ASYMPTOTIC QUANTIZATION OF THE ELECTROMAGNETIC FIELD AND THE RESULTING INEQUALITY FOR THE FINE STRUCTURE CONSTANT*

By A. Staruszkiewicz

Institute of Physics, Jagellonian University, Cracow**

(*Received May 30, 1990*)

Arguments are presented which allow us to derive the inequality $0 < e^2/\hbar c < \pi$ in a purely phenomenological way i.e. without assumptions about the dynamics of sources of the electromagnetic field.

PACS numbers: 14.80.Hv

1. *Introduction*

The problem I wish to consider can be formulated as follows: where the magnitude of the elementary charge comes from? This is a very old problem. Einstein [1] noted in 1909 that there are two fundamental velocities in physics: c and $e^2/\hbar$; this should be somehow explained. Sommerfeld [2] introduced the dimensionless combination $e^2/\hbar c$, henceforth called the fine structure constant, into his theory of the hydrogen atom and realized that the particular value of this constant has to be explained theoretically. His pupil Heisenberg took over this idea — see the recently published correspondence of Pauli [3]. It is amusing to learn that Heisenberg did some numerical guesswork on the fine structure constant and thought it good enough to be communicated to Bohr [3].

The true history of efforts to understand the fine structure constant will never be written, because these efforts were not successful and as a rule did not lead to a published work (see, however, the truly charming book by Lancelot Law Whyte [4]). Some important byproducts were, however, obtained, the most important being the Dirac relation $eg = 1/2$ [5]. This relation, properly understood, is a relation between the electric charge e and the quantized magnetic flux, at the (fictitious) end of which there is a magnetic monopole g. As such it is an important physical law, confirmed experimentally with a very high accuracy. It is evident, however, that Dirac's real objective was not the magnetic monopole (which

* Presented at the Session of Department of Theoretical Physics of JINR in Dubna, held in Cracow, May 23–25, 1988.

** Address: Instytut Fizyki, Uniwersytet Jagielloński, Reymonta 4, 30-059 Kraków, Poland.

no one needs) but the quantized electric monopole. This is seen, for example, from the following words by Dirac [5]:

"The theory leads to a connection ($eg = 1/2$) between the quantum of magnetic pole and the electronic charge. It is rather disappointing to find this reciprocity between electricity and magnetism, instead of a purely electronic quantum condition, such as $\hbar c/e^2 = 137$".

In this lecture I will not give you the solution of the problem; you cannot expect me to solve the problem which Einstein, Heisenberg and Dirac found intractable. I will present only results of my thinking on the problem which allow, among other things, to make the following quantitative statement on the magnitude of the elementary charge e:

$$0 < e^2/\hbar c < \pi.$$

This inequality is not very impressive, of course, but it is arrived at in a purely phenomenological way. I think I can say that in deriving this inequality I make no hypotheses. I use only the general laws of quantum mechanics and of Maxwell's theory. The general laws I will use are:
The Gauss law.
The free Maxwell equations.
The principle of gauge invariance.
The principle of Lorentz invariance.
The basic laws of quantum mechanics (especially positivity of norm or unitarity).

2. *The Gauss law*

The constant e is the electron charge measured at zero energy, for example in Millikan's experiment. The Gauss law

$$\operatorname{div} \vec{E} = 4\pi\varrho$$

says that the electric charge is determined completely by the electric field at the spatial infinity:

$$Q = \int \varrho dV = \frac{1}{4\pi} \int_{\Sigma} \vec{E} \cdot d\vec{\Sigma},$$

where Σ is a closed, two-dimensional, space-like surface at the space-like infinity. It follows, therefore, that to study the constant e as such i.e. as the electric charge at zero energy only the asymptotic ($1/r^2$) part of the electromagnetic field is needed.

This observation solves an important problem. We have all been taught Q.E.D. and know that one has to make a difference between the bare charge and the renormalized charge. This distinction is theoretically valid but the problem is complicated by the fact that one does not know how to calculate the proportionality factor. Worse than that, in standard calculations given in text-books this proportionality factor comes out infinite which makes the whole discussion pointless. The physical reason which makes it necessary to distinguish the bare charge and the renormalized charge is this: the electric field can create pairs i.e. the charge distribution which has to be added to the original one. However, this effect is small if $\hbar\omega/mc^2$ is small, where ω is the frequency of the field and m is the mass of the lightest particle (presumably this is the electron). If the system considered has the length scale l, then, by uncertainty principle, $\omega \sim 1/l \to 0$ for $l \to \infty$. This means that at the spatial infinity the electromagnetic field is too weak to produce pairs and we do not have to bother which charge we are talking about.

