PHYSICAL REVIEW D VOLUME 25, NUMBER 6 15 MARCH 1982

Fine-structure constant: Is it really a constant?

Jacob D. Bekenstein
Physics Department, Ben Gurion University of the Negev, Beer Sheva 84120, Israel
(Received 25 September 1981)

It is often claimed that the fine-structure "constant" α is shown to be strictly constant in time by a variety of astronomical and geophysical results. These constrain its fractional rate of change $\dot{\alpha}/\alpha$ to at least some orders of magnitude below the Hubble rate H_0. We argue that the conclusion is not as straightforward as claimed since there are good physical reasons to expect $\dot{\alpha}/\alpha << H_0$. We propose to decide the issue by constructing a framework for α variability based on very general assumptions: covariance, gauge invariance, causality, and time-reversal invariance of electromagnetism, as well as the idea that the Planck-Wheeler length (10^{-33} cm) is the shortest scale allowable in any theory. The framework endows α with well-defined dynamics, and entails a modification of Maxwell electrodynamics. It proves very difficult to rule it out with purely electromagnetic experiments. In a cosmological setting, the framework predicts an $\dot{\alpha}/\alpha$ which can be compatible with the astronomical constraints; hence, these are too insensitive to rule out α variability. There is marginal conflict with the geophysical constraints; however, no firm decision is possible because of uncertainty about various cosmological parameters. By contrast the framework's predictions for spatial gradients of α are in fatal conflict with the results of the Eötvös-Dicke-Braginsky experiments. Hence these tests of the equivalence principle rule out with confidence *spacetime* variability of α at any level.

I. INTRODUCTION

Early suggestions that the fine-structure constant α (or the elementary charge) varies with cosmological time were made by Dirac,[1] Teller,[2] and Jordan.[3] Dicke and Peebles[4] pointed out that the abundance ratios of Kr^{40} to Ar^{40} and of Rb^{87} to Sr^{87} in old ores and in meteorites constrain $\dot{\alpha}/\alpha$, the fractional rate of variation of α, to at least two orders of magnitude below the Hubble rate H_0.[5] They also suggested that better bounds should result from measurements of the Re^{187} to Os^{187} ratio, because the rate of the weak decay $Re^{187} \rightarrow Os^{187}$ is highly sensitive to the value of α. Later Dicke[6] showed that the Eötvös-Dicke experiments require any *spatial* variation of α caused by the Sun to be smaller, by orders of magnitude, than the corresponding variation in the solar gravitational potential. All this evidence against α variability was soon forgotten.

Interest in α variability reached a new peak following Gamow's proposal[7] (inspired by Dirac's large-numbers hypothesis) that α varies *at* the Hubble rate. Within weeks Gamow's $\alpha \propto t$ law had been newly ruled out. Peres[8] showed that it contradicts the excellent agreement between the list of nuclides found naturally and those expected to be β stable from nuclear mass systematics. Dyson[9] showed it to conflict with Re^{187} to Os^{187} ratios in old ores; these make sense only if $|\dot{\alpha}/\alpha| < 4\times 10^{-4} H_0$ according to a later review.[10] Bahcall and Schmidt[11] called attention to the good agreement between fine-structure splittings in radiogalaxy spectra and their laboratory values; this implies $|\dot{\alpha}/\alpha| < 10^{-2} H_0$ in conflict with Gamow's suggestion. Later work has strengthened the trend. A laboratory intercomparison of cesium and superconducting cavity clocks by Turneaure and Stein[12] has given the only known bound on the instantaneous $\dot{\alpha}/\alpha$: $|\dot{\alpha}/\alpha| < 8\times 10^{-2} H_0$. Wolfe, Brown, and Roberts[13] have extended the baseline of the astronomical test for $\dot{\alpha}$ by looking at a BL-Lacertae object at red-shift $z=0.52$; quasars make possible a further extension.[14] Finally, from an analysis of isotopic abundances of fission products in the Oklo "natural reactor" (a uranium vein in Gabon which spontaneously underwent a fission episode 1.8×10^9 years ago), Shlyakhter[15] derived the best constraint yet: $|\dot{\alpha}/\alpha| < 10^{-7} H_0$.

These impressive constraints are widely regarded as establishing perfect constancy of α. One often hears the argument that *if* α varied in time, it would be expected to vary roughly at the Hubble rate, which it does not; hence α is constant. There

are several objections to such reasoning. It implicitly assumes a power law in time for α, with an exponent not widely different from unity. Yet other behaviors are conceivable. For example, α could exhibit an asymptotic approach to a finite value, with little present variation (this is the typical behavior of G in a large class of gravitational theories[16,17]). The extant constraints might not be sufficient to rule this out. Evidently, in the absence of a concrete dynamical equation for α, the mentioned argument essentially begs the question. Further, the idea that the Hubble rate is the only characteristic rate for variable quantities in an expanding universe is problematic. The Hubble expansion is driven (via the gravitation equations) by the *total* energy density in the universe. But α is an electromagnetic quantity; its temporal variation, if any, should be driven by the density of some electromagnetic quantity. Since various electromagnetic energy densities are orders of magnitude below the rest energy density (at least in recent times), one has good reasons to expect $|\dot{\alpha}/\alpha|$ to be *much smaller* than H_0.

All this underscores the need for a *general* dynamical framework for α. Only within the context of such can it be clear if the sensitivity of a given experimental constraint is sufficient to yield firm conclusions.[16] And only within such a framework can one combine constraints on spatial and on temporal variability of α into unified evidence. Without it one hardly knows how to compare the merits of the two, let alone integrate them. The main goal of the present paper is the construction of just such a framework based on a few uncontroversial assumptions, and its application to the question, Do the extant constraints rule out all sorts of α variability? We answer affirmatively.

Our assumptions are discussed in Sec. II. In Sec. III we construct the Lagrangian giving the dynamics of the electromagnetic field and of α. An unspecified scale of length l enters into it. From the observed scale invariance of electromagnetism $l < 10^{-15}$ cm; also $l > 10^{-33}$ cm because the Planck-Wheeler length is the shortest conceivable length in physics. In Sec. IV we show that purely electromagnetic experiments are incapable of ruling out α variability. Temporal variability of α in our expanding universe is taken up in Sec. V. The framework predicts that in recent epochs $|\dot{\alpha}/\alpha| \approx kH_0$ with k probably in the range $10^{-4}-10^{-6}$. This small value is due partly to the mentioned smallness of electromagnetic energies and partly to the appearance in the expression for $\dot{\alpha}/\alpha$ of an integration constant which can act to suppress $\dot{\alpha}/\alpha$. As a result the constraints from intercomparison of clocks[12] and from the spectra of radio galaxies and quasars[11,13,14] do not suffice to rule out α variability. The geophysical constraints[10,15] come rather close to ruling it out; however, uncertainties about various cosmological parameters prevent us from reaching a firm conclusion. Predictions of the framework for spatial α variability are the theme of Sec. VI. We confirm Dicke's conclusion[6] that tests of the equivalence principle provide very strong constraints on α variability. The Dicke[6] and Braginsky[18] versions of the Eötvös experiment strongly rule out any α variability, temporal or spatial. Our conclusions are discussed in Sec. VII.

II. POSTULATES OF THE FRAMEWORK

The fine-structure constant α is the low-energy limit of the (renormalized) electromagnetic coupling constant. Being interested primarily in macroscopic phenomena, we adopt a classical description of the electromagnetic field and its interaction with (classical or quantum) matter.

Once the possibility that α depends on space and time is admitted, modification of standard Maxwell electrodynamics is inevitable. For example, adopting the usual relation between α and elementary charge, $\alpha = e^2/\hbar c$, we see that α variability implies e variability in units for which $\hbar$ and c are constant. But e variability seems to clash with charge conservation (the electron's charge depends on position) which follows from Maxwell's equations. Evidently, something in the accepted picture of electrodynamics must give way. What we need, then, is a set of assumptions to guide us in modifying Maxwell electrodynamics in a reasonable way. These should allow a general, model-independent, framework for α variability; at the same time they should respect generally accepted physical principles. We adopt the following assumptions and discuss each in turn.

