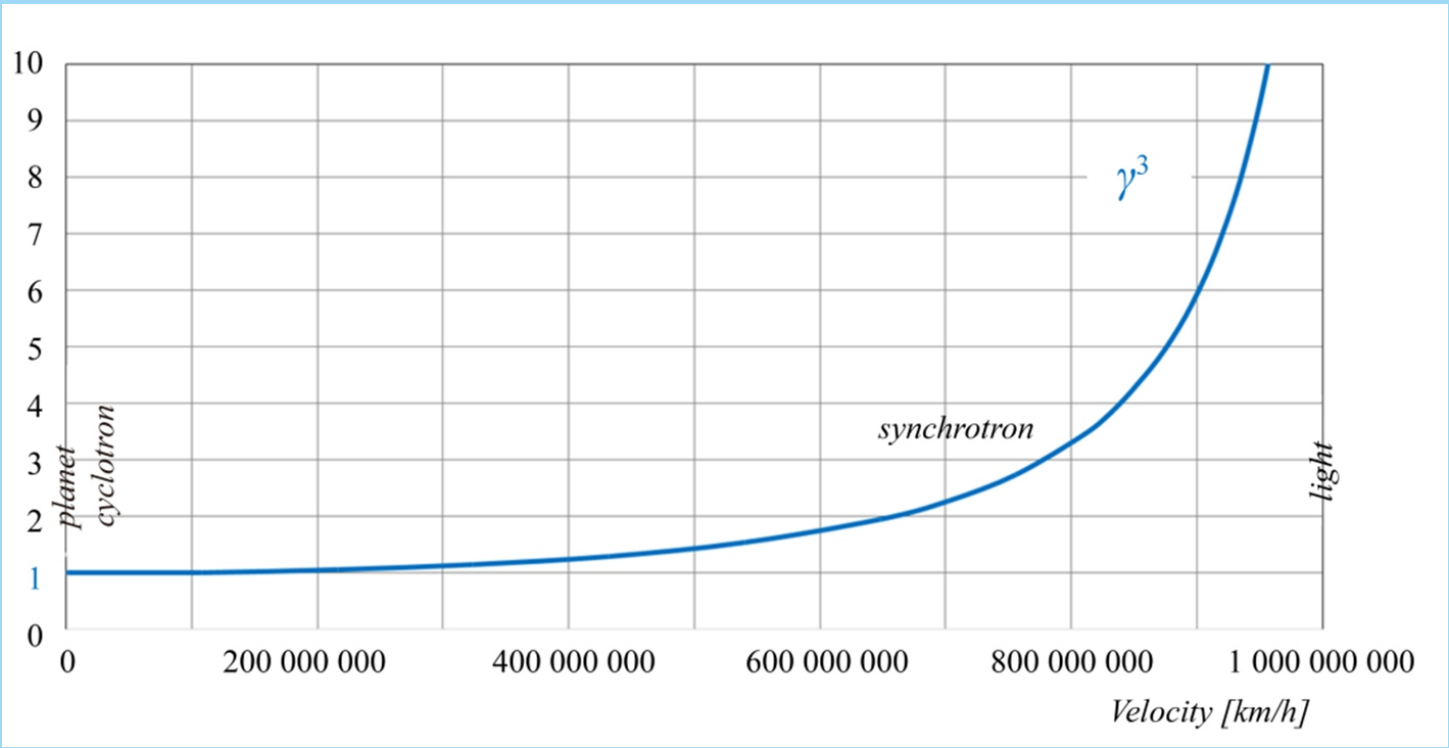
ISSN: 2153-1196 Vol. 7, No. 7, April 2016

Scientific Research Publishing

Journal of Modern Physics

Research on Gravitation, Astrophysics and Cosmology



ISSN: 2153-1196

www.scirp.org/journal/jmp

Journal Editorial Board

ISSN: 2153-1196 (Print) ISSN: 2153-120X (Online)
http://www.scirp.org/journal/jmp

Journal of Modern Physics, 2016, 7, 587-728
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jmp

Table of Contents

Volume 7 Number 7 April 2016

The figure on the front cover is from the article published in Journal of Modern Physics, 2016, Vol. 7, No. 7, pp. 656-661 by Olivier Serret.

Journal of Modern Physics (JMP)

Journal Information

SUBSCRIPTIONS

The *Journal of Modern Physics* (Online at Scientific Research Publishing, www.SciRP.org) is published monthly by Scientific Research Publishing, Inc., USA.

Subscription rates:

Print: $89 per issue.
To subscribe, please contact Journals Subscriptions Department, E-mail: sub@scirp.org

SERVICES

Advertisements
Advertisement Sales Department, E-mail: service@scirp.org

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.
E-mail: sub@scirp.org

COPYRIGHT

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: jmp@scirp.org

Journal of Modern Physics, 2016, 7, 587-610
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jmp
http://dx.doi.org/10.4236/jmp.2016.77061

On the Origin of Charge-Asymmetric Matter. I. Geometry of the Dirac Field

Alexander Makhlin

Rapid Research Inc., Southfield, MI, USA
Email: amakhlin@comcast.net

Received 25 February 2016; accepted 25 April 2016; published 28 April 2016

Abstract

This work presents new round of the author's pursuit for consistent description of the finite sized objects in classical and quantum field theory. Current paper lays out an adequate mathematical background for this quest. A novel framework of the matter-induced physical affine geometry is developed. Within this framework, (1) an intrinsic nonlinearity of the Dirac equation becomes self-explanatory; (2) the spherical symmetry of an isolated localized object is of dynamic origin; (3) the auto-localization is a trivial consequence of nonlinearity and wave nature of the Dirac field; (4) localized objects are split into two major categories that are clearly associated with the positive and negative charges; (5) of these, only the former can be stable as isolated objects, which explains the global charge asymmetry of the matter observed in Nature. In the second paper, the nonlinear Dirac equation is written down explicitly. It is solved in one-body approximation (in absence of external fields). Its two analytic solutions unequivocally are positive (stable) and negative (unstable) isolated charges. From the author's current perspective, the so for obtained results must be developed further and applied to various practical and fundamental problems in particle and nuclear physics, and also in cosmology.

Keywords

Dirac Field, Affine Geometry, Localization, Cosmological Charge Asymmetry

1. Introduction

This work addresses the long-standing puzzle of how the physical Dirac field of real matter becomes a finite-sized particle. This puzzle successfully withstood several major attacks undertaken since early 1930s in both classical and quantum contexts, for realistic fields of matter and for the *ad hoc* constructed effective field theories. The importance of a definitive conclusion goes far beyond the purely academic area. The present

How to cite this paper: Makhlin, A. (2016) On the Origin of Charge-Asymmetric Matter. I. Geometry of the Dirac Field. *Journal of Modern Physics*, **7**, 587-610. http://dx.doi.org/10.4236/jmp.2016.77061

uncertainty of an answer affects numerous studies in theoretical and experimental physics (e.g., quantum collisions of finite-size ultrarelativistic nuclei [1]) and reaches as far as the origin of the observable matter in the Universe. A dramatic difference between the charge asymmetry of the visible Universe and an apparent charge symmetry observed in transformations of elementary particles has never received a rational explanation. In mid-1960s, A.D. Sakharov [2] made an attempt to connect the cosmological charge asymmetry with the violation of the *CP*-invariance and nonequilibrium processes in the *early hot Universe*, but this hypothesis cannot be verified by a laboratory experiment. More specified (and exotic) scenarios were considered by A. Dolgov [3] [4], partially in connection with the problem of baryogenesis. For an extensive review with further references see Ref. [5].

The present study concludes that, for the Dirac field, *C* and *P* do not exist separately, and that both are intimately connected to inevitable localization of the Dirac field into finite-sized particles. Furthermore, it appears that *only positive charges are capable of stable auto-localization in real world*. The time scale and relative weight of all the underlying processes and/or mechanisms are not yet clear, but the Universe definitely had enough time to conduct such an experiment. Moreover, experimental studies of the last decade [6] revealed a surprising excess of positrons (and no excess of antiprotons) in the cosmic rays, which can be an indication that creation of the charge-asymmetric matter in the Universe is *an ongoing process*.

The present work was supposed to correct and augment the author's paper [7], which was focused mainly on the transient processes with localized particles. The accents have changed with the initial progress. In this work and then in paper [8], we pursue a somewhat narrower goal to find an exact auto-localized solution (a realistic Dirac particle), which could serve as an input for the study of transient processes. The problem is posed and solved in a novel framework of the *matter-induced affine geometry*, which deduces geometric relations in the space-time continuum from the dynamic properties of the Dirac field.

Framework is set in Section 2 by reviewing well-known algebraic identities between the bilinear Dirac forms (the Fierz identities). At any point in spacetime continuum (the principal differentiable manifold $\mathbb{M}$), there exist four fields of quadruples of these forms (*the Dirac currents*), which are linearly independent and Lorentz-orthogonal, and can serve as local algebraic basis for any four-dimensional vector space, including the infinitesimal displacements in coordinate space $\mathbb{R}^4$.

In Section 3 we use this basis of four Dirac currents as the Cartan's moving frame in spacetime and develop the technique of covariant derivatives for the vector and spinor fields.

Relying on results of Section 2 and Section 3, we meticulously derive in Section 4 various differential identities from the Dirac equations of motion. These identities are shown to be imperative for the geometry of the objects associated with the Dirac field to have a covariant form and be independent of coordinate background. We discover that coordinate lines and surfaces cannot be chosen by a fiat—the Dirac field cannot be embedded into a coordinate basis $\left(\partial/\partial x^{\mu}\right)$ (this observation had triggered the present work starting from [7], where the key argument regarding localization was found). In Section 5 the differential identities for the divergences and curls of the Dirac currents are written down in terms of components, and properties of the congruences of the Dirac currents are analyzed. All components of the connections are found as functions of the Dirac field. These two steps finalize the formal design of the *physical affine geometry*. There are only a few digressions regarding physical meaning of some equations, the most important of which is related to the existence of the matter-defined world time τ and the local time slowdown. The latter is the main physical mechanism behind the auto-localization. It appears that, in order to be compatible with the Dirac equation, its coordinate basis indeed cannot be holonomic.

The known connections made it possible to examine the properties of the admissible coordinate systems. Among four tetrad vector fields, we find in Section 6.1 two integrable subsets of three PDEs for the coordinate lines (two hypersurfaces with the corresponding normal congruences) and two two-dimensional surfaces. In Section 6.2 we study the internal geometry of these surfaces as submanifolds of $\mathbb{M}$. It appears that the two-dimensional surface of the constant "world time" and "radius" can be only spherical, which seems to be inevitable for an isolated stable object.

The general properties of coordinate surfaces in $\mathbb{M}$ (like their spherical symmetry and inherent stability) are discovered in the present paper without any assumptions on the nature of an ambient space or Dirac field. It appears that the main qualitative characteristic of the stationary Dirac object is the direction of the axial current, which can point only outward or inward. It must be clearly understood that *the locally defined notions of outward and inward are prerequisites for any reasonable discussion of the localization phenomenon*. The frame-

work of the matter-induced affine geometry not only ideally fits this goal but also explains the auto-localization, as it is seen in the real world, as an intrinsic property of the Dirac field.

This paper is continued in Ref. [8], where the capabilities of the matter-induced affine geometry are employed to address a specific problem of existence of the auto-localized Dirac waveforms. We begin with writing down the nonlinear Dirac equation and putting it in a practically solvable form. The localized configurations of the Dirac field are found analytically in the absence of external electromagnetic field. They require the Dirac spinor to have only up- or only down-components, when the axial current is pointing outward or inward, respectively. The up-mode is stable, has a bump of invariant density and the *negative* energy $E=-m$, while the down-mode is unstable, has a dip and the *positive* energy $E=+m$. At large spatial distances the invariant density has a universal vacuum unity value. Therefore, the two modes were (by a fortunate coincidence!) properly *interpreted* as positive and negative charges. The decay of unstable mode is due to the charged Dirac currents that naturally oscillate as $\mathrm{e}^{2im\tau}$, such a decay requiring only the presence of an external electromagnetic field. Possibly, these facts explain the vivid global charge (eventually, baryonic one) asymmetry in the Universe. Last section of paper [8] summarizes ideas, methods, current results and perspectives.

2. Vectors at a Point. Algebra of the Dirac Currents

1. Mathematical framework. We consider, as usually, the *mathematical spacetime* as a smooth four-dimensional manifold $\mathbb{M}$ so that every point P of $\mathbb{M}$ has an open neighborhood that can be mapped one-to-one onto an opened subset of points $\left(x^0,x^1,x^2,x^3\right)\in\mathbb{R}^4$. From the viewpoint of the differential topology, one has to start with scalar functions $f\left(\mathrm{P}\left(\lambda\right)\right)\subset\mathbb{M}$ on the curves $\mathrm{P}\left(\lambda\right)$ (determined by a map $\mathbb{R}^1\to\mathbb{M}$, $\lambda\in\mathbb{R}^1$) in order to build at each point $\mathrm{P}\in\mathbb{M}$ the linear space $T_p\left(\mathbb{M}\right)$ of tangent 4-vectors

$$\boldsymbol{h}_\lambda\left(\mathrm{P}\right)=\left(\mathrm{d}/\mathrm{d}\lambda\right)=\partial_{\boldsymbol{h}}=\left(h^\mu\partial_\mu\right)\Big|_{\mathrm{P}(\lambda)} \tag{2.1}$$

with the components h^μ with respect to the linearly independent vectors $\left(\partial/\partial x^\mu\right)_{\mathrm{P}}$ of the coordinate basis in $\mathbb{R}^4$.

Being defined via the mapping $\mathbb{R}^1\to\mathbb{M}$, a curve and its tangent vectors are invariant objects; only the components h^μ of a vector explicitly depend on a particular choice of coordinates in $\mathbb{R}^4$. Action of operator (2.1) on the functions $f\left(x\right)=x^\nu$ yields the system of ODEs for the unknown x^μ, $\mathrm{d}x^\nu/\mathrm{d}\lambda=h^\nu\left(x\right)$. It is said that u^μ are components of a vector if they are transformed as components $\mathrm{d}x^\mu$ of a displacement $\mathrm{d}x$.

Any four linearly independent vectors $\boldsymbol{h}_a=\mathrm{d}/\mathrm{d}\lambda^a=h_a^\mu\partial_\mu$, $a=0,1,2,3$, (with the non-degenerate matrix $\left|h_a^\mu\right|$, $\det\left|h_a^\mu\right|\neq 0$) can be used as the basis. Then there also exists the inverse matrix h_μ^a of the *1-forms* $\boldsymbol{h}^a$ so that $h_a^\mu h_\nu^a=\delta_\nu^\mu$ and $h_a^\mu h_\mu^b=\delta_a^b$. Since any quadruple $\boldsymbol{u}$ of numbers can be expanded over the basis $\boldsymbol{h}_a$, we have $\boldsymbol{u}=u^a\boldsymbol{h}_a=u^a h_a^\mu\partial_\mu=u^\mu\partial_\mu$. Therefore, $u^a=\boldsymbol{u}\boldsymbol{h}^a=h_\mu^a u^\mu$ and $\mathrm{d}\lambda^a=h_\nu^a\mathrm{d}x^\nu$, but in general, $\mathrm{d}\lambda^a$ are not the total differentials of any independent variables.

2. Physical framework. Basis of Dirac currents. In physical spacetime of special relativity points P are associated with events. The clocks of the net that register these events are synchronized by light signals; this results in Lorentz transformations between the coordinates of events measured by the nets of different inertial observers. Special relativity is based on independence of all physical processes from a particular choice of an inertial frame, and thus from the coordinate basis that is used to parameterize the events. As a matter of fact, the coordinate basis is built into a material reference frame, and thus is an invariant object.

All mathematical treatments of affine or Riemannian geometry start with an assumption of the independent tangent space with an arbitrarily oriented normal basis at every point of the continuum (differentiable manifold). While invariance with respect to the choice of coordinates x^μ is trivial, there cannot be absolute freedom of choosing tetrad vectors at every point—the components $h_a^\mu\left(x\right)$ of tetrad vectors must be continuous functions of the coordinates. Is there a way to endow the principal manifold $\mathbb{M}$ with basis of vector fields that would be invariant objects without reference to curves and/or derivatives at a point? *For the physical four-dimensional spacetime the answer is affirmative*, because there exists a matter field, the Dirac field $\psi\left(\mathrm{P}\right)$, a coordinate scalar, that provides such a basis at each point P of the manifold $\mathbb{M}$ and assigns the latter the status of a physical object. The algebraic descendants of the Dirac field are the vector-like objects, the so-called Dirac currents,

$$\boldsymbol{J}_{[0]}\left(\mathrm{P}\right)=j^a\left(\mathrm{P}\right)=\psi^+\alpha^a\psi,\quad \boldsymbol{J}_{[3]}\left(\mathrm{P}\right)=J^a\left(\mathrm{P}\right)=\psi^+\rho_3\alpha^a\psi,\quad \boldsymbol{J}_{[1]}\left(\mathrm{P}\right)=\Theta^a\left(\mathrm{P}\right),\quad \boldsymbol{J}_{[2]}\left(\mathrm{P}\right)=\Phi^a\left(\mathrm{P}\right), \tag{2.2}$$

of which the last two are the real and imaginary parts of the complex "matrix element" between the two charge-conjugated configurations, $\Lambda^a_{(+)} = (\psi_c)^+ \alpha^a \psi = \Theta^a + i\Phi^a$ and $\Lambda^a_{(-)} = (\psi)^+ \alpha^a \psi_c = \Theta^a - i\Phi^a$, where ψ_c is the charge-conjugate spinor.

The components J^a_A of the currents $\boldsymbol{J}_{(A)}(\mathrm{P})$ depend only on the Dirac field and on a particular choice of the matrices α^a at the point P. The numbers $\boldsymbol{J}_{[0]}(\mathrm{P})$ are the coordinate scalars but are dubbed components of the "vector current". Another four real numbers, $\boldsymbol{J}_{[3]}(\mathrm{P})$, are associated with the components of the "axial current". The idea to use $\Theta^a/\mathcal{R}$ and $\Phi^a/\mathcal{R}$ as the tetrad vectors was first spelled out in Ref. [9].

In these definitions, an explicit form of the Dirac matrices $\alpha^a = (1, \alpha^i) = (1, \rho_3\sigma^i)$, ρ_i and σ^i (a = 0, 1, 2, 3; $i = 1, 2, 3$), is not specified; it is only required that they satisfy commutation relations,

$$\alpha^a \beta \alpha^b + \alpha^b \beta \alpha^a = 2\beta\eta^{ab}, \quad \alpha^a \beta + \beta\alpha^a = 0,$$

and, in general, they are not just numeric matrices. One *can* resort to a particular set of numerical matrices α^a and β only in conjunction with the corresponding tetrad basis $\boldsymbol{h}_a$ [1].

3. Fierz identities. Completeness of the basis. It appears that the four quadruples, $\boldsymbol{J}_A(\mathrm{P})$ ($A = 0, 1, 2, 3$), along with the scalar $\mathcal{S}$ and pseudoscalar $\mathcal{P}$, satisfy the following identities[2],

$$\begin{aligned} &\mathcal{R}^2 \equiv j_a j^a = -\mathcal{J}_a \mathcal{J}^a = -\Theta_a \Theta^a = -\Phi_a \Phi^a = \mathcal{S}^2 + \mathcal{P}^2, \\ &j^a j^b - \mathcal{J}^a \mathcal{J}^b - \Theta^a \Theta^b - \Phi^a \Phi^b = \mathcal{R}^2 \eta^{ab}, \\ &j_a \mathcal{J}^a = j_a \Theta^a = j_a \Phi^a = \mathcal{J}_a \Theta^a = \mathcal{J}_a \Phi^a = \Theta_a \Phi^a = 0, \end{aligned} \tag{2.3}$$

where $\eta_{ab} = \mathrm{diag}(1, -1, -1, -1)$ is the Minkowski tensor (which was not contemplated to be here) and $j_a = \eta_{ab} j^b$,... The Dirac currents $\boldsymbol{J}_A(\mathrm{P})$ are almost always linearly independent[3]. In what follows, unless stated otherwise, we will consider only "regular" domains where $\mathcal{R}^2 > 0$ and use, instead of $\boldsymbol{J}_{(A)}$, the normalized currents $\boldsymbol{V}_A = \boldsymbol{J}_A/\mathcal{R}$. The matrix V^a_A is not degenerate and thus has an inverse matrix V^A_a,

$$V^a_A V^B_a = \delta^B_A, \quad V^a_A V^A_b = \delta^a_b. \tag{2.4}$$

By virtue of Equation (2.3), at every point P of the basic manifold $\mathbb{M}$ the currents $\boldsymbol{V}_A$ form a complete (in the sense of linear algebra) system of orthogonal (with respect to the " metric" η_{ab}) unit "vectors",

$$\eta_{ab} V^a_A V^b_B = \eta_{AB}, \quad \eta^{AB} V^a_A V^b_B = \eta^{ab}. \tag{2.5}$$

The vector $\boldsymbol{V}_0$ is timelike while the other three are spacelike. It is also straightforward to check the following identities,

$$V^a_A = \eta^{ab}\eta_{AB} V^B_b, \; V^A_a = \eta_{ab}\eta^{AB} V^b_B, \tag{2.6}$$

---

[1] Employing the Dirac matrices, we can define the four components of the " vector current", $j^a = \psi^+ \alpha^a \psi \equiv \bar{\psi}\gamma^a\psi$, the four components of the "axial current", $J^a = \psi^+ \rho_3 \alpha^a \psi \equiv \bar{\psi}\gamma^5\gamma^a\psi$, two "charged currents", $\Lambda^a_{(+)} = (\psi_c)^+ \alpha^a \psi = \Theta^a + i\Phi^a$ and $\Lambda^a_{(-)} = (\psi)^+ \alpha^a \psi_c = \Theta^a - i\Phi^a$, the "scalar" $\mathcal{S} = \psi^+ \rho_1 \psi \equiv \bar{\psi}\psi$ and " pseudoscalar" $\mathcal{P} = \psi^+ \rho_2 \psi \equiv -i\bar{\psi}\gamma^5\psi$. Well-known are the six components of the skew-symmetric "tensor" $\Sigma^{ab} = (1/2)\psi^+ \left[\alpha^a \rho_1 \alpha^b - \alpha^b \rho_1 \alpha^a\right]\psi$ (or its dual, $\overset{*}{\Sigma}_{ab} = (1/2)\epsilon_{abcd}\Sigma^{cd}$). All of them are interconnected by the so-called Fierz relations [10]. The charge-conjugated spinor is defined as $\psi_c = \boldsymbol{C}\psi^*$ with a real-valued matrix $\boldsymbol{C}$ (e.g., $\boldsymbol{C} = \rho_2\sigma^2$).

[2] This is a small subset of the Fierz identities that includes 28 basic relations and hundreds of derivable from them. They were studied in details in Ref. [10] as the basis for the mathematical *reconstruction theorem* [11] that states that Dirac spinor field can be uniquely restored via the Dirac currents (without any account for the dynamics). Within this approach it is possible to replace tetrad vectors of any coordinate system by an equivalent Dirac field thus simplifying various calculations [12]. Among the objects connected via the Fierz identities is present the skew-symmetric Σ^{ab}. The Σ^{ab} appears to be a combination of the skew-symmetric products $j^{[a}\mathcal{J}^{b]}$ and $\Theta^{[a}\Phi^{b]}$ and scalars. The author was not aware of this fact and wrongfully tried [7] to employ Σ^{ab} to build a substitute for the Θ^a and Φ^a.

[3] Indeed, the necessary and sufficient condition for the linear independence is that the system of linear equations, $\sum_A c^A \boldsymbol{J}_A = 0$, has only a trivial solution, $c^A = 0$; the latter is possible if and only if matrix that has these quadruples as its columns has a nonzero determinant, $\det\left|J^a_A\right| \neq 0$. The determinant of the 4×4 matrix $\left|J^a_A\right|$ equals $\det\left|J^a_A\right| = \mathcal{R}^4$, where $\mathcal{R}^2$ is the squared module of the complex number, $\mathcal{R}^2 = \left|\mathcal{S} + i\mathcal{P}\right|^2$. When $\mathcal{R}^2 > 0$ the four vectors $\boldsymbol{J}_A(\mathrm{P})$ are linearly independent and can serve as a basis of vector space over $\mathbb{M}$. The condition $\mathcal{R}^2 = 0$ is equivalent to two real equations, $\mathcal{S} = \mathcal{P} = 0$, which determine a singular two-dimensional surface in $\mathbb{R}^4$ (and thus on $\mathbb{M}$).

and also that the $V_{Ab}=\eta_{AB}V_b^B$ is the solution of the linear system, $V_a^AV_{Ab}=\eta_{ab}$. Therefore, all indices are moved up and down by the Minkowski η_{ab} or η_{AB}, which is nothing but a consequence of the Fierz identities. At every point $\mathrm{P}\in\mathbb{M}$, any quadruple of scalar fields $U^a(\mathrm{P})$, regardless of its origin, can be presented as a linear combination of the basic quadruples $V_A^a(\mathrm{P})$ determined by the Dirac field $\psi(\mathrm{P})$,

$$U^a(\mathrm{P})=u^A(\mathrm{P})V_A^a(\mathrm{P}),\ \ u^A(\mathrm{P})=U^a(\mathrm{P})V_a^A(\mathrm{P}), \tag{2.7}$$

where u^A are the components of the U^a with respect to the basis V_A^a.

4. An intermediate tetrad basis. The components $V_A^a(\mathrm{P})$ of a quadruple $\boldsymbol{V}_A(\mathrm{P})$ clearly cannot be associated with a tangent vector like (2.1) simply because the former are defined only in terms of the invariant components straight in the principal manifold $\mathbb{M}^4$ (!), while definition of the latter requires a reference to an arithmetic $\mathbb{R}^4$, and its components are not invariant. Despite being complete, the system V_A^a cannot immediately serve as a basis for the tangent vectors (2.1). Its completeness is purely algebraic by nature, while linear independence and completeness of the system $\boldsymbol{h}_a=h_a^\mu\partial_\mu$ is analytic and is always traced back to linear independence of the vectors of the basis $\left(\partial/\partial x^\mu\right)$ (the linear vector space over $\mathbb{R}^1: x^\mu=x^\mu(\lambda),\lambda\in\mathbb{R}^1$).

An invariant representation of vector $\boldsymbol{s}$ is possible only together with a system of the basic vectors $\boldsymbol{h}_a$; then it can be replaced by scalars, the tetrad components of the vector $\boldsymbol{s}$, $\boldsymbol{s}=\boldsymbol{h}_as^a, s^a=h_\mu^as^\mu$. Now, one can use (2.7) to expand the four scalars s^a over the system V_A^a

$$\boldsymbol{s}=\boldsymbol{h}_aV_A^as^A=\boldsymbol{e}_As^A,\ \ s^A=V_a^Ah_\mu^as^\mu \tag{2.8}$$

and interpret the quantities $e_A^\mu=h_a^\mu V_A^a$ as the components of such a vector $\boldsymbol{e}_A=e_A^\mu\partial_\mu$ in coordinate basis that the scalars $V_A^a=e_A^\mu h_\mu^a$ are the components of $\boldsymbol{e}_A$ in the basis $\boldsymbol{h}_a$. The system of ODEs for the unknown x^μ, $\mathrm{d}x^\mu/\mathrm{d}s^A=e_A^\mu$, defines the integral lines of the vector fields $\boldsymbol{e}_A$. It is also clear that the matrix $e_\mu^A=V_a^Ah_\mu^a$ is the inverse of matrix e_A^μ, *viz.* $e_A^\mu e_\nu^A=\delta_\nu^\mu$, and $e_A^\mu e_\mu^B=\delta_A^B$.

Let $\boldsymbol{s}$ in Equation (2.8) be one of the vectors of the basis $\boldsymbol{h}_a$ (or of the basis $\boldsymbol{e}_A$). Then $h_b^\mu=h_a^\mu V_A^ah_b^A$ and $e_B^\mu=e_A^\mu h_a^AV_B^a$, which results in

$$h_b^AV_A^a=\delta_b^a,\ V_B^ah_a^A=\delta_B^A. \tag{2.9}$$

Since $\det\left|V_A^a\right|=1$, the inverse matrix V_a^A is uniquely defined; therefore,

$$h_a^A(\mathrm{P})=V_a^A(\mathrm{P}),\ h_A^a(\mathrm{P})=V_A^a(\mathrm{P}). \tag{2.10}$$

The components of the tetrad vectors $\boldsymbol{h}_a(\mathrm{P})$ with respect to the basis $\boldsymbol{e}_A(\mathrm{P})$ must have invariant values (2.10). These equations together with normalization conditions (2.5) and unitarity, $\det\left|V_A^a\right|=1$, allow one to *interpret* $V_A^a(\mathrm{P})$ as the matrix of a local Lorentz rotation between the bases $\boldsymbol{e}_A(\mathrm{P})$ and $\boldsymbol{h}_a(\mathrm{P})$ with parameters that are determined by the Dirac field $\psi(\mathrm{P})$[4]. So far, as long as we are confined to a point, we must refrain from associating this rotation with the *physical Lorentz transformations of special relativity*.

Since $V_A^a(\psi)$ are immediately defined as the fields over entire manifold $\mathbb{M}$, we expect that if two systems, $\boldsymbol{e}_A(\mathrm{P})$ and $\boldsymbol{h}_a(\mathrm{P})$, do exist, they are isomorphic not only in tangent T_p but even as fields over $\mathbb{M}$. The question is whether the integral lines of the vector fields $\boldsymbol{h}_a(\mathrm{P})$ and/or $\boldsymbol{e}_A(\mathrm{P})$ can form a coordinate net.

5. An auxiliary fundamental tensor (not a metric). It takes simple algebra to verify that at the point $\mathrm{P}\in\mathbb{M}$ the objects

$$g_{\mu\nu}=\eta_{ab}h_\mu^ah_\nu^b=\eta_{AB}e_\mu^Ae_\nu^B,\ \ g^{\mu\nu}=\eta^{ab}h_a^\mu h_b^\nu=\eta^{AB}e_A^\mu e_B^\nu, \tag{2.11}$$

can be used to move the coordinate (Greek) indices up and down. Indeed, $g_{\mu\nu}h_b^\nu=\eta_{ca}h_\mu^ch_\nu^ah_b^\nu=\eta_{cb}h_\mu^c=h_{b\mu}$.

[4]Long ago, E. Cartan [13] pointed to a difficulty, *i.e.* there are no representations of the general linear group of transformations $GL(4)$ that are similar to spinor representations of the Lorentz group of rotations. From the physical standpoint this argument is marginal since Lorentz transformations are between the reference frames of inertial observers and not between different differentiable mappings $\mathbb{M}\to\mathbb{R}^4$. Cartan stated the following theorem, which *vetoed* spinors in Riemannian geometry:

"*With the geometric sense given to the word 'spinor' it is impossible to introduce spinors into classical Riemannian technique; i.e., having chosen an arbitrary system of co-ordinates x^μ for space, it is impossible to represent spinor by any finite number of components ψ_i such that ψ_i have covariant derivatives of the form $\psi_{i;\mu}=\partial_\mu\psi_i+\Gamma_{i\mu}^j\psi_j$, where $\Gamma_{i\mu}^j$ are determinate functions of x^μ.*" Of these two underscored reservations of Cartan, the first one was investigated by Ne'eman *et al.* [14], who proposed to overcome the veto by resorting to the infinite-dimensional representations of the Lorentz group. The present study explores the window, which is left open by the second reservation.

With $g_{\mu\nu}$ thus defined, we also have the formal relations

$$g_{\mu\nu}h_a^\mu h_b^\nu = \eta_{ab},\quad g_{\mu\nu}e_A^\mu e_B^\nu = \eta_{AB}, \tag{2.12}$$

which can be interpreted as orthonormality relations for the tetrad bases h_a^μ and e_A^μ if we *postulate* that this $g_{\mu\nu}$ determines *a metric* in coordinate basis. Indeed, by virtue of the *identities* (2.11) the equation,

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu = \eta_{ab}\mathrm{d}\lambda^a \mathrm{d}\lambda^b = \eta_{AB}\mathrm{d}S^A \mathrm{d}S^B, \tag{2.13}$$

determines an interval which is Euclidean locally and invariant with respect to the choice of the coordinate basis within a domain where $\mathcal{R} \neq 0$. Most likely, this is *not the metric* that governs propagation of signals at a larger scale. It is remarkable that Fierz identities determine a system of *unit vectors* even before a notion of length is introduced.

Finally, when $g_{\mu\nu}$ is defined according to (2.10) and $R^2 = 0$ then all four vectors $\boldsymbol{e}_A = V_A^a \boldsymbol{h}_a$, regardless of the tetrad $\boldsymbol{h}_a(x)$, which obviously does not have this property, also become lightlike on a two-dimensional surface, $\mathcal{S} = \mathcal{P} = 0$, in spacetime $\mathbb{M}^4$. Obviously, in this case matrix V_A^a has no inverse.

3. Vector and Dirac Fields in Spacetime. Analytic Preliminaries

From now on, we look at the $\psi_\sigma(\mathrm{P})$ as the physical Dirac field over four-dimensional manifold $\mathbb{M}$. The points $\mathrm{P} \in \mathbb{M}$ are mapped onto points $(x^0, x^1, x^2, x^3) \in \mathbb{R}^4$. The components ψ_σ are thought of as smooth functions of the arbitrarily parameterized points $x^\mu = (x^0, x^1, x^2, x^3)$ of the spacetime. So far, we have verified that the algebraic structure of bilinear forms of the Dirac field naturally contains an orthogonal quadruple of unit (with respect to Minkowski metric) vectors at a generic point. By the argument of algebraic completeness, this quadruple must be isomorphic to a basis of any four non-complanar tangent vectors $\boldsymbol{h}_a(\mathrm{P})$ in $\mathbb{M}$. In a coordinate space $\mathbb{R}^4$, the latter are transformed as $\mathrm{d}\mathbf{x}$, while the former are scalars. In $\mathbb{R}^4$, for a given fixed λ, we can consider $x^\lambda = \mathrm{const}$ as the equation of a coordinate hypersurface and the lines along which all coordinates, but x^λ, are constant as coordinate lines. Tangent vectors of these lines (which are gradients of the linear function $\varphi(x) = x^\lambda$) are $h_{(\lambda)}^\mu = \partial x^\mu / \partial x^\lambda = \delta_{(\lambda)}^\mu$. Their covariant counterparts, $h_\mu^{(\lambda)} = \partial x^\lambda / \partial x^\mu = \delta_\mu^{(\lambda)}$, are the gradient vectors and the system of equations $\partial_\mu x^{(\lambda)} = h_\mu^{(\lambda)}(x)$ is integrable, but there is no metric and no way to determine if its coordinate lines are orthogonal. One may replace x^μ by smooth functions of other coordinates y^μ, $x^\mu = f^\mu(y)$, thus redefining coordinate lines and surfaces, but such a change does not alter $\psi(x(y))$ and has nothing to do with "Lorentz rotations".

Thus, we have to account for two different kinds of invariance. One of them is the covariance, a trivial mathematical independence from the coordinate system. The second one is the invariance of the Dirac field as the matter, and it is dominant on every account, because any conceivable measurement requires the presence of the localized physical objects. In this section, we consider the Dirac field as a known function of coordinates and do not employ its equation of motion.

3.1. Dirac Currents as a "Moving Frame" in Spacetime

The Dirac field $\psi(\mathrm{P})$ is a coordinate scalar, but it naturally generates an affine centered vector space (spanned by the Dirac currents $\boldsymbol{e}_A(\mathrm{P})$) at P, which is similar to the tangent space T_p of the four-dimensional manifold $\mathbb{M}$ at P (spanned by the vectors $\partial_\mu\big|_\mathrm{P}$ or $\boldsymbol{h}_a(\mathrm{P})$). These currents constitute a complete basis, they are of unit length and orthogonal in the sense of Equation (2.5). The continuous *field of tetrad* $\boldsymbol{e}_A(\mathrm{P})$ is embedded into $\mathbb{M}$. Therefore, an infinitesimal change of the V_A^a (and, eventually, of the $\boldsymbol{e}_A$) from point P to point P′ is predetermined as,

$$\mathrm{d}V_A^a(\mathrm{PP}') = \varpi_A^B(\mathrm{PP}')V_B^a(\mathrm{P}). \tag{3.1}$$

Also predetermined is the derivative of the scalars V_A^a, $\partial_\mu V_A^a(\mathrm{P}) = \omega_{A\mu}^B(\mathrm{P})V_B^a(\mathrm{P})$, and it has a very simple meaning. For a given displacement $\mathrm{d}x^\mu$ in $\mathbb{R}^4$, the total change $\mathrm{d}V_A^a = \partial_\mu V_A^a(\mathrm{P})\mathrm{d}x^\mu$ can be expanded over a complete system $\boldsymbol{V}_B(\mathrm{P})$ with the coefficients $\varpi_A^B(\mathrm{PP}') = \omega_{A\mu}^B(\mathrm{P})\mathrm{d}x^\mu$. More precise is the directional derivative,

$$\nabla_{\boldsymbol{h}} V_A^a(\mathrm{P}) = \omega_{A\boldsymbol{h}}^B(\mathrm{P})V_B^a(\mathrm{P}), \tag{3.2}$$

along an arbitrary vector $\boldsymbol{h}$ in $\mathbb{M}$. By taking $\boldsymbol{h}=\boldsymbol{e}_D$, we immediately recognize the connections $\omega_{AD}^{B}(P)$, with the directional derivative, $\partial_D = e_D^{\mu}\partial_{\mu}$, along $\boldsymbol{e}_D \in \mathbb{M}$, as objects in principal manifold $\mathbb{M}$,

$$\nabla_D V_A^a(\mathrm{P}) = \omega_{AD}^{B}(\mathrm{P}) V_B^a(\mathrm{P}),\quad \nabla_D V_a^A(\mathrm{P}) = \omega_{BD}^{A}(\mathrm{P}) V_a^B(\mathrm{P}). \tag{3.3}$$

Then $\omega_{BD}^{A} = V_a^A\left(\nabla_D V_B^a\right) = -V_B^a\left(\nabla_D V_a^A\right)$. Since $\mathrm{d}\left(V_a^A V_B^a\right) = 0$ $\left[V_a^A V_B^a = \delta_B^A\right]$ we immediately conclude that

$$\eta_{BC}\omega_{AD}^{C} = \omega_{ABD} = -\omega_{BAD}, \tag{3.4}$$

viz., the $\omega_{ABD} \in \mathbb{M}$ is skew-symmetric in the first two indices.

3.2. Covariant Derivatives at a Point in $\mathbb{M}$

In what follows, we compute the covariant derivatives of the vector and spinor components with respect to different bases and establish their interrelation.