3. *The free Maxwell equations*

All charged particles are massive. We do not know why it is so but we are fairly certain that it is so. Schwinger [6] has shown that the vacuum of charged massless particles would be polarizable in a way inconsistent with the phenomenological validity of the Coulomb law. This means that at the spatial infinity the electric current must vanish (e.g. exponentially) and the electromagnetic field is free; it fulfils the free Maxwell equations

$$\partial_\lambda F_{\mu\nu} + \partial_\mu F_{\nu\lambda} + \partial_\nu F_{\lambda\mu} = 0,$$
$$\partial^\mu F_{\mu\nu} = 0.$$

These equations do not hold for the whole field, which is not known, but only for its asymptotic ($1/r^2$) part. Therefore, to proceed further, we have to separate the $1/r^2$ part of the field from the unknown rest. A clean way to do this was proposed by Gervais and Zwanziger [7]: to separate the asymptotic part of the field from the rest it is enough to consider only those solutions of free Maxwell's equations which are homogeneous of degree -2 functions of space-time coordinates: for each $\lambda > 0$ $F_{\mu\nu}(\lambda x) = \lambda^{-2} F_{\mu\nu}(x)$. Global solutions of Maxwell's equations with this homogeneity property are singular on the light cone $x \cdot x = 0$ but this is not important since we will use them only at the spatial infinity. The homogeneity condition breaks down the translational invariance; this is in accordance with the fact that the asymptotic part of the electromagnetic field is, in fact, a translationally invariant quantity.

Let x be the radius vector in the four-dimensional Cartesian basis in which the homogeneity condition holds good. Let us form two vectors

$$x^\mu F_{\mu\nu} \quad \text{and} \quad \tfrac{1}{2}\,\varepsilon_{\mu\nu\varrho\sigma} x^\nu F^{\varrho\sigma}.$$

Using the Maxwell equations and the homogeneity condition one finds that both these vectors are gradients of functions which satisfy the wave equation. One can show that these two functions determine the electromagnetic field which is thus seen to be a very simple object, determined completely by two homogeneous of degree zero solutions of the wave equation. We have

$$\begin{aligned} x^\mu F_{\mu\nu} &= x^\mu(\partial_\mu A_\nu - \partial_\nu A_\mu) \\ &= -A_\nu - \partial_\nu(x^\mu A_\mu) + \delta^\mu_\nu A_\mu, \end{aligned}$$

which gives the explicit expression for the first (electric) function

$$x^\mu F_{\mu\nu} = -\partial_\nu(x^\mu A_\mu).$$

4. *The principle of gauge invariance*

The principle of gauge invariance says that the vector potential A_μ should not appear in equations of mathematical physics alone but only in the linear combination

$$eA_\mu + \partial_\mu S,$$

where S is a scalar field which transforms under gauge transformations in such a way that the above vector is gauge invariant. An example is provided by the Hamilton-Jacobi equation for a particle of mass m and charge e in an external field ($c = 1$)

$$g^{\mu\nu}(\partial_\mu S + eA_\mu)(\partial_\nu S + eA_\nu) = m^2.$$

S is given by the integral

$$-m \int \sqrt{dx^\mu dx_\mu} - e \int A_\mu dx^\mu$$

calculated for the actual motion. Under the gauge transformation $A_\mu \to A_\mu + \partial_\mu f$, $S \to S - ef$ and the linear combination $eA_\mu + \partial_\mu S$ is seen indeed to be gauge invariant. In the quantum mechanics the classical action S, divided by $\hbar = 1 = c$, becomes the phase of the wave function. Thus interaction of the Klein-Gordon field $R \exp(iS)$ with the electromagnetic field A_μ is described by the action

$$-\frac{1}{16\pi}\int F_{\mu\nu}F^{\mu\nu}d^4x$$

$$+\tfrac{1}{2}\int [\partial_\mu R \partial^\mu R + R^2(\partial_\mu S + eA_\mu)(\partial^\mu S + eA^\mu) - m^2R^2]d^4x.$$