P1. *For constant α electromagnetism is Maxwellian and the coupling of the vector potential to matter is minimal.* This is a correspondence principle restraining us from introducing unnecessary modifications of standard physics.

P2. *Variations of α result from dynamics.* One must reject prescribed laws of variation. If α varies, the variation influences charged matter; charged matter should in turn influence α. Only

dynamics for α can incorporate this important feature.

P3. *Dynamics of electromagnetism and α are derivable from an invariant action.* We know of no other way to consistently incorporate the principle "action equals reaction" than to derive dynamical equations from an action principle. And no better way offers itself to guarantee relativistic invariance than to start from an invariant action.

P4. *The action is locally gauge invariant.* The importance of the gauge principle in contemporary physics cannot be overstated. Thanks to it there is now a unified picture of all the microscopic interactions. If we gave up demanding full gauge invariance, we would be modifying Maxwell's theory in an unreasonable way.

P5. *Electromagnetism is causal.* There is no *experimental* evidence for causality violation. Theoretically it is sometimes discussed, but only in rather esoteric situations. In our rather mundane problem causality should be enforced: dynamical equations should be hyperbolic and of at most second order, in order to forestall runaway solutions and other noncausality.

P6. *The electromagnetic action is time reversal invariant.* This is taken for granted in contemporary particle physics where any violations of C, P, or T invariance are blamed on the weak interaction, or the superweak facet of it.[19] Concrete evidence for P and T invariance of electromagnetism are the very tight upper bounds on the electric dipole moments of the proton and neutron.[20] Of course such evidence establishes T invariance only to some finite precision. We assume it is exactly true and thereby accept a risk.

P7. *The shortest scale of length which can enter into physical theory is the Planck-Wheeler length* $L_{PW}\equiv(G\hbar/c^3)^{1/2}\approx 10^{-33}$ *cm.* It has been argued that at shorter scales spacetime itself is not smooth, or even simply connected.[21] If so, it makes no sense to talk of smooth fields obeying differential equations on scales shorter than L_{PW}; introducing an even shorter length into a theory cannot have any palpable consequences. Another way of saying this is that to probe experimentally the predictions of a theory down the scale l requires use of particles of energy $E\approx\hbar c/l$. If $l<L_{PW}$, E is so large that its gravitational radius $2GE/c^4$ exceeds l. So the particle probe would tend to fall into a black hole of its own making, thus frustrating any attempt to learn from its scattering about the structure of the theory at scales $<L_{PW}$.

P8. *Gravitation is described by the metric of spacetime which satisfies Einstein's equations.* The aptness of the geometric description of gravitation is clear enough today. Einstein's equations are the simplest dynamics for the metric which have stood all experimental challenges to date. We adopt them, not out of blind confidence in their correctness, but because they allow a clean separation of the issue of α variability from that of G variability which exists in many competing gravitational theories. We do not assume the other half of general relativity—the strong equivalence principle—because it rules out α variability by definition.

III. DYNAMICS OF THE FRAMEWORK

A. Electrodynamics

We choose units of length, time, and mass so that $\hbar$, c, and G are spacetime constants; this can always be done. Since $\alpha=e^2/\hbar c$, where e is the electron charge, α variability means that e depends on the spacetime point. We expect the charges of all particle species to vary in exactly the same way. Otherwise charge ratios would vary, and there could not exist accurately neutral atoms (equal electron and proton charges) or hadrons (zero net quark charge), as observed. Thus every particle charge can be expressed in the form $e=e_0\epsilon(x^\mu)$, where e_0 is a constant characteristic of the particles and ϵ a dimensionless universal field. The particular split between e_0 and ϵ is our choice. We can rescale ϵ by a constant factor and make appropriate changes in all the e_0 so as to get the previous e's. Thus the theory governing ϵ should not fix its overall scale: it should be invariant under the change $\epsilon\rightarrow$ constant$\times\epsilon$. We shall find this constraint very useful.

Consider now the classical dynamics of a charged particle of rest mass m and charge $e_0\epsilon$ in flat spacetime. We may start with the Lagrangian[22,23]

$$L=-mc(-u^\alpha u_\alpha)^{1/2}+(e_0\epsilon/c)u^\alpha A_\alpha\ , \tag{1}$$

where $u^\alpha=dx^\alpha/d\tau$ is the four-velocity and τ is the proper time. We notice that L is Lorentz invariant and involves mimimal coupling. The appropriate gauge transformation law for A_α is

$$\epsilon A_\alpha\rightarrow\epsilon A_\alpha+\chi_{,\alpha}\ , \tag{2}$$

where χ is an arbitrary function: under (2) L changes only by the perfect derivative

$(e_0/c)d\chi/d\tau$. The minimal coupling in (1) is required (P1) only in the $\epsilon\to$ const limit. One might thus be tempted to introduce other couplings consistent with this, such as $\epsilon_{,\alpha}A^\alpha$ or $\epsilon_{,\alpha}\epsilon_{,\beta}A^\alpha u^\beta$. However, no single rule such as (2) involving a truly arbitrary function guarantees gauge invariance of the dynamics. Thus we exclude these modifications by P4.

We notice that the curl of A_α is not invariant under (2); it cannot be the physical electromagnetic field $F_{\alpha\beta}$. To identify $F_{\mu\nu}$ we turn to the Lagrange equations for (1):

$$\frac{d}{d\tau}[mu_\alpha+(e_0/c)\epsilon A_\alpha]=-m_{,\alpha}c^2+(e_0/c)(\epsilon A_\beta)_{,\alpha}u^\beta . \quad (3)$$

In writing (3) we have already taken into account the normalization relation

$$u_\alpha u^\alpha=-c^2 . \quad (4)$$

In addition we have regarded m as $m(x^\alpha)$; since the particle is charged, part of its rest energy is electromagnetic in nature and should depend on position through ϵ.[6] Simplifying (3) we get

$$d(mu_\alpha)/d\tau=-m_{,\alpha}c^2+(e_0/c)[(\epsilon A_\beta)_{,\alpha}-(\epsilon A_\alpha)_{,\beta}]u^\beta . \quad (5)$$

The first term on the right-hand side is the anomalous force first discussed by Dicke,[6] due to variation of ϵ. The second is a Lorentz-type force. Identifying it with the usual expression $(e_0\epsilon/c)F_{\alpha\beta}u^\beta$ we arrive at the definition

$$F_{\alpha\beta}=\epsilon^{-1}[(\epsilon A_\beta)_{,\alpha}-(\epsilon A_\alpha)_{,\beta}] . \quad (6)$$

This agrees with the Maxwellian version only for $\epsilon\to$ const. We note that $F_{\alpha\beta}$ is gauge invariant and invariant under rescaling ϵ by a constant.

To build the electromagnetic Lagrangian (P3) we have at our disposal $F_{\alpha\beta}$ and its dual

$$^*F^{\mu\nu}=\tfrac{1}{2}\epsilon^{\mu\nu\alpha\beta}F_{\alpha\beta} , \quad (7)$$

where $\epsilon^{\mu\nu\alpha\beta}$ is the Levi-Civita tensor. $^*F_{\mu\nu}$ shares the invariances of $F_{\mu\nu}$. One possible electromagnetic action is

$$S_{\rm EM}=-(16\pi)^{-1}\int F^{\mu\nu}F_{\mu\nu}(-g)^{1/2}d^4x , \quad (8)$$

which reduces to the Maxwell action for constant ϵ. The other possible Lagrangian density is $^*F^{\mu\nu}F_{\mu\nu}$. When ϵ is constant this is just a perfect divergence (recall that $^*F^{\mu\nu}{}_{;\nu}\equiv 0$). For this reason one never considers $^*F^{\mu\nu}F_{\mu\nu}$ in ordinary electromagnetism. When ϵ can vary, $^*F^{\mu\nu}{}_{;\nu}\neq 0$ and $^*F^{\mu\nu}F_{\mu\nu}$ is not a perfect divergence; one could thus include it in the action. However, $^*F^{\mu\nu}F_{\mu\nu}$ changes sign under time reversal so it is excluded by P6. We did not include a mass term $m^2A_\mu A^\mu$ since it violates P4. Likewise, we excluded terms of fourth or higher order in $F_{\mu\nu}$ from $S_{\rm EM}$ because they do not vanish for constant ϵ and thus violate P1. We are thus left with (8) as the general electromagnetic action satisfying our assumptions.