1. The Dirac tetrad. Starting from Equations (2.7) and (3.3) and following the Cartan's idea of a moving frame [15], we can compute the covariant derivative of the components of any vector $\boldsymbol{u}(\mathrm{P}) = u^A(\mathrm{P})\boldsymbol{V}_A(\mathrm{P})$,

$$\partial_D \boldsymbol{u}(\mathrm{P}) = \boldsymbol{V}_A(\mathrm{P})\left[\partial_D u^A(\mathrm{P}) + \omega_{BD}^{A}(\mathrm{P}) u^B(\mathrm{P})\right], \tag{3.5}$$

or, in terms of components with respect to the basis $\boldsymbol{e}_A$,

$$\nabla_D u^A = \partial_D u^A + \omega_{BD}^{A} u^B, \tag{3.6}$$

where $\mathrm{d}u^A = \partial_D u^A \mathrm{d}S^D$ are the relative changes of the components and $\nabla_D u^A \mathrm{d}S^D$ is their total change. We explicitly see that *the presence of the physical Dirac field over the principal manifold* $\mathbb{M}$ *immediately endows* $\mathbb{M}$ *with an affine connection*. It also provides a natural definition of parallel transport as a transformation that leaves the components u^A of a vector unchanged with respect to a local basis, even when the local tetrad (or a coordinate hedgehog) changes its orientation from point to point. Equation (3.3) is a special case of Equation (3.6) when $u^A = V_a^A$. Taking for u^A the components of the vector current, $j^A = V_a^A j^a = \psi^{+}\alpha^A\psi$, one can define the covariant derivative of the Dirac field *without leaving the principal manifold* $\mathbb{M}$. Indeed, assuming that

$$D_A\psi = \partial_A\psi - \Gamma_A\psi, \tag{3.7}$$

and comparing with Equation (3.6) one readily obtains the equation that determines the connection Γ_A [16],

$$\Gamma_D^{+}\alpha^A + \alpha^A\Gamma_D = -\omega_{BD}^{A}\alpha^B = V_B^a\left(\nabla_D V_a^A\right)\alpha^B, \tag{3.8}$$

where $\alpha^A = V_a^A\alpha^a$ and these matrices α^A, depending on ψ, must be considered as primary objects in $\mathbb{M}$.

2. Arbitrary tetrads. Knowing the affine connection in the basis of vectors $\boldsymbol{V}_A$, we can find it in any other basis $\boldsymbol{h}_a(P)$. Indeed, starting from Equation (3.6) we rewrite covariant derivative in terms of the basis vectors $\boldsymbol{h}_a$,

$$V_A^a V_d^D\left(\nabla_D u^A\right) = \partial_d u^a + \gamma_{bd}^{a} u^b \equiv \nabla_d u^a(\mathrm{P}), \tag{3.9}$$

where $\partial_d = h_d^{\mu}\partial_{\mu}$ and γ_{bd}^{a} stands for the expression, $\gamma_{bd}^{a} = V_B^a\partial_d V_b^B + \omega_{BD}^{A} V_A^a V_b^B V_d^D$. By virtue of Equations (2.9), we have $\gamma_{bd}^{a} = V_d^D\left[\nabla_D V_b^B\right]V_B^a = h_d^D\left[\nabla_D h_b^B\right]h_B^a$. Using Equation (3.3), we obtain (by definition, $\partial_D V_B^b = 0$; V_B^b is a matrix of Lorentz rotation),

$$\gamma_{bd}^{a} = \omega_{BD}^{A} V_A^a V_b^B V_d^D. \tag{3.10}$$

These invariants are nothing but the coefficients of rotation of the basic vectors $\boldsymbol{h}_a$ with respect to the basis $\boldsymbol{e}_A$. Conversely, the equation,

$$\omega_{BD}^{A} = -h_D^d\left[\partial_d h_b^A - \gamma_{bd}^{a} h_a^A\right]h_B^b \equiv h_D^d\left(\nabla_d h_B^b\right)h_b^A = V_D^d\left(\nabla_d V_B^b\right)V_b^A, \tag{3.11}$$

gives the coefficients of rotation ω_{BD}^{A} of the basic vectors $\boldsymbol{V}_A$ with respect to the basis $\boldsymbol{h}_a$.

3. Coordinate basis. In the coordinate picture, the basis vectors $\boldsymbol{h}_a(x)$ are assumed to be known in advance.

In this case, one can derive the covariant derivative as

$$h_\mu^d h_a^\nu \left(\nabla_d u^a\right) = \partial_\mu u^\nu + \Gamma_{\sigma\mu}^\nu u^\sigma \equiv \nabla_\mu u^\nu, \tag{3.12}$$

where $\Gamma_{\sigma\mu}^\nu$ stands for

$$\Gamma_{\sigma\mu}^\nu = h_b^\nu \partial_\mu h_\sigma^b + \gamma_{b\,d}^a h_\mu^d h_a^\nu h_\sigma^b \equiv h_\mu^d \left(\nabla_d h_\sigma^b\right) h_b^\nu. \tag{3.13}$$

and (because of the term with $\partial_\mu h_\sigma^b$) it is transformed as a connection under a change of the coordinates. Alternatively, we could start with $\nabla_\mu u^\nu = e_\mu^D e_A^\nu \left(\nabla_D u^A\right)$ (or just substitute $\gamma_{b\,d}^a$ from Equation (3.10)) and obtain another representation of *the same* connection $\Gamma_{\sigma\mu}^\nu$,

$$\Gamma_{\sigma\mu}^\nu = e_B^\nu \partial_\mu e_\sigma^B + \omega_{B\,D}^A e_\mu^D e_A^\nu e_\sigma^B \equiv e_\mu^D \left(\nabla_D e_\sigma^B\right) e_B^\nu, \tag{3.14}$$

which is now expressed via quantities that explicitly depend on the physical Dirac field. Finally, using Equations (12), we can invert the last two equations to obtain,

$$\gamma_{b\,d}^a = h_d^\mu \left(\nabla_\mu h_b^\nu\right) h_\nu^a,\ \omega_{B\,D}^A = e_D^\mu \left(\nabla_\mu e_B^\nu\right) e_\nu^A, \tag{3.15}$$

which is normally taken as an *ad hoc* definition of the coefficients of rotation of tetrad vectors when one prefers to stay in $\mathbb{R}^4$. Notably, Equations (3.15) and (3.3) determine the same $\omega_{B\,D}^A$, although Equation (3.3) apparently belongs to $\mathbb{M}^4$ and has nothing to do with the $\mathbb{R}^4$. This may be considered as an evidence that the vectors h_a^μ and the connections $\gamma_{b\,d}^a$ are the auxiliary quantities.

When h^μ is a vector and $g_{\nu\mu}(x)$ is a tensor (not necessarily determining a metric) then the covariant derivative $\nabla_\nu h^\mu$ with respect to $g_{\nu\mu}$ is also a tensor [17]. Using Equations (3.12) and (3.15), it is straightforward to check that if $g_{\nu\mu}(x)$ has the form (2.10) then $\nabla_\lambda g_{\nu\mu} = 0$. Indeed, since $\gamma_{bad} = -\gamma_{abd}$ we have

$$\nabla_\lambda g_{\nu\mu} = \eta^{ab}\left[h_{a\mu}\nabla_\lambda h_{b\nu} + h_{b\nu}\nabla_\lambda h_{a\mu}\right] = h_\mu^a h_\nu^b h_\lambda^d \left(\gamma_{bad} + \gamma_{abd}\right) = 0.$$

An idea of how to find this $g_{\nu\mu}$ practically, will become clear only in the next paper [8], where a concrete solution ψ is found. Starting from there, one can take the following path, $\psi \to V_A^a \to \omega_{ABC} \to \gamma_{abc} \to h_a^\mu \to e_A^\mu$ and, eventually, explicitly determine the $g_{\mu\nu}\left(\psi(x)\right)$.

4. Connections for the Dirac field. Starting from Equation (3.9) for the vector current $j^a = \psi^+ \alpha^a \psi$,

$$D_b j^a = \partial_b j^a + \gamma_{c\,b}^a j^c \equiv \nabla_b j^a,\ D_b j_a = \partial_b j_a - \gamma_{a\,b}^c j_c \equiv \nabla_b j_a. \tag{3.16}$$

or translating Equation (3.8) into the basis $\boldsymbol{h}_a$, it is straightforward to obtain the following equation for the matrix Γ_a [5]:

$$\Gamma_d^+ \alpha_a + \alpha_a \Gamma_d = \gamma_{a\,d}^b \alpha_b, \Gamma_d^+ \alpha^a + \alpha^a \Gamma_d = -\gamma_{b\,d}^a \alpha^b, \tag{3.17}$$

where $\Gamma_d = V_d^D \Gamma_D$, and nothing implies that α^a must be numerical matrices[6]. If we introduce $\alpha^\nu = h_a^\nu \alpha^a = e_A^\nu \alpha^A$ and $\Gamma_\mu = h_\mu^d \Gamma_d = e_\mu^D \Gamma_D$ and use (3.15), then Equations (3.8) and (3.17) can be rewritten entirely in $\mathbb{R}^4$,

$$\Gamma_\mu^+ \alpha^\nu + \alpha^\nu \Gamma_\mu = -\nabla_\mu \alpha^\nu, \tag{3.18}$$

Equations (3.17) and (3.18) indicate that the Dirac matrices α^a are covariantly constant with respect to the "connection" Γ_a of the Dirac field, $D_b \alpha^a = D_\mu \alpha^\mu = 0$. The same is true for other representations as well.

Either of Equations (3.8), (3.17) and (3.18) can be solved (algebraically) for the corresponding Γ_κ. The most general solution reads as

$$\Gamma_d = ieA_d + ig\rho_3 \mathcal{W}_d + (1/4)\gamma_{abd}\rho_1 \alpha^a \rho_1 \alpha^b, \tag{3.19}$$

where, so far, *e* and *g* are arbitrary constants. The term ieA_d in the connection (19) (or the field A_μ) is

[5] Indeed, multiplying both sides by $V_d^D V_A^a$ we will have in the r.h.s. $V_d^D V_A^a V_B^h \left(\nabla_D V_h^A\right)\alpha^B = -V_d^D V_A^a V_h^A \left(\nabla_D V_B^h\right)\alpha^B = -V_d^D \left(\nabla_D V_B^a\right) V_b^B \alpha^b = \gamma_{b\,d}^a \alpha^b$.

[6] This is straightforward to show, $\partial_b\left(\psi^+\alpha^a\psi\right) + \gamma_{c\,b}^a\left(\psi^+\alpha^a\psi\right) \equiv \nabla_b\left(\psi^+\alpha^a\psi\right) = D_b\left(\psi^+\alpha^a\psi\right) = \left(D_b\psi\right)^+\alpha^a\psi + \psi^+\left(D_b\alpha^a\right)\psi + \psi^+\alpha^a D_b\psi$
$= \partial_b\left(\psi^+\alpha^a\psi\right) - \psi^+\left(\Gamma_b^+\alpha^a + \alpha^a\Gamma_b\right)\psi + \psi^+\left[\Gamma_b^+\alpha^a + \alpha^a\Gamma_b + \gamma_{c\,b}^a\alpha^c\right]\psi$,
where $D_b\alpha^a = \partial_b\alpha^a + \Gamma_b^+\alpha^a + \alpha^a\Gamma_b + \gamma_{c\,b}^a\alpha^c = \nabla_b\alpha^a + \Gamma_b^+\alpha^a + \alpha^a\Gamma_b$.

unquestionably interpreted as the electromagnetic potential. The term $ig\rho_3\mathcal{W}_d$ (or field $\mathcal{W}_\mu$) could have been interpreted as another field that interacts with the axial current J^μ[7]. The connection (3.19) commutes with the matrix ρ_3, so that Equation (3.17) remains the same when $\alpha_a \to \rho_3\alpha_a$. So far, it neither commutes nor anti-commutes with ρ_1 and ρ_2, *viz.*

$$\Gamma_b^+\rho_1 + \rho_1\Gamma_b = 2g\rho_2\mathcal{W}_b, \quad \Gamma_b^+\rho_2 + \rho_2\Gamma_b = -2g\rho_1\mathcal{W}_b. \tag{3.20}$$

Similar formulae arise for the charge-conjugated connection. Since $\boldsymbol{C}\rho_1\boldsymbol{C}^{-1} = -\rho_1$ and $\left[\boldsymbol{C}\alpha^i\boldsymbol{C}^{-1}\right]^* = \alpha^i$,

$$\Gamma_b^+\alpha_a + \alpha_a\Gamma_b^{\boldsymbol{c}} = \gamma_{ab}^d\alpha_d - 2ieA_b\alpha_a, \quad \left[\Gamma^{\boldsymbol{c}}\right]_b^+\alpha_a + \alpha_a\Gamma_b = \gamma_{ab}^d\alpha_d + 2ieA_b\alpha_a. \tag{3.21}$$

The commutation relations for the Dirac matrices $\alpha^\mu = h_a^\mu\alpha^a$ and $\alpha^A = V_a^A\alpha^a$ are

$$\alpha^\mu\rho_1\alpha^\nu + \alpha^\nu\rho_1\alpha^\mu = 2\rho_1 g^{\mu\nu} \text{ and } \alpha^A\rho_1\alpha^B + \alpha^B\rho_1\alpha^A = 2\rho_1\eta^{AB}$$

in $\mathbb{R}^4$ and $\mathbb{M}$, respectively. We assume that the matrices α^a are associated with the basis $\boldsymbol{h}_a$ in the tangent T_p, while matrices α^A belong to the principal manifold $\mathbb{M}$. In what follows, we consider Dirac field as the primary matter field; covariant derivatives of its bilinear functions will be computed only using Equations (3.17)-(3.19).

5. Connections in different bases. Equations (3.10) and (3.11) are nothing but the well known formulae for transformation of a linear connection between two non-coordinate (anholonomic) bases. In these bases, all quantities are functions of the point P in the principal manifold $\mathbb{M}$, and thus independent of the coordinate basis in the $\mathbb{R}^4$. For example, we readily have the coordinate-independent equation of the parallel transport of a vector $\boldsymbol{u}$ along a vector $\boldsymbol{v} = \mathrm{d}/\mathrm{d}\lambda$, *viz.* $\nabla_{\boldsymbol{v}}\boldsymbol{u}(\mathrm{P}) = v^A\nabla_A\boldsymbol{u}(\mathrm{P}) = v^a\nabla_a\boldsymbol{u}(\mathrm{P}) = 0$.

If we omit indices and use the notation $\mathcal{U}$ for matrix V_A^a (as well as $\mathcal{U}^{-1}$ for V_a^A, Γ_κ for $\gamma_{b\kappa}^a$ and Ω_κ for $\omega_{B\kappa}^A$) then Equations (3.10) and (3.11) read as

$$\Gamma_\kappa = \mathcal{U}^{-1}\Omega_\kappa\mathcal{U} - \mathcal{U}^{-1}\partial_\kappa\mathcal{U}, \quad \Omega_\kappa = \mathcal{U}\Gamma_\kappa\mathcal{U}^{-1} - \mathcal{U}\partial_\kappa\mathcal{U}^{-1}, \tag{3.22}$$

which are the universal expressions[8] for all kinds of connections associated with local transformations. Equations (3.6) and (3.9), augmented by definition of the derivatives, $\partial_d = h_d^\mu\partial_\mu$ and $\partial_D = V_D^d\partial_d = e_D^\mu\partial_\mu$, are fixing the components of any vector with respect to the (moving) tetrads $\boldsymbol{e}_A$ and $\boldsymbol{h}_a$. The existence of the field of unitary matrix of the Lorentz transform $\mathcal{U} = \left|V_A^a\right|$ (and then of an affine connection ω_{BD}^A) appears to be an amazing consequence of the Fierz identities for bilinear forms of the Dirac field. Finally, it is straightforward to check that, once h_a^μ and e_A^μ are the components of vectors and γ_{bad} and ω_{BAD} are scalars, the connection $\Gamma_{\sigma\mu}^\nu$ transforms under a further change of the coordinates as

$$\frac{\partial x^\nu}{\partial x^{\nu'}}\Gamma_{\sigma'\mu'}^{\nu'} = \frac{\partial x^\sigma}{\partial x^{\sigma'}}\frac{\partial x^\mu}{\partial x^{\mu'}}\Gamma_{\sigma\mu}^\nu + \frac{\partial^2 x^\nu}{\partial x^{\sigma'}\partial x^{\mu'}},$$

which guarantees that the derivative $\nabla_\mu u^\nu$ transforms as a tensor. Transformations (3.10) and (3.11) are reduced to this formula when the tetrads are formed by the gradient vectors.

By definition, $\gamma_{abd} = \left(\nabla_d h_a^\kappa\right)h_{b\kappa}$, were index κ can belong to any of the bases. Therefore, Equation (3.19) has the required general form (3.22) and can be rewritten as $\Gamma_d = ieA_d + ig\rho_3\mathcal{W}_d + (1/4)\nabla_d\left(\rho_1\alpha^a h_a^\kappa\right)\rho_1\alpha^b h_{b\kappa}$ in

[7] In the early days of the Dirac theory, it was firmly established that $\mathcal{S} = \psi^+\rho_1\psi$ and $\mathcal{P} = \psi^+\rho_2\psi$ are Lorentz scalars, which, however, *does not guarantee that they are scalars with respect to the general coordinate transformations* of the group $GL(4)$. V. Fock [16] resorted to a specific choice of the Dirac matrices to demonstrate that $S^+\rho_1 S = \rho_1$ and $S^+\rho_2 S = \rho_2$ under special Lorentz transformations S. For now, we shall refer to the differential identity (4.4), $\nabla_\mu\mathcal{J}^\mu = 2m\mathcal{P}$; since $\mathcal{J}^\mu$ is a vector and $\nabla_\mu\mathcal{J}^\mu$ is a coordinate scalar, so are $\mathcal{P}$ and then $\mathcal{S}$ (due to the Fierz identity (2.3)). This argument is not geometric in its nature, because it relies on the equation of motion. Intriguing is that $\mathcal{P}$ and $\mathcal{S}$ are the coordinate scalars only due to equations of motion. At the moment, we have no convincing argument that would allow one to reject the presence of $\mathcal{W}_d$ in the Γ_d except that we have no experimental evidence that $\mathcal{W}$ exists as a physical field. Here, such an argument is reached later (with the reference to the equations of motion) from the physical (and then mathematical) requirement that nothing in physical manifold $\mathbb{M}$ or in coordinate space $\mathbb{R}^4$ can depend on a tetrad basis $\boldsymbol{h}_a$. For the sake of clarity, some equations will be ending with "$\cdots = 0, \{= \mathcal{W}_a\}$", until we reach Equations (4.16) and (4.21) and then prove that $\mathcal{W}_a = 0$.

[8] In general, κ is not a tetrad index.

tetrad basis and as $\Gamma_\mu = ieA_\mu + ig\rho_3 \mathcal{W}_\mu + (1/4)\nabla_\mu\left(\rho_1\alpha^\nu\right)\rho_1\alpha_\nu$ in the coordinate $\mathbb{R}^4$.

6. Symmetry of the connection $\Gamma^\nu_{\sigma\mu}$. If we naively assume that the Minkowski signature η_{ab} in Equations (2.4) and (2.5) determines the *local metric of an inertial reference frame at point* P (with local coordinates y^a) and that $g_{\nu\mu}(\mathrm{P})$ of Equations (2.10) is obtained by a local coordinate transformation of the η_{ab} then, being a tensor, the skew-symmetric part $\Gamma^\sigma_{[\mu\nu]}$ of the connection (the tensor of torsion) should be zero. This argument would require, in its turn, that the covariant tetrad vectors be the gradient vectors, $h^a_\mu(\mathrm{P}) = \partial y^a / \partial x^\mu\big|_{\mathrm{P}}$, which is by no means self-evident.

There is, however, another reason for the symmetry of $\Gamma^\sigma_{\mu\nu}$, which is hinted by the Cartan's method of moving frame. The field of tetrad $\boldsymbol{e}_A(\mathrm{P})$ belongs to $\mathbb{M}$ and can be used as a " moving frame" for all vectors $\boldsymbol{s} \in \mathbb{M}$, including the vectors $\mathrm{d}\mathbf{x}$ of infinitesimal displacements. Consider now a closed path $\mathbb{R}^1 \subset \mathbb{M}$ through the point P_0 and attach the " natural" tetrad $\boldsymbol{e}_A(\mathrm{P})$ to its points. Then every next point of the path has a position with respect to the tetrad of the previous point. Since the tetrad $\boldsymbol{e}_A(\mathrm{P})$ is changing from point to point, we have no other choice but to specify the transport of a vector as the parallel Fermi transport (in the sense that the components of a vector with respect to the local tetrad do not change) along the chosen path. We will be able to get back to P_0 (the image of the path in the moving frame will be closed) with the same $\psi(\mathrm{P}_0)$ and, therefore, *with the same tetrad* $\boldsymbol{e}_A(\mathrm{P})$ *and matrix* $V^a_A(\mathrm{P}_0)$, which is imperative, if and only if the components $\Gamma^\sigma_{\mu\nu}$ of the connection, as they are defined *in the coordinate basis* $\partial/\partial x^\mu$ *of the* $\mathbb{R}^4$, are symmetric in their subscripts. Then the torsion tensor vanishes, and only then will we be able to contract the entire path to the point $\mathrm{P}_0 \in \mathbb{M}$. Consequently, the following formulae,

$$\left(\partial_a\partial_b - \partial_b\partial_a\right)\psi = -\left(\gamma^d_{ab} - \gamma^d_{ba}\right)\partial_d\psi, \quad \left(\partial_A\partial_B - \partial_B\partial_A\right)\psi = -\left(\omega^D_{AB} - \omega^D_{BA}\right)\partial_D\psi. \tag{3.23}$$

can be confidently used for any coordinate scalar $\psi(x)$.

4. Differential Identities for the Dirac Currents

As it was pointed out above, Equations (3.6) and (3.9) with the predetermined coefficients of rotation fix the components of a vector with respect to an *a priori* arbitrary tetrad basis. One might expect that these equations can be trivially used to fix the components of any tensor field. However, the coefficients of rotation of the "geometric tetrad" h^μ_a and those of the tetrad e^μ_a of the normalized Dirac currents are interconnected by Equation (3.10). Hence, the dynamic can potentially limit a feasible choice of the basis h^μ_a. The coordinate system (coordinate lines) can be not arbitrary; not all coordinate variables can even have the meaning of coordinates. Therefore, it seems appropriate to postpone, for as long as possible, explicit use of any coordinate basis and treat the tetrad $e^\mu_{(A)}\left[\psi(x)\right]$ as an orthogonal moving frame [15]. An affine geometry will be constructive if and only if all the coefficients ω^B_{AC} of rotation of the tetrad $\boldsymbol{e}_A$ can be determined from the equations of motion.

In this section we show that this is indeed possible. There appears to be sufficient number of identities for the Dirac currents to completely determine the coefficients ω^B_{AC} and the connections Γ_B in the covariant derivative $D_B\psi = \left(\partial_B - \Gamma_B\right)\psi$. Therefore, from now on we are dealing with the physical material Dirac field that satisfies the Dirac equations of motion,

$$\alpha^a D_a\psi = -im\rho_1\psi, \quad \psi^+ D^+_a\alpha^a = im\psi^+\rho_1, \tag{4.1}$$

with the derivative $D_a = \partial_a - \Gamma_a$, connection Γ_a defined by Equation (3.19), and the mass parameter *m*. The latter is, for now, real, arbitrary and stands for the rate of mixing between the right and left components of the Dirac spinor. The equations of motion for the charge-conjugated spinor are

$$\alpha^a D^c_a\psi_c = -im\rho_1\psi_c, \; [\psi_c]^+ \overleftarrow{D}^{c+}_a\alpha^a = im\left[\psi_c\right]^+\rho_1, \tag{4.2}$$

where $D^c_a\psi^c = \left(\partial_a - \Gamma^c_a\right)\psi^c$ is the covariant derivatives of the charge-conjugate Dirac field, and Γ^c_a is given by Equations (3.21).

4.1. Divergences of the Dirac Currents

From the equations of motion (4.1) one immediately derives two well-known identities. Multiplying the Dirac equation by ψ^+ from the left and its conjugate by ψ from the right and taking their sum we readily obtain that

$$D_a j^a = \nabla_\mu j^\mu = 0. \tag{4.3}$$

This equation clearly indicates conservation of the *timelike* vector current (of probability) of the Dirac field. The second identity is obtained from the Dirac Equation (4.1), which is multiplied by ρ_3 from the left (and its conjugate from the right, and noting that $\rho_3\rho_1 = -\rho_1\rho_3 = i\rho_2$). It indicates that the *spacelike* axial current is not conserved,

$$D_a \mathcal{J}^a = \nabla_\mu \mathcal{J}^\mu = 2m\mathcal{P}, \tag{4.4}$$

and has the pseudoscalar density as a source. Since $\mathcal{P}$ is localized not less than $\mathcal{R}$, and the vector $\mathcal{J}^\mu$ is spacelike, it defines the radial direction. The existence of such a direction is a distinct characteristic of any localized object.

Similar identities can be derived for the vectors Θ^a and Φ^a of Section 2. Using Equations (3.21) and (4.2), we immediately arrive to covariant derivatives of the matrix elements $\Lambda_a^{(\pm)}$ as

$$D_b\Lambda_a^{(\pm)} = \partial_b\Lambda_a^{(\pm)} - \gamma_{ab}^d\Lambda_d^{(\pm)} \mp 2ieA_b\Lambda_a^{(\pm)} \equiv \nabla_b\Lambda_a^{(\pm)} \mp 2ieA_b\Lambda_a^{(\pm)}. \tag{4.5}$$

Though these vectors are complex and explicitly depend on the phase of ψ, this dependence is compensated in the covariant derivative (4.5) by the gauge transformation of the vector potential. The derivatives of Θ and Φ become

$$D_b\Theta_a = \nabla_b\Theta_a + 2eA_b\Phi_a,\ D_b\Phi_a = \nabla_b\Phi_a - 2eA_b\Theta_a. \tag{4.6}$$

The fields of complex currents $\Theta^a \pm i\Phi^a$ look like being "charged" with a charge $2e$. From the equations of motion (4.2) and using Equation (4.6), it is straightforward to get $D_a\Lambda^a_{(\pm)} = 0$ and, consequently,

$$\nabla_\mu\Theta^\mu = -2eA_\mu\Phi^\mu,\quad \nabla_\mu\Phi^\mu = 2eA_\mu\Theta^\mu. \tag{4.7}$$

Similarly to the vector of axial current, these vectors are not conserved due to electromagnetic potential A_μ.

4.2. Curls of the Dirac Currents

In order to access the differential identities for the curls of the Dirac currents one has to compute, using the equations of motion, the derivatives of the objects $T_a^a, P_a^a, \Theta_a^a, \Phi_a^a$, which are traces of tensors (objects), $T_b^a = \psi^+\alpha^a D_b\psi$, $P_b^a = \psi^+\rho_3\alpha^a D_b\psi$, $\Theta_b^a = \left(\psi^c\right)^+\alpha^a D_b\psi$ and $\Phi_b^a = \left(\psi^c\right)^+\overleftarrow{D}_b^{c+}\alpha^a\psi$, respectively. These tensors are neither real nor symmetric, and we are not concerned here about their physical interpretation.

1. T_ν^μ **—a tensor or not?** One would expect the absolute differential of T_b^a, being computed according to the Leibniz rule, be as follows,

$$D_c T_b^a = \partial_c T_b^a + \gamma_{dc}^a T_b^d - \gamma_{bc}^d T_d^a \equiv \nabla_c T_b^a. \tag{4.8}$$

and this expression would fix, similarly to Equations (3.9) and (3.12), the components of the tensor $T_\mu^\sigma = h_a^\sigma h_\mu^b T_b^a$ with respect to the tetrad $\boldsymbol{h}_a$. If this expectation turns out justified then the usual covariant derivative will be immediately reproduced as

$$\partial_\lambda T_\mu^\sigma + \Gamma_{\nu\lambda}^\sigma T_\mu^\nu - \Gamma_{\mu\lambda}^\nu T_\nu^\sigma = e_\lambda^c e_a^\sigma e_\mu^b \nabla_c T_b^a = \nabla_\lambda T_\mu^\sigma. \tag{4.9}$$

Contrary to the expectation of (4.8), the answer reads

$$\begin{aligned} D_c\left[\psi^+\alpha^a\vec{D}_b\psi\right] &= \partial_c\left[\psi^+\alpha^a\vec{D}_b\psi\right] - \psi^+\left[\Gamma_c^+\alpha^a + \alpha^a\Gamma_c\right]\vec{D}_b\psi \\ &= \partial_c\left[\psi^+\alpha^a\vec{D}_b\psi\right] + \gamma_{dc}^a\psi^+\alpha^d\vec{D}_b\psi, \end{aligned} \tag{4.10}$$

with the last term of Equation (4.8) missing, and no hope to recover the full *geometric* expression (4.9) of the covariant derivative of the tensor! Contracting here indices a and c and using equations of motion we would arrive at [7]

$$\partial_a T_b^a + \gamma_{ca}^a T_b^c - \gamma_{ba}^c T_c^a = -\gamma_{ab}^c T_c^a - i\psi^+\alpha^a\mathbb{D}_{ab}\psi. \tag{4.11}$$

with the normal covariant derivative in the l.h.s. The $\mathbb{D}_{ab}$ and an abnormal term $\gamma^{c}_{ab}T^{a}_{c}$ in the r.h.s. originate from the commutator of the covariant derivatives, $\left[\vec{D}_a,\vec{D}_b\right]$. Its real part is the Lorentz force, $Re\left[i\psi^{+}\alpha^{a}\mathbb{D}_{ab}\psi\right]=ej^{a}F_{ab}$ [7] [16][9].

2. Abnormal terms and how they restore the GL(4) covariance. The abnormal term enters another identity that follows from the Dirac equation, which arises after contracting indices *a* and *b* in Equation (4.10). On the one hand, we formally have (Cf. footnote[7]. The $S=\psi^{+}\rho_1\psi$ must be a scalar and the last term in the r.h.s. must be absent.)

$$D_c\left[\psi^{+}\alpha^{a}\vec{D}_a\psi\right]=\partial_c\left[\psi^{+}\alpha^{a}\vec{D}_a\psi\right]+\gamma^{a}_{bc}\psi^{+}\alpha^{b}\vec{D}_a\psi=-im\partial_c\left[\psi^{+}\rho_1\psi\right]+\gamma^{a}_{bc}\psi^{+}\alpha^{b}\vec{D}_a\psi. \tag{4.12}$$

On the other hand, by virtue of the Dirac equation, the first term on the r.h.s. of (4.12) becomes $\partial_c\left[-im\psi^{+}\rho_1\psi\right]$. Alternatively, one can immediately use the equations of motion on the l.h.s. and only then differentiate,

$$D_c\left[\psi^{+}\alpha^{a}\vec{D}_a\psi\right]=-imD_c\left[\psi^{+}\rho_1\psi\right]=-im\partial_c\left[\psi^{+}\rho_1\psi\right]+im\psi^{+}\left[\Gamma^{+}_{c}\rho_1+\rho_1\Gamma_c\right]\psi\,. \tag{4.13}$$

Comparing the last two equations and using (3.20), we finally find that the abnormal term $\gamma^{c}_{ab}T^{a}_{c}=h^{\lambda}_{b}\left(\nabla_{\lambda}h^{\nu}_{(a)}\right)h^{(a)}_{\sigma}T^{\sigma}_{\nu}$ vanishes (or at least can be expressed via abnormal field $\mathcal{W}_b$)

$$-\gamma^{c}_{ab}\cdot T^{a}_{c}=0,\quad\{=2mg\mathcal{P}\mathcal{W}_a\}, \tag{4.14}$$

thus restoring the covariance of Equation (4.11). Remarkably, the usual covariance in coordinate space is restored due to equations of motion. Equation (4.14) yields two nontrivial conditions on the structure of the Dirac currents as follows. The Ricci coefficients are real-valued and skew-symmetric in the first two indices. The r.h.s. of Equation (4.14) is real. Therefore, the imaginary part of Equation (4.14) reads as

$$\gamma_{acb}\,\mathrm{Im}\left(T_{ac}-T_{ca}\right)=\gamma_{acb}\left[D_c\left(\psi^{+}\alpha_a\psi\right)-D_a\left(\psi^{+}\alpha_c\psi\right)\right]=\gamma_{acb}\left(\nabla_c j_a-\nabla_a j_c\right)=0. \tag{4.15}$$

In order to facilitate further analysis of the real part of Equation (4.14), let us rewrite its l.h.s. in terms of the axial current. Using the dual representation of the axial current as $\epsilon^{stua}\mathcal{J}_a=i\psi^{+}\alpha^{s}\rho_1\alpha^{t}\rho_1\alpha^{u}\psi$, $(s,t,u,\neq)$ and employing the equations of motion we obtain,

$$D_u\epsilon^{stua}\mathcal{J}_a=-i\psi^{+}\alpha^{s}\vec{D}_t\psi+i\psi^{+}\alpha^{t}\vec{D}_s\psi-i\psi^{+}\overleftarrow{D}^{+}_{s}\alpha^{t}\psi+i\psi^{+}\overleftarrow{D}^{+}_{t}\alpha^{s}\psi\,,$$

where the r.h.s is four times the anti-symmetric Hermitian part of the energy momentum tensor. Therefore, the real part of Equation (4.14) reads as

$$(1/4)\,\gamma_{acb}\cdot\epsilon^{acst}\cdot\nabla_s\mathcal{J}_t=(1/2)\overset{\star}{\gamma}_{acb}\nabla_c\mathcal{J}_a=0\quad\{=2mg\mathcal{P}\mathcal{W}_b\}. \tag{4.16}$$

3. More non-tensors and abnormal terms. Next, consider the stress tensor $P^{a}_{b}=i\psi^{+}\rho_3\alpha^{a}D_b\psi$, mostly following the same protocol and starting from its covariant derivative. We find that

$$D_c\left[\psi^{+}\rho_3\alpha^{a}\vec{D}_b\psi\right]=\partial_c\left[\psi^{+}\alpha^{a}\vec{D}_b\psi\right]-\gamma^{a}_{dc}\psi^{+}\rho_3\alpha^{d}\vec{D}_b\psi. \tag{4.17}$$

Once again, the last term of Equation (4.8) is missing, and thus we have no confidence that the covariant derivative is a tensor. For the immediate purpose of this work, we only need the equations that emerge after contracting indices *a* and *b* in Equation (4.17),

$$D_c\left[\psi^{+}\rho_3\alpha^{a}\vec{D}_a\psi\right]=\partial_c\left[\psi^{+}\rho_3\alpha^{a}\vec{D}_a\psi\right]+\gamma^{a}_{bc}\psi^{+}\rho_3\alpha^{b}\vec{D}_a\psi. \tag{4.18}$$

By virtue of the Dirac equations, the first term in the r.h.s. becomes $\partial_c\left[m\psi^{+}\rho_2\psi\right]$. Alternatively, one can immediately use the equations of motion in the l.h.s. and only then differentiate (matrices ρ_3 and α^{a} commute),

[9]Three remarks are to be made here: 1) the Lorentz force in the r.h.s. allows one to associate the observable *j* and $\mathcal{R}$ with a variations of the charge density even without reference to the Maxwell equations. A uniform distribution is not distinct from vacuum; 2) if the basis h^{μ}_{a} were holonomic, *viz.* $\gamma^{c}_{ba}-\gamma^{c}_{ab}=0$, then there would have been no way to achieve the desired covariance. In fact, the abnormal term will vanish, but only if the nontrivial conditions (14) are met; 3) in general, $\mathbb{D}_{ab}=-(1/4)R_{abcd}\rho_1\alpha^{c}\rho_1\alpha^{d}+ieF_{ab}$, where R_{abcd} and F_{ab} are the Riemannian curvature and the electromagnetic field tensors, respectively.

$$D_c\left[\psi^+\rho_3\alpha^a\vec{D}_a\psi\right]=mD_c\left[\psi^+\rho_2\psi\right]=m\partial_c\left[\psi^+\rho_2\psi\right]+m\cdot 2g\mathcal{SW}_c\,. \tag{4.19}$$

Comparing the last two equations we finally get the equation,

$$-\gamma^{\ a}_{c\,b}\cdot P^c_a=-2igm\mathcal{SW}_b, \tag{4.20}$$

which is complementary to Equation (4.14). Since γ_{acb} is skew-symmetric in the first two indices, the imaginary part in the l.h.s. is due to $(1/2)\left[P_{ca}-P^+_{ca}\right]=(i/2)D_cJ_a$. Since the axial current is a vector, we can rewrite the imaginary part of the last equation as [C.f. footnote[7]],

$$(1/2)h^b_\mu\gamma_{acb}\nabla_c\mathcal{J}_a=(1/4)\eta^{ab}\left(\nabla_\mu h^\nu_a\right)h^\lambda_b\left(\nabla_\nu\mathcal{J}_\lambda-\nabla_\lambda\mathcal{J}_\nu\right)=0,\ \ \{=2mg\mathcal{SW}_b\} \tag{4.21}$$

which is dual to Equation (4.16). The skew-symmetric Hermitian part, $\left(P_{ca}+P^+_{ca}\right)-\left(P_{ac}+P^+_{ac}\right)$, must vanish since the r.h.s. of Equation (4.20) is an imaginary quantity. Since $\epsilon^{stua}j_a=i\psi^+\rho_3\alpha^s\rho_1\alpha^t\rho_1\alpha^u\psi$, $(s,t,u,\neq)$, this yields the equation,

$$\gamma_{acb}\left[\psi^+\rho_3\alpha_a\vec{D}_c\psi-\psi^+\overleftarrow{D}^+_c\alpha_a\rho_3\psi-\psi^+\rho_3\alpha_c\vec{D}_a\psi+\psi^+\overleftarrow{D}^+_a\alpha_c\rho_3\psi\right]=2\gamma_{acb}\epsilon^{acut}D_uj_t=4\overset{\star}{\gamma}_{acb}D_aj_c=0. \tag{4.22}$$

which is similar to Equation (4.16) and dual to Equation (4.15).

4. A full set of prerequisites for the covariance. Considered together, Equations (4.15) and (4.22) constitute a linear system of eight equations for the six unknowns, $\nabla_aj_b-\nabla_bj_a$. In general, the rank of its matrix equals 6. Therefore, it can only have a trivial solution. Since ∇_aj_b are the invariants of a true tensor, $\nabla_\mu j_\nu$, we have the tensor equation,

$$\nabla_\mu j_\nu-\nabla_\nu j_\mu=0\,. \tag{4.23}$$

Equations (4.16) and (4.21) constitute the system of 8 equations for 10 unknown quantities, $\nabla_{[\nu}\mathcal{J}_{\lambda]}$ and $\mathcal{W}_\mu$. These equations also explicitly depend on a choice of the auxiliary field of tetrad $\boldsymbol{h}_a(x)$, which is unacceptable. Insisting on independence as a physical (and then mathematical) requirement and realizing that $\mathcal{W}$ does not exist as a physical field, we must put $\mathcal{W}_a=0$ [10]. Then we have the system of 8 homogeneous equations for only 6 unknowns $\nabla_{[\nu}\mathcal{J}_{\lambda]}$ with a trivial solution,

$$\nabla_\nu\mathcal{J}_\lambda-\nabla_\lambda\mathcal{J}_\nu=0, \tag{4.24}$$

which is similar to Equations (4.23) that we had for the vector current.

