauge particles ?", *Phys. Lett. B***72** (1977) 117-120.

[2] H.A. Kramers and G.H. Wannier: "Statistics of the two-dimensional ferromagnet I" *Phys. Rev* **60** (1941) 252-276.

[3] A. D'Adda, R. Horsley and P. Di Vecchia: "Supersymmetric Monopoles and Dyons", *Phys. Lett. B***76** (1978) 298-302,

E. Witten and D.I. Olive: "Supersymmetry Algebras that Include Topological Charges", *Phys. Lett. B***78** (1978) 97-101,

H. Osborn: "Topological Charges for $N = 4$ Supersymmetric Gauge Theories and Monopoles of Spin 1", *Phys. Lett. B***83** (1979) 321-326.

[4] E. Witten: "Dyons of charge $e\theta/2\pi$", *Phys. Lett. B***86** (79) 283-287.

[5] M. F. Atiyah and N. J. Hitchin: "Low energy scattering of non-abelian monopoles", *Phys. Lett. A***107** (1985) 21-25.

[6] A. Sen: "Dyon-monopole bound states, self-dual harmonic forms on the multi-monopole moduli space, and $SL(2, \mathbb{Z})$ invariance in string theory" , *Phys. Lett. B***329** (94) 217-221.

[7] D.I. Olive: "Exact Electromagnetic Duality", *Nucl. Phys. B*(Proc Suppl)**45A** (1996) 88-102, *Nucl. Phys. B*(Proc Suppl)**46** (1996) 1-15, "Introduction to duality", in "Duality and Supersymmetric Theories" (Cambridge University Press 1999), 62-94.

[8] M. Alvarez and D.I. Olive: "The Dirac quantisation condition for fluxes on Four-manifolds", *Comm. Math. Phys.***210** (2000) 13-28, `hep-th/9906093`.

[9] M. Alvarez and D.I. Olive: "Spin and abelian electromagnetic duality on four-manifolds", `hep-th/0003155`.

[10] E. Verlinde: "Global aspects of electric-magnetic duality", *Nucl. Phys. B***455** (1995) 211-228.

[11] E. Witten: "On S-duality in abelian gauge theory", *Selecta Math (NS)***1** (1995) 383-410, `hep-th/9505186`.

[12] P.A.M. Dirac: "Quantised singularities in the electromagnetic field", *Proc. Roy. Soc.* **A133** (1931) 60-72.

[13] O. Alvarez: "Topological Quantisation and Cohomology" , *Comm. Math. Phys.* **100** (1985) 279-309

[14] E. Cremmer and B. Julia: "The $SO(8)$ Supergravity", *Nucl. Phys. B***159**, (1979) 141-212.

[15] M.K. Gaillard and B. Zumino: "Duality Rotations for Interacting Fields", *Nucl. Phys. B*, **193** (1981) 221-244.

[16] G. 't Hooft: "Magnetic monopoles in unified gauge theories", *Nucl. Phys. B***79** (74) 276-284.

[17] A.M. Polyakov: "Particle spectrum in quantum field theory", *JETP Lett* **20** (74) 194-195.

[18] R.P. Feynman and A.R. Hibbs: "Quantum Mechanics and Path Integrals", Chapter 10, (McGraw-Hill 1965), RP Feynman, "Statistical Mechanics", Chapter 3, (Benjamin/Cummings 1972).

[19] P. Fre: "Lectures on special Kähler geometry and electric-magnetic duality rotations", *Nucl. Phys.* (Proc Suppl)**45BC** (1996) 59-114, `hep-th/9512043`.

[20] M. Kalb and P. Ramond: "Classical direct interstring action", *Phys Rev* **D9** (1974), 2273-2284