B. Dynamics of α

$S_{\rm EM}$ is so simple in structure that variation of it with respect to ϵ does not give a proper dynamical equation for ϵ. By P2 and P3 we must thus introduce a separate action for ϵ. Evidently, the Lagrangian density should be constructed from the metric and the logarithmic gradient $\epsilon^{-1}\epsilon_{,\mu}$ which is invariant under rescaling of ϵ by a constant. One cannot introduce a function of ϵ by itself since it changes under rescaling. [For this reason also, we did not include such a function in (8).] Now, because ϵ is dimensionless, $\epsilon_{,\mu}\,\epsilon^{,\mu}\,\epsilon^{-2}$ does not have dimensions of a Lagrangian density (even if we introduce $\hbar$'s and c's at will). This led Meisels[24] to conclude that no satisfactory dynamics can be set up for a dimensionless coupling constant. Let us see if this pessimistic conclusion is unavoidable.

We can try the Lagrangian density $R\epsilon^{-2}\epsilon_{,\mu}\epsilon^{,\mu}$ or else $R^{\mu\nu}\epsilon^{-2}\,\epsilon_{,\nu}\epsilon_{,\mu}$, where $R^{\mu\nu}$ and R are, respectively, the Ricci tensor and scalar. These have, after multiplication by $\hbar c$, the correct dimensions. But they are problematical. The R or $R^{\mu\nu}$ are determined by Einstein's equations. The ϵ field is one of the sources of these equations, and up to third order derivatives of ϵ appear in the expressions for R or $R^{\mu\nu}$ (variation of R or $R^{\mu\nu}$ with respect to the metric brings two extra derivatives to bear on $\epsilon_{,\nu}\epsilon_{,\mu}$). Thus in its explicit form the differential equation for ϵ will contain third- and fourth-order derivatives as well as second. This contradicts P5 so we must exclude Lagrangians with curvature in them.

If a scale of length l is available, one can build a satisfactory action, namely,

$$S_\epsilon = -\tfrac{1}{2}\hbar c l^{-2}\int \epsilon^{-2}\epsilon_{,\mu}\epsilon^{,\mu}(-g)^{1/2}d^4x\ , \qquad (9)$$

where the factor $-\frac{1}{2}$ is conventional. Of course, introduction of l must clash at some level with scale invariance, an experimentally established feature of electromagnetism. However, the electromagnetic interaction has been probed only up to energies in the GeV range, so by the usual argument one can vouch for scale invariance only down to 10^{-15} cm or so. Provided l is shorter than this, no contradiction need arise. In fact we shall see in Sec. IV that the electric field of a truly point charge described by $S_{\rm EM}$ and S_ϵ is accurately Coulombic down to a distance $\lesssim l$. Hence scattering experiments with the present energies cannot reveal the presence of a $l < 10^{-15}$ cm, say. One may also recall the current view that electromagnetism merges with the weak interaction at energies equivalent to a length $\approx 10^{-16}$ cm, and both merge with the strong interaction at energies equivalent to $\approx 10^{-29}$ cm. Whatever the details of this unification, one should not be surprised if a scale of length makes its way into dynamics of the electromagnetic *coupling*. We do not try to guess l, but merely require (P6)

$$l \geq L_{\rm PW} \equiv \left[\hbar G/c^3\right]^{1/2} \approx 10^{-33}\ \text{cm}. \qquad (10)$$

We did not consider the more general Lagrangian density

$$-\tfrac{1}{2}\hbar c l^{-4} f\left[l^2\epsilon^{-2}\epsilon_{,\mu}\epsilon^{,\mu}\right]$$

with f an arbitrary function because, as shown in the Appendix, any choice of f, except that in (9), leads to causality violations, and must be excluded by (P5). Thus a Lagrangian quadratic in $\epsilon^{-1}\epsilon_{,\mu}$ is forced on us.

C. Structure of the equations

In order to write the dynamical equations for $F_{\mu\nu}$ and ϵ, we must specify the matter action. Being interested in the macroscopic manifestations of ϵ variability, we adopt a model of point classical particles for the matter. By "point" we mean small by macroscopic standards, but large compared to l. Since $l < 10^{-15}$, the model is a good one for ordinary matter composed of atoms, nuclei, electrons, ... For many point particles with masses m coupling to A_μ with strengths $e_0\epsilon$, the action may be written as[22] [see (1)]

$$S_m = \sum \int [-mc^2 + (e_0\epsilon/c)u^\mu A_\mu] \times \gamma^{-1}\delta^3[x^i - x^i(\tau)]d^4x\ , \qquad (11)$$

where $x^i(\tau)$ with $i=1,2,3$ describe the spatial track of the typical particle as a function of its proper time τ, $u^\mu = dx^\mu/d\tau$, while $\gamma = dx^0/cd\tau$ is the Lorentz factor. We regard each m as a function of ϵ because part of m is electromagnetic in nature and depends on the local strength of the electromagnetic interaction.

Variation of $S_{\rm EM} + S_\epsilon + S_m$ with respect to A_μ gives the electromagnetic field equations

$$(\epsilon^{-1}F^{\mu\nu})_{;\nu} = 4\pi j^\mu \qquad (12)$$

with

$$j^\mu = \sum (e_0/c\gamma)u^\mu(-g)^{-1/2}\delta^3[x^i - x^i(\tau)]\ . \qquad (13)$$

Because $\epsilon^{-1}F^{\mu\nu}$ is antisymmetric we have the identity $j^\mu{}_{;\mu} = 0$. The conserved charge, the sum $\sum e_0$ over particles, is thus distinct from the sum of the "charges" $e_0\epsilon$ which couple to A_μ in the action. Thus is charge conservation reconciled with variability of ϵ (or α). We note, also from (12), that Gauss's theorem only permits the determination of $\sum e_0$ from the distant electric field of a system (provided ϵ is known at the field points).

There exists an alternative view of the electrodynamics just discussed. One can take the view that the Lorentz force in (5) is the force on a *constant* charge e_0 of the field $\epsilon F_{\alpha\beta}$ (a curl). If $\epsilon F_{\alpha\beta}$ is interpreted as the electromagnetic field, then Eqs. (12) have just the form of Maxwell's equations in a *material* medium with dielectric constant ϵ^{-2} and permeability ϵ^2. That is, if time-space components of $\epsilon F^{\alpha\beta}$ are identified with the electric field $\vec{\rm E} = \epsilon^2\vec{\rm D}$, while space-space components are identified with the magnetic induction $\vec{\rm B} = \epsilon^2\vec{\rm H}$, then $\vec{\rm D}$ and $\vec{\rm H}$ satisfy the usual "macroscopic" Maxwell equations with the conserved charge and current as sources. This alternative description of variable α was actually discussed long ago by Dicke [Eqs. (26)–(28), Appendix 4 of Ref. 6] who also pointed out that α is to be identified, not with $e_0{}^2/\hbar c$, but with $e_0{}^2\epsilon^2/\hbar c$ as in our formalism. Variation of α can thus be described in two languages (variable charges in a homogeneous vacuum or constant charges in an inhomogeneous medium) with the same physical content. The difference between our formalism and Dicke's is in

the dynamics for ϵ. Dicke thinks of ϵ^{-2} as an ordinary scalar field;[6] our requirement that the overall scale of ϵ have no physical significance forces us to adopt scalar-field dynamics for $\ln\epsilon$ [Eq. (9)].