More identities are readily obtained along the same guidelines as Equation (4.14). Namely, duplicating (4.12)-(4.14), we compute $D_c\left[\left(\psi^c\right)^+\alpha^aD_a\psi\right]$ and $D_c\left[\left(\psi^c\right)^+\overleftarrow{D}^{c+}_a\alpha^a\psi\right]$ directly and using equations of motion. Adding up the results we obtain that

$$\gamma_{adc}D_a\left[\left(\psi^c\right)^+\alpha_d\psi\right]=\gamma_{adc}D_a\Lambda^{(+)}_d=0. \tag{4.25}$$

Computing in the same way the dual quantities, $D_c\left[\left(\psi^c\right)^+\rho_3\alpha^aD_a\psi\right]$ and $D_c\left[\left(\psi^c\right)^+\overleftarrow{D}^{c+}_a\rho_3\alpha^a\psi\right]$, we end up with

$$\overset{\star}{\gamma}_{adc}D_a\left[\left(\psi^c\right)^+\alpha_d\psi\right]=\overset{\star}{\gamma}_{adc}D_a\Lambda^{(+)}_d=0, \tag{4.26}$$

which once again is a system of 8 equations for six unknowns with only a trivial solution. Since γ_{adc} is skew-symmetric in the first two indices and is not zero, we arrive at

$$D_a\Lambda^{(\pm)}_b-D_b\Lambda^{(\pm)}_a=0, \tag{4.27}$$

which, by virtue of (4.6), results in

$$\nabla_\mu\Theta_\nu-\nabla_\nu\Theta_\mu=-2e\left(A_\mu\Phi_\nu-A_\nu\Phi_\mu\right),\ \ \nabla_\mu\Phi_\nu-\nabla_\nu\Phi_\mu=+2e\left(A_\mu\Theta_\nu-A_\nu\Theta_\mu\right). \tag{4.28}$$

The differential identities (4.15), (4.23) and (4.28) for the Dirac currents are written down in the covariant tensor form and can be transformed further into tetrad representation with respect to any tetrad. Therefore, it is indeed possible to overcome the Cartan's veto [C.f. footnote 4] relying on the second reservation in Cartan's statement.

[10] This accomplishes the proof of the statement outlined in the footnote7.

5. Dirac Field and Congruences of Curves

Each of four linear partial differential equations, $\partial_A f = e_A^\mu \partial_\mu f = 0$, determine a congruence of lines because it is equivalent to the system of three ODEs for unknown x^μ, $\mathrm{d}x^\mu / e_A^\mu(x) = \mathrm{d}S^A$, $\mu = 0,1,2,3$. The question is whether two or three of these PDEs can be solved together (if they form a complete system). The answer is encoded in the properties of the rotation coefficients ω_{AB}^C of the orthogonal net of the tetrad $\boldsymbol{e}_A$. These are not given *a priori*, but it is possible to find them as dynamic quantities. This is an immediate goal of this section. Technically, we will rely only on Equation (3.15),

$$\nabla_\mu\left(\mathcal{R}e_B^\nu\right) = e_B^\nu\partial_\mu\mathcal{R} + \mathcal{R}\nabla_\mu e_B^\nu,\quad \nabla_\mu e_B^\nu = e_\mu^D\omega_{BD}^A e_A^\nu,\quad \nabla_\mu e_B^\mu = \omega_{BA}^A = \omega_{B00} - \omega_{B11} - \omega_{B22} - \omega_{B33}. \tag{5.1}$$

5.1. Vector Current and Timelike Congruence

To analyze the lines of the vector current, the two obtained earlier equations, (4.3) and (4.23),

$$\nabla_\mu j_\nu - \nabla_\nu j_\mu = 0,\quad \nabla_\mu j^\mu = 0, \tag{5.2}$$

must be examined together. When the invariant density of the Dirac (spinor) matter is positive, $\mathcal{R} = \sqrt{j^2} > 0$, the vector field $j^\mu(x)$ is strictly timelike; its tangent unit vector is $e_{[0]}^\mu(x)$, $j^\mu = \mathcal{R}e_{[0]}^\mu$. Therefore, Equation (4.23) becomes

$$\nabla_\mu e_\nu^{[0]} - \nabla_\nu e_\mu^{[0]} + e_\nu^{[0]}\partial_\mu \ln\mathcal{R} - e_\mu^{[0]}\partial_\nu \ln\mathcal{R} = 0. \tag{5.3}$$

Contracting this equation with $e_A^\nu e_B^\mu$, $A,B = 1,2,3$ and using Equation (5.1) we find that

$$\omega_{0AB} - \omega_{0BA} = 0,\quad A,B = 1,2,3, \tag{5.4}$$

which is a necessary and sufficient condition for the congruence $e_{[0]}^\mu$ to be normal [17] [18]. Namely, there exists such a function, $\tau(x)$, that the vector field $e_\mu^{[0]}(x)$ is orthogonal to the family of surfaces $\tau(x) = \text{const}$,

$$\partial_\mu\tau(x) = f(x)e_\mu^{[0]}(x),\quad \mathrm{d}\tau = \partial_\mu\tau\mathrm{d}x^\mu = fe_\mu^{[0]}\mathrm{d}x^\mu = f\mathrm{d}S^0, \tag{5.5}$$

where $\tau(x)$ satisfies the complete system of three equations, $e_A^\mu(x)\partial_\mu\tau(x) = 0$, $A = 1,2,3$, and $f(x)$ is a coordinate scalar. Contracting Equation (5.3) with $e_{[0]}^\nu$ we get

$$\partial_\mu \ln\mathcal{R} = e_\mu^{[0]}\partial_{[0]}\ln\mathcal{R} - \omega_{B00}e_\mu^{(B)}, \tag{5.6}$$

where $\partial_A \ln\mathcal{R} = e_A^\mu\partial_\mu \ln\mathcal{R} = \partial\ln\mathcal{R}/\partial S^A$ is the derivative in the direction of the arc S^A. Contraction of Equation (5.3) with $e_{[0]}^\nu e_A^\mu$ yields

$$\partial_A \ln R = -\omega_{A00},\quad A = 1,2,3, \tag{5.7}$$

which indicates that congruences of lines, defined by the system of equations, $\mathrm{d}x^\mu/\mathrm{d}S^{[0]} = e_{[0]}^\mu$, must experience permanent bending (acceleration) whenever the invariant density $\mathcal{R}(x)$ of the Dirac field is not uniformly distributed. The spatial gradient of $\mathcal{R}(x)$ cannot vanish for any localized state.

Additional information can be extracted from Equation (4.3), $\nabla_\nu\left(\mathcal{R}e_{[0]}^\nu\right) \equiv \partial_{[0]}\mathcal{R} + \mathcal{R}\nabla_\nu e_{[0]}^\nu = 0$. Then, by definition,

$$\nabla_\nu e_{[0]}^\nu = \omega_{0A}^A = \omega_{101} + \omega_{202} + \omega_{303} = -\partial_{[0]}\ln\mathcal{R}. \tag{5.8}$$

Hence, we can rewrite (5.6) as

$$\partial_\mu \ln\mathcal{R} = -e_\mu^{[0]}\eta^{AB}\omega_{0AB} - \omega_{B00}e_\mu^B, \tag{5.9}$$

which shows that the r.h.s. of Equation (5.9), which contains only geometric objects, is a component of a gradient. Together with condition (5.4) this constitutes a necessary and sufficient condition that the function $\tau(x)$ defined by Equation (5.5) is an harmonic function [17],

$$\Box\tau = g^{\mu\nu}\nabla_\mu\nabla_\nu\tau = 0. \tag{5.10}$$

The parameter τ^* of $\tau(x) = \tau^* = \text{const}$ is the definition of the world time. For the harmonic function,

$\tau(x)$, the conditions of integrability for system (5.5) of partial differential equations reads as [17]

$$\partial_\mu \ln f = -e^{[0]}_\mu \eta^{AB} \omega_{0AB} - \omega_{B00} e^{(B)}_\mu .$$

Comparing it with (5.9) we find that $f(x) = \mathcal{R}$, so that the world time τ and the "proper time" $S^{[0]}$ are related by

$$\mathrm{d}\tau = \mathcal{R}\mathrm{d}S^{[0]} \tag{5.11}$$

Furthermore, since $f(x) = \mathcal{R}$ and system possesses the proper time, we can rewrite Equation (5.9) as $j_\mu(x) = \partial_\mu \tau(x)$, which could have been inferred directly from Equation (4.15). Then, the harmonic nature of $\tau(x)$ immediately follows from the current conservation, $\nabla_\mu j^\mu = 0$. Since $\mathrm{d}\tau$ is the total differential and the vector current $\boldsymbol{j}$ belongs, in fact, to the principal manifold $\mathbb{M}$, so does the interval of the world time τ,

$$\tau_2 - \tau_1 = \int_{x(\tau_1)}^{x(\tau_2)} j_\mu(x)\mathrm{d}x^\mu \ \left(= \int \mathcal{R}\mathrm{d}S^{[0]}\right), \tag{5.12}$$

and this interval does not depend on the path of integration (the time variable τ is a holonomic coordinate).

Now, we can draw the major conclusion: *The proper time, $S^{[0]}$, flows more slowly than the world time, τ, whenever Dirac matter has a magnified density.* Because of the wave nature of the Dirac field, its localization is inevitable. Since the congruence $e^\mu_{[0]}$ appeared to be normal, the hypersurfaces $\tau(x) = \tau^* = \mathrm{const}$ represent space at different times τ^*. The states can be considered stationary only with respect to τ; one can hope to find them only after replacing $i\partial/\partial S^0$ by $i\mathcal{R}\partial/\partial\tau$ in the operator of energy!

5.2. Axial Current and Radial Congruence

Here, we have to deal with the system of equations,

$$\nabla_\nu \mathcal{J}_\lambda - \nabla_\lambda \mathcal{J}_\nu = 0, \quad \nabla_\nu \mathcal{J}^\nu = 2m\mathcal{P}, \tag{5.13}$$

which is similar to Equations (5.2) that we had for the vector current. The only difference is that the axial current has a source $2m\mathcal{P}$. Since there is no flux of vector current in this direction (the amount of matter inside a closed surface remains the same), we associate the radial direction $e^\mu_{[3]}(x)$ with the axial current, $J^\mu = \mathcal{R}e^\mu_{[3]}$. Next, observe that by virtue of the Fierz identity (2.3), $\mathcal{R}^2 = \mathcal{S}^2 + \mathcal{P}^2$, we can parameterize, $\mathcal{S} = \mathcal{R}\cos\mathcal{Y}$, $\mathcal{P} = \mathcal{R}\sin\mathcal{Y}$. Then the second Equation (5.13) takes form

$$\nabla_\mu e^\mu_{[3]} + e^\mu_{[3]}\partial_\mu \ln\mathcal{R} = 2m\mathcal{P}/\mathcal{R} = 2m\sin\mathcal{Y}. \tag{5.14}$$

On the one hand, by definition, $\nabla_\mu e^\mu_{[3]} = \eta^{AB}\omega_{3AB} = \omega_{300} - \omega_{311} - \omega_{322}$. On the other hand, according to Equation (5.7), we have $e^\mu_{[3]}\partial_\mu \ln\mathcal{R} = \partial_{[3]}\ln\mathcal{R} = -\omega_{300}$. Substituting these expressions into Equation (5.14) we obtain an important relation,

$$\omega_{131} + \omega_{232} = 2m\sin\mathcal{Y}. \tag{5.15}$$

The first of Equations (5.13), being contracted with $e^\mu_A e^\nu_B$, yields

$$\omega_{3AB} - \omega_{3BA} = 0, \quad A,B = 0,1,2, A \neq B, \tag{5.16}$$

so that the congruence of lines $\boldsymbol{e}_{[3]}$ is normal and there exists such a family of hypersurfaces $\rho(x) = \rho^* = \mathrm{const}$ that

$$\partial_\mu \rho(x) = n(x) e^{[3]}_\mu(x), \tag{5.17}$$

where $\rho(x)$ satisfies the complete system of three equations, $e^\mu_A(x)\partial_\mu\rho(x) = 0$, $A = 0,1,2$, and $n(x)$ is a coordinate scalar. In the same way as before [cf. (5.6), (5.7)], contracting the first of Equations (5.13) with $e^\nu_{[3]}$ and $e^\mu_A e^\nu_{[3]}$, we will get

$$-\partial_\mu \ln\mathcal{R} = e_{[3]\mu}\partial_{[3]}\ln\mathcal{R} - \omega_{A33}e^A_\mu, \ \partial_A \ln\mathcal{R} = -\omega_{3A3}, \quad A = 0,1,2, \tag{5.18}$$

and this is compatible with the condition for integrability, $-\partial_\mu \ln n = e_{[3]\mu}\partial_{[3]}\ln n - \omega_{A33}e^A_\mu$, of the system (5.17) only when $n(x)/\mathcal{R}(x) = \mathrm{const}$. Next, we may compute the second derivative of ρ. Using Equation (5.7) and Equation (5.27) below, we arrive at

$$g^{\mu\nu}\nabla_\mu\nabla_\nu\rho=-\nabla_\mu\left(ne^\mu_{[3]}\right)=n\left[\omega^A_{3A}+\partial_{[3]}\ln n\right]=n\left[-3\partial_{[3]}\ln\mathcal{R}+\partial_{[3]}\ln n\right]=n\,\partial_{[3]}\ln\left(n/\mathcal{R}^3\right)$$

From here we find that if $n(x)=\mathcal{R}(x)$, then $\rho(x)$ is the solution of an inhomogeneous wave equation,

$$\Box\rho=g^{\mu\nu}\nabla_\mu\nabla_\nu\rho=2m\mathcal{P},\ \partial_\mu\rho=\mathcal{R}e^{[3]}_\mu=-\mathcal{J}_\mu,\ \ e^\mu_A\partial_\mu\rho=\mathcal{R}\delta^3_A, \tag{5.19}$$

for the " potential" ρ with the source density proportional to the mass parameter *m* of the Dirac equation and pseudoscalar density $\mathcal{P}$ (in static limit, it becomes the Poisson equation). Not surprisingly, this source is equal to the derivative of the invariant density in the direction of the axial current. If the invariant density was not changing in a "radial direction", the whole idea of a localized object would be vague. Similarly to (5.5) and (5.11), we have

$$\mathrm{d}\rho=\mathrm{d}x^\mu\,\partial_\mu\rho=\mathcal{R}e^{[3]}_\mu\mathrm{d}x^\mu=\mathcal{R}\mathrm{d}S^{[3]}. \tag{5.20}$$

From here, we conclude that the differential form $\mathcal{J}_\mu\mathrm{d}x^\mu$ is integrable and the "radial distance",

$$\rho_2-\rho_1=\int_{x(\rho_1)}^{x(\rho_2)}\mathcal{J}_\mu(x)\,\mathrm{d}x^\mu\ \left(=\int\mathcal{R}\mathrm{d}S^{[3]}\right), \tag{5.21}$$

does not depend on the integration path (the coordinate variable ρ is holonomic).

5.3. Congruences of the Angular Arcs

Here, we must deal with four equations (4.6) and (4.28). Taking $e^\mu_{[1]}=\Theta^\mu(x)/\mathcal{R}$, $e^\mu_{[2]}=\Phi^\mu(x)/\mathcal{R}$ (an alternative choice with $e^\mu_{[1]}\leftrightarrow e^\mu_{[2]}$ will be discussed later), starting from Equation (4.6), and duplicating the derivation of Equation (5.8) we arrive at the equations,

$$\omega_{212}+\omega_{313}=-2eA_{[2]},\quad \omega_{121}+\omega_{323}=+2eA_{[1]}.$$

Since by the second Equation (5.18) we have $\omega_{313}=-\partial_{[1]}\ln\mathcal{R}$ and $\omega_{323}=-\partial_{[2]}\ln\mathcal{R}$, these equations completely define ω_{212} and ω_{121},

$$\omega_{212}=\partial_{[1]}\ln\mathcal{R}-2eA_{[2]},\quad \omega_{121}=\partial_{[2]}\ln\mathcal{R}+2eA_{[1]}. \tag{5.22}$$

Putting further in Equations (4.28) $\Theta_\mu=\mathcal{R}e^\mu_{[1]}$ and $\Phi_\mu=\mathcal{R}e^\mu_{[2]}$, and duplicating the scheme of Equation (5.3)-(5.7), we obtain,

$$\omega_{1BA}-\omega_{1AB}=-2e\left(A_A\eta_{2B}-A_B\eta_{2A}\right),\ A,B\neq1;A\neq B, \tag{5.23}$$

$$-\omega_{1A1}-\partial_A\ln\mathcal{R}=2eA_{[1]}\eta_{2A},\ A=0,2,3, \tag{5.24}$$

$$\omega_{2BA}-\omega_{2AB}=2e\left(A_A\eta_{1B}-A_B\eta_{1A}\right),\ A,B\neq2;A\neq B, \tag{5.25}$$

$$-\omega_{2A2}-\partial_A\ln\mathcal{R}=-2eA_{[2]}\eta_{1A},\ A=0,1,3. \tag{5.26}$$

Giving index *A* in Equations (5.24) and (5.26) all possible values, we get the following constraints,

$$\omega_{101}=\omega_{202}=-\partial_{[0]}\ln\mathcal{R},\ \omega_{[3]}=\omega_{232}=-\partial_{[3]}\ln\mathcal{R}=m\sin\mathcal{Y}; \tag{5.27}$$

$$\omega_{212}=-2eA_{[2]}-\partial_{[1]}\ln\mathcal{R},\ \omega_{[2]}=2eA_{[1]}-\partial_{[2]}\ln\mathcal{R}. \tag{5.28}$$

Equations (5.28) and (5.22) are mutually compatible only when $\partial_{[1]}\ln\mathcal{R}=\partial_{[2]}\ln\mathcal{R}=0$, and

$$\omega_{010}=\omega_{313}=\partial_{[1]}\ln\mathcal{R}=0,\ \ \omega_{020}=\omega_{323}=\partial_{[2]}\ln\mathcal{R}=0, \tag{5.29}$$

i.e., when the vectors of the geodesic curvature ω_{0A0} and ω_{3A3} of the congruences [0] and [3] of the vector and axial currents have no projections on the lines of the congruences [1] and [2] of the charged currents.

[11]Having no metric, we assume here geodesic of an affine space, *i.e.* such a line $x^\mu(s)$ that its tangent vector, $p^\mu(s)=\mathrm{d}x^\mu/\mathrm{d}s$, is parallel transported (with respect to an affine connection $\Gamma^\mu_{\lambda\sigma}$ (3.14)) along the line, $\mathrm{d}p^\mu/\mathrm{d}s\propto p^\mu$. In our particular case of the tetrad vector $\boldsymbol{e}_3$, this amounts to $\nabla_3\boldsymbol{e}_3=\omega^A_{33}\boldsymbol{e}_A\propto\omega^3_{33}\boldsymbol{e}_3\equiv0$.

Together with the previously obtained Equations (5.8), (5.18) and (5.22), they give all ω_{ABA} in terms of derivatives of the invariant density and electromagnetic potentials. Namely, since $\omega_{303} = -\partial_{[0]} \ln \mathcal{R}$, we also have $\omega_{101} + \omega_{202} = 0$, which together with the first Equation (5.27) entails that

$$\partial_{[0]} \ln \mathcal{R} = 0, \quad \omega_{3A3} = 0, \quad A = 0,1,2. \tag{5.30}$$

The second of these equations means that the congruence [3] is geodesic[11]. Quite remarkably, this conclusion about static character of the configuration that satisfies Dirac equations of motion is reached only after all the differential identities are considered together. The additional constraints that follow from Equations (5.23) and (5.25), when indices *A* and *B* are given all possible values, are as follows,

$$\omega_{130} = \omega_{103}, \ \omega_{230} = \omega_{203}, \ \omega_{120} - \omega_{102} = 2eA_{[0]}, \tag{5.31}$$

$$\omega_{123} - \omega_{132} = 2eA_{[3]}, \quad \omega_{231} - \omega_{213} = 2eA_{[3]}. \tag{5.32}$$

Combined with the previous results (Equation (5.4), particularly) they yield,

$$\omega_{120} = 2eA_{[0]}, \ \omega_{012} = \omega_{021} = 0, \tag{5.33}$$

$$\omega_{3AB} = -\omega_{3BA}; A, B = 0,1,2; A \neq B. \tag{5.34}$$

The last of these equations is the necessary and sufficient condition for the congruences of lines $\boldsymbol{e}_{[0]}$, $\boldsymbol{e}_{[1]}$ and $\boldsymbol{e}_{[2]}$ being canonical of the congruence $\boldsymbol{e}_{[3]}$ [18]. This property appears to be yet another consequence of the Dirac equation of motion, which thus guarantees that the orthogonal tetrad is Fermi-transported. Finally, comparing Equations (5.16) and (5.34) we find that

$$\omega_{3AB} = 0, \quad A, B = 0,1,2; \ A \neq B. \tag{5.35}$$

5.4. Summary—Coefficients of Rotations That Completely Define the Matter-Induced Affine Geometry

By now, we have succeeded to find simple expressions for *all* coefficients ω_{ABC} of rotation of the basis $\boldsymbol{e}_A$ of the normalized Dirac currents. This is the last step in the design of the matter-induced affine geometry. From this point, one can rely on the common tools of the differential geometry. We can divide the not vanishing components of ω_{ABC} into two distinct groups:

1) Five geodesic curvatures (the ω_{ABC} with only two distinct indices),

$$\begin{aligned} &\omega_{030} = -\omega_{131} = -\omega_{232} = -m\mathcal{P}/\mathcal{R} = -m\sin\mathcal{Y}, \\ &\omega_{121} = +2eA_{[1]}, \quad \omega_{212} = -2eA_{[2]}. \end{aligned} \tag{5.36}$$

2) Only two of the ω_{ABC} with all three different indices are nonzero. These are

$$\omega_{120} = -\omega_{210} = 2eA_{[0]}, \quad \omega_{123} = -\omega_{213} = 2eA_{[3]} \tag{5.37}$$

3) The coefficients ω_{ABC}, which depend on the potential A_D, are of the same form

$$\omega_{12D} = 2eA_D, \tag{5.38}$$

so that presence of electromagnetic field causes rotation of the Dirac tetrad in the (12)—tangent plane. This inter-action makes it impossible, in general, to match Dirac equation with the all-orthogonal system of hypersurfaces[12].

It is essential that the only directional derivative that survived all constrains is $\partial_{[3]}\mathcal{R}$, and even it can be

[12]Keeping up with the promise given in Section 3, we compute, following Equation (3.10), the coefficients of rotation γ_{abd} of the basis $\boldsymbol{h}_a$.

$$\gamma_{abd} = \left[\omega_{12D} V_d^D V_a^{[1]} V_b^{[2]} + V_a^{[0]} V_b^{[3]} V_d^{[0]} \omega_{030} + V_a^{[1]} V_b^{[3]} V_d^{[1]} \omega_{131} + V_a^{[2]} V_b^{[3]} V_d^{[2]} \omega_{232}\right] - (b \leftrightarrow a).$$

Using Equations (5.36)-(5.37) and employing Equation (2.5) as, $V_a^{[0]} V_d^{[0]} - V_a^{[1]} V_d^{[1]} - V_a^{[2]} V_d^{[2]} = \eta_{ad} + V_a^{[3]} V_d^{[3]}$, we obtain,

$$\gamma_{abd}(\psi) = 2eA_d \left(V_a^{[1]} V_b^{[2]} - V_b^{[1]} V_a^{[2]}\right) + m\sin\mathcal{Y} \cdot \left(V_a^{[3]} \eta_{bd} - V_b^{[3]} \eta_{ad}\right). \tag{5.39}$$

expressed via pseudoscalar density. Therefore, the practical computation of the connection ω_{ABC} does not require any reference to a coordinate background. The congruence of integral lines of the vector field $\boldsymbol{e}_3$ is both normal and geodesic. This is the only geodesic of the principal manifold $\mathbb{M}$, and it is inherited by the hypersurfaces of the constant world time. The congruences $\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2$ constitute a canonical system with respect to the congruence $\boldsymbol{e}_3$. Therefore the entire tetrad is Fermi-transported along the the lines of the radial congruence $\boldsymbol{e}_3$. Equations (5.36)-(5.39) assume a localized configuration with maximum of invariant density in its interior and a naturally right-handed spatial trihedron $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$. If there is a minimum, then the signs of tetrad components A_B in coefficients of rotation (5.36)-(5.37) (*and only there*!) must be reverted.

6. Coordinate Surfaces and Coordinate Lines of the Dirac Field

Below, we attempt to find the submanifolds of the physical manifold $\mathbb{M}$, which can be mapped onto coordinate surfaces of the arithmetic $\mathbb{R}^4$. An advance knowledge of these surfaces will be critical for finding the auto-localized Dirac waveforms and then understanding their shape and internal field structure. If we denote the differential operators $e_A^\mu \partial_\mu$ as $\partial_{[A]}$ and introduce, for the sake of brevity, $Q \equiv \partial_{[3]} \ln \mathcal{R} = -m\mathcal{P}/\mathcal{R} = -m \sin \mathcal{Y}$, then an explicit calculation according to the second Equation (3.23),

$$\left(\partial_A \partial_B - \partial_B \partial_A\right)\psi = -\left(\omega_{AB}^D - \omega_{BA}^D\right)\partial_D \psi \equiv C_{AB}^D \partial_D \psi,$$

yields the following expressions for the Poisson brackets,

$$\begin{aligned}
&\left(\partial_{[0]}\partial_{[3]} - \partial_{[3]}\partial_{[0]}\right) f = -Q \cdot \partial_{[0]} f, && \text{(a)}\\
&\left(\partial_{[0]}\partial_{[1} - \partial_{[1]}\partial_{[0]}\right) f = -2eA_{[0]}\partial_{[2]} f, && \text{(b)}\\
&\left(\partial_{[0]}\partial_{[2]} - \partial_{[2]}\partial_{[0]}\right) f = 2eA_{[0]} \cdot \partial_{[1]} f, && \text{(c)}\\
&\left(\partial_{[1]}\partial_{[2]} - \partial_{[2]}\partial_{[1]}\right) f = 2eA_{[2]} \cdot \partial_{[2]} f + 2eA_{[1]} \cdot \partial_{[1]} f, && \text{(d)}\\
&\left(\partial_{[3]}\partial_{[1]} - \partial_{[1]}\partial_{[3]}\right) f = Q \cdot \partial_{[1]} f - 2eA_{[3]} \cdot \partial_{[2]} f, && \text{(e)}\\
&\left(\partial_{[3]}\partial_{[2]} - \partial_{[2]}\partial_{[3]}\right) f = Q \cdot \partial_{[2]} f + 2eA_{[3]} \cdot \partial_{[1]} f. && \text{(f)}
\end{aligned} \tag{6.1}$$

These expressions allow one to completely explore properties not only of the individual congruences and 3-d hypersurfaces but also of the 2-d surfaces. The latter is imperative as long as we aim at (and already have a hint of) dynamic localization of the Dirac field into finite-sized objects.

Some immediate observations are in order. Equations (6.1) are nothing but differential identities that express the integrability of the directional derivatives. From equations of motion we know that $\partial_A \mathcal{R} = 0$ for $A = 0,1,2$ and $\partial_{[3]}\mathcal{R} = -m\mathcal{P}$. Let us take in Equation (6.1) $f = \mathcal{R}$ and use Equations (5.29) and (5.30). Then from Equations (6.1.e,f) we have $\partial_{[1]}\mathcal{P} = 0$ and $\partial_{[2]}\mathcal{P} = 0$, while Equation (6.1.a) yields $\partial_{[0]}\mathcal{P} = 0$. Thus, we have even more constraints,

$$\partial_A \mathcal{R} = \partial_A \mathcal{S} = \partial_A \mathcal{P} = \partial_A \mathcal{Y} = 0, \quad A = 0,1,2. \tag{6.2}$$

At any point P of the principal manifold $\mathbb{M}$ all the scalars change only in the direction $\boldsymbol{e}_3$ of the axial current, and the rate of this change is determined by the product $m\mathcal{P}$.

6.1. Integrable Subsystems and Coordinate Surfaces in $\mathbb{R}^4$

Since we are aiming at the discovery of the localized solutions, a coordinate picture may become most appropriate, and it is useful to know in advance what the admissible coordinate net may look like. Solely for this purpose, we study here whether the congruences of the Dirac currents in $\mathbb{R}^4$ can form at least some of the four 3-d coordinate hypersurfaces and of the six 2-d coordinate surfaces. Once found, these surfaces will be studied in detail as submanifolds embedded into $\mathbb{M}^4$ endowed with the connections identified above.

1. Hypersurfaces $S_{(123)}$ and $S_{(120)}$. From visual inspection of the Poisson brackets (6.1), among the four equations, $e_A^\mu \partial_\mu f = 0$, there are two integrable systems of three equations that define two hypersurfaces and two integrable system of two equations that define two surfaces in the coordinate space $\mathbb{R}^4$. Namely, three com-

mutators between the $\partial_{[1]}$, $\partial_{[2]}$ and $\partial_{[3]}$ [Equations (6.1 d,e,f)] are the linear combinations of these operators alone. Therefore, the function $\tau(x)=\tau^*=\text{const}$ (as well as any function $f(\tau)$) is the first integral of the complete (Jacobian) system of three equations,

$$e^\mu_{[1]}\partial_\mu\tau=0, e^\mu_{[2]}\partial_\mu\tau=0 \text{ and } e^\mu_{[3]}\partial_\mu\tau=0. \tag{6.3}$$

The parameter τ^* enumerates the family of hypersurfaces $S_{(123)}$, which are spanned by the streamlines of the vector fields $e^\mu_{[1]}$, $e^\mu_{[2]}$ and $e^\mu_{[3]}$ and have $e^{[0]}_\mu$ as the normal. Equations (6.1 b,c,d) indicate that three equations,

$$e^\mu_{[1]}\partial_\mu\rho=0, e^\mu_{[2]}\partial_\mu\rho=0 \text{ and } e^\mu_{[0]}\partial_\mu\rho=0, \tag{6.4}$$

also constitute an integrable system with a first integral $\rho(x)=\rho^*=\text{const}$ (or any function $f(\rho)$); the latter represents hypersurfaces $S_{(120)}$ of the constant "radius" ρ when $\mathcal{R}^2>0$. These are spanned by the integral lines of the vector fields $e^\mu_{[1]}$, $e^\mu_{[2]}$ and $e^\mu_{[0]}$ and have $e^{[3]}_\mu$ as the spacelike normal.

2. Surfaces $S_{(12)}$ and $S_{(03)}$. Next, by Equation (6.1 d) the system of equations

$$e^\mu_{[1]}(x)\partial_\mu\vartheta=0,\ \ e^\mu_{[2]}(x)\partial_\mu\vartheta=0 \tag{6.5}$$

is integrable. Its two first integrals, $\theta_1(x)=c_1$ and $\theta_2(x)=c_2$, determine a two-dimensional surface $S_{(12)}$ spanned by the streamlines of the vector fields e^μ_1 and e^μ_2 having the normal vectors $\nu_\mu=\partial_\mu\vartheta=c_0e^0_\mu+c_3e^3_\mu$. The first integrals of the system (6.5) are known because both of its equations are satisfied by $\vartheta(x)=\tau(x)=\tau^*$ and $\vartheta(x)=\rho(x)=\rho^*$. Once $\tau(x)$ and $\rho(x)$ are algebraically independent, these are the two first integrals of the system (5), and the 2-d surface $S_{(12)}$ is uniquely fixed by the values of constants τ^* and ρ^*, which enumerate the surfaces of a constant "radius" ρ at a given "world time" τ.

Finally, according to Equation (6.1 a) the commutator between $\partial_{[0]}$ and $\partial_{[3]}$ is proportional to $\partial_{[0]}$. Therefore, the system of equations

$$e^\mu_{[0]}(x)\partial_\mu\varphi=0,\ \ e^\mu_{[3]}(x)\partial_\mu\varphi=0 \tag{6.6}$$

is integrable. It has two first integrals, $\phi_1(x)=C_1$ and $\phi_2(x)=C_2$, which determine a two-dimensional surface $S_{(03)}$ spanned by the streamlines of the vector fields $e^\mu_{[0]}$ and $e^\mu_{[3]}$. The two normal vectors $n_\mu=\partial_\mu\phi$ of these surfaces are the linear combinations $c_1e^{[1]}_\mu+c_2e^{[2]}_\mu$. One of the first integrals of the second Equation (6.6) is $\varphi(x)=\tau(x)$, *i.e.* we have $e^\mu_{[3]}(x)\partial_\mu\tau(x)=0$. Also, one of the first integrals of the first Equation (6.6) is $\varphi(x)=\rho(x)$, *i.e.* $e^\mu_{[0]}(x)\partial_\mu\rho(x)=0$. Since the congruences of integral lines of the fields $e^\mu_{[0]}$ and e^μ_3 are normal—(cf. Section 5), we have $\partial_\mu\tau(x)=\mathcal{R}e_{[0]\mu}(x)$ and $e^\mu_{[0]}(x)\partial_\mu\tau(x)=\mathcal{R}$, as well as $\partial_\mu\rho(x)=\mathcal{R}e_{[3]\mu}(x)$ and $e^\mu_{[3]}(x)\partial_\mu\rho(x)=-\mathcal{R}$. In terms of the new independent variables, $\tau=\tau\left(x^0,x^1,x^2,x^3\right), x^1, x^2, \rho=\rho\left(x^0,x^1,x^2,x^3\right)$, the system (6.6) immediately acquires the normal (Jacobian) form,

$$\begin{aligned} &-\mathcal{R}\frac{\partial\varphi}{\partial\rho}+e^1_{[3]}\frac{\partial\varphi}{\partial x^1}+e^2_{[3]}\frac{\partial\varphi}{\partial x^2}=0; \quad (a)\\ &\mathcal{R}\frac{\partial\varphi}{\partial\tau}+e^1_{[0]}\frac{\partial\varphi}{\partial x^1}+e^2_{[0]}\frac{\partial\varphi}{\partial x^2}=0. \quad (b)\end{aligned} \tag{6.7}$$

Its second equation is equivalent to the system of three ODEs,

$$\frac{\mathrm{d}\rho}{0}=\frac{\mathrm{d}\tau}{\mathcal{R}}=\frac{\mathrm{d}x^1}{e^1_{[0]}}=\frac{\mathrm{d}x^2}{e^2_{[0]}}, \tag{6.8}$$

which has three first integrals, $\rho=\rho^*=\text{const}$, $\phi_1=C_1$, $\phi_2=C_2$. In terms of the new independent variables, $\zeta^0=\tau,\zeta^3=\rho,\zeta^1=\phi_1,\zeta^2=\phi_2$, the system (6.7) reads as

$$-\mathcal{R}\frac{\partial\varphi}{\partial\rho}+\theta^1\frac{\partial\varphi}{\partial\zeta^1}+\theta^2\frac{\partial\varphi}{\partial\zeta^2}=0,\ \ \mathcal{R}\frac{\partial\varphi}{\partial\tau}=0, \tag{6.9}$$

where $\theta^i = \mathcal{R}\left(\partial\phi_i/\partial\rho\right) + e^1_{[3]}\left(\partial\phi_i/\partial x^1\right) + e^2_{[3]}\left(\partial\phi_i/\partial x^2\right)$. Since ϕ is independent of τ, we have one PDE in three variables, which is equivalent to the system of two ODEs. The variables τ and ρ form an orthogonal coordinate basis on every 2-d surface $S_{(03)}$ (enumerated by the values of ζ^1 and ζ^2).

6.2. Coordinate Surfaces as Submanifolds in $\mathbb{M}$

Conditions for simultaneous integrability of the PDEs for the streamlines of the Dirac currents prompted the existence of the (hyper)surfaces in $\mathbb{R}^4$ and, most importantly, in $\mathbb{M}$. Here, in order to understand their shape, we look at them as submanifolds of the principal manifold $\mathbb{M}$.

1. The method. For the sake of brevity, we will use the Latin capitals $H,\cdots,N = 0,1,2,3$ to label the entire tetrad basis $\boldsymbol{e}_H$ (or $\boldsymbol{V}_H$). In the context of the current work this is the basis of the ambient space. The capitals $P_t = (P,\cdots,U)$ will label the tangent tetrad vectors of a 3-d or 2-d submanifold. The capitals $A_n = (A,\cdots,F)$ will be used to label the normal vectors. Then the induced metric of a submanifold is $g_{PQ} = g_{\mu\nu}e^\mu_P e^\nu_Q = \eta_{PQ}$ and, by virtue of definition (2.11), the first quadratic form of the surface $S_{(P\cdots)}$ is (pseudo)-Euclidean, $\mathrm{d}s^2 = \eta_{PQ}\mathrm{d}S^P\mathrm{d}S^Q$.

Since we are interested in submanifolds that are spanned by the integral lines of the tetrad vectors, the Gauss and Weingarten decompositions of the covariant derivatives of tangent and normal (with respect to a submanifold) tetrad vectors immediately follow from Equations (3.2),

$$(\text{Gauss}) \quad \bar{\nabla}_R V^a_P = \omega^H_{PR}V^a_H = \sum_{Q_t}\omega^Q_{PR}V^a_Q + \sum_{A_n}\omega^A_{PR}V^a_A, \tag{6.10}$$

$$(\text{Weingarten}) \quad \bar{\nabla}_R V^a_A = \omega^H_{AR}V^a_H = \sum_{P_t}\omega^P_{AR}V^a_P + \sum_{B_n}\omega^B_{AR}V^a_B, \tag{6.11}$$

where all the ω's listed in Equations (5.36)-(5.37) are known explicitly[13]. The first term, ω^Q_{PR}, in the r.h.s. of the Gauss decomposition (6.10) is the connection of the intrinsic tangent space of the submanifold. The second term, $L^A_{PR} = \omega^A_{PR}$ (with two tangent and one normal indices), is the *second fundamental form* of the submanifold with respect to the normal V_A. The first term in Weingarten decomposition (6.11), $A^P_{AR} = \omega^P_{AR} = -\eta_{AC}\eta^{QP}\omega^C_{QR} = -\eta_{AC}\eta^{QP}L^C_{QR}$, (the *shape form* with two tangent and one normal indices) is similar to the second fundamental form in (6.10); both account for the rotation of the tetrad in the (*PA*) plane when it is displaced in a tangent direction $\boldsymbol{e}_R$. The second term of Equation (6.11), $D_{AC;R} = -D_{CA;R} = \eta_{CB}\omega^B_{AR}$, with two normal and one tangent indices, is the covariant derivative of the normal components of a vector in a tangent direction of the submanifold. It accounts for the rotation of the (*AB*)—plane of the two normals under infinitesimal displacement in tangent direction $\boldsymbol{e}_R$.

Now, since there is no question of how a submanifold is embedded into the ambient space with explicitly known tetrad vectors, we are in position to study the internal geometry of various *coordinate surfaces*, as submanifolds of the principal manifold $\mathbb{M}$. Besides the second fundamental form, we will use the Riemann curvature tensor in ambient space and in subspaces,

$$R^N_{HKL} = \partial_H\omega^N_{LK} - \partial_K\omega^N_{LH} + \omega^N_{LM}\left(\omega^M_{HK} - \omega^M_{KH}\right) + \omega^M_{LH}\omega^N_{MK} - \omega^M_{LK}\omega^N_{MH} \tag{6.12}$$

With these preliminaries, we are in the position to consider all subspaces on-by-one.