The principle of coupling between the electromagnetic field and its source, implicit in this expression, can be generalized as follows. We assume that an electrically charged system has always to have a degree of freedom S, called phase, which enters the action only as a part of the gauge invariant linear combination $eA_\mu + \partial_\mu S$:

$$\text{the total action} = -\frac{1}{16\pi}\int F_{\mu\nu}F^{\mu\nu}d^4x + W,$$

$$W = \int L(eA_\mu + \partial_\mu S, \ldots)d^4x.$$

(Dots denote the remaining degrees of freedom of the source). The phase S is the coordinate canonically conjugated with the total charge. Indeed, varying the action with respect to A_μ we have

$$\delta_A W = -\int j^\mu \delta\left(A_\mu + \frac{1}{e}\partial_\mu S\right)d^4x,$$

where j^μ is, by definition, the electric current. This means that varying with respect to S we obtain

$$\delta_S W = -\int j^\mu \frac{1}{e}\partial_\mu \delta S d^4x.$$

Therefore the momentum canonically conjugated with the phase S is

$$p_S = -\frac{1}{e} j_0.$$

The equal time canonical commutation relation is

$$\left[\frac{1}{e} j_0(x), S(y)\right]_{x^0=y^0} = i\delta(\vec{x}-\vec{y}).$$

Integrating over the hyperplane $x^0 = y^0$ we obtain

$$[Q, S(y)] = ie,$$

where

$$Q = \int j_0 d^3x.$$

Three comments will be useful.

The commutation relation $[Q, S] = ie$ is seen to be universally valid for every charged system. The canonical commutation relations for a gauge invariant system are known to be inconsistent with equations of motion. Our derivation of the relation $[Q, S] = ie$ passes over this difficulty. The difficulty, however, is real and has to be remembered.

The relation $[Q/e, S] = i$ might seem to resemble the relation between the square of the amplitude and the phase of the harmonic oscillator. Despite of this formal resemblance there is a basic difference: the relation between the square of the amplitude and the phase holds only classically and cannot be consistently implemented in the quantum mechanics [8] while the relation $[Q, S] = ie$ will be seen to be a meaningful mathematical statement. The difference comes from the fact that the square of the amplitude i.e. the energy is positive while the electric charge is not.

The basic idea of my work is to identify the phase with one of the two functions which determine the electromagnetic field at the spatial infinity and to use the relation $[Q, S] = ie$ to reduce the asymptotic electromagnetic field to a closed mechanical system. "Closed" means "described by an action integral" which, in consequence of this relation, will contain the constant e.

The phase is identified with the function

$$S(x) = -ex^\mu A_\mu(x).$$

I will make two observations to support this identification. The first one shows that it is esthetically appealing to make this identification. The second one falls only short of a proof that the function $-ex^\mu A_\mu(x)$ is indeed the phase of the field at the spatial infinity.

1. Consider a classical point particle moving in the field $F_{\mu\nu}$, which is homogeneous of degree -2. Multiplying the equations of motion

$$m\frac{d^2x^\mu}{ds^2} = eF^{\mu\nu}\frac{dx_\nu}{ds}$$

by x_μ and assuming that the potential A_μ is a function homogeneous of degree -1, one finds the first integral

$$mx^\mu \frac{dx_\mu}{ds} - ms + ex^\mu A_\mu(x) = \text{const.}$$

This integral exists because the potential is assumed to be homogeneous of degree -1; thus it is the relativistic generalization of the virial theorem in the classical mechanics. One sees that it is indeed appealing to identify the function $-ex^\mu A_\mu(x)$ with the phase. I make this observation because in the second one, which I consider a proof, there does arise the problem of the correct choice of a numerical factor.