We know experimentally that ϵ varies little over "earthly" distances and times. Thus, for fields $F^{\mu\nu}$ which do vary on such scales, the factor ϵ^{-1} in (12) can be taken out from under the derivative, and we recover, to a good approximation, Maxwell's equations with the $e_0\epsilon$ as sources. At this point we use our prerogative to choose the overall scale of ϵ so as to make $\epsilon=1$ at our present cosmological epoch, and far away from strong sources of the field ϵ (see Sec. IV.). With this choice the e_0 coincide with the usual charges of particles—the sources of Maxwell's equations— *at the present epoch.*

The dynamical equation for ϵ is obtained by varying $S_{\rm EM}+S_\epsilon+S_m$ with respect to $\ln\epsilon$. With the notation

$$\sigma=\sum mc^2\gamma^{-1}(-g)^{-1/2}\delta^3[x^i-x^i(\tau)]\ , \tag{14}$$

we have

$$\Box\ln\epsilon=\frac{l^2}{\hbar c}\left[\epsilon\frac{\partial\sigma}{\partial\epsilon}-\epsilon j^\mu A_\mu+\frac{1}{4\pi}\left[A_\mu F^{\mu\nu}\right]_{;\nu}\right]\ , \tag{15}$$

where $\Box$ denotes the covariant D'Alembertian. Substituting j^μ from (12) into (15) we obtain the alternative form

$$\Box\ln\epsilon=\frac{l^2}{\hbar c}\left[\epsilon\frac{\partial\sigma}{\partial\epsilon}-\frac{1}{8\pi}F^{\mu\nu}F_{\mu\nu}\right] \tag{16}$$

from which all reference to current has disappeared; this will be useful in Sec. V.

IV. ELECTROMAGNETIC TESTS

We now show that purely electrostatic laboratory experiments do not rule out the electrodynamics with variable α developed in Sec. III. One of the most accurate tests of Maxwell's equations is the classic Faraday "ice-pail" experiment. It is found that in the charge-free interior of a highly conducting cavity, there is no measurable electrostatic field, whatever the exterior situation. As is well known Gauss's law, together with the equipotential nature of the cavity walls predicts this vanishing of the electric field. A similar result holds in the variable α electrodynamics, as we now show.

Working in flat spacetime we identify F^{0i}, $i=1,2,3$ with the electric field vector $\vec{\rm E}$. Setting $\mu=0$ in (12) we get in the cavity interior

$$\vec{\nabla}\cdot(\epsilon^{-1}\vec{\rm E})=0\ . \tag{17}$$

Multiplying (17) by ϵA_0, integrating over the interior volume of the cavity, and integrating by parts we are left with the surface term

$$\oint \epsilon A_0\epsilon^{-1}\vec{\rm E}\cdot d\vec{\rm S}\ , \tag{18}$$

where the integral extends over the walls. Since these conduct well, ϵA_0 must be strictly constant in (18); otherwise, $\epsilon\vec{\rm E}=\vec{\nabla}(\epsilon A_0)$ would not vanish in the walls, and a large current would flow. Thus, we may take ϵA_0 out of the integral (18). The remaining integral must vanish by (17) and Gauss's theorem. The volume integral is then

$$\int \vec{\rm E}^2 dV=0 \tag{19}$$

which proves that $\vec{\rm E}$ vanishes identically inside the cavity, in harmony with the ice-pail experiment.

Does the proposed electrodynamics predict observable departure from Coulomb's law? To answer this we calculate the *exterior* field of a spherical distribution of charge in flat space. We exclude magnetic monopoles; thus we allow only a spherically symmetric radial electric field. Outside the charges (17) is solved by

$$\vec{\rm E}=\hat{r}\epsilon Q/r^2\ , \tag{20}$$

where Q is a constant. By Gauss's law one interprets Q as $\sum e_0$ over the particles in the charge distribution (see Sec. III). In order to calculate ϵ we substitute (20) into the right-hand side of (16). Again, outside the charges, for a static situation we get the equation

$$r^{-2}\partial_r[r^2(\partial_r\ln\epsilon)]=(4\pi\hbar c)^{-1}l^2Q^2\epsilon^2r^{-4} \tag{21}$$

since $F^{\mu\nu}F_{\mu\nu}=-2\vec{\rm E}^2$.

By appropriate transformations we are able to solve (21). Here we only give the result:

$$\epsilon(r)=\sec[lQ(4\pi\hbar c)^{-1/2}r^{-1}]\ . \tag{22}$$

This can be checked by direct substitution. We observe that as $r\to\infty$, $\epsilon\to1$ in harmony with our choice of scale. At large distances [when $r>>lQ(4\pi\hbar c)^{-1/2}$], $\vec{\rm E}$ has a purely Coulombic form: $\vec{\rm E}=\hat{r}Qr^{-2}$. This accords with the success of Coulomb's law in experimental studies. If we are dealing with an elementary charge

$(Q/\sqrt{\hbar c} \cong 1/\sqrt{137})$, then we see that the Coulombic behavior is already accurate (argument of secant small compared to unity) for $r \gtrsim l$. Since leptons are known to have an accurately Coulombic field (apart from vacuum polarization corrections) down to distances 10^{-15} cm, we conclude that $l < 10^{-15}$ cm, in harmony with our earlier claim.

Let us now consider a spherical *macroscopic* distribution of charge of radius R. If $l < 10^{-15}$ cm, can one ever detect departures from the r^{-2} law outside such a distribution? This would require that $lQ(4\pi\hbar c)^{-1/2}R^{-1}$ be not much smaller than unity, i.e., that the *formal* electrostatic potential Q/R be no smaller than about $\sqrt{\hbar c}/l$. If $l = 10^{-16}$ cm, the required potential is at least 10^{10} V (1 esu = 300 V). For laboratory-sized charged objects, R is in the range $10^{-2} - 10^{3}$ cm, so there would be fields of strength at least 10^{7} V/cm present. Such fields exceed by far the critical field for air breakdown (3×10^{4} V/cm), so that the required experiments would have to be carried out in high vacuum (which is not usually done). Even this option would prove useless if $l \lesssim 10^{-23}$ cm since the required fields could never be set up; they would exceed the vacuum's breakdown field (10^{14} V/cm). As it is, potentials of 10^{7} V obtainable in Van de Graaf accelerators are about as large as can be had in the laboratory, but they fall short of our requirements.

Departures from spherical symmetry do not mend matters. For a collection of point charges Q_i, the solution of (17) is

$$\vec{E} = \epsilon \sum Q_i \vec{r}_i r_i^{-3} , \tag{23}$$

where $\vec{r}_i$ is the vector from Q_i to the field point. Instead of (21) we now have

$$\nabla^2 \ln\epsilon = (4\pi\hbar c)^{-1} l^2 \vec{E}^2 \tag{24}$$

outside the sources. As direct substitution and use of (17) shows, the solution of (23)—(24) is

$$\epsilon = \sec\left[l(4\pi\hbar c)^{-1/2} \sum Q_i r_i^{-1} \right] . \tag{25}$$

Thus despite the nonlinearity of (23)—(24), there is a sense in which the effects of separate charges can be superimposed. For $\vec{E}$ to depart from a superposition of Coulomb fields, ϵ must depart from unity. From (25) it is clear that this requires the potential $\sum Q_i/r_i$ to be large in our earlier sense. For the reasons already mentioned we again conclude that electrostatic experiments have trouble distinguishing between the variable α electrodynamics and Maxwell electrodynamics.