2. The hypersurface $S_{(123)}$ represents space at a given time. It has three spacelike tangent vectors $\boldsymbol{V}_P$, $(P = 1,2,3)$, and a single timelike normal vector $\boldsymbol{V}_0$. The coefficients of the single second fundamental form are $L^0_{12} = L^0_{23} = L^0_{31} = 0$ and $L^0_{11} = L^0_{22} = L^0_{33} = -\partial_{[0]}\ln\mathcal{R} = 0$. The second fundamental form, $\mathrm{II}^0 = L^0_{PQ}\mathrm{d}S^P\mathrm{d}S^Q$, is proportional to first fundamental form, $\mathrm{I}_{(123)} = \eta_{PQ}\mathrm{d}S^P\mathrm{d}S^Q = -\left(\mathrm{d}S^1\right)^2 - \left(\mathrm{d}S^2\right)^2 - \left(\mathrm{d}S^3\right)^2$ of the $S_{(123)}$,

$$\mathrm{II}^0 = \partial_{[0]}\ln\mathcal{R}\cdot\left[-\left(\mathrm{d}S^1\right)^2 - \left(\mathrm{d}S^2\right)^2 - \left(\mathrm{d}S^3\right)^2\right] = 0. \tag{6.13}$$

Therefore, the $S_{(123)}$ is a totally umbilical submanifold[14] with zero mean normal curvature $H = \partial_{[0]}\ln\mathcal{R} = 0$. The latter means that $S_{(123)}$ is a totally geodesic submanifold; it inherits its sole geodesic $\boldsymbol{e}_3$ from the ambient $\mathbb{M}$. From the perspective of the ambient space, the hypersurface $S_{(123)}$ has no curvature, it is extrinsically flat.

[13] In mathematical literature the Gauss and Weingarten formulae are written down as $\bar{\nabla}_X Y = \nabla_X Y + h(X,Y)$ and $\bar{\nabla}_X\xi = -A_\xi X + D_X\xi$, respectively. Here, X,Y are tangent and ξ is normal to the submanifold.

[14] All points of which are umbilical. A point is called umbilical if all principal curvatures at this point are equal.

The extrinsic part vanishes together with the connections $\omega_{[0]} = \omega_{202} = -\omega_{303} = -\partial_{[0]} \ln \mathcal{R} = 0$. The intrinsic Riemann curvature of the $S_{(123)}$ has six different (modulo sign) components; it is given by the terms of (6.12) with all indices in tangent space of the $S_{(123)}$,

$$\begin{aligned} R^t_{1212} &= 2e\left(\partial_{[1]}A_{[2]} - \partial_{[2]}A_{[1]}\right) - 4e^2\left(A^2_{[1]} + A^2_{[2]}\right) + Q^2 = 2eF_{12} + Q^2, \\ R^t_{1313} &= R^t_{2323} = \partial_{[3]}Q - Q^2, \;\; R^t_{1213} = 2eA_{[1]}Q, \; R^t_{1232} = -2eA_{[2]}Q, \; R^t_{1323} = 4eA_{[3]}Q, \end{aligned} \tag{6.14}$$

where $F_{BC} = e^\mu_A e^\nu_C F_{\mu\nu} = \partial_B A_C - \partial_C A_B + \left(\omega^D_{BC} - \omega^D_{CB}\right)A_D$ coincide, by appearance, with the tetrad components of the electromagnetic field tensor rewritten in the basis $\boldsymbol{e}_A$. It should be remembered that all the A_C here came from the components of the Ricci coefficients of rotation (5.38).

3. The hypersurface $S_{(120)}$ represents the surface of a given "radius" at all times. It has two spacelike and one timelike tangent vectors $\boldsymbol{V}_P$, $(P = 0,1,2)$, and a single spacelike normal vector $\boldsymbol{V}_3$. The coefficients of the second fundamental form are $L^3_{12} = L^3_{20} = L^3_{30} = 0$ and $L^3_{11} = L^3_{22} = -L^3_{00} = \partial_{[3]} \ln \mathcal{R} = Q$. The second fundamental form, $\mathrm{II}^3 = L^3_{PQ}\mathrm{d}S^P\mathrm{d}S^Q$, is proportional to the first fundamental form $\mathrm{I}_{(120)} = \eta_{PQ}\mathrm{d}S^P\mathrm{d}S^Q = \left(\mathrm{d}S^0\right)^2 - \left(\mathrm{d}S^1\right)^2 - \left(\mathrm{d}S^2\right)^2$ of the $S_{(120)}$,

$$\mathrm{II}^3 = -Q \cdot \left[\left(\mathrm{d}S^0\right)^2 - \left(\mathrm{d}S^1\right)^2 - \left(\mathrm{d}S^2\right)^2\right] = -Q \cdot \mathrm{I}_{(120)}. \tag{6.15}$$

Therefore, the hypersurface $S_{(120)}$ is also a totally umbilical submanifold with the mean curvature $H = -Q = m\mathcal{P}/\mathcal{R} > 0$. By virtue of Equations (6.2), the vector of (mean) geodesic curvature H is constant and parallel throughout every hypersurface $S_{(120)}$.

The intrinsic part of the Riemann curvature of the hypersurface $S_{(120)}$ has only the following components,

$$R^t_{1212} = 2e\left(\partial_{[1]}A_{[2]} - \partial_{[2]}A_{[1]}\right) - 4e^2\left(A^2_{[1]} + A^2_{[2]}\right) = 2eF_{12}, \tag{6.16}$$

identical with those of $S_{(12)}$. The extrinsic parts are due to $\omega_{131} = \omega_{232} = -\omega_{030} = -Q$, *i.e.*, the connections that contain normal component $\boldsymbol{e}_3$,

$$R^n_{1212} = -R^n_{1010} = -R^n_{2020} = Q^2 = m^2 \sin^2 \mathcal{Y}. \tag{6.17}$$

Since congruences $\boldsymbol{e}_0$, $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are canonical with respect to the normal congruence $\boldsymbol{e}_3$, their lines are the lines of curvature of the hypersurface $S_{(120)}$. If at some point of $S_{(120)}$ we have $\mathrm{II}^3 = L^3_{PQ}\mathrm{d}S^P\mathrm{d}S^Q = 0$, then the directions of $\boldsymbol{e}_0$, $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ become the asymptotic directions.

4. Surface $S_{(12)}$ is the surface of a given "radius" at a given time and can be viewed as a hypersurface of either $S_{(123)}$ or $S_{(120)}$ with the normals $\boldsymbol{e}_3$ or $\boldsymbol{e}_0$, respectively. It has two spacelike tangent vectors $\boldsymbol{V}_P$, $(P = 1,2)$, and two normal vectors $\boldsymbol{V}_A$, $(A = 0,3)$, timelike $\boldsymbol{V}_0$ and spacelike $\boldsymbol{V}_3$. Accordingly, there are two second fundamental forms, $\mathrm{II}^0 = L^0_{PQ}\mathrm{d}S^P\mathrm{d}S^Q$ and $\mathrm{II}^3 = L^3_{PQ}\mathrm{d}S^P\mathrm{d}S^Q$, with the following coefficients $L^0_{12} = L^3_{12} = 0$, $L^0_{11} = L^0_{22} = 0$, $L^3_{11} = L^3_{22} = \partial_{[3]} \ln \mathcal{R} = Q$. The first fundamental form of $S_{(12)}$ is $\mathrm{I}_{(12)} = \eta_{PQ}\mathrm{d}S^P\mathrm{d}S^Q = -\left(\mathrm{d}S^1\right)^2 - \left(\mathrm{d}S^2\right)^2$, and the two second fundamental forms are

$$\mathrm{II}^0 = 0, \;\; \mathrm{II}^3 = -2Q \cdot \mathrm{I}_{(12)}. \tag{6.18}$$

Therefore, the 2-d surface $S_{(12)}$ is a totally umbilical submanifold with the mean curvature $H = m\mathcal{P}/\mathcal{R} = m\sin\mathcal{Y} > 0$, which is determined by the Dirac field within principal manifold $\mathbb{M}$. The Gaussian curvature $K = \sqrt{k_1 k_2} = m\sin\mathcal{Y}$ is positive. *Such a surface can only be the sphere with the radius of curvature* $\kappa = 1/H$ [19] [20]. (It is a plane, when $\kappa = 0$, but then $\mathcal{R}$ must be uniform and $\mathcal{P} = 0$. Here, the spherical shape is a dynamic symmetry since it originates from equations of motion.). Nearly the most important property of submanifolds $S_{(12)}$ follows from the compatibility conditions (5.29) and Equation (6.2), which indicate that the invariant densities $\mathcal{R}, \mathcal{S}, \cdots$ are constant along every 2-d surface $\tau = \tau^* = \mathrm{const}$, $\rho = \rho^* = \mathrm{const}$. The mean curvature H is constant along $S_{(12)}$ as well. The normal connection for this submanifold can be only due to the components ω^3_{01} and ω^3_{02} of the connection ω^B_{AR}, but these vanish identically, $D_{03;1} = D_{03;2} = 0$, so that both normal vector fields (and the mean curvature vector) are parallel with respect to the tangent displacements along $S_{(12)}$, $D_R H = 0$. The Riemann curvature of $S_{(12)}$ has only one component, R_{1212} and it can be decomposed in two parts. The intrinsic one, R^t_{1212}, is given by the terms of (6.12) with all indices in tangent space of $S_{(12)}$. The only nonzero connections here are $\omega_{212} = -2eA_{[2]}$ and $\omega_{121} = 2eA_{[1]}$, so that sectional curvature

of the $S_{(12)}$,

$$R^t_{1212} = 2e\left(\partial_{[1]}A_{[2]} - \partial_{[2]}A_{[1]}\right) - 4e^2\left(A^2_{[1]} + A^2_{[2]}\right) = 2eF_{12}, \tag{6.19}$$

is entirely due to the tangent tetrad components of the electromagnetic field A_μ. The extrinsic part, R^n_{1212}, is due to the connections $\omega_{131} = \omega_{232} = -Q$ from the second fundamental form and

$$R^n_{1212} = L^0_{11}L^0_{22} - L^3_{11}L^3_{22} = -\left(\partial_{[3]}\ln\mathcal{R}\right)^2 = -Q^2. \tag{6.20}$$

5. The surface $S_{(03)}$ represents a given "angular direction" at all "radial" distances and at all times. It has one spacelike and one timelike tangent vectors $\boldsymbol{V}_P$, $(P=0,3)$, and two spacelike normal vectors $\boldsymbol{V}_A$, $(A=1,2)$. Here, we also have two second fundamental forms, $\mathrm{II}^1 = L^1_{PQ}\mathrm{d}S^P\mathrm{d}S^Q$ and $\mathrm{II}^2 = L^2_{PQ}\mathrm{d}S^P\mathrm{d}S^Q$, with the following coefficients $L^1_{03} = L^2_{03} = 0$, $L^1_{00} = -L^1_{33} = 0$, $L^2_{00} = -L^2_{33} = 0$. The first fundamental form of the $S_{(03)}$ is $\mathrm{I}_{03} = \left(\mathrm{d}S^0\right)^2 - \left(\mathrm{d}S^3\right)^2$ and both second fundamental forms are just zero, $\mathrm{II}^1 = \mathrm{II}^2 = 0$.

The submanifold $S_{(03)}$ is totally umbilical with the mean curvature $H=0$, and as such is a totally geodesic submanifold. The shape form of $S_{(03)}$ is zero. The normal connection for the coordinate surface $S_{(03)}$ (and only for this surface) does not vanish,

$$\nabla_R\boldsymbol{e}_1 = -2eA_R\boldsymbol{e}_2,\ \nabla_R\boldsymbol{e}_2 = 2eA_R\boldsymbol{e}_1,\ R=0,3, \tag{6.21}$$

solely due to the external potential A_μ, $D_{12;0} = -2eA_{[0]}$, $D_{12;3} = -2eA_{[3]}$. A displacement in the directions of $\boldsymbol{e}_0$ and $\boldsymbol{e}_3$, rotates the tetrad in plane (12). The Riemannian sectional curvature of the $S_{(03)}$ is induced by an ambient space,

$$R^t_{0303} = -\partial_{[3]}Q + Q^2. \tag{6.22}$$

6.3. Coordinate Lines

According to Equation (6.2), system (6.5) of PDEs admits, along with the first integrals $\tau(x) = \tau^*$ and $\rho(x) = \rho^*$ of hypersurfaces $S_{(123)}$ and $S_{(120)}$, respectively, the first integrals $\vartheta(x) = \mathcal{R}(x) = c_R$, $\vartheta(x) = \mathcal{P}(x) = c_P$ and $\vartheta(x) = \mathcal{S}(x) = c_S$, which must be functions of the former ones, and *vice versa*,

$$\mathcal{S}(\mathrm{P}) = \mathcal{S}(\rho,\tau), \mathcal{P}(\mathrm{P}) = \mathcal{P}(\rho,\tau), \text{and } \tau = \tau(\mathcal{S},\mathcal{P}), \rho = \rho(\mathcal{S},\mathcal{P}) \tag{6.23}$$

being, ultimately, the known functions of the Dirac field $\psi(\mathrm{P})$. Potentially, one can obtain the functions τ and ρ purely algebraically,without even solving system (6.5) of PDEs. Every 2-d surface $S_{(12)}$ is fixed not only by the constants τ^* and ρ^*, but also, e.g., by c_S and c_P, which indicates that surface $S_{(12)}$ belongs to the principal manifold $\mathbb{M}$ without any reference to a coordinate $\mathbb{R}^4$. These observations are compli- mentary to the main idea of this work that Dirac field naturally determines the moving frame. Here, *the two scalars, e.g., $\mathcal{S}$ and $\mathcal{P}$, can replace the coordinates τ and ρ* (similarly to the hodograph transformation in hydrodynamics). From Equation (6.2) with tetrad index $A=0$ one can see that neither of the scalars $\mathcal{S},\mathcal{P},\mathcal{R}$ depends on the time variable τ (or S^0). Therefore, these quantities depend only on the radial variable ρ (or, equivalently, on the affine parameter $\sigma = S^3$).

1. Radial lines. When a geodesic line is given in the parametric form, $x^\mu = x^\mu(\sigma)$, the unit tangent vector is $e^\mu_{[3]} = \mathrm{d}x^\mu/\mathrm{d}\sigma$. The affine parameter of the radial geodesic lines is $\sigma = S^3$, but it differs from the parameter ρ of the hypersurfaces $\rho(x) = \rho^* = \mathrm{const}$, which determines distance (5.21) at some moment of the *world time* τ (5.12). In terms of the variable σ, the ODE for geodesic line with the tangent vector $e^\mu_{[3]}$ is

$$e^\mu_{[3]}\nabla_\mu e^\lambda_{[3]} = e^\lambda_A\omega^A_{33} = 0 = \frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}\nabla_\mu\left(\frac{\mathrm{d}x^\lambda}{\mathrm{d}\sigma}\right) = \frac{\mathrm{d}^2x^\lambda}{\mathrm{d}\sigma^2} + \Gamma^\lambda_{\nu\mu}\frac{\mathrm{d}x^\nu}{\mathrm{d}\sigma}\frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}, \tag{6.24}$$

where the connection $\Gamma^\lambda_{\nu\mu}$ is defined by Equation (3.14). The ODE for a geodesic line $x^\lambda(\rho)$ in terms of the physical variable ρ that can be obtained by means of a simple transformation, $\mathrm{d}x^\mu/\mathrm{d}\sigma = \mathcal{R}\,\mathrm{d}x^\mu/\mathrm{d}\rho$, and reads as

$$\frac{\mathrm{d}^2x^\lambda}{\mathrm{d}\sigma^2} + \Gamma^\lambda_{\nu\mu}\frac{\mathrm{d}x^\nu}{\mathrm{d}\sigma}\frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma} = -\frac{\mathrm{d}\ln\mathcal{R}}{\mathrm{d}\rho}\frac{\mathrm{d}x^\lambda}{\mathrm{d}\rho} = \frac{m\mathcal{P}}{\mathcal{R}^2}\frac{\mathrm{d}x^\lambda}{\mathrm{d}\rho} = \frac{m\mathcal{P}}{\mathcal{R}}e^\lambda_{[3]}, \tag{6.25}$$

where the r.h.s. does not contain derivatives of the Dirac field and it clearly manifests that the (not unit) tangent

vector $\mathrm{d}x^\lambda/\mathrm{d}\rho$ and its change are parallel along the "radial" geodesic curve.

2. The lines of the world time. The acceleration of the unit tangent vector of the lines of the vector current j^μ is

$$e^\mu_{[0]}\nabla_\mu e^\lambda_{[0]} = e^\lambda_A\omega^A_{00} = e^\lambda_{[3]}\omega^3_{00} = m\left(\mathcal{P}/\mathcal{R}\right)e^\lambda_{[3]}, \tag{6.26}$$

and it has only the radial component (precisely the same as radial geodesic (6.25)), which equals in magnitude but has opposite sign with respect to the mean curvature vector of surface $S_{(12)}$ and hypersurface $S_{(120)}$. The ODE for the trajectory $x^\lambda(\tau)$ reads as

$$\frac{\mathrm{d}^2x^\lambda}{\mathrm{d}\tau^2} + \Gamma^\lambda_{\nu\mu}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} = m\mathcal{P}\frac{\mathrm{d}x^\lambda}{\mathrm{d}\rho} = \frac{m\mathcal{P}}{\mathcal{R}}e^\lambda_{[3]}. \tag{6.27}$$

Obviously, the line of the vector current that passes through a point with the radial coordinate ρ^* never leaves the the surface $\rho = \rho^* = \text{const}$. Therefore, there is no flux of the charge density $\mathcal{R}$ in the outside direction, which is an indirect but indisputable evidence of localization.

3. The coordinate net over $S_{(12)}$. Finally, the lines of the Dirac currents Θ^μ and Φ^μ are also bound to the surface $\rho = \rho^* = \text{const}$. Indeed, for the curves $x^\lambda = x^\lambda\left(S^1\right)$ and $x^\lambda = x^\lambda\left(S^2\right)$ we have

$$\begin{aligned} e^\mu_{[1]}\nabla_\mu e^\lambda_{[1]} &= e^\lambda_A\omega^A_{11} = e^\lambda_{[3]}\omega^3_{11} + e^\lambda_2\omega^2_{11} = -m\left(\mathcal{P}/\mathcal{R}\right)e^\lambda_{[3]} - 2eA_{[1]}e^\lambda_{[[2]]}, \\ e^\mu_{[2]}\nabla_\mu e^\lambda_{[2]} &= e^\lambda_A\omega^A_{22} = e^\lambda_{[3]}\omega^3_{22} + e^\lambda_{[1]}\omega^1_{22} = -m\left(\mathcal{P}/\mathcal{R}\right)e^\lambda_{[3]} + 2eA_{[2]}e^\lambda_{[1]}, \end{aligned} \tag{6.28}$$

so that they have the same normal component of the mean curvature vector, and they are bent within surface $S_{(12)}$ even when the components $A_I = e^\mu_I A_\mu \neq 0$, $I = 1,2$.

To summarize, *all the currents passing in a tangent direction through a point on hypersurface $S_{(120)}$ of a given radius ρ^* never leave this surface*.

7. Conclusions

The (hyper)surfaces emerging from the Dirac equation and differential identities for the Dirac currents point to a fairly simple geometric structure of the lines and surfaces of the admissible coordinate net. These surfaces are built into the Dirac matter and completely determined by the latter. We will extensively refer to their properties in the second part [8] of this work. They will be used to write down the exact nonlinear Dirac equations and to find their analytic solutions, which represent a finite-sized stable particle. These solutions will necessarily be localized and have a spherical symmetry. This symmetry is not contemplated as a property of the ambient space. Within the framework of the matter-induced affine geometry, *the spherical symmetry is the property of a solution, and thus is a dynamic symmetry*.

A general discussion of the method, its results and perspectives is postponed till the last section of the Ref. [8].

References

[1] Makhlin, A. (2001) *Physical Review C*, **64**, Article ID: 064904. http://dx.doi.org/10.1103/PhysRevC.64.064904

[2] Sakharov, A.D. (1967) *JETP Letters*, **5**, 24-27.

[3] Dolgov, A.D. (2007) Cosmological Charge Asymmetry and Rare Processes in Particle Physics. *Les Rencontres de Physique de La Vallee d'Aoste*, 4-10 March 2007, Aosta Valley, Italy, 5 p. arXiv:0706.1229 [hep-ph]

[4] Dolgov, A.D. (2010) *Physics of Atomic Nuclei*, **73**, 588-592. http://dx.doi.org/10.1134/S1063778810040022

[5] Dolgov, A.D. (2015) Antimatter in the Universe and Laboratory. *The European Physical Journal Conferences*, **95**, Article ID: 03007.

[6] Serpico, P.D. (2012) *Astroparticle Physics*, **39-40**, 2-11. http://dx.doi.org/10.1016/j.astropartphys.2011.08.007

[7] Makhlin, A. (2010) Localization, CP-Symmetry and Neutrino Signals of the Dirac Matter. arxiv:1005.2693 [math-ph]

[8] Makhlin, A. (2016) *Journal of Modern Physics*, **7**, 662-679. http://dx.doi.org/10.4236/jmp.2016.77066

[9] Takabayasi, T. (1958) *Il Nuovo Cimento* (1955-1965), **7**, 118-121. http://dx.doi.org/10.1007/BF02746891

[10] Takahashi, Y. (1983) *Journal of Mathematical Physics*, **24**, 1783. http://dx.doi.org/10.1063/1.525896

[11] Takahashi, Y. (1982) *Physical Review D*, **26**, 2169. http://dx.doi.org/10.1103/PhysRevD.26.2169
Crawford, J.P. (1985) *Journal of Mathematical Physics*, **26**, 1439. http://dx.doi.org/10.1063/1.526906
[12] Israel, W. and Nester, J.M. (1981) *Physics Letters A*, **81**, 259. http://dx.doi.org/10.1016/0375-9601(81)90951-8
[13] Cartan, E. (1966) The Theory of Spinors. Hermann, Paris.
[14] Ne'eman, Y. (1978) *Annales de l'Institut Henri Poincaré, Section A*, **28**, 369.
Hehl, F.W, Lord, E.A. and Ne'eman, Y. (1978) *Physical Review D*, **17**, 428.
http://dx.doi.org/10.1103/PhysRevD.17.428
[15] Cartan, E. (2001) Riemannian Geometry in an Orthogonal Frame. World Scientific, Singapore.
[16] Fock, V. (1929) *Zeitschrift für Physik*, **57**, 261-277. http://dx.doi.org/10.1007/BF01339714
[17] Eisenhart, L.P. (1926) Riemannian Geometry. Princeton University Press, Princeton.
[18] Levi-Civita, T. (1926) The Absolute Differential Calculus. Blackie & Son Ltd., London and Glasgow. (Dover, 1977)
[19] Stoker, J.J. (1969) Differential Geometry. Wiley Interscience, Hoboken.
[20] O'Neill, B. (1966) Elementary Differential Geometry. Academic Press, Cambridge, Massachusetts.

Journal of Modern Physics, 2016, 7, 611-626
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jmp
http://dx.doi.org/10.4236/jmp.2016.77062

The Shell Model of the Universe: A Universe Generated from Multiple Big Bangs

Tower Chen*, Zeon Chen

Unit of Mathematical Sciences, College of Natural and Applied Sciences, University of Guam, UOG Station, Mangilao, Guam, USA
Email: *tower_c@Yahoo.com, zeon_chen@yahoo.com

Received 21 February 2016; accepted 25 April 2016; published 28 April 2016

Abstract

The Current Standard Model of the Universe asserts that the universe is generated from a single Big Bang event followed by inflation. There is no center to this universe, hence, no preferential reference frame to describe the motions of celestial objects. We propose a new, Shell Model of the Universe, which contends that the universe is created from multiple, concentric big bangs. Accordingly, that origin presents itself as a unique, preferential reference frame, which furnishes the simplest description of the motions of galaxies in the cosmos. This is similar in manner to how planetary motion is more straightforwardly described via a sun-centered Solar System rather than an earth-centered one. The appeal of the Shell Model of the Universe lies in its simplistic ability to resolve the paradox of quasars, explain the variability in Hubble's Constant, and solve the problematic accelerated expansion of the universe.

Keywords

Big Bang, Variability in Hubble's Constant, Paradox of Quasars, Problematic Accelerated Expansion of the Universe

1. Introduction

1.1. Hubble's Law and Constant

In 1929, Edwin Hubble discovered a linear relationship between the recession velocities of remote galaxies, v', and the distances to those galaxies, d. His findings suggested that our universe was expanding. The relation,

$$v' = H_0 d \tag{1}$$

*Corresponding author.

How to cite this paper: Chen, T. and Chen, Z. (2016) The Shell Model of the Universe: A Universe Generated from Multiple Big Bangs. *Journal of Modern Physics*, **7**, 611-626. http://dx.doi.org/10.4236/jmp.2016.77062

is known today as Hubble's Law, where Hubble's Constant, H_0, is a measure of the cosmic expansion rate. To obtain a definitive value for H_0, it is necessary to be able to precisely measure v' and d. Astronomers are capable of accurately calculating a celestial body's recession velocity by measuring the object's spectral redshift, z, and utilizing the relativistic Doppler effect relation,

$$v' = \frac{(z+1)^2 - 1}{(z+1)^2 + 1} c \tag{2}$$

where z is defined to be $\frac{\lambda - \lambda_0}{\lambda_0}$, λ_0 is defined to be the wavelength of a spectral band as measured in the rest frame, and λ is defined to be the wavelength of the corresponding spectral band as measured in a moving frame. Determining intergalactic distances has always presented itself as the more difficult challenge, especially when gauging comparatively large ones. However, recent advances in technology, the aid of the Hubble Space Telescope, and the development of new measuring techniques have allowed astronomers to measure distances out to 400 Mpc [1]. Capitalizing on the recent progression, the Hubble Space Telescope Key Project to Measure the Hubble Constant has obtained a value of $H_0 = 72 \pm 2 \pm 7 \text{ km} \cdot \text{sec}^{-1} \cdot \text{Mpc}^{-1}$ [2]. Presently, Hubble's Law stands as one of the fundamental pillars of modern day astrophysics. Roughly 70 years later, this observation is still helping to shape the current theories in astronomy.

1.2. Standard Candle

Spurred by curiosity and concepts from Hubble's Law, astronomers sought a means of probing into the early expansion history of the universe, which entailed studying relatively distant galaxies. The more remote an object is, the greater the amount of time required for the light to reach us, and thus the further one would be looking back into time. To conduct this probe into the past, it was necessary to find a means of measuring distances to extremely remote galaxies. This could easily be achieved by employing a "standard candle," which is defined to be any distinguishable class of astronomical objects of known intrinsic brightness that can be defined over a wide distance range [3]. After determining a stellar object's intrinsic brightness (or luminosity), L, its distance, d, can be straightforwardly calculated using the simple relation,

$$L = 4\pi d^2 f \tag{3}$$

where f is the apparent brightness of the celestial body. The quest to find such a "standard candle" was fulfilled with the suggestion to use the intensely studied type Ia supernovae.

1.3. Expectations

Astronomers jumped at the chance to employ this newly discovered astronomical tool, feverishly searching for distant type Ia supernovae to probe into the universe's expansion history. By studying remote members of this class of stellar objects, scientists expected to find the expansion rate of the universe to be decreasing over time. Their expectations were largely influenced by the standard model of the universe at that time, which stated that our mass-dominated universe arose following the Big Bang and inflation [3]. Under that model, as time progressed, the mass generated during the birth of the universe should have served to slow the expansion of the universe, a consequence of the attractive, decelerating gravitational effects. Astronomers were hoping to confirm this theory by finding Hubble's Constant of the past, H_p, which is related galaxies further away from us, to be greater than Hubble's Constant of the present, H_0, which is related to galaxies closer to us.

1.4. The Search and the Unexpected

After studying a number of distant type Ia supernovae, astronomers discovered these objects to be at greater distances, d_p, than expected given their recession velocities, v'_p. This implied that these objects had moved further over time than had been anticipated. After applying Hubble's Law to their analyses, they discovered Hubble's Constant of present to be greater than Hubble's Constant of the past, $H_p < H_0$, signifying an accelerated expansion of the universe over time! This led astronomers to revise the standard model of the universe by the addition of "dark energy," which presently dominates the decelerating effects of matter and serves to accelerate the expansion of the universe. Currently, the best fitting mathematical models used to simulate the universe show

mass energy density to be within a factor of two of "dark energy" density, implying that we are witness to a unique era where the universe is transitioning from matter dominance to "dark energy" dominance. It seems almost unreasonably coincidental that mankind just happened to be studying the expansion of the universe and that technology had advanced just enough to be able to conduct this investigation, during this special transformation period. Many prominent astrophysicists believe this fortuitous set of events signify that there is some fundamental physics that is missing from the Current Standard Model of the Universe [3].

1.5. Quasar Paradox

In addition to helping shape the Current Standard Model of the Universe, Hubble's Law also has implications to many other aspects of astronomy. In particular, this simple concept has affected how astronomers view the nature of quasars. Quasars, often referred to as quasi-stellar objects or QSO's for short, were first discovered in 1963[1]. Their most intriguing aspect lies in their enormously high redshifts, which by Hubble's Law implies that they are receding away from us at extremely high relative velocities. The exceedingly large recession velocities of quasars imply that they are at distances of 5 to 10 billion light years from the earth. Furthermore, the apparent brightness of a QSO at such enormous separations would imply an energy output of 100 times that of the entire Milky Way Galaxy generated by an object roughly the size of our Solar System! There is no simple explanation for these phenomena, and it is proposed that matter falling into very massive black holes is the mechanism whereby such enormous amounts of energy are energy.

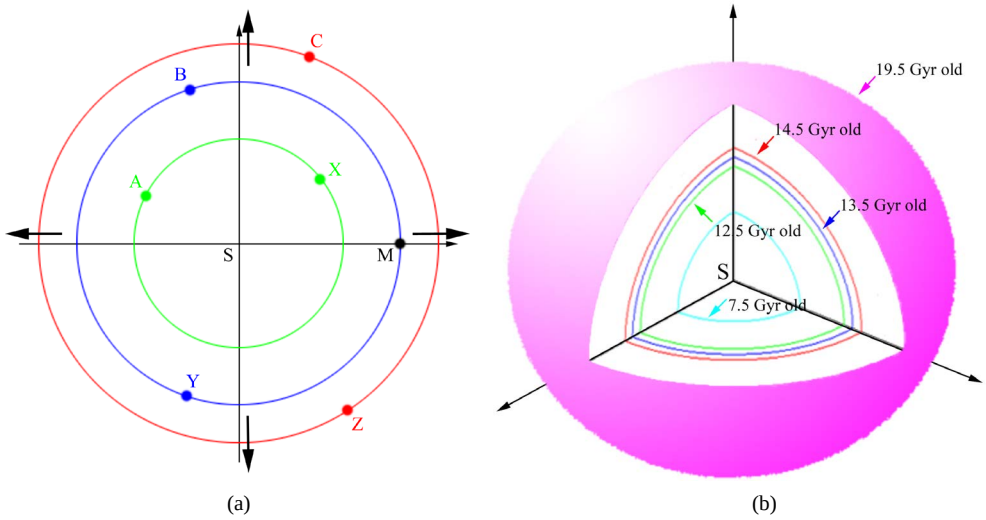
1.6. The Shell Model of the Universe

The variations in Hubble's Constant, the accelerated expansion of the universe, and the tremendous power output of quasars are all explained by the Current Standard Model of the Universe using rather complicated mathematical models. The Shell Model of the Universe was developed to provide a new, alternative framework for interpreting astronomical observations. This unconventional model asserts that the universe is comprised of numerous radially-expanding, concentric, galactic shells, each the result of a big bang. **Figure 1**, including **Figure 1(a)** and **Figure 1(b)**, shows that all galaxies residing on the same shell are at a corresponding age. The contention is that our universe is conglomerate of numerous, unequally-matured shells created from multiple big bangs with a common origin and not simply a center-less collection of masses generated by a single Big Bang per the Current Standard Model of the Universe. **Figure 1(a)** represents the cross-sectional view of the Shell Model of the Universe.

1.7. Valuable Astronomical Observations

In order for the framework provided by the Shell Model of the Universe to be relevant and useful for discussing astronomical phenomena, it must be based on real observations. Because no galaxies observed have exhibited blueshifts, only redshifts, the multiple big bangs must have a common center. One piece of data that is crucial to the construction of this model is the largest observed redshift of $z = 6.4$, which belongs to a quasar discovered by the Sloan Digital Sky Survey [4]. Based on this information and a couple of simple assumptions, the expansion velocity of the concentric, galactic shells can be calculated to be $0.762c$. Other applicable pieces of data emerge from observations that certain aged stars in our galaxy are approximately 12.5 billions years old and that the universe as a whole is roughly 14 billions years of age [5] [6]. This places the age of our galactic shell between 12.5 and 14 billion years. Note that a value of 13.5 billion years will be selected for discussion purposes. Because the universe was confirmed to be flat by evidence from the microwave background radiation [7], interstellar distances can be gauged using straight lines with no need for non-Euclidean geometry. The Shell Model that will be used to analyze the quasar paradox, the variability of Hubble's Constant, and the accelerated expansion of the universe will consist of five shells: 1) the shell containing our galaxy, which is at an age of 13.5 billion years, 2) a shell one billion years younger than ours at an age of 12.5 billion years, 3) a shell one billion years older than ours at an age of 14.5 billion years, 4) a shell six billion years younger than ours at an age of 7.5 billion years, and 5) a shell six billion years older than ours at an age of 19.5 billion years (see **Figure 1(b)**).

[1]The terms QSO's and quasars are used interchangeably here. It should be noted that some astronomers define quasars, *i.e.* quasi-stellar radio sources, to be the radio emitting subset of QSO's. Semantics should not be the focus.



(a) (b)

Figure 1. (a) This figure represents the cross-sectional view of a three-shelled model of the universe. All three shells are expanding radially outward at the same velocity. S designates the center of the universe, and Galaxy M represents our galaxy. Galaxies B and Y reside on the same shell as us and are at the same age as our galaxy. Galaxies A and X are located on an inner shell with respect to our galaxy, while Galaxies C and Z are on an outer shell. (b) This is a representation of the Shell Model of the Universe in 3 dimensions. This working model will consists of five shells: 1) the shell containing our galaxy, which is at an age of 13.5 billion years, 2) a shell one billion years younger than ours at an age of 12.5 billion years, 3) a shell one billion years older than ours at an age of 14.5 billion years, 4) a shell six billion years younger than ours at an age of 7.5 billion years, and 5) a shell six billion years older than ours at an age of 19.5 billion years, which are represented by blue, green, red, cyan, and magenta, respectively.

2. Constructing a Shell Model of the Universe

For simplicity, this model ignores the decelerating effects of gravitation and assumes that the five galactic shells all have the same, constant radial expansion velocity of $0.762c$.

2.1. Determining the Expansion Velocity of the Universe

This section will outline the methodology used to arrive at a value of $0.762c$ for the expansion velocity of the universe. As was mentioned earlier, the largest observed redshift of $z = 6.4$ [2] belongs to a quasar. By substituting that value into the aforementioned Doppler relation stated in (2), we can determine that the quasar is receding away from us at the tremendous relative velocity of $v' = 0.964c$. Imagine a typical, spherical balloon being inflated. Perceptively, different points on the balloon's surface will recede from the mouth of the balloon at unequal rates with the end located directly opposite the distention point retreating most rapidly. Of all celestial objects, quasars have the highest relative velocities with respect to us. The Shell Model of the Universe logically places them on the other half of the shell nearly completely opposite the Milky Way. Because all galactic shells are assumed to maintain the same expansion velocity, quasars could hypothetically reside on any of them, the crucial factor being that their movements are oriented in a direction completely opposite ours along the diagonal connecting our galaxy to the center (see **Figure 2**).

Utilizing this conception, the calculation of the expansion velocity of the galactic shells, *i.e.* the expansion velocity of the universe, becomes a straightforward task. **Figure 3**, including **Figure 3(a)** and **Figure 3(b)**, shows the movements of Milk Way Galaxy and quasars with respect the reference frame located at the center of the big bangs. This computation employs the relativistic velocity transformations, which states that if an inertial reference frame M moves with a velocity of u relative to a reference frame S and an object moves with a velocity of v in S, then this object moves with a velocity of v' with respect to M, where

$$v'_x = \frac{v_x - u}{1 - uv_x/c^2} \tag{4a}$$

[2] An even higher value may have been observed recently or maybe be observed in the future, but the conceptual basis for this calculation remains the same.

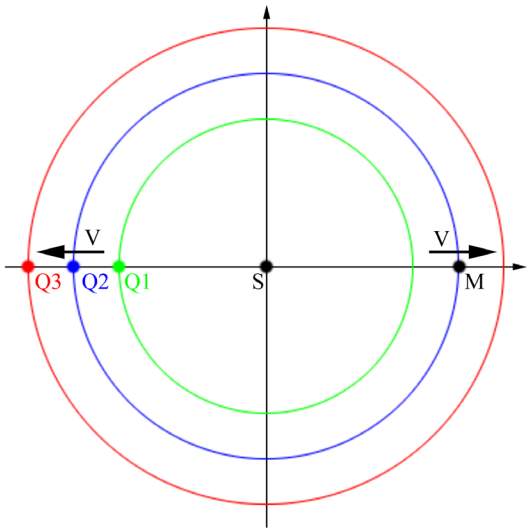


Figure 2. M designates our Milk Way Galaxy. Quasars (Q1, Q2, and Q3) can be found in one of three places. They can either reside on the same shell as our galaxy, on an inner shell relative to our galaxy, or on an outer shell relative to our galaxy. All shells are expanding radially outward at V. The quasars' locations on the shells are such that they are moving in a direction opposite to ours.

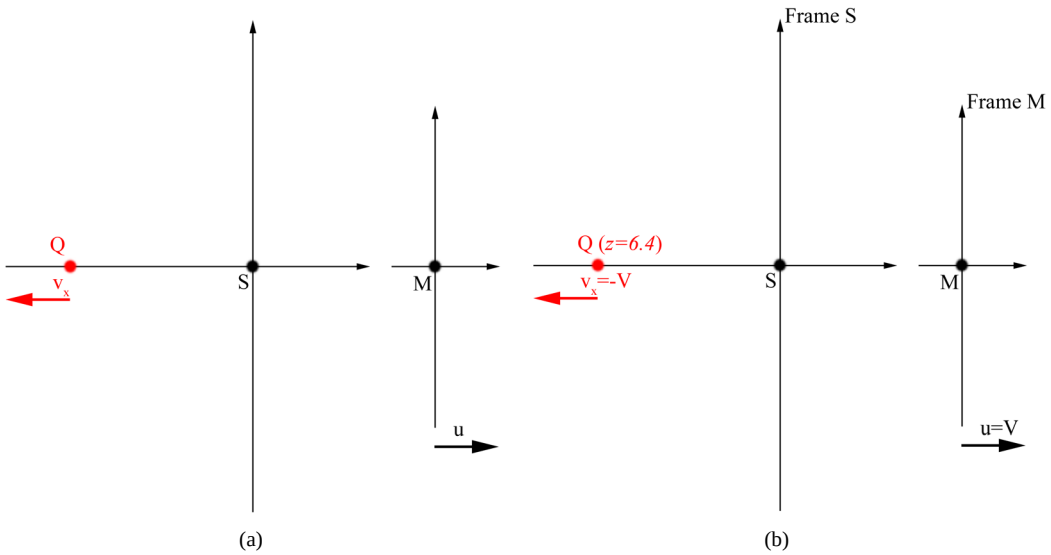


Figure 3. (a) S designates the center of the multiple big bangs, while M represents our Milky Way Galaxy. Frame M moves with velocity u with respect to S, and Q travels with velocity v_x with respect to Frame S. u is parallel to v_x. (b) Our galaxy M and quasar Q are both receding from the center of the universe M at the expansion velocity of the universe V. Because the highest observed redshift for a quasar is $z = 6.4$ discovered by the Sloan Digital Sky Survey, that quasar is assumed to be in an orientation directly opposite ours, *i.e.* $u = -v_x$.

$$v'_y = \frac{v_y\sqrt{1-u^2/c^2}}{1-uv_x/c^2} \tag{4b}$$

$$v'_z = \frac{v_z\sqrt{1-u^2/c^2}}{1-uv_x/c^2}. \tag{4c}$$

In these formulae, u is oriented parallel to v_x (see **Figure 3(a)**).