2. The phase can be associated with each field $A_\mu(x)$ by means of the formula

$$S(x) = -e \int A_\mu(x-y) j^\mu(y) d^4 y,$$

where

$$\partial_\mu j^\mu(y) = \delta^{(4)}(y)$$

and we agree to ignore all surface integrals. There is only one Lorentz invariant solution to the last equation:

$$j_\mu(y) = \partial_\mu \frac{1}{4\pi} \delta(y \cdot y),$$

where $(1/4\pi)\delta(yy)$ is the half-retarded, half-advanced Green's function of the wave equation. This means that there is only one Poincaré invariant way to associate a phase with a given field:

$$S(x) = -\frac{e}{4\pi} \int A_\mu(x-y) \partial^\mu \delta(y \cdot y) d^4 y.$$

Assume now that $A_\mu(x)$ is a homogeneous of degree -1 solution of Maxwell's equations. It is difficult to calculate the last integral for a generic field of this kind but one can argue in this way. The asymptotic (at the spatial infinity) part of radiation produced in a scattering process, in which charged particles take a part, is surely close to being generic. The asymptotic part of radiation is known to be equal to

$$A_\mu(x) = \Theta(-xx) \sum_n e_n \frac{u_{n\mu}}{r(u_n)},$$

where $r^2(u) = (ux)^2 - (uu)(xx)$, u_n is the asymptotic (at the time-like infinity) four-velocity of the n'th particle and

$$\sum_n e_n = 0.$$

For this potential the Poincaré invariant phase can be calculated term by term with the result

$$S(x) = -\tfrac{1}{2}\, e x^{\mu} A_{\mu}(x).$$

This result is completely unambiguous but I maintain that the factor 1/2 has to be removed, because it reflects a contribution from the surface term which we have agreed to ignore in the original definition. (This is a mathematical detail which has no influence on the qualitative features of the theory. I do think that the factor 1/2 has to be removed. Should we need it, however, we can restore it with completely clear conscience.) The phase $S(x) = -ex^{\mu}A_{\mu}(x)$ is a gauge invariant quantity, if we agree to add to $A_{\mu}(x)$ only gradients of homogeneous of degree zero functions, which is natural. This sounds paradoxical; the paradox is resolved, however, by the remark that the gauge has been fixed — by the homogeneity condition — only at the spatial infinity, in a way similar e.g. to the way one fixes the Coulomb potential for the Coulomb field.

5. *The Lorentz invariance*

We have seen that the electric part of the electromagnetic field at the spatial infinity is equivalent to a scalar field $S(x) = -ex^{\mu}A_{\mu}(x)$ which is homogeneous of degree zero, thus lives on the three-dimensional de Sitter hyperboloid and fulfils there the wave equation. The action for such a field is

$$\text{const} \int d^3\xi \sqrt{g}\, g^{ik} \frac{\partial S}{\partial \xi^i} \frac{\partial S}{\partial \xi^k},$$

where ξ's are internal (angular) variables on the de Sitter hyperboloid. I propose to fix the constant in this way: the scalar field on the three-dimensional de Sitter hyperboloid has a spherically symmetric solution with a nonvanishing total charge. The spherically symmetric solution is the only one which does not have its counterpart among global solutions of Maxwell's equations, because a global solution of Maxwell's equations cannot be spherically symmetric. This additional solution allows us to introduce the electric charge into the theory in a completely clean, model independent way. I propose to choose the constant in the action so that the previously established relation $[Q, S] = ie$ is canonical i.e. follows from the canonical commutation relations. This fixes uniquely the constant in the action and the whole theory is summed up in the expressions for the total action

$$\frac{1}{8\pi e^2} \int d^3\xi \sqrt{g}\, g^{ik} \frac{\partial S}{\partial \xi^i} \frac{\partial S}{\partial \xi^k}$$

and the total charge

$$Q = -\frac{1}{4\pi e} \int_{\Sigma} \frac{\partial S}{\partial \xi^i}\, d\Sigma^i,$$

Σ being an arbitrary Cauchy surface in the de Sitter hyperboloid. This theory is per se an interesting object of theoretical inquiry. It is made important, however, by the claim, which I tried to substantiate above, that the constant e^2 is physically identical with the fine structure constant. The vacuum in this theory has to satisfy the condition

$$\langle 0|Q = 0, \quad Q|0\rangle = 0.$$

This seems natural enough but it is interesting to note that this condition actually follows from the Lorentz invariance of the vacuum. The vacuum is Lorentz invariant if it is annihilated by all components of the angular momentum and the centre of mass motion:

$$\langle 0|M_{\mu\nu} = 0, \quad M_{\mu\nu}|0\rangle = 0.$$

The integrals $M_{\mu\nu}$ do contain the total charge Q but do not contain the coordinate S (only its derivatives) which has the consequence that the condition

$$\langle 0|Q = 0, \quad Q|0\rangle = 0$$

is necessary for the Lorentz invariance of the vacuum. A further consequence is that the coordinate canonically conjugated with the total charge has to be periodic, otherwise the vacuum will not be normalizable. Assuming the period to be 2π — a natural, in fact necessary assumption for the phase — we obtain the charge quantization

$$Q = ne, \quad n = 0, \pm 1, \pm 2, \ldots$$

as a consequence of the Lorentz invariance of the vacuum.