V. TEMPORAL VARIATIONS OF α

We now wish to calculate temporal variations of α or ϵ in a cosmological setting. We assume the universe to be homogeneous and isotropic in the large over the epochs of interest. Adopting a Robertson-Walker metric with expansion factor $a(t)$, we write the equation (16) for ϵ as

$$(a^3 \dot{\epsilon}/\epsilon)^{\cdot} = -\frac{a^3 l^2}{\hbar c}\left[\epsilon \frac{\partial \sigma}{\partial \epsilon} + \frac{1}{4\pi}(\vec{E}^2 - \vec{B}^2) \right] , \tag{26}$$

where a dot stands for a time derivative, and $\vec{E}$ and $\vec{B}$ are the fields measured by a commoving observer. Since we are not interested in spatial variations of ϵ resulting from graininess in the contents of the Universe, the source term in the right-hand side of (26) is to be understood as averaged over a large volume.

The Universe is filled with electromagnetic radiation, but this does not contribute to the source of ϵ in (26). We recall that for a plane electromagnetic wave the invariant $\vec{E}^2 - \vec{B}^2$ vanishes.[25] An incoherent distribution of waves (i.e., blackbody radiation) is a linear superposition of plane waves with no definite phase relations. Because of this last point, a space average of $\vec{E}^2 - \vec{B}^2$ lacks any cross terms between different waves. Thus, incoherent radiation does not contribute to (26). The incoherence also eliminates cross terms between fields of the waves and field of material particles. Thus *only matter* is a source of ϵ.

Both terms in the square brackets in (26) are Lorentz invariants. In evaluating the contribution of a particle to the source, one can thus pass to its rest frame.[26] Averaging such a contribution over a *large* volume V_0 in the rest frame of the particle gives, by (14), the source term

$$\left[\epsilon(\partial m/\partial \epsilon)c^2 + (4\pi)^{-1} \int_{V_0} (\vec{E}^2 - \vec{B}^2) dV \right] V_0^{-1} . \tag{27}$$

Now, one can represent the electromagnetic part of the energy mc^2, $m_{\mathrm{EM}}c^2$, by

$$m_{\mathrm{EM}}c^2 = (8\pi)^{-1} \int_{V_0} (\vec{E}^2 + \vec{B}^2) dV \tag{28}$$

so that the source can be written as

$$\left[\epsilon(\partial m/\partial\epsilon)c^2+2m_{\rm EM}c^2-(2\pi)^{-1}\int_{V_0}\vec{B}^2dV\right]V_0^{-1}\,. \tag{29}$$

Since magnetic energy is a very tiny fraction of the total electromagnetic energy in ordinary nonrelativistic matter, the magnetic term in (29) can be neglected. Further, the bulk of $m_{\rm EM}c^2$ is Coulombic in nature, so to a good approximation $m_{\rm EM}\propto\epsilon^2$. It follows that $\epsilon(\partial m/\partial\epsilon)c^2\approx 2m_{\rm EM}c^2$. Thus the quantity in square brackets in (26) is essentially *four* times the density of Coulomb energy.

The bulk of rest mass in the Universe is in the form of protons and neutrons, free and bound in nuclei; these are the main contributors to the Coulomb energy density. Electrons, because they are so light, contribute negligibly. So do atomic and interionic interactions. The Coulomb energy of the proton can be estimated in two ways. If we extrapolate the nuclear mass formula[27,28] to $Z=1$, $A=1$ we get $m_{\rm EM}c^2\approx 0.7$ MeV. If we model the proton by a uniformly charged sphere of radius 1 fm, we get $m_{\rm EM}c^2\approx 0.86$ MeV. Both estimates should be on the small side since the charge in the proton is strongly concentrated.[27] But even if it were concentrated within the proton's Compton length (which it is not), we should have $m_{\rm EM}c^2\approx\alpha mc^2\approx 7$ MeV. Thus, to within a factor of 2, $m_{\rm EM}c^2\approx 3$ MeV $\approx 3\times10^{-3}mc^2$. The neutron's rest energy exceeds the proton's by 1.29 MeV. It is generally agreed that this splitting is purely electromagnetic in origin.[27] Thus we estimate for the neutron $m_{\rm EM}c^2\approx 4.3$ MeV.

Most neutrons in the Universe have been incorporated into nuclei, primarily into He^4 which makes up about a quarter by mass of matter in the Universe. The mass formula indicates that the Coulomb contribution to the *binding* energy of the He^4 nucleus is ≈ 1.8 MeV, or 0.45 MeV per nucleon. Thus a nucleon in He^4 contributes only $\approx 15\%$ more Coulomb energy than a free nucleon. Nucleons in other abundant nuclei do better. For example, in O^{16} and Fe^{56} they contribute 40% and 70% more Coulomb energy than when free. However, all nuclei apart from H^1 and He^4 account for only 2% of all matter by mass.[28,29] Thus, when weighted by the appropriate abundances, the average nucleon's Coulomb energy exceeds the free proton contributions by only 10%. We may thus rewrite (26) as

$$(a^3\dot{\epsilon}/\epsilon)^{\cdot}=-a^3\zeta(l^2/\hbar c)\rho_m c^2\,, \tag{30}$$

where ρ_m is the total rest mass density of matter, and $\zeta\approx 1.3\times10^{-2}$.

We now argue that ζ may be taken as constant in (30). For it to change significantly there must be a significant change in the relative abundance of H^1 and other nuclei. It might be argued that nucleosynthesis in stars has been accomplishing just such a change. However, as we now show, the effect is negligible. All transformation of H^1 into heavier nuclei proceeds via the process $4H^1\rightarrow He^4$ which releases 2×10^{-5} erg per proton converted. This energy is the principal source of optical luminosity of galaxies. A recent estimate[30] sets the present mean luminosity density at 1.6×10^{-32} erg sec^{-1} cm^{-3} corresponding to a proton conversion rate of 8×10^{-28} sec^{-1} cm^{-3}. A *minimal* estimate for the nucleon number density in the Universe follows from considerations of primordial He^4 synthesis[31]: 1×10^{-7} cm^{-3}. Some 90% of these are protons.[28] Thus the time scale for proton conversion is $\gtrsim 10^{20}$ sec; that for significant changes in ζ is similar. But the time scale for changes of ρ_m or a^3 in (30) is $(3H_0)^{-1}\approx 10^{17}$ sec. Thus in (30) ζ is highly constant when compared with the other quantities. It changes $\lesssim 0.3\%$ per Hubble time H_0^{-1}. However, galaxies, when young, should have been much brighter than today.[32] Thus, constancy of ζ can safely be assumed only as far back as $z\approx 1$. Before then galaxies may have converted protons much faster than our estimate suggests, and ζ might have varied substantially.

Since ρ varies as a^{-3} in an expanding Universe, we can immediately integrate (30); assuming ζ constant

$$\dot{\epsilon}/\epsilon=-\zeta(l^2c/\hbar)\rho_m(t-t_C)\,, \tag{31}$$

where t_C is an integration constant. To estimate $\dot{\epsilon}/\epsilon$ at the present time t_0, we may as usual express ρ_m as a multiple Ω_{m0} of the critical density $\rho_C\equiv(3/8\pi G)H_0^2$, where G is the gravitation constant. Since $(G\hbar/c^3)^{1/2}$ is just the Planck-Wheeler length $L_{\rm PW}$, we get the suggestive relation

$$(\dot{\epsilon}/\epsilon)_{t=t_0}=-1.6\times10^{-3}(l/L_{\rm PW})^2\Omega_{m0}H_0^2\tau\,, \tag{32}$$

where $t\equiv t_0-t_C$ and we have set $\zeta=1.3\times10^{-2}$. An upper bound on $\dot{\epsilon}/\epsilon$ in terms of the unknowns l and τ follows by taking for H_0 the largest value discussed today, $H_0=100$ km/Mpc$=1.02\times10^{-10}$ yr^{-1}, and for Ω_{m0} the large value 0.2 obtained by assuming that all gravitating matter is nucleonic (an exaggeration), and that the specific luminosity of great clusters of galaxies equals that of the typi-

cal galaxy.[31] Both assumptions tend to overestimate Ω_{m0}. Thus

$$|(\dot{\epsilon}/\epsilon_0)_{t=t_0}| \leq 3.2\times 10^{-24}(l/L_{\rm PW})^2 |\tau| y^{-2} . \tag{33}$$

Turneaure and Stein[12] have set the laboratory bound $|\dot{\epsilon}/\epsilon| = \frac{1}{2}|\dot{\alpha}/\alpha| < 5\times 10^{-12}$ yr^{-1}. We do not know τ but since our equation is valid for times of order t_0, it would be strange indeed if $|\tau|$ were orders of magnitude larger than t_0. If $|\tau| < t_0$ and $l/L_{\rm PW} < 10$, or if $|\tau| < 0.1 H_0^{-1} \approx 0.1\ t_0$ and $(l/L_{\rm PW}) < 40$ there is no contradiction between (33) and the experimental bound. Evidently, in the absence of a concrete value for l, the experiment *cannot* rule out α variability.