In this model, S will designate the reference frame located at the center of the big bangs. M will represent the Milky Way's (our) reference frame, which travels along the positive x-direction at the expansion velocity of $u = V$ with respect to frame S. Because the aforementioned quasar with the highest redshift has a z-value of 6.4 discovered by the Sloan Digital Sky Survey, we will place this object directly opposite our galaxy. As a consequence, the velocity vector of the quasar with respect to S can be decomposed as follows: $v_x = -V, v_y = 0, v_z = 0$ (see **Figure 3(b)**). Making the appropriate substitutions yields the resulting transformation:

$$v'_x = \frac{-2V}{1+V^2/c^2} \tag{5a}$$

$$v'_y = 0 \tag{5b}$$

$$v'_z = 0. \tag{5c}$$

Thus, for a relative velocity of $v' = v'_x = 0.964c$, the expansion velocity of the universe can be determined to be $V = 0.762c$. This value will be utilized throughout the remainder of this discussion.

2.2. Determining the Age of the Galactic Shells

In this section, we focus our attention on determining the age and the number of galactic shells, correspondingly, the time in between and abundance of big bangs. Hubble's Constant, H_0, has been estimated to be between $65\ \text{km}\cdot\text{sec}^{-1}\cdot\text{Mpc}^{-1}$ and $75\ \text{km}\cdot\text{sec}^{-1}\cdot\text{Mpc}^{-1}$, but for sake of simplicity a value of $70\ \text{km}\cdot\text{sec}^{-1}\cdot\text{Mpc}^{-1}$ will be used. The traditional units used to express Hubble's Constant can be converted to other units using dimensional analysis to yield an H_0 value of $7.15\times10^{-5}\ \text{Myr}^{-1}$. Hubble Time or H_0^{-1} is often associated with the age of the universe, and a Hubble's Constant value of $70\ \text{km}\cdot\text{sec}^{-1}\cdot\text{Mpc}^{-1}$ translates to an age of approximately 14 billion years. Indeed, other methods have confirmed the age of the universe to be more or less that value [5]. Additionally, the age of old stars in our galaxy have been estimated at 12.5 billion years [6], which places the age of our Milky Way Galaxy between 12.5 and 14 billion years. For discussion purposes, it would not be unreasonable to approximate the age of the shell containing our galaxy at 13.5 billion years. The other four shells in this working model of the universe were assigned ages of 7.5, 12.5, 14.5, and 19.5 billion years. These numbers have been chosen to provide a wide range of values for analysis. These figures and even the shell count will undoubtedly need to be refined to fit astronomical data, but as a first estimate, they will serve to provide valuable insight. With this, the foundation has been laid to begin construction of a working Shell Model of the Universe.

2.3. Constructing a Working Shell Model of the Universe

The model that will be employed later for discussion purposes asserts that the universe was created 19.5 billion years ago, when the first big bang generated the outermost of the five shells in our universe[3]. Accompanying the first big bang was the first inflation that sent the mass engendered flying radially outwards at very rapid speeds. As the effects of inflation subsided, the outermost shell settled at the current expansion velocity of $V = 0.762c$.

Five billion years after that first cataclysmic event, another big bang generated the second to the outermost shell. Following a period of inflation, that shell also settled at its current expansion velocity of $V = 0.762c$. This process was repeated again one billion years after the second big bang (six billion years after the first big bang) to produce the shell that would eventually give rise to our Milky Way Galaxy. The second to the innermost and the innermost shells were created by the same processes seven and twelve billion years, respectively, after the first big bang. Finally, 7.5 billion years after the most recent big bang, we arrive at the current state of our universe. All five shells have assumed a constant expansion velocity of $V = 0.762c$ following the relatively brief inflationary period that each separately underwent[4].

Figure 4 depicts the generation of this 5 shelled model of our universe up until its present state. We have made numerous references to the ages of the galactic shells. However, at first glance, there seems to be no inclusion of a dimension in our figures to account for time. Here we contend that, indeed, time can be depicted in a single diagram concurrently with the three dimensions of space. Our assertions are founded upon one guiding principle: movement is the most fundamental quantity/unit; only when we have movement do we have the concepts of time

[3]Again, it should be recognized that this is an incomplete model that should be subjected to further refinement. Nevertheless, as a working model, it provides ripe ground for discussion.

[4]Again, for simplicity, the effects of gravity are not factored in. They should be considered in more refined versions of this model.

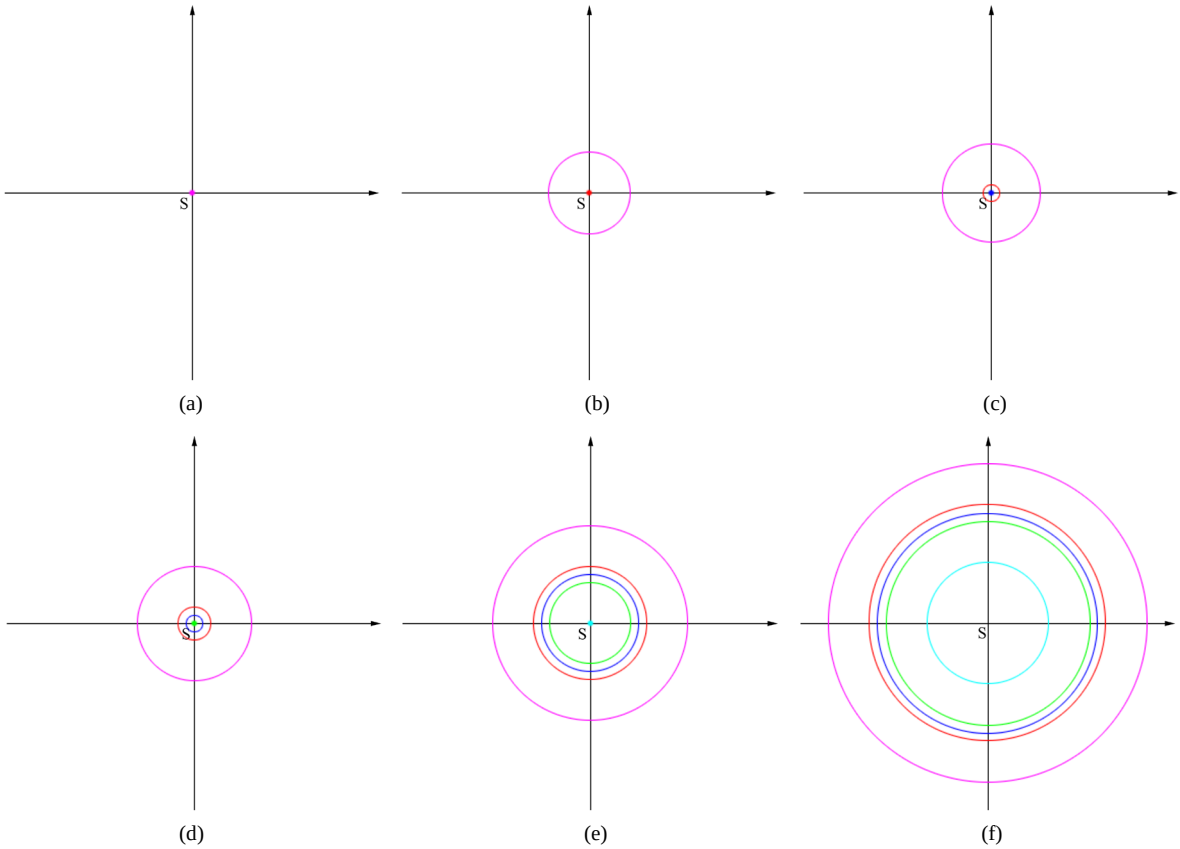


Figure 4. (a) The universe as described by the Shell Model of the Universe began when the first big bang generated the outermost shell (magenta). (b) After the outermost shell has expanded for 5 billion years at velocity V, the second big bang created the second the outermost shell (red). (c) After the outermost shell has expanded for an additional 1 billion years at velocity *V*, the third big bang created third shell (blue), which is also our shell. (d) After the outermost shell has receded from the center for a total of 7 billion years at velocity *V*, the fourth big bang conceived the fourth shell (green). (e) 12 billion years after the creation of the universe, the final big bang gave rise to the inner most shell (cyan). (f) 7.5 billion years after the conception of the final shell, our universe has arrived at its present state.

elapsing and space being occupied [8]. The first big bang generated a sizeable amount of energy in the form of light along with the cosmic mass it created. One of Einstein's postulates of Special Relativity states that the speed of light is constant. Bearing that in mind along with the simple relation: *distance* = *speed* × *time*, then the distance traveled by the light generated during the first big bang is, in essence, a timepiece recording the age of the universe. The expansion velocity of the shells in our model of the universe maintains a one-to-one correspondence with that light by sustaining a constant value of $V = 0.762c$ with respect to an observer at the center of the universe. Hence, the distance traveled by the expanding shells can also be used to record the passage of time. It should be recognized that none of the galactic shells have sustained that constant expansion velocity throughout their maturation. This is especially true during inflation, when the expansion speeds far exceed that value [5]. However, the duration of the inflationary periods is comparatively small to the timescale we are working with, which makes the approximation within reason.

3. Interpreting Astronomical Observations Using the Shell Model of the Universe

Now that a working Shell Model of the Universe has been constructed, we will proceed to examine its utility in interpreting astronomical observations.

3.1. Relative Velocity of Galaxy X as a Function of ∠XSM

In the proceeding discussion, we will establish a geometric relationship for determining a galaxy's relative ve-

locity with respect to us. It may be helpful to make references to **Figure 5** throughout this discussion. In this figure, our Milky Way Galaxy is designated by the letter M. Galaxy B lies on the same shell as us, while Galaxies A and C reside on a younger shell one and an older one, respectively. Additionally, Galaxies A, B, and C are all on the same radial vector, such that $\angle ASM = \angle BSM = \angle CSM = \theta$, and all assumed to expanding radially outward at the expansion rate of the universe, V. To calculate the relative velocity, v', of any galaxy X with respect to us, the relativistic velocity transformations are again employed and summarized as follows:

$$v'_x = \frac{v_x - u}{1 - uv_x/c^2} = \frac{(V\cos\theta) - V}{1 - V(V\cos\theta)/c^2} \tag{6a}$$

$$v'_y = \frac{v_y\sqrt{1-u^2/c^2}}{1 - uv_x/c^2} = \frac{(V\sin\theta)\sqrt{1-V^2/c^2}}{1 - V(V\cos\theta)/c^2} \tag{6b}$$

$$v'_z = \frac{v_z\sqrt{1-u^2/c^2}}{1 - uv_x/c^2} = 0 \tag{6c}$$

$$|v'| = \sqrt{v'^2_x + v'^2_y + v'^2_z} = \sqrt{v'^2_x + v'^2_y}. \tag{6d}$$

As a concrete example, a value of 30˚ will be substituted for θ along with the previously calculated value for the expansion velocity of $V = 0.762c$. From this, a value of $0.537c$ is obtained for v'. This means that under this given model, all galaxies that form a 30˚ angle with the center of the universe and the Milky Way will be receding from us at a relative velocity of $0.537c$, a value that can be measured from the redshift. Indeed, the relative velocity of any galaxy with respect to our reference frame, remains solely a function of the expansion velocity of the universe and the angle formed by the body, the origin of the big bangs, and our galaxy.

3.2. Location of Galaxy X, Where Light Currently Being Received Was Emitted

To determine the location of Galaxy X, where the light we are currently receiving was emitted from, three general cases must be considered. There is the first case of a galaxy residing on the same shell as ours, the second case

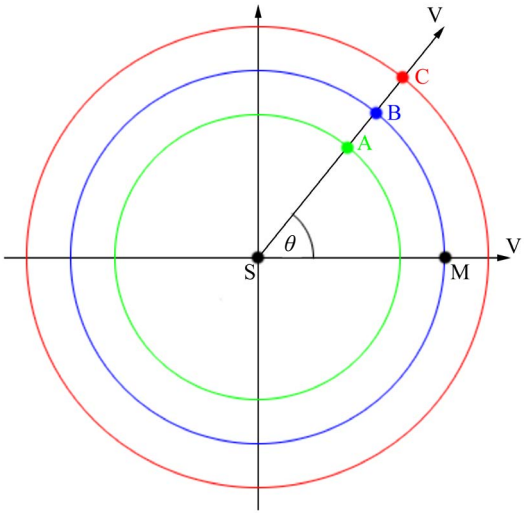


Figure 5. Our galaxy is designated by the letter M. Galaxy B resides on the same shell as us, while Galaxy A resides on an inner shell with respect to ours and Galaxy C resides on an outer shell with respect to ours. All galaxies are receding radially from S at the same expansion velocity V. $\angle ASM = \angle BSM = \angle CSM = \theta$. The relative velocities of A, B, and C with respect to M are all the same and are only functions of V and θ, where $|v'| = \sqrt{v_x^2 + v_y^2}$ and $v_x = \frac{(V\cos\theta) - V}{1 - V(V\cos\theta)/c^2}$ and $v_y = \frac{(V\sin\theta)\sqrt{1-V^2/c^2}}{1 - V(V\cos\theta)/c^2}$.

of a galaxy residing on an inner shell with respect to ours, and the third case of a galaxy residing on an outer shell with respect to ours.

3.2.1. Case 1: Galaxy B Resides on the Same Shell as Our Galaxy

This section will draw upon numerous references to **Figure 6**. In this figure, our Milky Way Galaxy is designated by the letter M and is currently at an age of t_0. Galaxy B, which is on the same shell as our galaxy, is also at an age of t_0. Light presently emitted by B has not yet reach us at M, as light itself has a finite speed and B and M are spacially separated. We are, however, receiving light that is t years old from B′, from a time when Galaxy B was only at an age of $t_0 - t$. Time is accounted for by placing B's at a radius of

$$r = V\left(t_0 - t\right) \tag{7}$$

and B and M at a radius of

$$R = Vt_0 \tag{8}$$

consistent with principles stating that movement is the most fundamental quantity/unit; only when we have movement, we have the concepts of elapsing and space being occupied, discussed earlier in Section 2.3.

When we take a gander up at the night sky and see Galaxy B, we are not looking at Galaxy B at its present location. Rather, we are looking at Galaxy B of the past and light that is t years old from a time when Galaxy B was still at B′. Thus, when astronomers measure the distance to Galaxy B, they are actually measuring the distance to B′ and not to Galaxy B′s current location, which is presently unobservable. Here we make the distinction that the measured astronomical distances are actually the distances traveled by the light before hitting our eyes or instrument (the light-traveled distances, d) and not the actual distance to the celestial object's current position. Computing d is straightforward, as it is simply a product of the speed of light and the time traveled by the light, *i.e.*

$$d = ct\,. \tag{9}$$

This concept is extremely important, because it is this the accurate measurement of this value that sparked the search for a standard candle. It was also the employment of this value in calculating Hubble's Constants of the past and present that led to the conclusion that our universe was undergoing an accelerated expansion.

Another important relation,

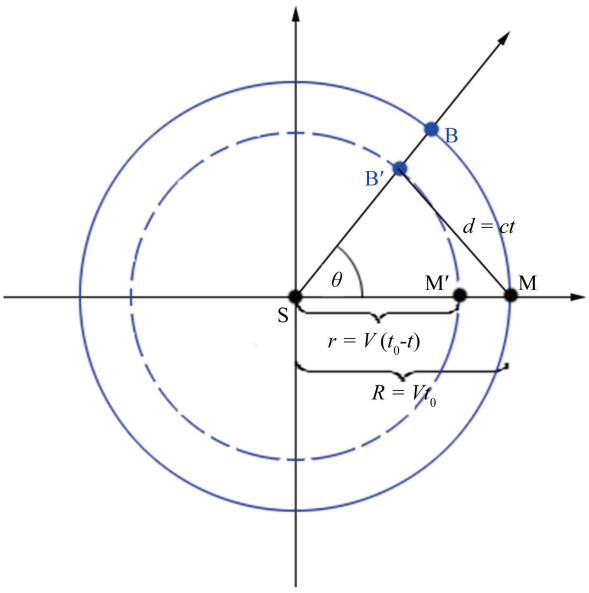


Figure 6. Our Milky Way Galaxy is designated by M, while S represents the center of the universe. Galaxy B resides on the same shell as us, which is at an age of t_0. M is currently receiving light that is t years old from a time, $t_0 - t$, when Galaxy B was still at B′. Time is accounted for by placing B′ at a radius of r and B and M at a radius of R. The distance from M to B′ is d.

$$\cos\theta = \frac{R^2 + r^2 - d^2}{2Rr} = \frac{\left(Vt_0\right)^2 + \left[V\left(t_0 - t\right)\right]^2 - \left(ct\right)^2}{2\left(Vt_0\right)\left[V\left(t_0 - t\right)\right]} \tag{10}$$

can be derived from this diagram using the law of cosines. Using this formula, we are able to calculate *t*, given the expansion velocity of the universe, the current age of our shell, and $\theta = \angle BSM = \angle B'SM$, which will correspond to a particular redshift. After obtaining a value for *t*, the Cartesian coordinates of B' can be computed using the following relations.

$$x = r\cos\theta \tag{11a}$$

$$y = r\sin\theta. \tag{11b}$$

As an example, if the age of our shell is $t_0 = 13.5$ Gyr, the expansion velocity of the universe is $V = 0.762c$, and the angle formed by the trajectory vectors of Galaxy M and B is $\theta = 30^\circ$, then the time it has taken for the light to reach our eyes from B′ is $t = 6.09$ Gyr from a distance of $d = 6.09$ Glyr. Hence, in our survey of the sky, Galaxy B appears to be at B′ $(x, y) = (4.89\text{ Glyr}, 2.82\text{ Glyr})$ from relation (11a-b) when in reality it has already moved to the position B $(x = R\cos\theta, y = R\sin\theta) = (8.91\text{ Glyr}, 5.15\text{ Glyr})$. In **Figure 7(a)**, the coordinates of B′ were determined for a variety of θ values and graphed as the blue line.

3.2.2. Case 2: Galaxy A Resides on an Inner Shell with Respect to Ours

It will be helpful to refer to **Figure 8** during this portion of the discussion. Our Milky Way Galaxy, designated by the letter M, presently resides in a shell at an age of t_0, while Galaxy A presently belongs to a younger, inner shell at an age of $t_0 - T$. Our galaxy, M, is presently receiving light emitted by Galaxy A, when it was at the point A′, while the shell containing the galaxy was at an age of $t_0 - T - t$. The radius to A′ is

$$r = V\left(t_0 - T - t\right), \tag{12}$$

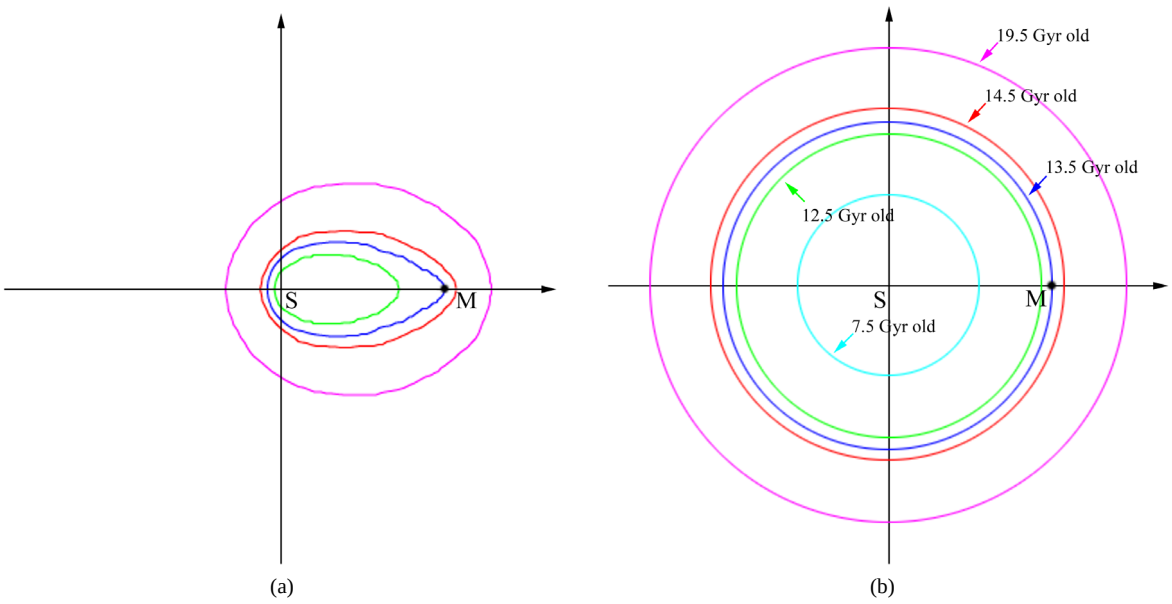


Figure 7. (a) This is a representation the Shell Model of our universe as currently seen from our galaxy, M, which resides in a shell of age t_0 = 13.5 Gyr. S is the center of our universe. The blue line represents our perspective of galaxies that share the same shell as ours; the green line, a shell 1 billion years younger than ours; the red line, a shell 1 billion years older than ours, and the magenta line, a shell 6 billion years older than ours. (b) This is representation of the 5-shelled model of our universe. his defers from (a) in that this is from the perspective of an omniscient observer who is able to instantaneously view the entirety of the universe. The blue line represents galaxies that share the same shell as ours, the green line, a shell 1 billion years younger than ours; the red line, a shell 1 billion years older than ours, and the magenta line, a shell 6 billion years older than ours. This figure also has a cyan line, which represents a shell 6 billion years younger than ours. It is not present in (a), because the light from galaxies in that shell has not reached us yet.

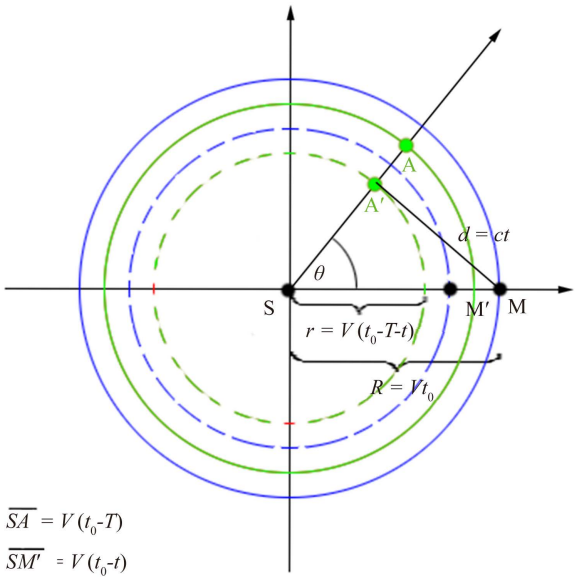


Figure 8. Our Milky Way Galaxy, which is designated by the letter M, is at an age of t_0. S represents the center of the universe. Galaxy A, which resides on a shell T years younger than ours, is at an age of $t_0 - T$. M is presently receiving light that is t years old when Galaxy A was still at A′. Time is accounted for by placing A′ at a radius of r and M at a radius of R. The distance measured to A′ from M is d.

and the radius of the shell in which M is located is given by the same relation stated in (8).

Because it has taken the light emitted when Galaxy A was at position A' a time of t to reach the observer at M, the light-traveled distance is $d = ct$. t can be calculated using a relation derived from the law of cosines:

$$\cos\theta = \frac{R^2 + r^2 - d^2}{2Rr} = \frac{(Vt_0)^2 + \left[V(t_0 - T - t)\right]^2 - (ct)^2}{2(Vt_0)\left[V(t_0 - T - t)\right]} \tag{13}$$

Reasonable values for V and t_0 have already been established, and θ can be correlated to the object's recession velocity. All that remains in solving for t is making reasonable estimates[5] for T. The coordinates of A' are given by (11a-b).

If $t_0 = 13.5$ Gyr, $V = 0.762c$, $\theta = 30^\circ$, and the inner shell that A resides on is $T = 1000$ Myr younger than our shell, then the time it has taken for the light to reach us from A′ is t = 6.96 Gyr from a distance of d = 6.96 Glyr. This places A′ at the coordinates $(4.89\text{ Glyr}, 2.82\text{ Glyr})$, although Galaxy A has long since moved to the position $\left(V(t_0 - T)\cos\theta, V(t_0 - T)\sin\theta\right) = (8.25\text{ Glyr}, 4.76\text{ Glyr})$. Similar calculations were performed for a number of θ values and plotted as the green line in **Figure 7(a)**. An identical set of computations were performed for $T = 6000$ Myr and an array of θ's. However, negative values were obtained for t, implying that no light from the 6 billion-year history of this younger shell has yet reached our galaxy.

3.2.3. Case 3: Galaxy C Resides on An Outer Shell with Respect to Ours

In this final case under consideration, it would be informative to refer to **Figure 9**. Again, our Milky Way Galaxy is designated by the letter M and presently resides on a shell at the age of t_0, while Galaxy C presently belongs to an older, outer shell at an age of $t_0 + T$. Our galaxy is currently receiving light that is t years old from C′, from a time when Galaxy C was at an age of $t_0 + T - t$. Time is denoted by placing C′ at a radius of

$$r = V(t_0 + T - t) \tag{14}$$

and M at a radius given by (8).

It has taken the light emitted by Galaxy C when it was at C′ a total of t years to reach an observer at M. Hence,

[5]Fitting values to actual data should provide better estimates for T.

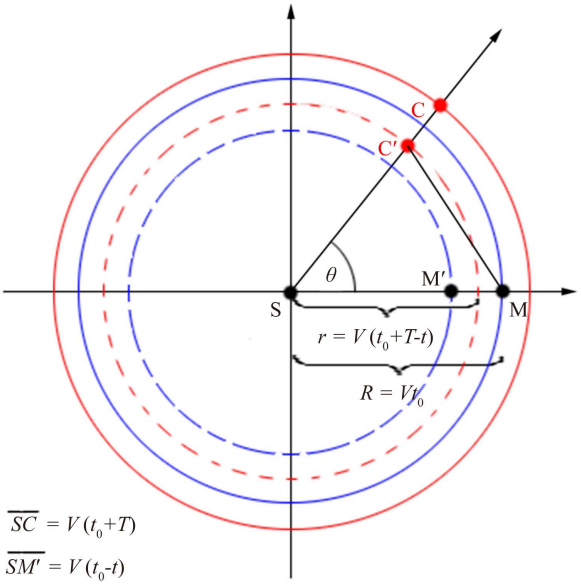


Figure 9. Our Milky Way Galaxy is represented by the letter M and is at an age of t_0. S is the center of the multiple big bangs. Galaxy C resides on a shell T years older than ours and is at an age of $t_0 + T$. M is currently receiving light that is t years old from a time when Galaxy C was still at C′. Time is designated by placing C′ at a radius r and M at a radius R. The distance to C′ from M is d.

the distance from M to C′ is, again, simply $d = ct$. Similar to the other two cases, a law of cosines relation:

$$\cos\theta = \frac{R^2 + r^2 - d^2}{2Rr} = \frac{(Vt_0)^2 + \left[V(t_0 + T - t)\right]^2 - (ct)^2}{2(Vt_0)\left[V(t_0 + T - t)\right]} \tag{15}$$

can be derived to solve for t. Again, values for the expansion velocity of the universe (V) and the age of our shell (t_0) have already been reasonably determined, all that remains to solve for t is inputting appropriate values for T for a broad range of θ's.

For example, if $t_0 = 13.5$ Gyr, $V = 0.762c$, $\theta = 30^\circ$, and the outer shell that C resides on is $T = 1000$ Myr older than our shell, then the time it has taken for the light to reach us from C' is $t = 5.55$ Gyr from a distance of $d = 5.55$ Glyr. Although Galaxy C is now currently at the position $\left(V(t_0 + T)\cos\theta, V(t_0 + T)\sin\theta\right) = (9.57\text{ Glyr}, 5.52\text{ Glyr})$, we perceive it to be at C' $(5.90\text{ Glyr}, 3.41\text{ Glyr})$. Similarly calculated values were obtained for a variety of θ's and diagramed as the red line in **Figure 7(a)**. The magenta line in this figure represents similar computations performed for an outer shell $T = 6000$ Myr older than ours.

Assuming that our universe can be modeled by a Shell Model of the Universe consisting of five shells: 1) the shell containing our galaxy, which is at an age of 13.5 billion years, 2) a shell one billion years younger than ours at an age of 12.5 billion years, 3) a shell one billion years older than ours at an age of 14.5 billion years, 4) a shell six billion years younger than ours at an age of 7.5 billion years, and 6) a shell six billion years younger than ours at an age of 19.5 billion years, **Figure 7(a)** represents the universe as presently seen by us, here in the Milky Way. This is to be contrasted with the actual state of the universe shown in **Figure 7(b)**, which is from the vantage point of an observer sitting at the center of the big bangs, taking a record of motions of objects in the universe. Notice that from our perspective, we are unaware of the existence of the innermost galactic shell, whose light has yet to reach our eyes.

3.3. Hubble's Law and the Ratio of $\frac{v'}{d}$

The previous sections provided us with the means of constructing a view of the universe based our limited, his-

torically-based perspective (see **Figure 7(a)**). Additionally, it furnished us with the tools necessary to take a closer look at Hubble's Law and the ratio of v'/d. Assuming the expansion velocity of the universe to be a constant $V = 0.762c$, it was shown in relation (6a-d) that the relative velocity v' of any galaxy with respect to our reference frame, remains solely function of the angle, θ, formed between the velocity vectors of that galaxy and ours. Thus, assuming no change in the expansion rate, the relative velocity of all galaxies with respect to us should remain the same, because the angle formed by the velocity vectors does not change.

Additionally, we derived methods for determining the light-traveled distance, $d = ct$, which is the distance that astronomers are interested in measuring using standard candles, by examining the following three cases: 1) the light source is from a galaxy that presently resides on the same shell as us, 2) the light source is from a galaxy that presently resides on an inner shell relative to ours, and 3) the light source is from a galaxy that presently resides on an outer shell relative to ours. To solve for *t*, we employ the law of cosines relations stated in Equations (10), (13) and (15) respectively, for the three cases.

Suppose Galaxies A, B, and C are all expanding along the same line, such that their trajectory vectors make a 30 degree angle, θ, with our velocity vector. With respect to our galactic shell, Galaxy A presently resides on an inner shell and Galaxy B resides on the same shell, while Galaxy C resides on an outer shell (see **Figure 5**). The age difference between our shell and the inner and outer shells is 1 billion years, *i.e.* $T = 1\,\text{Gyr}$. Using the previously calculated values for expansion velocity of the universe, $V = 0.762c$, and the age of our shell, t_0 = 13.5 Gyr, we can arrive at and compare the ratio of v'/d for the three cases. There are the first case of galaxy B residing on the same shell as our galaxy, the second case of galaxy A residing on an inner shell with respect ours, and galaxy C residing on an out shell with respect to ours. Because θ is the same for Galaxies A, B, and C, the recession velocities of these galaxies are all $v' = 0.537c$.

3.3.1. Case 1: Galaxy B, which Presently Resides on the Same Shell as Our Galaxy

Solving the first law of cosines Equation (10) for *t* and multiplying by *c*, we obtain a value of $d = 6.09\,\text{Glyr}$, which means that the source of light is 6.09 Glyr away. Dividing this galaxy's recession velocity of $v' = 0.536c$ by this *d* value, we obtain a ratio of

$$\frac{v'}{d} = \frac{8.80\times10^{-5}}{\text{Myr}} = 86.1\,\text{km/sec/Mpc} \tag{16}$$

for Galaxy B at B′.

3.3.2. Case 2: Galaxy A, Which Presently Resides on An Inner Shell Relative to Our Galaxy

By solving for *t* in (13) and multiplying by the speed of light, we calculate the light-traveled distance to be $d = 6.96\,\text{Glyr}$. This corresponds to

$$\frac{v'}{d} = \frac{7.70\times10^{-5}}{\text{Myr}} = 75.3\,\text{km/sec/Mpc} \tag{17}$$

for Galaxy A at A′.

3.3.3. Case 3: Galaxy C, Which Presently Resides on An Outer Shell Relative to Our Galaxy

Similarly, by deriving *t* from Equation (15) and multiplying by *c*, we arrive at a value of $d = 5.55\,\text{Glyr}$. Dividing the previously calculated value of $v' = 0.536c$ by *d*, we obtain a ratio of

$$\frac{v'}{d} = \frac{9.66\times10^{-5}}{\text{Myr}} = 94.5\,\text{km/sec/Mpc} \tag{18}$$

for Galaxy C at C′.

Because the recession velocity for all three of these galaxies is the same, the only factor in the variability of $\frac{v'}{d}$ is the distance, *d*. There is trend in the ratios of $\frac{v'}{d}$, *i.e.* "Hubble's Constant," with those galaxies furthest away yielding the smallest values and those at the closest proximity producing the largest numbers. These results imply that as we look further and further back into time, *i.e.* we look at more and more distant galaxies, "Hubble's

Constant" decreases. Thus, as the universe has aged, the ratio of $\frac{v'}{d}$ has increased. Under the Current Standard Model of the Universe, the resulting increasing progressing in "Hubble's Constant" with the passage of time implies an accelerated expansion of the universe. The ever elusive dark energy is purported to be the driving force behind this increase. However, throughout this discussion and the history of the Shell Model of the Universe, it has been assumed that the expansion velocity of the universe has remained a constant, $V = 0.762c$. Thus, the variation in the ratio of $\frac{v'}{d}$ can be attributed to the model under which the astronomical phenomena are interpreted. Dark energy has no place in this new model.

The appeal to this new Shell Model of the Universe over the Current Standard Model of the Universe lies in its simplicity and its ability to straightforwardly address the quasar paradox, the variation in "Hubble's Constant," and the purported accelerated expansion of the universe. Consideration of this new model would seriously call to question not only the current model but more fundamentally, Hubble's Law. Just from the example above, we see that there is no one-to-one correspondence between v' and d, as stated by "Hubble's Relation": $v' = H_0 d$. Undoubtedly, this new Shell Model of the Universe requires refinement. However, as a rough first model, it provides invaluable insight and seriously challenges current prevailing theories. For instance, the enormously large redshifts of quasars are a result of their orientation with respect to us and are not at the extensive distances implied by "Hubble's Law." Because those phenomena are not as far away as previously imagined, there is no enormous energy output to explain by matter falling into black holes.

4. Predictions of the Shell Model of the Universe

A useful model should not only be able to explain current phenomena but should also be able to make predictions.

4.1. Youngest, Hypothetical, Visible Shell

Previously, in **Figure 7(a)**, it was shown that we are not able to see galaxies on an inner shell that is 6 billion years younger than ours. We cannot calculate a ratio for $\frac{v'}{d}$ from an inner shell at an age 7.5 billion years, because we cannot gather data from things we cannot see. In the subsequent discussion, we will calculate where that cutoff is, *i.e.* what is the youngest a shell can be for us to be able to see it. Looking at **Figure 8**, two properties should be satisfied by this hypothetical inner shell: 1) the light that is just reaching us shows the shell right as it is being born, *i.e.*

$$r = V\left(t_0 - T - t\right) = 0 \tag{19}$$

and 2) that light has traveled the distance that we are from the origin of the multiple big bangs, *i.e.*

$$R = d \tag{20}$$

From equations (8) and (9), we saw that $R = Vt_0$ and $d = ct$, respectively. Hence, by substitution, we find that

$$T = \left(\frac{c - V}{V}\right) t_0 . \tag{21}$$

In our current working model $V = 0.762c$ and $t_0 = 13.5\,\text{Gyr}$, therefore, we find that $T = 3.213\,\text{Gyr}$. Thus, we are only able to see a hypothetical shell that is younger than ours by a maximum of 3.2 billion years. The unobservable mass in younger shells could explain some of the lost mass necessary to reach the critical density for the Big Crunch. Likewise, if the galaxies in outer shells of our universe have died a very long time ago, we may also be unable to see them, because the last of the light they emitted has already passed us by, but they will still contributors, adding to the density required for the Big Crunch. Furthermore, additional mass could potentially be generated with more and more big bangs.

4.2. Maximal Redshift Value That Cannot Be Exceeded by Objects in Line of Sight

The final prediction that will be discussed will utilize **Figure 10**. The line connecting our galaxy M to the origin of the big bangs will be designated as the x-axis. Our mathematical model predicts that for any angle γ, we may

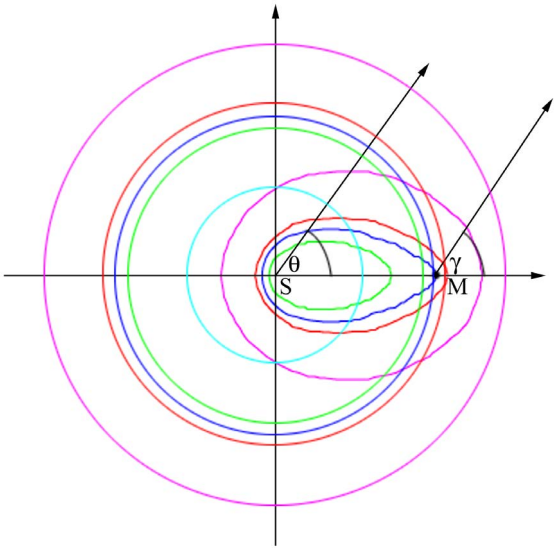


Figure 10. The redshift, *z*, of any galaxy is a function of the expansion velocity of the universe, *V*, and the angle, θ. Assuming *V* to be constant, the redshift becomes solely a function of θ. If an observer looks to the cosmos at a particular line of sight, γ, they should never see a galaxy with a redshift higher than $z_\gamma = \sqrt{\frac{1+v'/c}{1-v'/c}} - 1$, where $|v'| = \sqrt{v_x^2 + v_y^2}$ and $v_x = \frac{(V\cos\theta) - V}{1 - V(V\cos\theta)/c^2}$ and $v_y = \frac{(V\sin\theta)\sqrt{1 - V^2/c^2}}{1 - V(V\cos\theta)/c^2}$. This is because the ray that forms the angle γ never intersects the ray that forms θ.

observe galaxies with redshifts less that or equal to z_γ but no larger than z_γ, where

$$z_\gamma = \sqrt{\frac{1+v'/c}{1-v'/c}} - 1 \tag{22}$$

and v' is given by (6a-d). It was proved earlier that recession velocity and analogously, redshifts, are only functions of the expansion velocity *V* and θ. Thus, with a value for *V* determined, this relation is true, because the upper ray of γ never intersects with the upper ray of θ. This ray only intersects with the rays of angles less than θ, which corresponds to redshift values less than z_γ. This means that if you observe the cosmos at a particular line of sight, γ, you should never see a galaxy with a redshift higher than z_γ. On a similar note under this Shell Model of the Universe, high redshift objects, *i.e.* quasars, should all be clustered around the same patch of sky, at $\gamma \approx 180^\circ$.