6. *The positivity of norm*

The theory described in the previous Section reveals an intricate dependence on the magnitude of the coupling constant e^2. To give an example of a nontrivial dependence we derive the inequality $0 < e^2 < \pi$, mentioned at the beginning, as the positivity of norm condition for a class of charged states. Consider the vacuum expectation value

$$\langle 0|e^{iS(k)}e^{-iS(l)}|0\rangle,$$

where k and l are two distinct points in the de Sitter space-time i.e. two distinct space-like directions. To simplify calculations and the procedure of renormalization assume that the "time" i.e. the third (hyperbolic) angle tends to infinity. Using the vacuum definition due to Chernikov and Tagirov [9] one finds (after an obvious renormalization, the renormalization does not involve the constant e^2, which is fixed)

$$\langle 0|e^{iS(\vec{k})}e^{-iS(\vec{l})}|0\rangle = \text{const}\, |\vec{k}-\vec{l}|^{-2e^2/\pi},$$

$\vec{k}$ and $\vec{l}$ being unit, space-like parts of k and l. Therefore the scalar product of two charged states of the form

$$\int d^2\vec{k} f(\vec{k}) e^{-iS(\vec{k})}|0\rangle$$

will have as a kernel the expression

$$|\vec{k}-\vec{l}|^{-2e^2/\pi}$$

i.e. it will be the scalar product for the supplementary series of unitary representations of the Lorentz group of Gelfand, Graev and Vilenkin [10]. This product is known to be positive definite only for $0 < e^2/\pi < 1$ (see [10], page 260, for the proof).

The theory leads in a natural way to several special functions for which the segment $0 < e^2 < \pi$ appears to be distinguished. This can be seen for the function

$$f(z) = 4\pi \int_0^\infty d\lambda \,\mathrm{sh}^2 \lambda \exp\left[-2z(\lambda \coth \lambda - 1)\right],$$

$$z = e^2/\pi,$$

which I introduced in [11]. This function can be extended analytically over the whole z plane by means of the identity

$$\int_0^\infty d\lambda \,\mathrm{sh}^2 \lambda e^{-2z\lambda \coth \lambda}$$

$$= -\tfrac{1}{2} z^3 \sum_{n=-\infty}^{\infty} \frac{(n-z)^{n-2}}{(n+z)^{n+2}},$$

valid for $z > 1$. Points $z = 0$ and $z = 1$ are seen to be poles of the function $f(z)$, a fact not immediately obvious from the definition of $f(z)$.

REFERENCES

[1] A. Einstein, *Phys. Zeit.* **10**, 185 (1909).
[2] A. Sommerfeld, *Atombau und Spektrallinien*, Friedr. Vieweg, Braunschweig 1969, page 272.
[3] W. Pauli, *Scientific Correspondence*, Vol. II, Springer Verlag, Berlin 1985, page 366.
[4] L. L. Whyte, *The Atomic Problem, A Challenge to Physicists and Mathematicians*, Allen and Unwin, London 1961.
[5] P. A. M. Dirac, *Proc. R. Soc.* **A133**, 60 (1931).
[6] J. Schwinger, *Selected Papers*, Reidel, Dordrecht 1979, p. 337.
[7] J. Gervais, D. Zwanziger, *Phys. Lett.* **B94**, 389 (1980).
[8] R. G. Newton, *Ann. Phys.* (*N. Y.*) **124**, 327 (1980).
[9] N. A. Chernikov, E. A. Tagirov, *Ann. Inst. Henri Poincaré* **94**, 109 (1968).
[10] I. M. Gelfand, M. I. Graev, N. Ya. Vilenkin, *Generalized Functions*, Vol. V, Moscow 1962, in Russian.
[11] A. Staruszkiewicz, *Ann. Phys.* (*N. Y.*) **190**, 354 (1989).