We now turn to the evolution of ϵ in the past. Since $\rho_m \propto (1+z)^3$, where z is the red-shift, we may integrate (31) by using the standard relation between "lookback time" $\Delta t \equiv t_0 - t$ and red-shift[33]

$$z = H_0\Delta t + (1+\tfrac{1}{2}q_0)H_0^2\Delta t^2 + O(\Delta t^3) , \tag{34}$$

where q_0 is the deceleration parameter. We get

$$\rho_m(t) = (3/8\pi G)\Omega_{m0}H_0^2[1+3H_0\Delta t + 3(2+\tfrac{1}{2}q_0)H_0^2\Delta t^2 + O(\Delta t^3)] . \tag{35}$$

Recalling our convention $\epsilon(t_0)=1$ and assuming $|\epsilon - 1|$ to be small we get

$$\epsilon(t) - 1 = -1.6\times 10^{-3}(l/L_{\rm PW})^2\Omega_{m0}H_0^2\Delta t F(\Delta t,\tau) , \tag{36}$$

$$F(\Delta t,\tau) \equiv \tau[1+\tfrac{3}{2}H_0\Delta t + (2+\tfrac{1}{2}q_0)H_0^2\Delta t^2] - \tfrac{1}{2}\Delta t - H_0\Delta t^2 + O(\Delta t^3) . \tag{37}$$

From observations of fine-structure splitting in radio galaxies, Bahcall and Schmidt[11] concluded that at red-shift $z=0.2$, $|\epsilon(t)-1| = \frac{1}{2}|\alpha(t)/\alpha(t_0) - 1| < 1\times 10^{-3}$. For the currently discussed values, $0.05 < q_0 < 0.5$, $z=0.2$ corresponds to $H_0\Delta t \approx 0.17$. Since $\Omega_{m0} \leq 0.2$, (36) gives

$$|\epsilon(z=0.2) - 1| \leq 5.4\times 10^{-5}(l/L_{\rm PW})^2 |1.32 H_0\tau - 0.11| . \tag{38}$$

Thus, if $|\tau| < t_0$ while $l/L_{\rm PW} < 3$, or if $|\tau| < 0.1 H_0^{-1} \approx 0.1 t_0$ while $l/L_{\rm PW} < 8$, the prediction is below the observational bound. From observations of fine-structure splitting in a BL-Lac object, Wolfe, Brown, and Roberts[13] deduce that at $z=0.52$, $|\epsilon - 1| < 0.03$. Now $z=0.52$ corresponds to $H_0\Delta t \approx 0.37$ for $0.05 < q_0 < 0.5$. With $\Omega_{m0} \leq 0.2$ (36) gives

$$|\epsilon(z=0.52) - 1| \leq 1.1\times 10^{-4}(l/L_{\rm PW})^2 (1.86 H_0\tau - 0.32) . \tag{39}$$

Thus, if $|\tau| < t_0$ and $l/L_{\rm PW} < 10$ or if $|\tau| < 0.1 H_0^{-1}$ while $l/L_{\rm PW} < 20$, the prediction is again below the observational bound. Thus the astronomical observations do *not* rule out α variability unless one is willing to assume that the shortest conceivable scale is more than an order-of-magnitude larger than $L_{\rm PW}$.

From the Re^{187}, Os^{187} abundances in ores, Dyson[9,10] deduced that 4.5×10^9 yr ago $|\epsilon - 1| < 5\times 10^{-5}$. With $H_0 \approx 1.02\times 10^{-10}$ yr^{-1} and $\Omega_{m0} \approx 0.2$, (36) predicts

$$|\epsilon(\Delta t = 4.5\times 10^9 \text{yr}) - 1| \approx 1.5\times 10^{-4}(l/L_{\rm PW})^2 |2.2 H_0\tau - 0.44| . \tag{40}$$

Even for $l \approx L_{\rm PW}$ this is consistent with Dyson's constraint only if $\tau = H_0^{-1}(0.2 \pm 0.15)$. Shlyakhter[15] deduced from the isotopic abundances in the Oklo "natural reactor" that 1.8×10^9 yr ago $|\epsilon - 1| < 5\times 10^{-9}$. The appropriate prediction is made by (38) since $\Delta t = 1.8\times 10^9$ yr corresponds closely to $H_0\Delta t = 0.17$. With $\Omega_{m0} = 0.2$ and $l \approx L_{\rm PW}$ it is consistent with Shlyakhter's constraint only if $\tau = H_0^{-1}(0.08 \pm 7\times 10^{-5})$. The two determinations of τ are consistent; however, this is no longer so if $l/L_{\rm PW} \geq 1.3$. Thus Dyson and Shlyakhter constraints *together* force l down to near its minimum allowable value; in this sense they *marginally* rule out α variability. However, this conclusion is strongly conditional on the assumed cosmological parameters. Consideration of He^4 synthesis in the big bang suggests[31] that values of $\Omega_{m0}H_0^2$ a factor 25 times smaller than we have assumed are quite possible (recall, Ω_{m0} refers to nuclear matter only). With these low values $l/L_{\rm PW}$ need no longer be near unity. Also, small $\Omega_{m0}H_0^2$ allows the very narrow range of values for

τ deduced from Shlyakhter's constraint to broaden considerably. Thus the suggestive implication that Shlyakhter's result requires the present time t_0 to be a very special one ($t_0 - t_C$ very precisely determined), or α not to vary, looses much of its force. We conclude that only when Ω_{m0} is determined more accurately will it be clear if the geophysical constraints provide a strong case against α variability.

VI. SPATIAL VARIATIONS OF α

In the laboratory there should be spatial gradients of α due to nearby matter with electromagnetic structure. The relevant equation is (16). If one is interested only in macroscopic gradients, one may average its source over volumes containing many atoms. As shown in Sec. V, one can then replace the quantity in parentheses in (16) by $\zeta\rho_m c^2$, where ζ is a number of order 10^{-2}. Actually ζ varies somewhat with the composition of the source. Thus the ζ appropriate to the sun (predominantly H) is somewhat smaller than that appropriate to the earth (mostly heavy elements). In a static situation (16) becomes

$$\nabla^2 \ln\epsilon = \zeta (l/L_{\rm PW})^2 G c^{-2} \rho_m \tag{41}$$

[recall $L_{\rm PW} = (G\hbar/c^3)^{1/2}$].

Our convention for ϵ is that $\epsilon \to 1$ at infinity. Hence $\ln\epsilon$ obeys the same boundary condition as gravitational potential ϕ. In fact, by comparing (41) with the Poisson equation, we have

$$\ln\epsilon = (4\pi c^2)^{-1} \zeta (l/L_{\rm PW})^2 \phi \ . \tag{42}$$

That any fractional variation of α (or 2 $\ln\epsilon$ for $\epsilon \approx 1$) should be proportional to ϕ was conjectured by Dicke.[34,6] He estimated the proportionality constant as α/c^2. Since we know now that $l \sim L_{\rm PW}$, we see that this is about right.