5. Conclusion

This new Shell Model of the Universe has been constructed to provide simpler explanations to astronomical phenomena. This model has parsimoniously addressed the quasar paradox, the variability of "Hubble's Constant," and the purported accelerated expansion of the universe, something which the Current Standard Model of the Universe has had limited success with. In science, Ockham's Razor reigns supreme.

References

[1] Freedman, W., *et al*. (2001) *Astrophysical Journal*, **553**, 47. http://dx.doi.org/10.1086/320638

[2] Freedman, W. and Turner, M. (2003) *Reviews of Modern Physics*, **75**, 1433. http://dx.doi.org/10.1103/RevModPhys.75.1433

[3] Perlmutter, S. (2003) *Physics Today*, **2003**, 53.

[4] Ruderman, G. (2005) Three Distant Quasars Found at Edge of Universe. Sloan Digital Sky Survey. http://www.sdss.org/news/releases/20030109.quasar.html

[5] Britt, R. (2005) Astounding Findings' Pin down Age of Universe, Birth of First Stars. http://www.space.com/scienceastronomy/map_discovery_030211.html

[6] Sneden, C. (2001) *Nature*, **409**, 673-375. http://dx.doi.org/10.1038/35055646

[7] de Bernardis, P., *et al.* (2000) *Nature*, **404**, 955-959. http://dx.doi.org/10.1038/35010035

[8] Chen, T. and Chen, Z. (2011) The Paper "Advantages of Three-Dimensional Space-Time Frames". Journal of Frontiers in Science, Sciences Academic Publisher 2011.

Journal of Modern Physics, 2016, 7, 627-641
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jmp
http://dx.doi.org/10.4236/jmp.2016.77063

Scientific Research Publishing

Zeeman-Like Topologies in Special and General Theory of Relativity

Ravindra Saraykar[1], Sujatha Janardhan[2]

[1]Department of Mathematics, R T M Nagpur University, Nagpur, India
[2]Department of Mathematics, St. Francis De Sales College, Nagpur, India
Email: ravindra.saraykar@gmail.com, sujata_jana@yahoo.com

Received 18 January 2016; accepted 25 April 2016; published 28 April 2016

Abstract

This is a short review article in which we discuss and summarize the works of various researchers over past four decades on Zeeman topology and Zeeman-like topologies, which occur in special and general theory of relativity. We also discuss various properties and inter-relationship of these topologies.

Keywords

Zeeman Topology, Fine Topologies on Minkowski Space, Zeeman-Like Topologies in General Relativity, Homeomorphism Group, Lorentz Group, Conformal Group, Topological Properties

1. Introduction

In special as well as general theory of relativity, space-time models are usually taken as differentiable manifolds. The main reason for representing a space-time as a topological space which is also a differentiable manifold is that we need space-time to have a well-defined topological dimension and we can talk about curves and their tangent vectors, and neighbourhoods to develop a causal theory of space-time. This is achieved by assuming a pseudo-metric structure on a space-time manifold which enables us to define time-like, null and space-like vectors and corresponding curves. In general theory of relativity, metric also determines the geometry and curvature of space-time which represents the gravitational field. In special theory of relativity, Minkowski space M is usually given the topology of real 4-dimensional Euclidean space.

According to Zeeman [1], this topology is not physically reasonable for two reasons: first, the 4-dimensional Euclidean topology is locally homogeneous, whereas Minkowski space M is not, because to every point in M, there is an associated light cone which separates space-like vectors from time-like vectors. Secondly, the group of all homeomorphisms of 4-dimensional Euclidean space is vast and is of no physical significance. So, he proposed a new topology for Minkowski space, which is now well-known as Zeeman topology. This is defined

How to cite this paper: Saraykar, R. and Janardhan, S. (2016) Zeeman-Like Topologies in Special and General Theory of Relativity. *Journal of Modern Physics*, **7**, 627-641. http://dx.doi.org/10.4236/jmp.2016.77063

as the finest topology on *M* which induces 3-dimensional Euclidean topology on every space axis and 1-dimensional Euclidean topology on every time axis. Zeeman proved that this topology has the following physically reasonable properties: Firstly, this topology is not locally homogeneous, and light cone through any point can be derived from the topology. Secondly, the group of all homeomorphisms of this topology is generated by the inhomogeneous Lorentz group and dilatations.

Zeeman also proved that the topology on a light ray induced from this fine topology is discrete. This means that every function on the light cone is continuous, as every function will be continuous if the domain space has discrete topology. In quantum field theory also, we face similar difficulties regarding "real" space-time topology, where we talk frequently about continuous wave functions and fields, but we really do not know the meaning of that because "real" topology of space-time is unknown. However, studies have been dedicated to the topological properties of function spaces, such as spaces of quantum fields, but the study of proper space-time topology which is the most important space of all Physics, remains incomplete. Here, we need a topology in which known quantum quantities such as classical paths on which integrations are to be performed in Feynman's formalism, or Green's functions are continuous. We note that if a function is continuous on a space with topology *T*, it will be continuous in any refinement of *T*.

We also note that Zeeman topology is a refinement of $E^4 = M$ with 4-dimensional Euclidean topology, but a function which is continuous in Zeeman topology could be discontinuous in the Euclidean topology. Thus, the procedures by which physical quantities such as Green's functions and S-matrix elements, defined on space-time, are transformed by, say, analytic continuation into analogous quantities on E^4, will put constraints on possible topologies on space-time.

On mathematical side, *M* with Zeeman topology is not a normal topological space , as proved by Dossena [2] and hence it can not be a differentiable manifold, since, by definition, a differentiable manifold is Hausdorff and paracompact as a topological space, and hence normal.

After Zeeman published his paper in 1967, it attracted attention of some of the relativists cum mathematicians and they proved a number of results which are refinements over Zeeman's work. Modified results about Zeeman- and Zeeman-like topologies were published in the context of both special as well as general theory of relativity. Most remarkable are the results by S. Nanda [3]-[5], G. Williams [6], R. Göbel [7] [8], Hawking-King-McCarty [9], Malament [10] and Lindstrom [11] proved in 1970's. S.G. Popvassilev [12] generalized some of these results to R^n. Around 2005, researchers started gaining renewed interest in this field, and further interesting results were published by D.H. Kim [13], G. Dossena [2], G. Agrawal and S. Shrivastava [14] [15], G. Agrawal and Soami P. Sinha [16] and R. Low [17]. In fact R. Low extended the results of G. Agrawal and S. Shrivastava [14] to any dimension and also to general curved space-times. He used simpler arguments which do not require the use of Zeno sequences. We reproduce this proof for the sake of completeness.

Since Zeeman topology and other fine topologies defined in special and general theory of relativity in above works have many interesting properties, we discuss these properties and also discuss inter-relationships among these topologies. Most important and remarkable of these results are the results proved by R. Göbel and G. Dossena. Göbel proved that the group of all homeomorphisms of a space-time of general relativity with Zeeman-like topology is the group of all homothetic transformations. And Dossena proved that the first homotopy group of Zeeman topology for Minkowski space is non-trivial and contains uncountably many subgroups isomorphic to Z. In particular, this topology is not simply connected. Lindstrom generalized the results of Göbel and gave a sequence of Zeeman-like topologies which are in the ascending order of fineness.Thus, in Section 2, we describe Zeeman topology and other fine topologies on Minkowski space and discuss their properties. We also discuss t-topology, s-topology and A-topology introduced by Nanda [3]-[5] and studied in details by G. Agrawal and S. Shrivastava [14] [15]. In Section 3, we describe path topology of Hawking-King-McCarty (HKM topology), and improvements by Malament [10], Fullwood [18] and D.H. Kim [13]. We also discuss properties of HKM topology proved recently by R. Low. In Section 4, we describe the work of Göbel on Zeeman-like topologies defined on space-time of general relativity and discuss the results proved by him. We also remark on the work of other researchers, especially that by Lindstrom [11] and Mashford [19].

2. Zeeman- and Zeeman-Like Topologies on Minkowski Space

2.1. Zeeman Topology

We begin this section with definition of *Zeeman topology* as given in Dossena [2]. Let *M* denote 4-dimensional

Minkowski space-time and M_0 denote the associated 4-dimensional real vector space equipped with a non-degenerate symmetric bilinear form g of signature (−, +, +, +). In M_0, vector axes are either space-like hyperplanes passing through the origin or straight time-like lines passing through the origin. We denote by $\mathcal{A}_0$ the set of vector axes, and the set $p + A_0$ with $p \in M$ and $A_0 \in \mathcal{A}_0$, is an axis. We denote the set of axes by $\mathcal{A}$. The Zeeman topology, denoted by *Z*, is by definition, the finest topology on *M* with the property that it induces the affine space natural topology on every axis. *M* endowed with *Z* is denoted by M^Z.

A set *U* is open in M^Z if and only if for every $A \in \mathcal{A}, U \cap A$ is open in A^E. Here, A^E is the set *A* with natural topology *i.e*. Euclidean topology. As proved in Zeeman [1] and Dossena [2], the homeomorphism group of M^Z is generated by the Lorentz group, translations and dilatations.We denote this group by *G*.

Physically speaking, the Zeeman topology M^Z is defined as the finest topology on a space-time such that its induced topology on world lines of freely falling test particles with positive rest mass, and on space-like hypersurfaces, is locally Euclidean. Zeeman topology is not as nice as manifold topology, e.g. it is not a normal topological space. On the other hand it has many physically interesting properties: The Zeeman topology does not provide any geometric information along a light ray. Mathematically the topology induced by the Zeeman topology on a light cone is discrete. Secondly, there are many unphysical world lines, e.g. bad trips (cf Penrose [20]).

2.2. t-Topology, s-Topology and A-Topology

If we interpret continuity of a world line with respect to Zeeman topology, *world lines* are automatically physically realistic, namely, piecewise geodesics which are future directed and time-like with finitely many edges. Hence a world line is the orbit of a freely falling test particle within the gravitational field with a finite number of collisions. This result is a well known basic assumption for a kinetic theory in general relativity (cf Ehlers [21]).

Moreover if we allow the Zeeman topology to depend on a gravitational field as well as on the Maxwell field, it is possible to derive the corresponding result for charged particles as we discuss below.

In addition to above discussion, we also note that the group of all homeomorphisms of a space-time with its manifold topology is neither of interest for physics nor for mathematics since it is vast and it reflects no information of space-time. However, the group of all homeomorphisms of a space-time *M* with respect to its Zeeman topology M^Z coincides with its group of all homothetic transformations, *i.e*. homeomorphisms are isometries or isometries upto a constant factor. Thus homeomorphisms are proper symmetry transformations of the space-time. As proved in Zeeman [1], for a Minkowski space, the homothetic transformations are Lorentz transformations or dilatations of Minkowski space. Hence the homeomorphism group of Minkowski space under Zeeman topology is its Weyl group, which is generated by Lorentz transformations and linear dilatations.

After Zeeman published his paper in 1967, the first paper by other researcher on this topic was that of S. Nanda [3] in 1971 followed by another one in 1972 [4]. Nanda [3] proved one of the Zeeman's conjecture that the group of homeomorphisms of the finest topology on Minkowski space which induces three dimensional Euclidean topology on every space-like plane is the group *G*. To prove this conjecture, Nanda, like Zeeman, studied chronology preserving and causality preserving mappings and used the notion of Zeno sequences. He defines two topologies, space topology and s-topology with a fine distinction that space topology is strictly finer than s-topology. We recall definitions of these topologies as it would facilitate us to understand other work on fine topologies and compare it with the work of Nanda and Zeeman. As noted above, the space topology on M is defined as the finest topology with respect to which the induced topology on every space-like hyperplane is Euclidean. Let M^S and M^E denote Minkowski space *M* equipped with space topology and Euclidean topology. Then space topology is finer than Euclidean topology and hence Hausdorff.

Let $Q(x) = x_0^2 - x_1^2 - x_2^2 - x_3^2$ where $x = (x_0, x_1, x_2, x_3) \in M$. Then $Q(x)$ denotes the Minkowski quadratic form. We denote by $C^N(x)$, $C^T(x)$ and $C^S(x)$ the following cones at *x*:

Light cone or null cone at *x* : $C^N(x) = \{y / Q(y - x) = 0\}$,

Time-like cone at $x : C^T(x) = \{y / y = x \text{ or } Q(y - x) > 0\}$,

Space-like cone at $x : C^S(x) = \{y / y = x \text{ or } Q(y - x) < 0\}$.

Let $C^{NT}(x) = C^N(x) \cup C^T(x)$.

Furthermore, let $N_\epsilon^E(x)$ denote Euclidean ϵ-neighbourhood of *x* given by $N_\epsilon^E(x) = \{y / \rho(x, y) < \epsilon\}$, ρ

being the Euclidean metric, and let

$$N_\epsilon^s(x) = N_\epsilon^E(x) \cap C^S(x).$$

Then the topology generated by the family $\{N_\epsilon^s(x) / \epsilon > 0\}$ of local neighbourhoods at x which induces three dimensional Euclidean topology on every space-like hyperplane is s-topology as defined by Nanda, and we denote Minkowski space with this topology by M^s. Then M^S is strictly finer than M^s. After proving a series of lemmas about chronology preserving homeomorphisms, Nanda [3] proves that the group of homeomorphisms of M^S is G. In the subsequent paper, Nanda [4] defines t-topology in a similar way:

Let $N_\epsilon^t(x) = N_\epsilon^E(x) \cap C^T(x)$.

Then t-topology is defined as the topology which has the family $\{N_\epsilon^t(x) / \epsilon > 0\}$ as a local base of neighbourhoods at each point x of M. M equipped with this topology is denoted by M^t. In [4], Nanda proves another version of Zeeman's conjecture, namely that the group of homeomorphisms of M^t is the group G (Theorem 1 [4]). Furthermore, he also proves that the group of homeomorphisms of M^s is also group G (Theorem 2 [4]). If M and M' are Minkowski spaces, (or space-times of general relativity) then the mapping $f : M \to M'$ with the property that both f and f^{-1} preserve chronological order is known in the literature as chronal isomorphism (cf. P.S. Joshi [22]) Similarly, if both f and f^{-1} preserve causal order, then f is called causal isomorphism or simply a causal map. Such maps are extensively studied in the literature as cone preserving mappings (see for example, Garcia-Parrado and Senovilla [23] and S. Janardhan and R.V. Saraykar [24], and references therein).

Williams [6] studies other Zeeman-like topologies on the Minkowski space and derives homeomorphism groups for these topologies. We summarize below the results proved by Williams. It is interesting to note that C^1 subgroup of homeomorphism group of some of these fine topologies is the same as G.

Here M^E is M with natural topology as above and so are $C^N(x)$, $C^T(x)$ and $C^S(x)$. Let M^{F_i} denote the set of finest topologies such that the restrictions of the identity mapping of M^E onto each M^{F_i} to time-like and space-like lines are homeomorphisms. Williams proves that there is a unique such finest topology. M with this topology is denoted by M^F. The fine topology M^F is defined as follows :

Topology M^F is the topology on M generated by the local base of open neighbourhoods $\{N_\epsilon^F(x), x \in M, \epsilon > 0\}$ at x. Here $N_\epsilon^E(x)$ is defined as

$$N_\epsilon^F(x) = N_\epsilon^E(x) \cap C^{ST}(x) \quad \text{where} \quad C^{ST}(x) = C^S(x) \cup C^T(x).$$

He further proves that the group of homeomorphisms of M^F is the conformal group of Minkowski space. This is in fact the group generated by the Lorentz group, translations and dilatations, and thus, it is the same as G.

Williams further describes two more fine topologies for M and describes their homeomorphism groups. The first of these topologies is M^T. A physically significant topology for M is the finest topology such that the restrictions of the identity mapping of M^E onto M^{F_i} to time-like lines are homeomorphisms. In this topology the relative topology along space-like lines is discrete. M^T is Minkowski space with this fine topology. Group of C^1-homeomorphisms of M^T is the conformal group which is again same as G.

Following the argument in Nanda [3], though it can be proved that M^T is strictly finer than t-topology, homeomorphism groups of both these topologies are the same and their topological properties are also similar.

Second of these topologies is M^L. M^L is the unique finest topology such that the restrictions of the identity mapping of M^E onto M^L to straight lines are homeomorphisms. Also, there exists a unique such finest topology and that it is strictly finer than M^E. It is weaker than the two previous topologies discussed here. The line sequence introduced here is however a Zeno sequence and any homeomorphic image of I must be piecewise linear. (Here I is the closed unit interval.) Thus the group of C^1 homeomorphisms of M^L does preserve straight lines and is thus a subgroup of the projective group on R^4. In fact, group of C^1-homeomorphisms of M^L is the projective group which is generated by full linear group and translations. Thus it coincides with homeomorphism group of M^E. This work resembles the work of S. Nanda [3]-[5]. In fact, in the third paper [5], Nanda defines yet another fine topology on the Minkowski space, called A-topology and derives its homeomorphism group. He also compares his results with those of Williams. The A-topology is defined as follows:

Definition 2.1. A-topology: The A-topology on M is defined to be the finest topology on M with respect to which the induced topology on every time-like line and light-like line is one-dimensional Euclidean and the induced topology on every space-like hyperplane is three-dimensional Euclidean.

Thus A-topology is strictly finer than the Euclidean topology.

2.3. Williams M^F Topology and Other Topologies

The topology M^F suggested by Williams on Minkowski space is characterized by the property that the induced topology on time-like and space-like lines is Euclidean and that it is the finest such topology on M having this property. This topology differs significantly from the A-topology (or from Zeeman's fine topology) in its group of homeomorphisms. Williams has proved that the C^1-subgroup of homeomorphisms of this topology is G. Without the C^1-condition, the result may not be valid. Nanda proves, by using Zeno sequence method, that the group of homeomorphisms of A-topology is also same as G. Furthermore, as remarked by Nanda [5], if $f: I \to M^F$ is a continuous map, then $f(I)$ is a connected union of time-like and (or) space-like intervals. This is in contrast with the result for A-topology where $f(I)$ is a connected union of finite number of time-like and (or) null intervals. If, however, f is assumed to be order-preserving, then it follows that $f(I)$ is a connected union of time-like intervals representing the path of an inertial particle under a finite number of collisions. This excludes the path of photons. Thus A-topology is significantly different from William's topology in this respect.

Popvassilev [12] generalized the concept of Zeeman-like fine topologies to R^n and proved that these topologies are non-regular. Since these topologies are Hausdorff, it follows that they are not normal. This property was proved by Dossena [2] in a different way by using Urysohn Lemma.

S. Nanda and H.K. Panda [25] define yet another topology on Minkowski space. This is a non-Euclidean topology, namely order topology generated by the positive cone at origin and its translates. They prove that it is non-compact, non-Hausdorff but path-wise connected. Moreover, it has the property that every loop based at a point is homotopic to the constant loop at that point. Thus, this topology is simply-connected. This is contrary to the non-simply connected nature of M^Z, M^t and M^s.

2.4. Contributions by Dossena, Agrawal and Shrivastava

We now discuss the work of Dossena [2] and G. Agrawal and S. Shrivastava [14] [15] where many interesting topological properties of M^Z, M^t and M^s have been proved, including non-simply connectedness when restricted to two dimensional Minkowski space.

As defined in the begining of this section, Dossena presents Zeeman topology M^Z in the language of affine spaces and proves that Zeeman topology is separable, non-first countable and non-trivial. We discuss below the results proved by Dossena in some details, especially for two dimensional Minkowski space.

For two dimensional Minkowski space with topologies M^E and M^Z, Dossena gives characterization of the sets $\Sigma \subseteq M$ on which M^E and M^Z induce the same topology. *i.e.* $\Sigma \cap M^E = \Sigma \cap M^Z$. To prove this, he uses the concept of Zeno sequences. Furthermore, in this two dimensional case, he gives characterization of compact subsets of M^Z. We summarize these results below:

Lemma 2.1. A compact subset of M^Z is compact in M^E.

Lemma 2.2. Let X be a Hausdorff topological space and let $(x_n)_{n\in N}$ be a sequence of distinct points of X converging to x. Then x is the unique limit point for the set $\{x_n\}_{n\in N}$. In particular, every x_j is an isolated point for $\{x_n\}_{n\in N}$.

Lemma 2.3. Every Zeno sequence admits a subsequence whose image is a non closed, discrete subset of M^E, closed in M^Z.

Theorem 2.4. A compact subset K of M^Z contains no images of Zeno sequences.

This is true for A-topology also, as proved by Nanda [5].

Theorem 2.5. For a subset $K \subset M$, the following are equivalent:

1) K is compact in M^Z.

2) K is compact in M^E and contains no completed images of Zeno sequences.

3) K is covered by a finite family $(A_j)_{j=1,\cdots,J}$ of axes such that for each $j = 1, \cdots, J$ the set $A_j \cap K$ is compact in A_j^E.

We now discuss countability properties of M^Z.

We choose an orthonormal frame of reference $(o, (e_i)_{i=0,\cdots,k})$. Then every $p \in M$ is identified by its coordinates $\{p^i\}_{i=0,\cdots,k}$, such that $p = o + \sum_{i=0}^{k} p^i e_i$.

Clearly M^E is separable (so are all finite-dimensional affine spaces endowed with their natural topology). A countable dense subset Q of M^E can be constructed by choosing an orthonormal frame of reference and defining Q as the set of points in M with rational coordinates.

Then we have the following proposition:

Proposition 2.6. For every orthonormal frame of reference, the above-mentioned set Q is also dense in M^Z. Thus M^Z is separable.

Corollary 2.7. The cardinality of the set $\mathcal{C}\left(M^Z, R\right)$ of all real continuous functions on M^Z is at most equal to 2^{χ_0}, where χ_0 is the cardinality of Natural numbers.

Proposition 2.8. M^Z is not first countable at any point.

Zeeman [1] has sketched the proof of the result that M^Z is not normal. As noted earlier, Dossena gives another proof of the same result using Urysohn lemma. Thus, we have:

Theorem 2.9. M^Z is not normal and hence not metrizable.

For a path-connected topological space X, $\pi_1(X)$ denotes the fundamental group or first homotopy group of X. The following is the most remarkable result proved by Dossena:

Theorem 2.10. $\pi_1\left(M^Z\right)$ is nontrivial and possesses uncountably many subgroups isomorphic to Z. In particular, M^Z is not simply connected. For details of proofs we refer the reader to Dosssena [2].

A topological study of the n-dimensional Minkowski space, M^n, with t-topology, denoted by M^t, has been carried out by G. Agrawal and S. Shrivastava [14]. Path-topology defined by Hawking, King and Mc Carthy [9] on a space-time of general relativity will be discussed in Section 3. If we restrict this topology to four dimensional Minkowski space, then it comes out to be identical with t-topology. Non-simply connectedness of M^t, compact sets of M^t, and subsets of M that have the same subspace topologies induced from the Euclidean and t-topologies are also discussed in this paper.

t-topology for four dimensional Minkowski space has been defined above. Similar definition follows for M^n also. Thus $U \subset M$ is open with respect to t-topology if and only if for each $x \in U$ there exists some $N_\epsilon^t(x)$ such that $N_\epsilon^t(x) \subset U$.

It thus follows that $N_\epsilon^E(x)$ and $N_\epsilon^t(x)$ are open in M with t-topology, $N_\epsilon^E(x)$ is open in M with Euclidean topology, while $N_\epsilon^t(x)$ is not open in M with Euclidean topology. Hence $\left\{N_\epsilon^t(x) / x \in M, \epsilon > 0\right\}$ is a basis for the t-topology and the t-topology is strictly finer than the Euclidean topology on M.

s-topology can be defined similarly on R^n.

Summarizing, we have the following:

The collection $\left\{N_\epsilon^t(x) / x \in M, \epsilon > 0\right\}$ being a basis for the path topology on four-dimensional Minkowski space, the path topology on four-dimensional Minkowski space is same as the t-topology. It thus follows that the four-dimensional Minkowski space with t-topology is Hausdorff, path connected, separable, first countable, not second countable, not countably compact, not Lindelof, not regular, not normal and hence is not compact, not locally compact, not paracompact, not mertizable, and not locally n-Euclidean.

2.5. Other Works

Other works on Zeeman-like topologies include that of Struchiner and Rosa [26] and Domiaty [27] [28]:

Struchiner and Rosa [26] study Zeeman topology in Kaluza-Klein and Gauge theories. They generalize the notion of Zeeman topology by using the projection theorem of Kaluza-Klein theories, and this remains valid for any gauge fields. Here, the authors consider differential geometric frame work of fiber bundles and define Zeeman topology in the total space of fiber bundle. From this, they obtain a topology in the base manifold for which the continuous curves correspond to motions of charged particles in the base manifold. It would be interesting to see the generalizations of typical gauge theoretical ideas when the space-time has such a topology.

Domiaty [27] [28] considers yet another topology on Lorentz manifolds. This topology is in a certain sense the space-like version of an analogous result for the Hawking-King-McCarthy path topology which has been discussed below. The space topology is the finest topology on a Lorentz manifold, which induces the manifold topology on every space-like hypersurface. As proved in these papers, its geometric significance comes from the fact that its full homeomorphism group is the group of all conformal diffeomorphisms.

Finally, we remark that even though Zeeman topology on Minkowski space has several advantages over the standard topology, it has some drawbacks also. These are as follows:

1) A three dimensional section of simultaneity has no meaning in terms of physically possible experiments.

Also, the use of straight time like lines in defining M^Z suggests that M^Z from the beginning has been equipped with information involving inertial observers, so that occurrence of linear structure is not surprising.

2) The isometry and conformal groups of M^Z are physically significant but same thing is not clear about homothecy group of M^Z.

3) The set of M^Z-continuous paths does not incorporate accelerating particles moving under forces in curved lines.

4) M^Z is not first countable and hence it is difficult to handle.

Keeping these drawbacks in mind, Hawking, King and Mc Carthy [9] defined another topology called path topology on a space-time of general relativity. We now discuss, below, this topology and its properties. We also discuss other related topologies as studied by Kim [13] and Low [17] and their inter-relationships with HKM topology.

3. Path Topology of Hawking, King and Mc Carthy (HKM) and Other Related Topologies

Here, we consider a space-time of general relativity which is assumed to be connected, Hausdorff, paracompact, C^∞ real four-dimensional manifold V without boundary, with a C^∞-Lorentz metric and associated pseudo-Riemannian connection. V is also assumed to be time-orientable *i.e.* V admits a non-vanishing time-like vector field.

The path topology $\mathcal{P}$ of V is defined as follows:

$\mathcal{P}$ is the finest topology satisfying the requirement that the induced topology on every time-like curve coincides with the topology induced from $\mathcal{V}$, where $\mathcal{V}$ is the given manifold topology on V.

Thus if a set $E \subset V$ is $\mathcal{P}$-open, for every time-like curve γ, there is an $O \in \mathcal{V}$ with $E \cap \gamma = O \cap \gamma$. Conversely, if E satisfies this condition, it is $\mathcal{P}$-open and $\mathcal{P}$ is the largest collection of such sets. Obviously, if $O \in \mathcal{V}$, then $O \in \mathcal{P}$.

HKM show that $\mathcal{P}$ is strictly finer than $\mathcal{V}$, but however $\mathcal{P}$ is not comparable to Zeeman topology.

Let $T_p(V)$ denote the tangent space of $p \in V$ and $exp : T_p(V) \to V$ be the exponential mapping. Then there is an open neighbourhood N of the origin of $T_p(V)$ such that $U = exp(N)$ is an open convex neighbourhood of $p \in V$. Let $\epsilon > 0$ be sufficiently small so that the Euclidean open ball B of radius ϵ, with centre at origin, is contained in N. Then $B_u(p,\epsilon) = exp(B)$. For any open set V, define $C(p,V) = I^+(p,V) \cup I^-(p,V)$, $K(p,V) = C(p,V) \cup \{p\}$ and for an open convex normal neighbourhood U of p, define $L_u(p,\epsilon) = B_u(p,\epsilon) \cap K(p,U)$. ($B_u(p,\epsilon) = expB$). Then we have the following :

Proposition 3.1. Sets of the form $K(p), K(p,U)$ and $L_u(p,\epsilon)$ are $\mathcal{P}$-open.

$K(p,U)$ is not open in the manifold topology $\mathcal{V}$ because $p \in K(p,U)$ has no $\mathcal{V}$-nbd contained in $K(p,U)$. Thus $\mathcal{P}$ is strictly finer than $\mathcal{V}$.

Theorem 3.2. $L_u(p,\epsilon)$ forms a basis for the topology $\mathcal{P}$.

This property has no analogue in the finer topologies $\mathcal{T}$. $\mathcal{P}$-continuous paths are characterized as follows:

Theorem 3.3. A path $\gamma : F \to V$ is $\mathcal{P}$-continuous if and only if it is a Feynman Path.

Theorem 3.4. $\mathcal{P}$ is first countable and separable. $\mathcal{P}$ is Hausdorff, path connected and locally path connected and hence locally connected. However, $\mathcal{P}$ is not regular, normal, locally compact or paracompact.

Furthermore, HKM determine the group of $\mathcal{P}$-homeomorphisms and prove that it is the group of smooth conformal diffeomorphisms.

To begin with, they prove the following:

Proposition 3.5. $\mathcal{P}$-homeomorphisms h take time-like curves to time-like curves.

This has been proved for strongly causal space-times. It is done by singling out a subclass of $\mathcal{P}$-continuous curves which coincides with time-like curves.

After proving a series of results, HKM prove the following important theorem:

Theorem 3.6. A $\mathcal{P}$-homeomorphism h is a smooth conformal diffeomorphism. This leads to the description of $\mathcal{P}$-homeomorphisms of M.

Theorem 3.7. The group of $\mathcal{P}$-homeomorphisms of M coincides with the group of smooth conformal diffeomorphisms of M.

Finally, HKM give an example of a manifold for which the group of smooth conformal diffeomorphisms is strictly larger than the homothecy group. We note here that for Minkowski space, the two groups are equal.

For more details and proofs, we refer the reader to HKM [9].

Malament [10] improved the results of [9] in the sense that the condition of strong causality is no longer necessary. We now discuss briefly the work of Malament [10]:

Main result of this paper is the following:

Suppose we consider two space-times (V,g) and (V',g') and a bijection $f:V\to V'$, where both *f* and f^{-1} preserve continuous time-like curves. This means, if $\gamma:I\to V$ is a continuous time-like curve in (V,g) then $f\circ\gamma:I\to V'$ is a continuous time-like curve in (V',g'). Similar condition holds for f^{-1}. Then *f* must be homeomorphism. Thus the class of continuous time-like curves in a space-time determines its topology. By Hawking's theorem, *f* will then be a smooth conformal isometry.

Brief summary of the proof is as follows:

If *f* preserves all continuous curves, then *f* would be continuous. Given any sequence $\{p_n\}$ converging to *p*, one could find a continuous curve "threading" all the p_n in sequence and then *p*. Its image would have to be a continuous curve threading all the $f(p_n)$ in sequence and then $f(p)$. Hence $f(\{p_n\})$ would converge to $f(p)$. Under the hypotheses under consideration, this construction can only be applied to sequence $\{p_n\}$ which converge chronologically to *p*. The problem is with those sequences $\{p_n\}$ which converge to *p* but are locally space-like related to *p*.

The idea to overcome this difficulty is as follows:

To show that *f* is continuous at *p*, one proves that one may assume that *f* is continuous over a `nice-looking' region near *p*. Then one uses continuous null geodesic segments in this region to characterize the convergence of points to *p*. This then leads one to the required result because continuous null geodesics in this region are necessarily preserved by *f*. For technical details, we refer the reader to Malament [10]. HKM-topology is an improvement over Zeeman topologies in the sense that it removes many unpleasant features of those topologies.

Fullwood [18] modified the HKM topology and defined a new topology $\bar{\mathcal{P}}$ as follows :

$q\gg p$ if and only if $q\in I^+(p)$. Then $p\in I^-(q)$ and $p\ll q$. We denote $I^+(q)\cup I^-(q)$ by $I(q)$.

Then, define $\langle p,q,r\rangle = I^+(p)\cap I(q)\cap I^-(r)\cup\{q\}$ for $p\ll q\ll r$ in *V*.

Now, let $\mathcal{B}=\{B: B=\langle p,q,r\rangle \text{ for some } p\ll q\ll r \text{ in } V\}$.

Then, $\mathcal{B}$ forms a base for a topology which is denoted by $\bar{\mathcal{P}}$.

Fullwood proves that if the space-time *V* is future and past distinguishing, then the topology $\bar{\mathcal{P}}$ coincides with HKM $\mathcal{P}$-topology. More precisely, he proves the following theorem:

Theorem 3.8. The following three conditions are equivalent upon a space-time manifold:

1) $\bar{\mathcal{P}}=\mathcal{P}$ *i.e.*, the topology $\bar{\mathcal{P}}$ is equivalent to the Path topology; 2) the distinguishing condition holds on *V*, and 3) *V* is $\bar{\mathcal{P}}$-Hausdorff.

Do-Hyung Kim [13] proved that the path topology of Hawking, King, and McCarthy can be extended to the causal completion of a globally hyperbolic Lorentzian manifold. The suggested topology $\mathcal{T}$ is defined only in terms of chronological structures and $\mathcal{T}$ is finer than the extended Alexandrov topology denoted by $\bar{\mathcal{A}}$. It is also shown that a $\mathcal{T}$-homeomorphism induces a conformal isomorphism and a homeomorphism in $\bar{\mathcal{A}}$. Let $\bar{V}$ denote causal completion of *V*. Then $\mathcal{T}$ is defined on $\bar{V}$ as follows:

Definition 3.1. $U\subset\bar{V}$ is $\mathcal{T}$-closed if every time-like sequence that converges has a limit in *U* and $W\subset\bar{V}$ is $\mathcal{T}$-open if its complement is $\mathcal{T}$-closed.

Proposition 3.9. The above family of open sets define a new topology $\mathcal{T}$ on $\bar{V}$.

Proposition 3.10. The topology $\mathcal{T}$ on $\bar{V}$ is finer than the extended Alexandrov topology $\bar{\mathcal{A}}$ on $\bar{V}$.

Since $\mathcal{T}$ is finer than $\bar{\mathcal{P}}$ and $\bar{\mathcal{P}}$ is Hausdorff, it can be concluded that $\mathcal{T}$ is a Hausdorff topology on $\bar{V}$.

Corollary 3.11. $I^-(\gamma)$ is also an end point of a time-like curve $\gamma:(a,b)\to V$ in $\mathcal{T}$-topology.

The construction of $\mathcal{T}$-topology on the causal completion extends the $\mathcal{P}$-topology on *V* by use of the sequential convergence.

Furthermore, Kim studies homeomorphisms with respect to topology $\mathcal{T}$. To understand the results in this direction, let *V* and *N* be two space-times and let $\bar{V}$ and $\bar{N}$ be their causal completions. Then we have the following definition:

Definition 3.2. A bijection $f:\bar{V}\to\bar{N}$ is a chronological isomorphism if $x\ll y\Leftrightarrow f(x)\ll f(y)$ and antichronological isomorphism if $x\ll y\Leftrightarrow f(y)\ll f(x)$. Likewise, a bijection $f:\bar{V}\to\bar{N}$ is a causal isomorphism if $x\le y\Leftrightarrow f(x)\le f(y)$ and anticausal isomorphism if $x\le y\Leftrightarrow f(y)\le f(x)$. A bijection

$f:\bar{V}\to\bar{N}$ is a conformal isomorphism if f is both (anti) chronological isomorphism and (anti) causal isomorphism. In a Lorentzian manifold, it is known that the causal isomorphism and the chronological isomorphism are equivalent. The topology $\mathcal{T}$ is defined only in terms of chronological relations and so any chronological isomorphism $f:\bar{V}\to\bar{N}$ induces a $\mathcal{T}$-homeomorphism. The chronological isomorphism has the same effects on the $\bar{\mathcal{A}}$-topology.We also have the following:

Proposition 3.12. If *V* and *N* are globally hyperbolic and $f:\bar{V}\to\bar{N}$ is either a chronological isomorphism or an antichronological isomorphism, then *f* is an $\bar{\mathcal{A}}$-homeomorphism.

Theorem 3.13. If $f:\bar{V}\to\bar{N}$ is a $\mathcal{T}$-homeomorphism, then *f* is either a chronological isomorphism or an antichronological isomorphism.

Theorem 3.14. If $f:\bar{V}\to\bar{N}$ is a $\mathcal{T}$-homeomorphism, then *f* is a conformal isomorphism.

Since $\mathcal{T}$ is finer than $\bar{\mathcal{A}}$, by combining proposition 3.8 and theorem 3.9, we have the following theorem.

Theorem 3.15. A $\mathcal{T}$-homeomorphism induces an $\bar{\mathcal{A}}$-homeomorphism.

Also if $f:V\to N$ is a $\bar{\mathcal{P}}$-homeomorphism, then *f* is a conformal isomorphism. If, in addition, both *V* and *N* are strongly causal, the manifold topologies are the same as the Alexandrov topologies since the Alexandrov topology is defined only in terms of a chronological relation. In other words, a $\bar{\mathcal{P}}$-homeomorphism induces an $\mathcal{V}$-homeomorphism. By the above theorem, this is indeed the case in the path topology of the causal completion. Thus, the extended Alexandrov topology is natural to the causal completion. The $\bar{\mathcal{P}}$-topology mentioned here is that defined in Fullwood [18], and the causal completion of space-times mentioned in the discussion above is in the sense of Budic and Sachs [29].

Such bijective mappings have also been studied by Domiaty [27] [28]. These mappings are defined in such a manner that they leave the class of space-like paths invariant. Homeomorphisms with respect to *S*-topology defined by Nanda [4] are called *S*-homeomorphisms. Domiaty proved that if (V,g) and (V',g') are Lorentz manifolds and if $f:V\to V'$ is a bijection, then *f* is a *S*-homeomorphism if and only if *f* and f^{-1} preserve space-like paths. Furthermore, after proving a series of lemmas, he proves that if *f* and f^{-1} preserve space-like paths,then *f* is a manifold-homeomorphism (*V*-homeomorphism). There is a substantial literature on causality-preserving maps (causal maps) or cone-preserving maps in special as well as general theory of relativity. See, for example, a review article by Sujatha Janardhan and R.V. Saraykar [24] and references therein. If we denote homeomorphisms with respect to path-topology (HKM-topology) by $\mathcal{T}$-homeomorphisms, then every *S*-homeomorphism is a $\mathcal{T}$-homeomorphism. Since (ref. Kim [13]) a $\mathcal{T}$-homeomorphism is a smooth conformal diffeomorphism, it follows, by combining results of Domiaty and Kim, that every *S*-homeomorphism is also a smooth conformal diffeomorphism. (This has been noted by Domiaty [27] [28] Theorem 2.) This result improves the result by Göbel [7] which was proved for strongly causal Lorentz manifolds.

More recently Huang [30] proved the result: Let (V,g) be a strongly causal space-time, dim $V\geq 3$. Let $f:V\to V$ be a bijection such that images and pre-images of null geodesics (as point sets) are null geodesics. Then *f* is a homeomorphism and hence by Hawking's theorem, a conformal transformation. This generalizes the result proved by Jan Peleska [31]. Define a local distance function on convex normal neighbourhoods by $\phi(p,q)=g\left(exp_p^{-1}q,exp_p^{-1}q\right)$ Then every homeomorphism *f* which locally preserves these functions is an isometry. If (V,g) has indefinite signature and *f* locally preserves distance zero, then it is a conformal diffeomorphism.

The physical meaning of the condition used in this theorem is that images and pre-images of paths which photons travel between emission and absorption should again be such paths.

Coming to the topological properties of Zeeman-like topologies on Minkowski space *M* again, we note the Theorem proved by Dossena, namely, two dimensional Minkowski space is not simply connected. Its first homotopy group contains uncountably many subgroups isomorphic to *Z*.G. Agrawal and S. Shrivastava [13] proved similar result for t-topology. Both these proofs use the notion of Zeno sequences introduced by Zeeman. Robert Low [17] recently gave a proof for the same result for n-dimensional Minkowski space with Zeeman topology without using Zeno sequences. For the sake of completeness, we reproduce the proof of this important theorem below.

Theorem 3.16. A space-time *V*, equipped with the path topology is not simply connected or locally simply connected. Furthermore, no two closed continuous curves in *V* with distinct images are homotopic.

Proof: Let c_1 and c_2 be curves in *V* with distinct images, let $h:I\times I\to V$ such that $h(0,.)$ is c_1 and $h(1,.)$ is c_2, and let *T* be some time-like two-plane and π be the associated projection such that the projections of c_1 and c_2 to *T* are distinct. Now neither of πoc_1 nor πoc_2 can be space-filling, for then we

already have an open set in T containing infinitely many points in some space-like surface and in the image of πoh. R. Low then considers the intersection of this open set with some surface of constant time and argues to conclude that there must be some point x in T round which c_1 and c_2 have different winding numbers. Since c_1 and c_2 are closed curves in T, x has an open neighbourhood in T which lies in the image of πoh, and again we obtain a contradiction. Hence, if c_1 and c_2 are closed continuous maps from S^1 to V with distinct images, then c_1 and c_2 are not homotopic in the path topology. Moreover the fundamental group of V with the path topology is as large as possible, since two continuous loops are only homotopic if one is a re-parameterisation of the other. Also, the above result is true in case of a general Lorentz manifold. The general space-time V can be embeded in a pseudo-Euclidean space of appropriate dimension, and arguing as above, by projecting to some suitable time-like plane in the pseudo-Euclidean space, we can obtain the same result.

Here, it will not be out of place to mention that Sorkin and Woolgar [32] introduced the concept of K-causality with the aim that it should be possible to derive the causal structure from order relation and topological structure. Some results in this direction were proved by S. Janardhan and R.V. Saraykar [33]. Later, after a good deal of efforts, Minguzzi [34] proved that *Stable causality* is equivalent to *K-causality*. In the description of path-topology above, if analogously, if we replace a time-like curve by a K-causal curve which is compact, connected and linearly ordered, then we can define *K-causal topology* on V, denoted by $\mathcal{K}$ as follows:

We specify closed sets of $\mathcal{K}$ as follows:

$\tilde{F}$ is a $\mathcal{K}$-closed subset of V if $\tilde{F} \cap \gamma = F \cap \gamma$ for some closed $F \subseteq V$, in the manifold topology and $\mathcal{K}$ is the finest such topology. If F is closed in V, with respect to $\mathcal{V}$, then F is closed with respect to $\mathcal{K}$ also. Thus $\mathcal{K}$ is finer than $\mathcal{V}$. For a detailed discussion of K-causal curves in K-causal space-time, we refer the reader to S. Janardhan and R.V. Saraykar [31] and Minguzzi [32] and references therein.

4. Zeeman-Like Topologies in General Relativity

In this section, we describe and discuss the work of Göbel [7] [8], Lindstrom [11] and others on Zeeman-like topologies defined on a space-time of general relativity. In particular, Göbel [7] has proved the result that two space-times are homeomorphic with respect to its Zeeman topology if and only if they are isometric. This shows that it is possible to determine the metric of a space-time from its Zeeman topology.

We start with definitions of Zeeman topologies as given by Göbel [7] and discuss their main properties.

Let $(V, \mathcal{T})$ denote a differentiable manifold with an underlying manifold topology $\mathcal{T}$. The most general setting for Zeeman topologies is the following:

Let Σ be a set of subsets of V. Then a subset $X \subset V$ belongs to $Z = Z(\Sigma, \mathcal{T})$ iff $X \cap Y$ is open within the topological space $Y = (Y, \mathcal{T}_Y)$ with its induced topology $\mathcal{T}_Y$, for all $Y \in \Sigma$. -------($\star$)

Then (V, Z) is the space V provided with the Zeeman topology $Z = Z(\Sigma, \mathcal{T})$ generated by $(\Sigma, \mathcal{T})$. Thus the topology Z is the finest topology $\mathcal{F}$ on V such that $\mathcal{F}_Y = \mathcal{T}_Y$ for all $Y \in \Sigma$.

On Minkowski space this topology coincides with the topology Z defined by Zeeman mentioned above, for two specially chosen systems Σ which are significant for special relativity. Since $\mathcal{T}$-open subset of V always satisfies condition ($\star$), Z is always finer than $\mathcal{T}$.

Further Göbel defines a *Special system* $\Sigma = (\Gamma, \Delta)$ of V as follows:

$\Sigma = (\Gamma, \Delta)$ is called a special system of V if there is a locally finite covering $\mathcal{U}$ of V by neighbourhoods U, such that $\Gamma = \bigcup_{U \in \mathcal{U}} \Gamma_U$ and $\Delta = \bigcup_{U \in \mathcal{U}} \Delta_U$ where $\Gamma_U = \{X \in \Gamma / X \subset U\}$ and $\Delta_U = \{X \in \Delta / X \subseteq U\}$ have the following properties:

1) If $X \in \Sigma = \Gamma \cup \Delta$, then X is a closed subset of V.

2) If $X \in \Gamma_U, Y \in \Gamma_V$ and $|X \cap Y| = \infty$ then $X \cap V = Y \cap U$ for all $U, V \in \mathcal{U}$. (Here $|A|$ denotes cardinality of A)

3) If $p, q \in U \in \mathcal{U}$ and $\Gamma_U(p, q) = \{X \in \Gamma_U / p, q \in X\}$ is infinite, then $p = q$.

4) We have $|X \cap Y| \leq 1$ for all $X \in \Gamma$ and $Y \in \Delta$. ----------($\star\star$)

With this definition, the following results follow:

Proposition 4.1. Let Σ be a special system of V and $f : [0,1] \to V$ be a 1-1 map which is Γ-directed at $p \in f([0,1])$. Then f is a piecewise Γ-curve at p if f is continuous at p with respect to the Zeeman topology Z.

(A curve f is called Γ-directed at $p \in f([0,1])$ if there is a neighbourhood U of p defined by ($\star\star$) such that if $p \neq q \in U \cap f([0,1])$, then $\Gamma_U(p, q) \neq \phi$. f is called a piecewise Γ-curve at $p = f(a)$ if there are

b, c with $0 \le b < a < c \le 1$ such that $f([b,a]) \subseteq X$ and $f([a,c]) \subseteq Y$ for some $X, Y \in \Gamma_U$).

Proposition 4.2. If Σ is a special system of V and f is a Z-continuous curve which is Γ-directed at each point $p \in f([0,1])$, then f is a piecewise Γ-curve.

This implies the following:

Proposition 4.3. For a manifold $(V, \mathcal{T})$ with an affine connection, following two statements are equivalent:

1) the curve f is a piecewise geodesic *i.e.* f is a broken geodesic line with a finite number of edges.

2) the 1-1 map $f : [0,1] \to V$ is continuous with respect to the Zeeman topology Z.

Göbel then restricts Zeeman topology on a space-time and studies Zeeman topology by incorporating electromagnetic fields. To state the results proved by Göbel in this situation, we need to understand certain notations:

Let V denote a space-time for general relativity and F be a given electromagnetic field on V. An electric charge q_p of a test particle p has its absolute value bounded by a number depending on F, and mass m_p of this particle ($m_p > 0$) is bounded by a number depending on the gravitational field. Since the charge-spectrum Q and mass spectrum W are discrete, there are finitely many possible values $q_p \in Q$ and $m_p \in W$ for test par- ticles p. We assume the presence of charge free test particles so that $O \in Q$. If $Q = 0$, we allow the mass spectrum W to be arbitrarily > 0. Under these conditions, there are covering $\mathcal{U}$ and $\mathcal{B}$ which are locally finite, so that there are only finitely many world lines of freely falling test particles in $U \in \mathcal{U}$ from $p \in U$ to $q \in U$ if $p \ne q$. For $U \in \mathcal{U}$, let Γ_{qU}^m be the set of all world lines of freely falling test particles and let Δ_U be all closed space-like C^1-hypersurfaces of $\bar{V}$. (Here $W \in \mathcal{B}$ such that there is one and only one $U(W) \in \mathcal{U}$ which contains the closure $\bar{W}$ of W.) The corresponding system Σ_Q^W is then a special system of W. Then the following result holds:

Proposition 4.4. If V is a space-time with a given external electro-magnetic field F and a world line f, the following statements are equivalent:

1) f is continuous with respect to the Zeeman topology $Z(\Sigma_Q^W, \mathcal{T})$.

2) f is a chain of finitely many connected world lines of freely falling charged test particles.

If $F = 0$, then Z-continuous world lines are future directed time-like piecewise geodesic lines. For simplicity, we denote $Z(\Sigma_Q^W, \mathcal{T})$ by Z_R. Then open sets with respect to Z_R are described as follows:

A subset Y of V is open with respect to Z_R iff $Y \cap U$ is open in $(U, \mathcal{T}_U)$ for the following subsets U of V:

(I) U is an arbitrary closed space-like hypersurface contained in a simple region of V.

(II) U is the world line of an arbitrary charged test particle p freely falling in the gravitational and the electro-magnetic field within a simple region of V.

If $Q = 0$, then condition (II) is equivalent to

(II)' U is an arbitrary time-like geodesic in a simple region of V.

If U is a simple neighbourhood of p then let $U^*(p) = (U \setminus \mathcal{C}_U(p)) \cup \{p\}$.

Lemma 4.5. The set $U^*(p)$ defined above is a $Z(\sum_Q^W, \tau)$-neighbourhood of p.

Göbel then proves an important result that

Proposition 4.6. The topology induced by Z_R on a light cone is discrete.

Thus we do not have any geometric information along a light ray.

The main theorem of Göbel [7] is the following (which he proves in the last section of his paper).

Theorem 4.7. Let h be a mapping from space-time V onto a space-time V'. The following are equivalent:

1) h is a homeomorphism with respect to Zeeman topology Z.

2) h is a homothetic transformation.

Unusual property of Zeeman topology is that homeomorphism characteristic of h implies its differentiability as well as its "linearity", since h is an isometric map "up to scaling". Thus we can state this property in the following forms:

Theorem 4.8. The space-times V and V' are homeomorphic with respect to Zeeman topology if and only if they are isometric (up to a constant positive factor).

Theorem 4.9. The group of all homeomorphisms with respect to the Zeeman topology coincides with the group of all homothetic transformations of space-time V onto itself.

Thus Zeeman topology contains all information about the metric.

We again note here that (locally) causal maps defined by Göbel [7] in Section 2 and described in Section 5 are

similar to causal maps of García-Parrado and Senovilla [23], and subsequently similar to K-causal maps described and studied by Sujatha Janardhan and R.V. Saraykar [31].

As far as Minkowski space-time is concerned, Zeeman [1] has suggested other topologies on it. Göbel generalized some of the results which hold for these topologies. Following remarks are in order about these topologies:

Remark 1. The topology Z_1 defined by Zeeman is now well-known as t-topology studied by Nanda [4]. The induced topology on any space axis is discrete. Under this topology, Göbel has generalized this result as follows:

Theorem 4.10. Let $f: I \to (V, Z_1)$ be a continuous map of the unit interval I into V (endowed with Z_1-topology). If f is strictly order preserving, *i.e.* $x < y$ implies $f(x) < f(y)$ (*i.e.* the vector $f(y) - f(x)$ is time like), then the image $f(I)$ is a piecewise linear path, consisting of a number of intervals along time axis.

Further, this topology has a physically attractive feature as follows:

If $f: I \to V$ be an embedding (not necessarily order preserving), then $f(I)$ is a piecewise linear path along time axes, zig-zagging with respect to time orientation like the Feynman track of an electron.

Hawking, King and Mc Carthy [9] has defined *Feynman path* mathematically precisely as follows:

Let $K(p,U)$ denote $I^+(p,U) \cup I^-(p,U) \cup \{p\}$ where U denotes an open convex normal neighbourhood of p. A path $\gamma: I \to V$ is a *Feynman path* if γ is continuous and for each $t_0 \in I$, there is an open connected neighbourhood U of t_0, and an open convex normal neighbourhood U of $p = \gamma(t_0)$ such that $\gamma(U) \subseteq K(p,U)$.

A locally one-one Feynman path is then a Feynman track mentioned above.

Let G denote the group of automorphisms of V given by

1) the Lorentz group of all linear maps leaving quadratic form Q invariant

2) translations and

3) dilatations.

Every element of G either preserves or reverses the partial ordering "<" mentioned above. These features have been studied in details by Nanda, Dossena and Kim.

Remark 2. The topology Z_2 defined by Zeeman is well-known as s-topology studied by Nanda [3] [4]. The induced topology on any time axis is discrete. Homeomorphism group of this topology was determined by Nanda thus proving another version of Zeeman conjecture. Topological properties of t-topology and s-topology have been studied by *G*. Agrawal and S. Shrivastava [14] [15] as mentioned in Section 2.

Remark 3. The topology Z_3 defined by Zeeman is same as Williams Topology M^F. As proved by Williams [6], this topology possesses the following properties:

1) It is not locally homogeneous and the light cone through any point can be deduced from it.

2) The group of all homeomorphisms with respect to Z_3 is generated by inhomogeneous Lorentz group and dilatations.

3) It induces the 3-dimensional Euclidean topology on every space axis and the 1-dimensional Euclidean topology on every time axis.

For the proof of these properties, we refer the reader to Williams [6] and Zeeman [1]. However, this topology does not satisfy the theorem mentioned above. Nevertheless, the group of homeomorphisms of (V, Z_3) is G. Thus although (V, Z_3) has a countable base of neighbourhoods for each point, it is physically less attractive than Z. Such topologies can also be described on a general space-time following Göbel's method.

Ulf Lindstrom [11] re-examined the separating topology studied in earlier works. Using methods and ideas in papers by Göbel, Hawking, King and McCarthy, he introduced a new class of topologies $\{S_{nm}\}$. The topology $\{S_{nm}\}$ is the finest which induces Euclidean topology on time-like C^n- and space-like C^m-curves. A relation between $\{S_{nm}\}$ and some topologies studied by Göbel is derived—For an arbitrary space-time the group of homeomorphisms is shown to be the smooth conformal diffeomorphism group. The restriction to strongly causal space-times employed in earlier work is no longer necessary. We note that Lindstrom topology reduces to Williams C^1 topology M^F for $m = n = 1$ on Minkowski space. Group of C^1 homeomorphisms is C^1 conformal diffeomorphisms as noted in Section 2.

Finally, we add a comment about the work of Mashford [19]: As is well-known, a space-time in the general theory of relativity is a Lorentz manifold modeled on 4-dimensional Euclidean space, which is locally a Minkowski space. Mashford [19] constructs a tangent bundle whose base space is not a Lorentz manifold, but is a set Y of events which is equipped with an acyclic signal relation $\sim\!\!\rightarrow$ and the $\sim\!\!\rightarrow$ structure of Y is locally

that of Minkowski space with Zeeman topology. Moreover, the *piecing together* maps are smooth in an appropriate sense. The parent space E is the tangent bundle TY of Y. Mashford then proves that this bundle has, as structure group, the group of linear causal automorphisms of Minkowski space, which coincides with the group G of Lorentz transformations along with translations and dilatations which has been discussed in Section 2.

5. Conclusions

In this article, we have given a short review of Zeeman- and Zeeman-like fine topologies on Minkowski space and space-time of general relativity. We have avoided giving detailed proofs of the results mentioned, otherwise the article would have become lengthy. To the best of our knowledge, we have reviewed most of the research work which appeared on this topic since the first paper was published by Zeeman in 1967. To get a consolidated view about definitions and the main properties of these topologies like their homeomorphism groups and topological properties, we give two tables summarizing definitions and their properties:

Definitions and properties of fine topologies on Minkowski space refer **Table 1** and fine topologies on space-times of general relativity refer **Table 2**. Whereas fine topologies have interesting topological properties and their homeomorphism groups are physically useful, however it is true that manifold structure is not compatible with fine topologies. This is because, topologically, a manifold is second countable, Hausdorff and paracompact, and hence normal and metrizable, whereas fine topologies are not, in general, normal (and hence not metrizable). Moreover, it is also true that unless differential structure is there, we can not define notions of connection and curvature and hence fine topologies may not be useful in discussing Einstein field equations in general theory of relativity. Finally, we would like to refer to a paper by A. Heathcote [35], where it has been argued that the suggestions for replacement of manifold topology with fine topology misrepresent the significance of the manifold topology and overstate the necessity for a finer topology. He claims to have given a

Table 1. Definitions and properties of fine topologies on Minkowski space.

Sr. No	Fine topology	Homeomorphism group	Topological properties
1	Zeeman topology M^Z (1967): Finest topology which induces three dimensional Euclidean topology on every space-axis and one dimensional Euclidean topology on every time-axis	G = Lorentz group with translations and dilatations	Dossena (2007): neither locally compact nor Lindelof, not normal, separable but not first countable, path-connected but not simply connected
2	s-topology M^s: Nanda (1971): Finest topology which induces three dimensional Euclidean topology on every space-like hypersurface	G	G.Agrawal and S. Shrivastava (2012): separable, first countable, path-connected, not regular, not metrizable, not second countable, noncompact, and non-Lindelof, not simply connected
3	t-topology M^t: Nanda (1972): Finest topology which induces one dimensional Euclidean topology on every time-like line	G	G.Agrawal and S. Shrivastava (2009): separable, first countable, path-connected, not regular, not metrizable, not second countable, not locally compact, not simply connected
4	A-topology M^A: Nanda (1979): Finest topology which induces one dimensional Euclidean topology on every time-like line and light-like line and three dimensional Euclidean topology on every space-like hypersurface	G	G.Agrawal and Soami Pyari Sinha (2014): separable, not first countable, connected and path-connected, not normal, not metrizable, Not comparable with t-topology nor with s-topology
5	Fine topologies M^F by Williams (1974): Finest topology which induces one dimensional Euclidean topology on every time-like line and space-like line	Conformal group of Minkowski space whose C^1 subgroup is G	Hausdorff, separable, first countable, but not regular and hence not metrizable
6	M^L: Finest topology which induces one dimensional Euclidean topology on every straight line	C^1 homeomorphisms form projective group generated by full linear group and translations	Weaker than M^F and M^T, Hausdorff, separable and first countable, not regular and hence not metrizable

Table 2. Fine topologies on space-times of general relativity.

Sr. No	Fine topology on space-time of GR	Diffeomorphism Group	Topological properties
1	HKM-path topology described by Hawking-King-McCarty (1976)	Conformal diffeomorphisms	Hausdorff, path connected and locally path connected, first countable, separable, but not normal or locally compact
2	Extended HKM-topology (Kim, 2006)	Conformal isomorphism group	Finer than Alexandrov topology
3	S-topology on Lorentz manifolds (Domiaty, 1985)	Conformal C^{∞} -diffeomorphisms	Hausdorff, first countable and separable, not regular and hence not metrizable, path connected and locally path connected
4	Zeeman -like fine topology in general relativity described by Göbel (1976)	Homeomorphism group with respect to Zeeman-like topology is the group of all homothetic transformations of V	Strongly causal space-times
5	Lindstrom (1978): Finest topology S_{nm} that induces the topology as a submanifold on time-like C^{n} -curves and on space-like C^{m} -curves	Group of Conformal C^{∞} -diffeomorphisms or group of all homothetic transformations of V	Space-time need not be strongly causal

realist view of space-time topology. Other philosophical issues about space-time have been discussed by D. Dieks and M. Redel in two volumes [36] [37].

References

[1] Zeeman, E. (1967) *Topology*, **6**, 161-170. http://dx.doi.org/10.1016/0040-9383(67)90033-X

[2] Dossena, G. (2007) *Journal of Mathematical Physics*, **48**, 113507. http://dx.doi.org/10.1063/1.2804758

[3] Nanda, S. (1971) *Journal of Mathematical Physics*, **12**, 394-401. http://dx.doi.org/10.1063/1.1665602

[4] Nanda, S. (1972) *Journal of Mathematical Physics*, **13**, 12-15. http://dx.doi.org/10.1063/1.1665841

[5] Nanda, S. (1979) *Journal of the Australian Mathematical Society*, **21**, 53-64. http://dx.doi.org/10.1017/S0334270000001910

[6] Williams, G. (1974) *Cambridge Philosophical Society*, **76**, 50 -509.

[7] Gobel, R. (1976) *Communications in Mathematical Physics*, **46**, 289-307. http://dx.doi.org/10.1007/BF01609125

[8] Gobel, R. (1976) *Journal of Mathematical Physics*, **17**, 845-853. http://dx.doi.org/10.1063/1.522984

[9] Hawking, S., King, A. and McCarthy, P. (1986) *Journal of Mathematical Physics*, **17**, 174-181. http://dx.doi.org/10.1063/1.522874

[10] Malament, D. (1977) *Journal of Mathematical Physics*, **18**, 1399-1404. http://dx.doi.org/10.1063/1.523436

[11] Lindstrom, U. (1977) Further Considerations on the Separating Topology for the Space-Times of General Relativity. Technical Report, Stockholm University (Sweden) Institute of Physics, NTIS Issue Number 197905, 20 p.

[12] Popvassilev, G. (1994) *Mathematica Pannonica*, **5**, 105-110.

[13] Kim, D. (2006) *Journal of Mathematical Physics*, **47**, Article ID: 072503. http://dx.doi.org/10.1063/1.2218981

[14] Agrawal, G. and Shrivastava, S. (2009) *Journal of Mathematical Physics*, **50**, Article ID: 053515. http://dx.doi.org/10.1063/1.3129188

[15] Agrawal, G. and Shrivastava, S. (2012) *ISRN Mathematical Physics*, **2012**, Article ID: 896156.

[16] Agrawal, G. and Sinha, S.P. (2014) Properties of A—Topology. Preprint.

[17] Low, R. (2010) *Classical and Quantum Gravity*, **27**, 107001. http://dx.doi.org/10.1088/0264-9381/27/10/107001

[18] Fullwood, D. (1992) *Journal of Mathematical Physics*, **33**, 2232-2241. http://dx.doi.org/10.1063/1.529644

[19] Mashford, J. (1981) *Journal of Mathematical Physics*, **22**, 1990-1993. http://dx.doi.org/10.1063/1.525145

[20] Penrose, R. (1972) Techniques of Differential Topology in Relativity. SIAM, Philadelphia. http://dx.doi.org/10.1137/1.9781611970609

[21] Ehlers, J. (1971) General Relativity and Kinetic Theory. In: Sachs, R.K., Ed., *Relativita generale a cosmologia*, Academic Press, New York, 1-70.

[22] Joshi, P.S. (1993) Global Aspects in Gravitation and Cosmology. Clarendon Press, Oxford.

[23] Garcia-Parrado, A. and Senovilla, J. (2003) *Classical and Quantum Gravity*, **20**, 625-664. http://dx.doi.org/10.1088/0264-9381/20/4/305

[24] Janardhan, S. and Saraykar, R.V. (2013) *Gravitation and Cosmology*, **19**, 42-53. http://dx.doi.org/10.1134/S0202289313010052

[25] Nanda, S. and Panda, H.K. (1975) *International Journal of Theoretical Physics*, **12**, 393-400. http://dx.doi.org/10.1007/BF01808166

[26] Struchiner, I. and Rosa, M. (2005) On Zeeman Topology in Kaluza-Klein and Gauge Theories. arXiv:math-ph/0504071v1

[27] Domiaty, R. (1985) *General Relativity and Gravitation*, **17**, 1165-1176. http://dx.doi.org/10.1007/BF00773622

[28] Domiaty, R. (1985) *Topology and Its Applications*, **20**, 39-46. http://dx.doi.org/10.1016/0166-8641(85)90033-1

[29] Budic, R. and Sachs, R.K. (1974) *Journal of Mathematical Physics*, **15**, 1302-1309. http://dx.doi.org/10.1063/1.1666812

[30] Huang, W. (1998) *Journal of Mathematical Physics*, **39**, 1637-1641.

[31] Peleska, J. (1984) *Aequationes Mathematicae*, **27**, 20-31. http://dx.doi.org/10.1007/BF02192656

[32] Sorkin, R. and Woolgar, E. (1996) *Classical and Quantum Gravity*, **13**, 1971-1994. http://dx.doi.org/10.1088/0264-9381/13/7/023

[33] Janardhan, S. and Saraykar, R.V. (2008) *Pramana-Journal of Physics*, **70**, 587-601.

[34] Minguzzi, E. (2009) *Communications in Mathematical Physics*, **290**, 239-248. http://dx.doi.org/10.1007/s00220-009-0794-4

[35] Heathcote, A. (1988) *British Journal for the Philosophy of Science*, **39**, 247-261. http://dx.doi.org/10.1093/bjps/39.2.247

[36] Dieks, D. and Redei, M., Eds. (2006) The Ontology of Spacetime. In: *Philosophy and Foundations of Physics Series*, Vol. 1, Elsevier, Amsterdam.

[37] Dieks, D. and Redei, M. (2008) The Ontology of Spacetime II. In: *Philosophy and Foundations of Physics Series*, Vol. 4, Elsevier, Amsterdam.

Journal of Modern Physics, 2016, 7, 642-655
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jmp
http://dx.doi.org/10.4236/jmp.2016.77064

The Three Postulates of the Theory of Everything

Ding-Yu Chung

Utica, Michigan, USA
Email: dy_chung@yahoo.com

Received 18 February 2016; accepted 25 April 2016; published 28 April 2016

Abstract

The three postulates of the posited dynamic and reversible theory of everything are: 1) the oscillating M-theory postulate for the oscillating matter structure, 2) the digital transitional Higgs-reversed Higgs fields postulate for the digital space structure, and 3) the reversible multiverse postulate for all physical laws and phenomena. The posited theory of everything based on the three postulates explains cosmology, the composition (baryonic matter, dark matter, and dark energy) in the universe, the periodic table of elementary particles (quarks, leptons, and bosons), the galaxy evolution, superconductivity, black hole, thermodynamic, and quantum mechanics. Oscillating M-theory is derived from oscillating membrane-string-particle whose space-time dimension number oscillates between 11D and 10D and between 10D and 4D. Space-time dimension number between 10 and 4 decreases with decreasing speed of light, decreasing vacuum energy, and increasing rest mass. The digital transitional Higgs-reversed Higgs fields are derived from digital attachment-detachment spaces which couple to particles. Under spontaneous symmetry breaking, the coupling of massless particle to zero-energy attachment space (the space for mass) produces the transitional nonzero-energy Higgs field-particle composite which under spontaneous symmetry restoring produces massive particle on zero-energy attachment space with the longitudinal component. The opposite of attachment space is detachment space as the space for kinetic energy and the nonzero-energy reverse Higgs field. The combination of n units of attachment space (denoted as 1) and n units of detachment space (denoted as 0) brings about the three digital structures: binary partition space $(1)_n(0)_n$, miscible space $(1 + 0)_n$, and binary lattice space $(1\ 0)_n$ to account for quantum mechanics, special relativity, and the force fields, respectively. In the third postulate, all physical laws and phenomena are permanently reversible in the multiverse, and temporary irreversible entropy increase is allowed. Our universe is an asymmetrical dual positive-energy-negative-energy universe where the positive-energy universe on attachment space absorbed the interuniversal void on detachment space to result in the combination of attachment space and detachment space, while the negative-energy universe did not absorb the interuniversal void, resulting in temporary irreversible entropy increase through reversibility breaking, symmetry violation, and low entropy beginning. Guided by the reversible negative-energy universe,

How to cite this paper: Chung, D.-Y. (2016) The Three Postulates of the Theory of Everything. *Journal of Modern Physics*, **7**, 642-655. http://dx.doi.org/10.4236/jmp.2016.77064

our dual universe is a globally reversible cyclic dual universe.

Keywords

The Theory of Everything, M-Theory, Higgs Field, Reverse Higgs Field, Multiverse, Cosmology, Matter Structure, Space Structure, Entropy, Thermodynamic, Cyclic Universe, Interuniversal Void

1. Introduction

In this paper, the posited theory of everything involves the dynamic and reversible process of continuingly changing matters and spaces in the reversible multiverse. There are three postulates of the posited dynamic and reversible theory of everything. The first postulate is the oscillating M-theory postulate for the oscillating matter structure. In conventional M-theory, space-time dimensional number (D) is fixed. As a result, the observed 4D results from the compactization of the extra space dimensions in 11D M-theory. However, there is no experimental proof for compactized extra space dimensions, and there are numerous ways for the compactization of the extra space dimensions [1]. In the first postulate, oscillating M-theory is derived from oscillating membrane-string-particle whose space-time dimension number oscillates between 11D and 10D and between 10D and 4D. There are no extra space dimensions and compactization. Space-time dimension number between 10 and 4 decreases with decreasing speed of light, decreasing vacuum energy, and increasing rest mass. 4D has zero vacuum energy. The theoretical calculated masses based on the oscillating M-theory postulate for all elementary particles including quarks, leptons, gauge bosons, the Higgs boson, and the knees-ankles-toe in cosmic rays are in good agreements with the observed values [2]-[4]. Different universes at different times in the multiverse have different space-time dimension numbers.

The second postulate of the theory of everything is the digital transitional Higgs-reversed Higgs fields postulate for the digital space structure. In the convention model, the Higgs field is a nonzero-energy field that is hypothesized to exist permanently in the universe. The problem with such nonzero-energy field is the cosmological constant problem from the huge gravitational effect by the nonzero-energy Higgs field [5]. In the conventional model, the nonzero-energy Higgs field is derived from zero-energy ground state space under spontaneous symmetry breaking. In this paper, to avoid the cosmological problem from the huge gravitational effect by the nonzero-energy Higgs field is to make the Higgs field a transitional field which exists momentarily and to make zero-energy ground state space a permanent zero-energy ground state space. In the second postulate for the digital space structure, such zero-energy ground state space is zero-energy "attachment space" which attaches particles to account for the longitudinal component, mass, and reversible movement. Unlike the conventional model, attachment space actively couples to massless particle. Under spontaneous symmetry breaking, the coupling of massless particle to zero-energy attachment space produces the transitional nonzero-energy Higgs field-particle composite which under spontaneous symmetry restoring produces massive particle on zero-energy attachment space with the longitudinal component. The opposite of attachment space is zero-energy detachment space which detaches particles to account for irreversible kinetic energy. Under spontaneous symmetry breaking, the coupling of massive particle to zero-energy detachment space produces the transitional nonzero-energy reverse Higgs field-particle composite which under spontaneous symmetry restoring produces massless particle on zero-energy detachment space without the longitudinal component. The combination of n units of attachment space (denoted as 1) and n units of detachment space brings about the three digital structures: binary partition space $(1)_n(0)_n$, miscible space $(1 + 0)_n$, and binary lattice space $(1\ 0)_n$ to account for quantum mechanics, special relativity, and the force fields, respectively [6] [7]. Different universes at different times in the multiverse have different spaces. In the second postulate, the digital transitional Higgs-reversed Higgs fields are digital attachment-detachment spaces which couple to particles.

The third postulate of the theory of everything is the reversible multiverse postulate for all physical laws and phenomena. In the second law of thermodynamics, the entropy (a measure of the disorder of a system) of an isolated system can increase, but not decrease. In other words, the entropy of a closed system will never decrease into the future. There are two mysteries about this irreversible entropy increase as described in "From Eternity to Here: The Quest for the Ultimate Theory of Time" by Sean Carroll [8]. Firstly, this irreversible en-

tropy increase of a macroscopic collection of particles is different from all microscopic reversible processes where for every allowed process there exists a time-reversed process that is also allowed. Secondly, the universe started with the very low entropy state as the inflation-Big Bang in a very small space, not with the high entropy state near equilibrium state in a large space. Such low entropy beginning is a mystery. In this paper, the mysteries of the irreversible entropy increase are explained by the third postulate, the reversible multiverse postulate. In the reversible multiverse postulate, all physical laws and phenomena are permanently reversible, and temporary irreversibility of entropy increase is allowed through reversibility breaking, symmetry violation, and low entropy beginning [7] [9] [10]. The temporary irreversible entropy increase is shown in our universe. Our universe is the dual asymmetrical positive-energy-negative-energy universe where the positive-energy universe on attachment space absorbed the interuniversal void on detachment space to result in the combination of attachment space and detachment space, and the negative-energy universe did not absorb the interuniversal void. Irreversible kinetic energy from detachment space is the source of irreversible entropy increase, so the positive-energy universe is irreversible, while the negative-energy universe without irreversible kinetic energy from detachment space is reversible. As a result, our whole dual universe is reversible.

The posited theory of everything based on the three postulates explains: 1) cosmology [7] [9]-[11], 2) the composition (baryonic matter, dark matter, and dark energy) in the universe [9] [11] [12], 3) the periodic table of elementary particles [2]-[4] for all elementary particles (quarks, leptons, gauge bosons, the Higgs boson, and the knees-ankles-toe in cosmic rays), 4) the galaxy evolution [13] [14], 5) superconductivity [15], and 6) black hole [16] [17], 7) thermodynamics [7], and 8) quantum mechanics [6] [7]. The theoretical results are in good agreement with the observations, and the calculated values from cosmology, the composition in the universe, and the periodic table of elementary particles are in good agreement with the observed values.

The purpose of this paper is to describe all the three postulates of the theory of everything. Section 2 describes the oscillating reversible M-theory postulate for the oscillating matter structure. Section 3 explains the digital transitional Higgs-reverse Higgs fields postulate for the digital space structure. Cosmology and the reversible multiverse postulate for all physical laws and phenomena are explained in Section 4.

2. The Oscillating M-Theory Postulate for the Oscillating Matter Structure

In this paper, the posited theory of everything is the dynamic and reversible process of continuingly changing matters and spaces in the reversible multiverse. Instead of remaining fixed, space-time dimension numbers of matters oscillate. Instead of being passive, spaces actively couple to particles. There are two different spaces whose digital combinations are different under different conditions. In the midst of continuously changing matters and spaces, the multiverse is simple and neat, because the multiverse is permanently reversible to exclude all irreversible physical laws and phenomena. However, temporary irreversible entropy increase is allowed through reversibility breaking, symmetry violation, and low entropy beginning. We live in the universe with such temporary irreversible entropy increase.

The first postulate of the theory of everything is the oscillating M-theory for the oscillating matter structure. M-theory with eleven-dimensional membrane is an extension of string theory with ten-dimensional string, in contrast to the observed 4D. In conventional M-theory with fixed space-time dimension number, the explanation of the hidden extra space dimensions is the compactization of the extra space dimensions, so space-time appears to be 4D. In the first postulate, oscillating M-theory is derived from oscillating membrane-string-particle whose space-time dimension number oscillates between 11D and 10D and between 10D and 4D dimension by dimension reversibly. There is no compactization. Matters in oscillating M-theory include 11D membrane (2_{11}) as membrane (denoted as 2 for 2 space dimensions) in 11D, 10D string (1_{10}) as string (denoted as 1 for 1 space dimension) in 10D, and variable D particle ($0_{4 \text{ to } 11}$) as particle (denoted as 0 for 0 space dimension) in 4D to 11D.

As described previously [7], the QVSL (quantum varying speed of light) transformation transforms both space-time dimension number (D) and mass dimension number (d). In the QVSL transformation, the decrease in the speed of light leads to the decrease in space-time dimension number and the increase of mass in terms of increasing mass dimension number from 4 to 10,

$$c_{\mathrm{D}} = c/\alpha^{\mathrm{D}-4}, \tag{1a}$$

$$E = M_0 \cdot \left(c^2/\alpha^{2(\mathrm{D}-4)}\right) \tag{1b}$$

$$= \left(M_0 / \alpha^{2(\mathrm{d}-4)} \right) \cdot c^2. \tag{1c}$$

$$c_{\mathrm{D}} = c_{\mathrm{D}-n} / \alpha^{2n}, \tag{1d}$$

$$M_{0,\mathrm{D},\mathrm{d}} = M_{0,\mathrm{D}-n,\mathrm{d}+n} \alpha^{2n}, \tag{1e}$$

$$\mathrm{D},\mathrm{d} \xrightarrow{\mathrm{QVSL}} \left(\mathrm{D} \mp n\right), \left(\mathrm{d} \pm n\right) \tag{1f}$$

$$E_{\mathrm{vacuum,D}} = E - M_{0,\mathrm{D}} c^2, \tag{1g}$$

where c_{D} is the quantized varying speed of light in space-time dimension number, D, from 4 to 10, c is the observed speed of light in the 4D space-time, α is the fine structure constant for electromagnetism, E is energy, M_0 is rest mass, D is the space-time dimension number from 4 to 10, d is the mass dimension number from 4 to 10, n is an integer, and E_{vacuum} = vacuum energy. For example, in the QVSL transformation, a particle with 10D4d is transformed to a particle with 4D10d from Equation (1f). Calculated from Equation (1e), the rest mass of 4D10d is $1/\alpha^{12} \approx 137^{12}$ times of the mass of 10D4d. In terms of rest mass, 10D space-time has 4d with the lowest rest mass, and 4D space-time has 10d with the highest rest mass. Rest mass decreases with increasing space-time dimension number. The decrease in rest mass means the increase in vacuum energy ($E_{\mathrm{vacuum,D}}$), so vacuum energy increases with increasing space-time dimension number. The vacuum energy of 4D particle is zero from Equation (1g). 11D membrane and 10D string are equal in the speed of light, rest mass, and vacuum energy. Since the speed of light for >4D particle is greater than the speed of light for 4D particle, the observation of >4D superluminal particles by 4D particles violates casualty. Thus, >4D particles are hidden particles with respect to 4D particles. Particles with different space-time dimensions are transparent and oblivious to one another, and separate from one another if possible.