As discussed by Dicke,[6] gradients of α give rise to "anomalous" acceleration. Basically, the total force acting on a *neutral* body of mass M bearing electromagnetic energy $E_{\rm EM}$ is, by energy conservation,

$$\begin{aligned} \vec{\rm F} &= -M\vec{\nabla}\phi - \vec{\nabla} E_{\rm EM} \\ &= -M\vec{\nabla}\phi - (\partial E_{\rm EM}/\partial\epsilon)\vec{\nabla}\epsilon \ . \end{aligned} \tag{43}$$

Here we have neglected a fractional correction of order ϕ/c^2 to the last term. Since $E_{\rm EM}$ is for the most part Coulombic, one can replace $\partial E_{\rm EM}/\partial E$ by $2E_{\rm EM}\epsilon^{-1}$ (see Sec. V). By substituting (42) into (43) we get for the acceleration $\vec{\rm F}/M$

$$\vec{a} = -[1 + (2\pi)^{-1}\zeta (l/L_{\rm PW})^2 (E_{\rm EM}/Mc^2)]\vec{\nabla}\phi \ . \tag{44}$$

The second term in the square brackets varies from body to body, giving rise to an anomalous composition-dependent acceleration. Tests of the equivalence principle search for just such a term. In the typical experiment the accelerations of two substances are compared and a bound is set on the relative difference, i.e., on

$$D = (2\pi)^{-1}\zeta (l/L_{\rm PW})^2 \Delta (E_{\rm EM}/Mc^2) \ . \tag{45}$$

Bessel[35] used a pendulum to establish that $|D| < 2\times 10^{-5}$ for various pairs of substances. The source in this case was the earth (Fe, Si, C, O, $\cdots$) for which we estimate roughly $\zeta \approx 1\times 10^{-2}$. We also estimate $\Delta(E_{\rm EM}/Mc^2) \approx 1\times 10^{-3}$ (see below). Bessel's constraint then tells us that $l/L_{\rm PW} < 3.5$. Thus, this ancient test of the equivalence principle sets a bound on l as tight as any of the astronomical bounds (Sec. V), but one free of ambiguity related to the uncertain values of Ω_{m0} and t_C. The 1922 torsion balance test by Eötvös, Pekár, and Fekete[36] gave $|D| < 5\times 10^{-9}$ for various pairs. Using again $\zeta \approx 1\times 10^{-2}$ and $\Delta(E_{\rm EM}/Mc^2) \approx 1\times 10^{-3}$ we infer that $l/L_{\rm PW} < 10^{-1}$. This result argues against α variability. Evidently, experiments sensitive to spatial variation of α are the only ones sensitive enough to decide the question. In order to make the most of the data from modern tests of the equivalence principle, we now turn to some fine points.

In his pioneering analysis of the relevance of the experiments for α variability,[6] Dicke calculated $E_{\rm EM}$ from the Coulomb interaction energies between nucleons alone. Evidently atomic and lattice energies can safely be neglected by comparison. But at first sight it seems that self-energies of nucleons cannot, so let us include them. If ϵ_e, ϵ_p, and ϵ_n denote the self-electromagnetic energies of electron, proton, and neutron, respectively, we should have for an atom of atomic and mass numbers Z and A

$$\begin{aligned} E_{\rm EM}(Z,A) = {} & Z(\epsilon_e + \epsilon_p) + (A-Z)\epsilon_n \\ & + 7.7\times 10^{-4} \bar{m} c^2 Z^2 A^{-1/3} \ , \end{aligned} \tag{46}$$

where $\bar{m}$ is the atomic mass unit ($\frac{1}{12}$ of the mass of the C^{12} atom); the last term in (46) is the Coulomb term from the nuclear mass formula.[28] To a sufficiently good approximation $M(Z,A) = A\bar{m}$ (the maximal error is 0.3%). Hence, for nuclei specified by Z_1, A_1 and Z_2, A_2,

$$\left[\frac{E_{\rm EM}}{Mc^2}\right]_1-\left[\frac{E_{\rm EM}}{Mc^2}\right]_2=\left[\frac{Z_2}{A_2}-\frac{Z_1}{A_1}\right]\frac{\Delta}{\bar{m}c^2}+7.7\times10^{-4}\left[\frac{{Z_1}^2}{{A_1}^{4/3}}-\frac{{Z_2}^2}{{A_2}^{4/3}}\right], \tag{47}$$

where $\Delta\equiv\epsilon_n-\epsilon_p-\epsilon_e$. Evidently $\epsilon_e<m_ec^2=0.51$ MeV. Also, the 1.29 MeV mass splitting between neutron and proton is known to come exclusively from electromagnetic effects: $\epsilon_n-\epsilon_p=1.29$ MeV. Thus $8.4\times10^{-4}<\Delta/\bar{m}c^2<1.4\times10^{-3}$. In most cases the two terms in (47) are opposite in sign. When one nucleus is heavy and the other light, as in the actual experiments, the correction due to self-electromagnetic energies is small (typically 5%); this justifies Dicke's approximation.

Roll, Krotkov, and Dicke[37,6] showed that the sun accelerates gold and aluminum masses equally to great accuracy. Their formal result was $D=(0.96\pm1.04)\times10^{-11}$ (1σ interval). Thus at the 95% confidence level $D<2.7\times10^{-11}$. For gold and aluminum the difference (47) is 2.7×10^{-3}. We get a definite lower bound on ζ of the sun by regarding it as composed only of hydrogen, and taking the minimal value 0.7 MeV for ϵ_p (see Sec. V). This gives $\zeta>3\times10^{-3}$. Putting all these into (45) we come up with the 95% confidence bound $l/L_{\rm PW}<5\times10^{-3}$. Braginsky and Panov[18] established with 95% confidence that $D=(0.3\pm0.9)\times10^{-12}$ for platinum and aluminum. For these the difference in (47) is also about 2.7×10^{-3}. With our earlier bound on ζ we get (95% confidence) $l/L_{\rm PW}<1\times10^{-3}$. Thus, the Eötvös-Dicke-Braginsky experiments strongly rule out the framework for α variability developed here from assumptions P1–P8. Because these are reasonable assumptions we now know with some confidence that α is a parameter, not a dynamical variable. It undergoes *no* spatial or temporal changes whatsoever.

VI. CONCLUSIONS AND CAVEATS

Experimental constraints on variation of α cannot by themselves rule out variability. A theoretical framework capable of making specific predictions is required to judge the relevance of any particular constraint. Just such a framework has been developed under very general assumptions; it necessarily entails a modification of Maxwellian electrodynamics. A characteristic length l enters into it. An experimental constraint rules out α variability of any kind if it is in clear conflict with predictions of the framework for l no shorter than the fundamental length 10^{-33} cm. By this criterion neither the astronomical constraints nor those from laboratory intercomparison of clocks are sufficiently sensitive to rule out variability. The geophysical constraints are marginally sensitive, but uncertainty about various cosmological parameters precludes firm conclusions. The constraints from laboratory tests of the equivalence principle are very sensitive and strongly rule out all α variability.

Our approach divorces electromagnetism from the weak and strong interactions. Yet, it is suspected that all three are unified at high energies. Thus it would be more logical to introduce a spacetime variable coupling constant into the unified gauge theory of the three interactions and recover electromagnetism with a variable α after symmetry breaking. What guarantee have we that our scheme would emerge? Although only a full treatment can decide the question, we would argue that the very generality of our assumptions precludes other outcomes *if,* in fact, electromagnetism separates cleanly from the other interactions in the presence of a variable coupling constant. A full treatment would be worthwhile, not only to verify this point, but because it would make it clear whether our conclusions about α carry through to the constants of the weak and strong interactions, or whether further experiments are needed to establish their strict constancy.[38]

ACKNOWLEDGMENTS

The author is grateful to Freeman Dyson, Amnon Meisels, and Jim Peebles for helpful discussions and to Robert Dicke for drawing attention to his earlier formulation of the variable-α problem. This work was supported by the Rector's Fund, Ben Gurion University, Israel.

APPENDIX

Consider the Lagrangian density

$$\mathcal{L}=-\tfrac{1}{2}f(\psi_{,\mu}\psi^{,\mu}), \tag{A1}$$

where $f(x)$ is an arbitrary function. The factor $-\frac{1}{2}$ is conventional; ψ represents $l\ln\epsilon$, and for simplicity we have dropped l. For what f's is (A1)

consistent with causality? We show that only for $f(x) \propto x$.

The wave equation corresponding to (A1) is

$$f' g^{\mu\nu} + 2f'' \psi^{,\mu} \psi^{,\nu} \psi_{;\mu\nu} = \text{source} , \tag{A2}$$

where f' and f'' denote first and second derivatives with respect to the *argument*. We consider the evolution of small disturbances $\delta\psi$ on the background $(g_{\mu\nu}, \psi)$. The characteristics $S(x^\alpha) = \text{const}$ of the corresponding linearized equation obey

$$f' g^{\mu\nu} S_{,\mu} S_{,\nu} + 2f''(\psi^{,\mu} S_{,\mu})^2 = 0 . \tag{A3}$$

Now, by causality characteristics cannot be spacelike surfaces (information transmitted outside the light cone) so $g^{\mu\nu} S_{,\mu} S_{,\nu} \geq 0$. Consider the option $g^{\mu\nu} S_{,\mu} S_{,\nu} > 0$. It follows from (A3) that f' and f'' always have opposite signs. This rules out maxima, minima, or inflection points of $f(x)$. Hence, either (a) $f' > 0$ and $f'' < 0$ for all x, or else (b) $f' < 0$ and $f'' > 0$ for all x. We also want f' and f'' to be well behaved in order that the wave equation be well defined. This is especially true near $x = 0$ because for easily realizable gradients $| l^2 \epsilon^{-2} \epsilon_{,\alpha} \epsilon^{,\alpha} |$ should be small if, as we know, l is a very short length. Thus for small $| x |$ we must have $f' \cong \alpha + \beta x^\gamma$ for constant α, β and γ.[39] For case (a) above we need $\alpha, \beta > 0$ ($f' > 0$ for $x > 0$), $\gamma =$ ratio of even to odd integers (f' real and positive for $x < 0$), and $\gamma < 0$ ($f'' < 0$ for $x > 0$). But then $f'' > 0$ for $x < 0$ and $f', f'' \to \infty$ at $x = 0$, so this alternative is excluded. Case (b) is excluded in a similar manner.

We return to the alternative $g^{\mu\nu} S_{,\mu} S_{,\nu} = 0$ (null characteristics). The field $\psi_{,\mu}$ can be quite complicated and need not lie in the null surfaces of the spacetime. In fact if $\psi_{,\mu}$ is timelike ($\epsilon_{,\mu}$ in an expanding Universe), it cannot lie in null surfaces. Hence, in general $\psi^{,\mu} S_{,\mu} \neq 0$. It follows from (A3) that $f'' = 0$. Now, $\psi^{,\mu} \psi_{,\mu}$ need not be constant on a characteristic. The implication is that $f(x) = ax + b$ for constant a, b. Of course b does not generate any dynamics for ψ and so can be dropped.

[1] P. A. M. Dirac, Nature 139, 323 (1937); Proc. R. Soc. London A165, 199 (1938).

[2] E. Teller, Phys. Rev. 73, 801 (1948).

[3] P. Jordan, *Schwerkraft and Weltall* (Vieweg, Braunschweig, 1955); Z. Phys. 157, 112 (1959).

[4] R. H. Dicke and P. J. E. Peebles, Phys. Rev. 128, 2006 (1962).

[5] Whenever necessary we adopt the value $H_0 = 100$ km/Mpc $= 1.02 \times 10^{-10}$ yr^{-1}.

[6] R. H. Dicke, *The Theoretical Significance of Experimental Relativity* (Gordon and Breach, New York, 1965).

[7] G. Gamow, Phys. Rev. Lett. 19, 759 (1967).

[8] A. Peres, Phys. Rev. Lett. 19, 1293 (1967).

[9] F. J. Dyson, Phys. Rev. Lett. 19, 1291 (1967).

[10] F. J. Dyson, in *Aspects of Quantum Theory*, edited by A. Salam and E. Wigner (Cambridge University Press, London, 1972).

[11] J. N. Bahcall and M. Schmidt, Phys. Rev. Lett. 19, 1294 (1967).

[12] J. P. Turneaure and S. R. Stein, in *Atomic Masses and Fundamental Constants, Vol. 5*, edited by J. H. Sanders and A. H. Wapstra (Plenum, New York, 1976).

[13] A. M. Wolfe, R. L. Brown, and M. S. Roberts, Phys. Rev. Lett. 37, 179 (1976).

[14] M. P. Savedoff, Nature 178, 689 (1956); J. Bahcall, W. Sargent, and M. Schmidt, Astrophys. J Lett. 149, L11 (1967); A. D. Tubbs and A. M. Wolfe, *ibid.* 236, L105 (1980).

[15] A. I. Shlyakhter, Nature 264, 340 (1976).

[16] J. Bekenstein, Comments Astrophys. 8, 89 (1979).

[17] J. Bekenstein and A. Meisels, Phys. Rev. D 22, 1313 (1980); Astrophys. J. 237, 342 (1980).

[18] V. B. Braginsky and V. I. Panov, Zh. Eksp. Teor. Fiz. 61, 873 (1972) [Sov. Phys.—JETP 34, 463 (1972)].

[19] D. C. Cheng and G. K. O'Neill, *Elementary Particle Physics* (Addison-Wesley, Reading, 1979).

[20] I. S. Altarev *et al.*, Nucl. Phys. A341, 269 (1980); E. A. Hinds and P. G. H. Sanders, Phys. Rev. A 21, 480 (1980); W. B. Dress, P. D. Miller, J. M. Pendlebury, D. Perrin, and N. F. Ramsey, Phys. Rev. D 15, 9 (1977).

[21] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973), p. 1191.

[22] L. D. Landau and E. M. Lifshitz, *Classical Theory of Fields* (Pergamon, Oxford, 1962), 2nd edition.

[23] Our conventions are signature $-+++$; Greek (Latin) indices are spacetime (space) indices; the summation convention applies; semicolon denotes covariant derivative; g denotes determinant of metric $g_{\mu\nu}$.

[24] A Meisels, dissertation, Ben Gurion University, 1979 (unpublished).

[25] This Maxwellian theorem applies to an excellent approximation because ϵ varies little; see Sec. III.

[26] For several particles this approach still gives the correct answer if interactions are negligible. If not there will be an error of order v^2/c^2 times the interac-

tion energy density.
[27]E. Segre, *Nuclei and Particles* (Benjamin, New York, 1965).
[28]K. R. Lang, *Astrophysical Formulae* (Springer, New York, 1974).
[29]L. Spitzer, *Physical Processes in the Interstellar Medium* (Wiley, New York, 1978).
[30]M. Davis, M. J. Geller, and J. Huchra, Astrophys. J. 221, 1 (1978).
[31]K. A. Olive, D. N. Schramm, G. Steigman, M. S. Turner, and J. Yang, Astrophys. J. 264, 557 (1981).
[32]B. Tinsley, in *Physical Cosmology,* edited by R. Balian, J. Adouze, and D. N. Schramm (North-Holland, Amsterdam, 1980).
[33]S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972), p. 442.
[34]R. H. Dicke, Science 129, 621 (1959).
[35]F. W. Bessel, Poggendorff's Ann. 25, 401 (1832).
[36]R. V. Eötvös, V. Pekár, and E. Fekete, Ann. Phys. (Leipzig) 68, 11 (1922).
[37]P. G. Roll, R. Krotkov, and R. H. Dicke, Ann. Phys. (N.Y.) 26, 442 (1964).
[38]For bounds on their variability see C. M. Will, in *General Relativity,* edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, 1979).
[39]Other conceivable behaviors such as $\ln|x|$ and $\exp(-1/x^2)$ are singular and lead to varying signs for f' and f''.