As described previously [7], in oscillating M-theory, there are two different reversible oscillations: the oscillation between 11D and 10D and the oscillation between 10D and 4D. The reversible oscillation between 11D membrane and 10D string is described in Section 4.The transformation during the oscillation between 10D particle and 4D involves the stepwise two-step transformation: the QVSL transformation and the varying supersymmetry transformation from 10D4d to 4D4d. The QVSL transformation involves the transformation of space-time dimension, D whose mass increases with decreasing D for the decrease in vacuum energy. The varying supersymmetry transformation involves the transformation of the mass dimension number, d whose mass decreases with decreasing d for the fractionalization of particle, as follows.

stepwise two-step varying transformation

$$\begin{aligned} &(1) \quad \mathrm{D},\mathrm{d} \xleftrightarrow{\mathrm{QVSL}} \left(\mathrm{D} \mp 1\right), \left(\mathrm{d} \pm 1\right) \\ &(2) \quad \mathrm{D},\mathrm{d} \xleftrightarrow{\text{varying supersymmetry}} \mathrm{D}, \left(\mathrm{d} \pm 1\right) \end{aligned} \tag{2}$$

The repetitive stepwise two-step transformation between 10D4d and 4D4d as follows.

$$10\mathrm{D4d} \leftrightarrow 9\mathrm{D5d} \leftrightarrow 9\mathrm{D4d} \leftrightarrow 8\mathrm{D5d} \leftrightarrow \cdots \leftrightarrow 4\mathrm{D5d} \leftrightarrow 4\mathrm{D4d} \tag{3}$$

In this two-step transformation, the transformation from 10D4d to 9D5d involves the QVSL transformation as in Equation (1d). Calculated from Equation (1e), the mass of 9D5d is $1/\alpha^2 \approx 137^2$ times of the mass of 10D4d. The transformation of 9D5d to 9D4d involves the varying supersymmetry transformation. In the normal supersymmetry transformation, the repeated application of the fermion-boson transformation carries over a boson (or fermion) from one point to the same boson (or fermion) at another point at the same mass. In the "varying supersymmetry transformation", the repeated application of the fermion-boson transformation carries over a boson from one point to the boson at another point at different mass dimension number in the same space-time number. The repeated varying supersymmetry transformation carries over a boson B_d into a fermion F_d and a fermion F_d to a boson B_{d-1}, which can be expressed as follows

$$M_{\mathrm{d,F}} = M_{\mathrm{d,B}} \alpha_{\mathrm{d,B}}, \tag{4a}$$

$$M_{\mathrm{d}-1,\mathrm{B}} = M_{\mathrm{d,F}} \alpha_{\mathrm{d,F}}, \tag{4b}$$

where $M_{\mathrm{d,B}}$ and $M_{\mathrm{d,F}}$ are the masses for a boson and a fermion, respectively, d is the mass dimension number, and $\alpha_{\mathrm{d,B}}$ or $\alpha_{\mathrm{d,F}}$ is the fine structure constant that is the ratio between the masses of a boson and its fermionic

partner. Assuming α's are the same, it can be expressed as

$$M_{\mathrm{d,B}} = M_{\mathrm{d+1,B}}\alpha_{\mathrm{d+1}}^{2} \tag{4c}$$

The mass of 9D4d is $\alpha^2 \approx (1/137)^2$ times of the mass of 9D5d through the varying supersymmetry transformation. The transformation from a higher mass dimensional particle to the adjacent lower mass dimensional particle is the fractionalization of the higher dimensional particle to the many lower dimensional particles in such way that the number of lower dimensional particles becomes

$$N_{\mathrm{d-1}} = N_{\mathrm{d}}/\alpha^{2} \approx N_{\mathrm{d}}(137)^{2} \tag{4d}$$

The fractionalization also applies to D for 10D4d to 9D4d, so

$$N_{\mathrm{D-1}} = N_{\mathrm{D}}/\alpha^{2} \tag{4e}$$

Since the supersymmetry transformation involves translation, this stepwise varying supersymmetry transformation leads to a translational fractionalization, resulting in the cosmic expansion. Afterward, the QVSL transformation transforms 9D4d into 8D5d with a higher mass. The two-step transformation repeats until 4D4d, and then reverses stepwise back to 10D4d for the cosmic contraction. The oscillation between 10D and 4D results in the reversible cyclic fractionalization-contraction for the reversible cyclic expansion-contraction of the universe which does not involve irreversible kinetic energy.

3. The Digital Transitional Higgs-Reverse Higgs Fields Postulate for the Digital Space Structure

In the conversion model, under spontaneous symmetry breaking, zero-energy ground state space turns into the nonzero-energy scalar Higgs Field which exists permanently in the universe. The problem with such nonzero-energy field is the cosmological constant problem from the huge gravitational effect by the nonzero-energy Higgs field [5]. The coupling of massless particle to the Higgs field produces the transitional nonzero-energy Higgs field-particle composite which under spontaneous symmetry restoring produces the massive particle with the longitudinal component on zero-energy ground state space without the Higgs field as follows.

$$\begin{aligned}&\text{zerro-energy groud state space} \xrightarrow{\text{spontaneous symmetry breaking}} \text{nonzero-energy scalar Higgs field}\\ &\xrightarrow{\text{massless particle}} \left[\text{the transitional nonzero- energy Higgs field-particle composite}\right]\\ &\xrightarrow{\text{spontaneous symmetry restoring}} \text{massive particle with the longitudinal component on}\\ &\text{zero-energy ground state space without the Higgs field}\end{aligned} \tag{5}$$

In this paper, to avoid the cosmological problem from the huge gravitational effect by the nonzero-energy Higgs field is to make the Higgs field a transitional field which exists momentarily and to make zero-energy ground state space a permanent zero-energy ground state space which exists permanently in the universe. In the second postulate for the digital structure, such zero-energy ground state space is zero-energy "attachment space" which attaches particles to account for the longitudinal component, mass, and reversible movement. Unlike the conventional model, attachment space actively couples to massless particle. Under spontaneous symmetry breaking, the coupling of massless particle to zero-energy attachment space produces the transitional nonzero-energy Higgs field-particle composite which under spontaneous symmetry restoring produces massive particle on zero-energy attachment space with the longitudinal component without the Higgs field as follows.

$$\begin{aligned}&\text{massless particle} + \text{zero-energy attachment space} \xrightarrow{\text{spontaneous symmetry breaking}}\\ &\left[\text{the transitional non-zero energy Higgs field-particle composite}\right]\\ &\xrightarrow{\text{spontaneous symmetry restoring}} \text{massive particle with the longitudinal component}\\ &\text{on zero-energy attachment space without the Higgs field}\end{aligned} \tag{6}$$

The opposite of attachment space is zero-energy detachment space which detaches particles to account for irreversible kinetic energy. Unlike the conventional model, detachment space actively couples to massive particle. Under spontaneous symmetry breaking, the coupling of massive particle to zero-energy detachment space produces the transitional nonzero-energy reverse Higgs field-particle composite which under spontaneous symme-

try restoring produces massless particle on zero-energy detachment space without the longitudinal component without the reverse Higgs field as follows.

$$\begin{aligned}&\text{massive particle} + \text{zero-energy detachment space} \xrightarrow{\text{spontaneous symmetry breaking}}\\&\left[\text{the transitional nonzero-energy reverse Higgs field- particle composite}\right]\\&\xrightarrow{\text{spontaneous symmetry restoring}} \text{massless particle without the longitudinal}\\&\text{compnenent on zero-energy detachment space without the reverse Higgs field}\end{aligned} \tag{7}$$

For the electroweak interaction in the Standard model where the electromagnetic interaction and the weak interaction are combined into one symmetry group, under spontaneous symmetry breaking, the coupling of the massless weak W, weak Z, and electromagnetic A (photon) bosons to zero-energy attachment space produces the transitional nonzero-energy Higgs fields-bosons composites which under partial spontaneous symmetry restoring produce massive W and Z bosons on zero-energy attachment space with the longitudinal component without the Higgs field, massless A (photon), and massive Higgs boson as follows.

$$\begin{aligned}&\text{massless WZ} + \text{zero-energy WZ attachment space} + \text{massless A}\\&+\text{zero-energy A attachment space A} \xrightarrow{\text{spontaneous symmetry breaking}}\\&\left[\text{the transitional nonzero-energy WZ Higgs field -WZ composite}\right]\\&+\left[\text{nonzero-energy A Higgs field -A composite}\right] \xrightarrow{\text{partial spontaneous symmetry restoring}}\\&\text{massive WZ with the longitudinal component on attachment space without}\\&\text{the Higgs field} + \text{massless A} + \text{the nonzero energy massive Higgs boson}\end{aligned} \tag{8}$$

From the periodic table of elementary particles, the theoretical calculated mass of the Higgs boson is 128.8 GeV in good agreements with the observed 125 or 126 GeV [3]. In terms of mathematical expression, the conventional permanent Higgs field model and the posited transitional Higgs field model are identical. The interpretations of the mathematical expression are different for the permanent Higgs field model and the transitional Higgs field model. The transitional Higgs field model avoids the cosmological problem in the permanent Higgs field model.

As shown later, our universe is the dual asymmetrical positive-energy-negative-energy universe where the positive-energy universe on attachment space absorbed the interuniversal void on detachment space to result in the combination of attachment space and detachment space, and the negative-energy universe did not absorb the interuniversal void. The combination of n units of attachment space as 1 and n units of detachment space as 0 brings about three different digital space structures: binary partition space, miscible space, or binary lattice space as below.

$$\begin{aligned}&(1)_n \quad + \quad (0)_n \xrightarrow{\text{combination}} (1)_n(0)_n, \quad (1+0)_n, \quad \text{or} \quad (1\,0)_n\\&\text{attachment space} \quad \text{detachment space} \quad \text{binary partition space,} \quad \text{miscible space,} \quad \text{binary lattice space}\end{aligned} \tag{9}$$

Binary partition space, $(1)_n(0)_n$, consists of two separated continuous phases of multiple quantized units of attachment space and detachment space. In miscible space, $(1 + 0)_n$, attachment space is miscible to detachment space, and there is no separation of attachment space and detachment space. Binary lattice space, $(1\ 0)_n$, consists of repetitive units of alternative attachment space and detachment space.

Binary partition space is the space for wavefunction in quantum mechanics. In wavefunction,

$$\left|\Psi\right\rangle = \sum_{i=1}^{n} c_i \left|\phi_i\right\rangle \tag{10}$$

Each basis element, $\left|\phi_i\right\rangle$, has both attachment space and detachment space as binary partition space. Neither attachment space nor detachment space is zero in binary partition space for a basic element. The measurement in the uncertainty principle in quantum mechanics is essentially the measurement of attachment space size and momentum from the detachment space in binary partition space: large momentum from detachment space has small non-zero attachment space size, while large attachment space size has low non-zero momentum from detachment space. In binary partition space, an entity is both in constant motion as wave for detachment space and in stationary state as a particle for attachment space, resulting in the wave-particle duality.

In binary partition space, for every detachment space, there is its corresponding adjacent attachment space. Thus, no part of the mass-energy can be irreversibly separated from binary partition space, and no part of a different mass-energy can be incorporated in binary partition space. Binary partition space represents coherence as wave function. Binary partition space is for coherent system. Any destruction of the coherence by the addition of a different mass-energy to the mass-energy causes the collapse of binary partition space into miscible space. The collapse is a phase transition from binary partition space to miscible space.

$$\underset{\text{binary partition space}}{(0)_n\,(1)_n} \xrightarrow{\text{collapse}} \underset{\text{miscible space}}{(0+1)_n} \tag{11}$$

The information in miscible space is contributed by the miscible combination of both attachment space and detachment space, so information can no longer be non-localized. Any value in miscible space is definite. All observations in terms of measurements bring about the collapse of wavefunction, resulting in miscible space that leads to eigenvalue as definite quantized value. Such collapse corresponds to the appearance of eigenvalue, E, by a measurement operator, H, on a wavefunction, Ψ.

$$H\Psi = E\Psi \tag{12}$$

In miscible space, attachment space is miscible to detachment space, and there is no separation of attachment space and detachment space. In miscible space, attachment space contributes zero speed, while detachment space contributes the speed of light. For a moving massive particle consisting of a rest massive part and a massless part, the massive part with rest mass, m_0, is in attachment space, and the massless part with kinetic energy, K, is adjacent to detachment space. The combination of the massive part in attachment space and massless part in detachment leads to the propagation speed in between zero and the speed of light. To maintain the speed of light constant for a moving particle, the time (t) in moving particle has to be dilated, and the length (L) has to be contracted relative to the rest frame.

$$\begin{aligned} t &= t_0 \Big/ \sqrt{1-\upsilon^2/c^2} = t_0\gamma, \\ L &= L_0 \,/\gamma, \\ E &= K + m_0c^2 = \gamma m_0c^2 \end{aligned} \tag{13}$$

where $\gamma = 1\Big/\left(1-v^2/c^2\right)^{1/2}$ is the Lorentz factor for time dilation, and length contraction, E is the total energy, and K is the kinetic energy.

Bounias and Krasnoholovets [18] propose that the reduction of dimension can be done by slicing dimension, such as slicing 3 space dimension object (block) into infinite units of 2 space dimension objects (sheets). As shown in Section 4, the positive-energy 10D4d particle universe as our observable universe with high vacuum energy was transformed into the 4D10d universe with zero vacuum energy at once, resulting in the inflation. During the Big Bang following the inflation, the 10d (mass dimension) particle in attachment space denoted as 1 was sliced by detachment space denoted as 0. For example, the slicing of 10d particle into 4d particle is as follows.

$$\underset{\text{10d particle}}{1_{10}} \xrightarrow{\text{slicing}} \underset{\text{4d core particle}}{1_4} \qquad \underset{\text{binary lattice space}}{\sum_{\mathrm{d}=5}^{10}\left(0_4\,1_4\right)_{n,\mathrm{d}}} \tag{14}$$

where 1_{10} is 10d particle, 1_4 is 4d particle, d is the mass dimension number of the dimension to be sliced, n as the number of slices for each dimension, and $(0_4\,1_4)_n$ is binary lattice space with repetitive units of alternative 4d attachment space and 4d detachment space. For 4d particle starting from 10d particle, the mass dimension number of the dimension to be sliced is from d = 5 to d = 10. Each mass dimension is sliced into infinite quantized units (n = ∞) of binary lattice space, $(0_4\,1_4)_\infty$. For 4d particle, the 4d core particle is surrounded by 6 types (from d = 5 to d = 10) of infinite quantized units of binary lattice space. Such infinite quantized units of binary lattice space represent the infinite units (n = ∞) of separate virtual orbitals in a gauge force field, while the dimension to be sliced is "dimensional orbital" (DO), representing a type of gauge force field. The mass-energy in each dimensional orbital increases with the number of dimension number, and the lowest dimension orbital with d = 5 has the lowest mass-energy [2] [11]. 10d particle was sliced into six different particles: 9d, 8d, 7d, 6d, 5d, and 4d equally by mass. Baryonic matter is 4d, while dark matter consists of the other five types of particles (9d, 8d, 7d, 6d, and 5d).

$$\begin{aligned} &10\text{D4d} \xrightarrow{\text{the inflation}} 4\text{D10d} \xrightarrow{\text{the Big Bang}} \text{baryonic matter } (4\text{D4d}) \\ &+\text{dark matter } (4\text{D5d, 4D6d, 4D7d, 4D8d, 4D9d}) + \text{kinetic energy} \end{aligned} \tag{15}$$

The mass ratio of dark matter to baryonic matter is 5 to 1. At 72.8% dark energy, the calculated values for baryonic matter and dark matter (with the 1:5 ratio) are 4.53% (=(100 – 72.8)/6) and 22.7% (=4.53 × 5), respectively, in excellent agreement with observed 4.56% and 22.7%, respectively [11] [19]. The dimensional orbitals of baryonic matter provide the base for the periodic table of elementary particles to calculate accurately the masses of all elementary particles, including quarks, leptons, gauge bosons, the Higgs boson, and the knees-ankles-toe in cosmic rays [2]-[4]. The calculated masses of all elementary particles are in good agreement with the observed values. For examples, the calculated mass of top quark and the Higgs boson are 176.5 GeV and 128.8 GeV in good agreement with the observed 173.34 GeV and 125 or 126 GeV, respectively.

The lowest dimensional orbital is for electromagnetism. Baryonic matter with maximum number of gauge force fields (dimensional orbitals) is the only one with the lowest dimensional orbital for electromagnetism. With higher dimensional orbitals, dark matter does not have this lowest dimensional orbital [6] [12]. Without electromagnetism, dark matter cannot emit light, and is incompatible to baryonic matter with electromagnetism, like the incompatibility between oil and water. Derived from the incompatibility between dark matter and baryonic matter, the modified interfacial gravity (MIG) between homogeneous baryonic matter region and homogeneous dark matter region to separate baryonic matter region and dark matter region explains galaxy evolution and unifies the CDM (Cold Dark Matter) model, MOG (Modified Gravity), and MOND (Modified Newtonian Dynamics) [13] [14]. The digital space structure based on the combination of binary partition space and binary lattice space explains superconductivity [15] and superstar without singularity to replace black hole with singularity [16] [17]. Singularity is permanently irreversible by losing information permanently, forbidden in the reversible multiverse.

4. Cosmology and the Reversible Multiverse Postulate for All Physical Laws and Phenomena

In the reversible multiverse postulate, all physical laws and phenomena are permanently reversible, and temporary irreversibility of entropy increase is allowed through reversibility breaking, symmetry violation, and low entropy beginning. The reversible multiverse postulate can be explained by cosmology. The multiverse has been studied extensively. For example, Brian Greene [20] described the nine types of the multiverse which produce complicated collections of universes. The reversible multiverse postulate posits a simple version of the multiverse. The posited simple multiverse is the reversible multiverse that excludes any permanently irreversible physical laws and phenomena. In the reversible multiverse, the allowed universes have to be reversible cyclic universes with permanently reversible physical laws and phenomena, resulting in only limited types of allowed universes. Temporary irreversible universes are allowed. One irreversible phenomenon which is not allowed is the collision of expanding universes. The collision of expanding universes which have the inexhaustible resource of space-time to expand is permanently irreversible due to the impossibility to reverse the collision of expanding universes. To prevent the collision of expanding universes, every universe is surrounded by the interuniversal void that is functioned as the permanent gap among universes. The interuniversal void has zero-energy, zero space-time, and zero vacuum energy, and detachment space only, while universe has nonzero-energy, the inexhaustible resource of space-time to expand, zero or/and non-zero vacuum energy, and attachment space with or without detachment space. The detachment space of the interuniversal void has no space-time, so it cannot couple to particles with space-time in universes, but it prevents the advance of expanding universes into the interuniversal void to avoid the collision of expanding universes.

A zero-sum energy dual universe of positive-energy universe and negative-energy universe can be created in the zero-energy interuniversal void, and the new dual universe is again surrounded by the interuniversal void to avoid the collision of universes. Under symmetry, the new positive-energy universe and the new negative-energy universe undergo mutual annihilation to reverse to the interuniversal void immediately. Our universe is the dual asymmetrical positive-energy-negative-energy universe where the positive-energy universe on attachment space absorbed the interuniversal void on detachment space to result in the combination of attachment space and detachment space, and the negative-energy universe did not absorb the interuniversal void. Within the positive-energy universe, the absorbed detachment space with space-time can couple to particles in the positive-

energy universe to result in massless particles with irreversible kinetic energy. The formation of our universe involves symmetry violation between the positive-energy universe and the negative energy universe. Irreversible kinetic energy from detachment space is the source of irreversible entropy increase, so the positive-energy universe is locally irreversible, while the negative-energy universe without irreversible kinetic energy from detachment space is locally reversible. The locally reversible negative-energy universe guides the reversible process of the dual universe. As a result, our whole dual universe is globally reversible. Our dual universe is the globally reversible cyclic dual universe as shown in Equation (16) and **Figure 1** for the evolution of our universe.

$$\begin{aligned}&\text{the dual positive energy} - \text{negative energy universe} \xrightarrow{\text{symmetry breaking} - \text{reversibility breaking}}\\ &\text{the locally irreversible positive energy universe} + \text{the locally reversible negative energy universe}\\ &\to \text{the globally reversible cyclic dual universe}\end{aligned} \tag{16}$$

The four reversible steps in the globally reversible cyclic dual universe are 1) the formation of the 11D membrane dual universe, 2) the formation of the 10D string dual universe, 3) the formation of the 10D particle dual universe, and 4) the formation of the asymmetrical dual universe.

1) The formation of the 11D membrane dual universe

As described previously [7] [9]-[11], the reversible cyclic universe starts in the zero-energy interuniversal void, which produces the dual universe of the positive-energy 11D membrane universe and the negative-energy 11D membrane universe as in **Figure 1**. In some dual 11D membrane universes, the 11D positive-energy membrane universe and the negative-energy 11D membrane universe coalesce to undergo annihilation and to return to the interuniversal void as in **Figure 1**.

2) The formation of the 10D string dual universe

Under the reversible oscillation between 11D and 10D, the positive-energy11D membrane universe and the negative-energy 11D membrane universe are transformed into the positive-energy 10D string universe and the negative-energy 10D string universe, respectively, as in **Figure 1**. The positive-energy 11D membrane universe is transformed into the positive-energy 10D string universe as in Equations (17a) and (17b).

$$\begin{aligned}&\text{The RS1 Membrane Transformation}\\ &\text{step}\,1: 2_{11} \xrightarrow{\text{from 11D membrane to 10D string}} 1_{10} \text{ in the 11D AdS space}\\ &\text{step}\,2: 2(1_{10}) \xrightarrow{\text{the close string vibration}} 1_{10}\,0_{10} = 1_{10}\,g_e \text{ in the 11D AdS space}\end{aligned} \tag{17a}$$

$$2(2_{11}) \xleftrightarrow{\text{the close string and the open string vibrations}} (s1_{10})\,g_e \tag{17b}$$

where 2_{11} is membrane (denoted as 2) in 11D, s is the pre-strong force, 1_{10} is string (denoted as 1) in 10D, 0_{10} is particle (denoted as 0) in 10D, AdS is anti-de Sitter, and g_e is the external graviton.

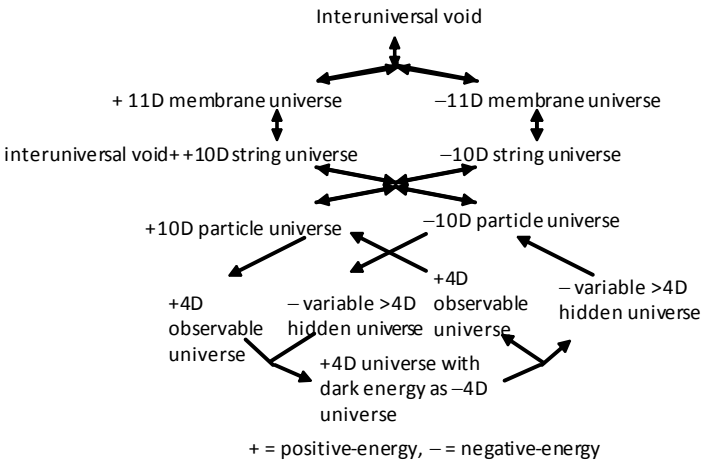


Figure 1. The globally reversible cyclic dual universe.

According Randall and Sundrum, the RS1 (Randall-Sundrum model 1) in an anti-de Sitter (AdS) space consists of one brane with extremely low graviton's probability function and another brane with extreme high graviton's probability function [21] [22]. The formation of the 10D string dual universe involves the RS1. As shown in Equation (17a), one of the possible membrane transformations from the 11D membrane to the 10D string is the RS1 membrane transformation which involves two steps. In the Step 1, the extra spatial dimension of the 11D membrane in the transformation from the 11D membrane to the 10D string becomes the spatial dimension transverse to the string brane in the bulk 11D anti-de Sitter space [21]. This transformation is derived from the transformation from membrane to string. In the transformation from the two-dimensional membrane to the one-dimensional string, the extra spatial dimension of the two-dimensional membrane on the x-y plane becomes the x-axis transverse to the one-dimensional string on the y-axis in the two-dimensional x-y space. In the Step 2, for the RS1 membrane transformation, two string branes are combined into the combined string brane. The external 10D particles generated by the close string vibration of the combined string brane are the 10D external gravitons which form the external graviton brane as the Gravity brane (Planck Plane) in the RS1 of the Randall-Sundrum model [21] [22]. As in the RS1 of the Randall-Sundrum model, the two branes with equal mass-energy in the 11D anti-de Sitter space are the string brane with weak gravity and the external graviton brane with strong gravity. The weak gravity in the string brane is the predecessor of the observed weak gravity generated during the Big Bang [2] [9]. The external graviton in the external graviton brane is the predecessor of a part of the observed dark energy [11]. The 10D string brane and the 10D external graviton brane correspond to the predecessors of the observed universe (without dark energy) and a part of observed dark energy, respectively [2] [9]. The reverse transformation from 10D to 11D is the RS1 string transformation.

In Equation (17b), the particles generated from the 10D open string vibration are the 10D particles for the pre-strong force (denoted as *s*) in addition to the external graviton from the close string vibration in the 11D AdS. According to Maldacena, the AdS/CFT correspondence is a correspondence between quantum gravity in AdS space and quantum field theory of conformal field theory (CFT) in one dimension lower [23]. The AdS/CFT correspondence describes Equation (17b) as the correspondence between the external graviton in the 11D AdS and the pre-strong force of 10D CFTin one dimension lower. The pre-strong force is the same for all strings without positive or negative sign. This pre-strong force is the prototype of the observed strong force generated during the Big Bang [2] [9].

In the negative universe through symmetry, the 11D anti-membrane (2_{-11}) is transformed into10D antistring (1_{-10}) with external anti-graviton g_e and the pre-strong force s as follows.

$$2\left(2_{-11}\right) \leftrightarrow \left(s1_{-10}\right)\overline{g_e} \tag{18}$$

The dual universe of the positive-energy 10D string universe with n units of $(1_{10})_n$ and the negative-energy 10D string universe with n units of $(1_{-10})_n$ is as follows.

$$\left(\left(s1_{10}\right)g_e\right)_n\left(\overline{g_e}\left(s1_{-10}\right)\right)_n \tag{19}$$

There are four equal regions: the positive-energy 10D string universe, the external graviton, the external anti-graviton, and the negative-energy 10D string universe.

Some dual 10D string universes return to the dual 11D membrane universes under the reversible oscillation between 11D and 10D. Alternatively, under symmetry violation as in the case of our universe, the positive-energy 10D string universe absorbs the interuniversal void, while the negative-energy 10D string universe does not absorb the interuniversal void. The interuniversal void has zero vacuum energy. In our universe, the absorption of the interuniversal void by the positive-energy 10D string universe forced the positive-energy 10D universe with high vacuum energy to be transformed into the universe with zero vacuum energy that was the vacuum energy of the 4D universe. However, the transformation from 10D to 4D was not immediate, because the strings had to be 10D, and it could not be transformed into 4D, therefore, strings had to be transformed into particles that allowed the change of its dimension number freely to accommodate the transformation from the 10D universe to the 4D universe driven by the absorption of the interuniversal void.

3) The formation of the 10D particle dual universe

As described previously [9], the transformation of strings into particles came from the emergence of positive charge and negative charge that allowed the mutual annihilation of positively charged 10D strings and negatively charged 10D antistrings in the 10D string universes to produce positively charged 10D particles and negatively

pre-charged 10D antiparticles in the 10D particle universes as follows.

$$\left(\left(s0_{10}e^{+}e^{-}0_{-10}s\right)g_{e}\right)_{n}\left(\overline{g_{e}}\left(s0_{10}e^{+}e^{-}0_{-10}s\right)\right)_{n}, \tag{20}$$

where s and e are the pre-strong force and the pre-charged force in the flat space, g_e is the external graviton, g_e is the external graviton, and $0_{10}0_{-10}$ is the particle-antiparticle. There are four equal regions: the 10D positive-energy particle universe, the external graviton, the 10D negative-energy particle universe, and the external anti-graviton. The emergence of positive charge and negative charge provides the prototype of the observed electromagnetic force with charge generated during the Big Bang [2] [9].

4) The formation of the asymmetrical dual universe

The formation of our current universe follows immediately after the formation of the 10D particle dual universe through the asymmetrical dimensional oscillations, leading to the asymmetrical dual universe. The 10D positive-energy universe was transformed immediately into the 4D positive-energy particle universe with zero vacuum energy. The 10D negative-energy particle universe undergoes the stepwise dimension number oscillation between 10D and 4D. Without absorbing the interuniversal void, the external graviton and the anti-graviton also undergo the stepwise dimension number oscillation between 10D and 4D. The result is the asymmetrical dual universe consisting of the four equal regions of the 4D positive-energy particle universe, the variable D external graviton, the variable D negative-energy particle universe, and the variable D external anti-graviton. The asymmetrical dual universe is manifested as the asymmetry in the weak interaction in our observable universe as follows.

$$\begin{aligned}&\text{the 4D positive-energy particle universe and the external graviton}\\&\left(\left(s0_{4}e^{+}w^{+}e^{-}w^{-}0_{-4}s\right)g_{e}\right)_{n}\\&\text{the variable D negative-energy particle universe and the external anti-graviton}\\&\left(\overline{g_{e}}\left(s0_{4\text{ to }10}e^{+}w^{+}e^{-}w^{-}0_{-4\text{to}-10}s\right)\right)_{n}\end{aligned} \tag{21}$$

where s, g_e, g_e, e, and w are the strong force, external graviton, external anti-graviton, electromagnetism, and weak interaction, respectively for the observable universe, and where 0_40_{-4} and $0_{4\text{ to }10}\ 0_{-4\text{ to }-10}$ are 4D particle-antiparticle for the 4D positive-energy particle universe and variable D particle-antiparticle for the variable D negative-energy particle universe, respectively. For our asymmetrical dual universe, the step 3 for the transformation of 10D string into 10D particle had to be followed by the step 4, so the electromagnetic interaction from the step 3 was unified with the weak interaction from the step 4 to become the electroweak interaction, which was generated during the Big Bang [2] [9].

a) The formation of the 4D positive-energy particle universe

The formation of 4D positive-energy particle universe involved the two-step transformation: 1.the inflation and 2, the Big Bang. In the first step, the inflation is the transformation from 10D4d to 4D10d immediately. Calculated from Equation (1e), the rest mass of 4D10d is $1/\alpha^{12} \approx 137^{12}$ times of the mass of 10D4d, resulting in the first step of the inflation as the rapid expansion of space from the high vacuum energy 10D4d to the zero vacuum energy 4D10d as follows.

$$\begin{aligned}&1.\ \text{the inflation}\\&10\text{D4d}\xrightarrow{\text{quick QVSL transformation}}\text{4D10d}\end{aligned} \tag{22}$$

In the second step of the transformation, the Big Bang is a two-step process. The first step is the coupling of detachment space and the massive particles on attachment space in the positive-energy universe that absorbed the interuniversal void on detachment space. The result is the total conversion to generate massless particles on detachment space in the positive-energy universe and the external attachment space surrounded the positive-energy universe as described in Equation (7). In the second step, the coupling of attachment space and the massless particles in the positive-energy universe that absorbed the external attachment space surrounded the positive-energy universe. The partial conversion resulted in massive particles such as weak bosons, leptons, the Higgs boson, and massless particles such as photon. The second step is described in Equation (8) through the Higgs mechanism. The irreversible kinetic energy resulted from detachment space started the positive-energy universal expansion. The positive-energy universe has the combination of attachment space and detachment as follows.

2. The Big Bang

$$\begin{aligned}&1.\ \text{massive particles on attachment space} + \text{detachment space} \xrightarrow{\text{total conversion}}\\ &\text{massless particles on detachment space} + \text{the external attachment space}\\ &2.\ \text{massless particles on detachment space} + \text{the external attachment space} \xrightarrow{\text{partial conversion}}\\ &\text{massless particles} + \text{massive particles} + \text{detachment space} + \text{attachment space} + \text{the Higgs boson}\end{aligned} \tag{23}$$

The presence of irreversible kinetic energy induces irreversible entropy increase. In the Boltzmann formula of thermodynamic, the absolute entropy S of an ideal gas to the quantity W, which is the number of the arrangements of particles corresponding to a given macroscopic collection of particles:

$$S = k_B \ln W \tag{24}$$

where k_B is the Boltzmann's constant. The Boltzmann formula shows the relationship between entropy and the number of ways the atoms or molecules of a thermodynamic system can be arranged. The various atoms or molecules have different positions and momenta for irreversible entropy increase, because the increase in the number of different arrangements of particles in a macroscopic collection of particles requires the movements of individual particles in a macroscopic collection of particles. (There is no entropy increase in a single microscopic particle.) Individual momenta from kinetic energy are required for irreversible entropy increase in a macroscopic collection of particles. In other words, kinetic energy transforms a macroscopic collection of particles from one way of the arrangement of particles (order) into many ways of the arrangements of particles (disorder), and the process from order to disorder is irreversible in an isolated macroscopic collection of moving particles.

In our universe, the interuniversal void on detachment space was by the 10D positive-energy string universe which was very small. (The rest mass of 4D10d is $1/\alpha^{12} \approx 137^{12}$ times of the mass of 10D4d.) To have exactly reversible absorption-desorption of the interuniversal void for the reversible dual universe, the absorption and desorption have to be uniform. The space of the universe where the absorption-desorption occurs has to be small enough for the uniform absorption-desorption. The 10D string was small mass to allow the uniform absorption-desorption for reversible absorption-desorption of the interuniversal void, resulting in low entropy beginning. (The subsequent irreversible absorption of the interuniversal void is forbidden.) In the reversible multiverse postulate, all physical laws and phenomena are permanently reversible, and temporary irreversibility of entropy increase is allowed through reversibility breaking, symmetry violation, and low entropy beginning. Our 4D positive-energy particle universe is an example of irreversibility of entropy increase through reversibility breaking, symmetry violation, and low entropy beginning.

b) The formation of the variable D negative-energy particle universe

The formation of the variable D negative-energy particle universe involves the stepwise two-step transformation: the QVSL transformation and the varying supersymmetry transformation from 10D4d to 4D4d. (The particles in the 10D dual particle universe are 10D4d.) The QVSL transformation involves the transformation of space-time dimension, D. The repetitive stepwise two-step transformation from 10D4d to 4D10d as follows.

$$\begin{aligned}&10\text{D4d} \to 9\text{D5d} \to 9\text{D4d} \to 8\text{D5d} \to \cdots \to 4\text{D5d} \to 4\text{D4d}\\ &\mapsto \text{hidden} \quad \text{dark} \quad \text{universe} \leftarrow\mapsto \text{dark} \quad \text{energy} \leftarrow\end{aligned} \tag{25}$$

The variable D negative-energy particle universe consists of two periods: the hidden variable D negative-energy particle universe and the dark energy universe. The hidden variable D negative-energy particle universe composes of the >4D particles. As mentioned before, particles with different space-time dimensions are transparent and oblivious to one another, and separate from one another if possible. Thus, >4D particles are hidden and separated particles with respect to 4D particles in the 4D positive-energy particle universe (our observable universe). The hidden variable D negative-energy particle universe with D > 4 and the observable universe with D = 4 are the "parallel universes". The 4D particles transformed from hidden > 4D particles in the variable D negative-energy particle universe are observable dark energy for the 4D positive-energy particle universe, resulting in the accelerated expanding universe. Since the variable D negative-energy particle universe does not have detachment space, the presence of dark energy is not different from the presence of the cosmological constant. According to the theoretical calculation based on the asymmetrical dual universe, dark energy started in 4.47 billion years ago in agreement with the observed 4.71 ± 0.98 billion years ago [11]. Our asymmetrical dual universe consists of the four equal regions of the 4D positive-energy particle universe, the variable D external graviton, the variable D

negative-energy particle universe, and the variable D external anti-graviton, so the percentage the variable D area is 75%, three out of four regions, as the maximum percentage of dark energy. In terms of quintessence, such dark energy can be considered the tracking quintessence [23] [24] from the variable D area with the space-time dimension number as the tracker.

After the maximally connected universe, 4D dark energy transforms back to >4D particles that are not observable. The removal of dark energy in the observable universe results in the stop of accelerated expansion and the start of contraction of the observable universe. The end of dark energy starts another "parallel universe period". Both hidden universe and observable universe contract synchronically and equally. Eventually, the Big Crush and the two-step deflation occur in the 4D positive-energy particle universe. In the first step of the deflation, the 4D positive-energy particle universe loses all detachment space, kinetic energy, light, cosmic radiation, and force fields as dimensional orbitals, resulting in returning to 4D10d. In the second step of the deflation, without irreversible kinetic energy, the reversible direct dimension number oscillation resumes to transform the low vacuum energy 4D10d into the high vacuum energy 10D4d for the rapid contraction of space. Meanwhile, hidden >4D particles-antiparticles in the hidden universe transform into 10D4d particles-antiparticles. The dual universe can undergo another cycle of the light-dark dual universe. On the other hand, both universes can undergo the reverse charge transformation to become the 10D dual string universe, which in turn can return to the 11D dual membrane universe that in turn can return to the zero-energy universe as **Figure 1**.

5. Summary

The posited theory of everything is the dynamic and reversible process of continuingly changing matters and spaces in the reversible multiverse. Instead of remaining fixed, space-time dimension numbers of matters oscillate. Instead of being passive, spaces actively couple to particles. There are two different spaces whose digital combinations are different under different conditions. In the midst of continuously changing matters and spaces, the multiverse is simple and neat, because the multiverse is permanently reversible to exclude all irreversible physical laws and phenomena. However, temporary irreversible entropy increase is allowed through reversibility breaking, symmetry violation, and low entropy beginning. We live in the universe with such temporary irreversible entropy increase.

The dynamic and reversible theory of everything consists of the three postulates as 1) the oscillating M-theory postulate for the oscillating matter structure, 2) the digital transitional Higgs-reversed Higgs fields postulate for the digital space structure, and 3) the reversible multiverse postulate for all physical laws and phenomena. Oscillating M-theory is derived from oscillating membrane-string-particle whose space-time dimension number oscillates between 11D and 10D and between 10D and 4D. Space-time dimension number between 10 and 4 decreases with decreasing speed of light, decreasing vacuum energy, and increasing rest mass. The digital transitional Higgs-reversed Higgs fields are derived from digital attachment-detachment spaces which couple to particles. Under spontaneous symmetry breaking, the coupling of massless particle to zero-energy attachment space (the space for mass) produces the transitional nonzero-energy Higgs field-particle composite which under spontaneous symmetry restoring produces massive particle on zero-energy attachment space with the longitudinal component. The opposite of attachment space is detachment space as the space for kinetic energy and the nonzero-energy reverse Higgs field. The combination of n units of attachment space (denoted as 1) and n units of detachment space (denoted as 0) brings about the three digital structures: binary partition space $(1)_n(0)_n$, miscible space $(1 + 0)_n$, and binary lattice space $(1\ 0)_n$ to account for quantum mechanics, special relativity, and the force fields, respectively. In the third postulate, all physical laws and phenomena are permanently reversible in the multiverse, and temporary irreversible entropy increase is allowed.

Our universe is an asymmetrical dual positive-energy-negative-energy universe where the positive-energy universe on attachment space absorbed the interuniversal void on detachment space to result in the combination of attachment space and detachment space, while the negative-energy universe did not absorb the interuniversal void, resulting in temporary irreversible entropy increase through reversibility breaking, symmetry violation, and low entropy beginning. Guided by the reversible negative-energy universe, our dual universe is a globally reversible cyclic dual universe. The four reversible steps in the globally reversible cyclic dual universe are 1) the formation of the 11D membrane dual universe, 2) the formation of the 10D string dual universe, 3) the formation of the 10D particle dual universe, and 4) the formation of the asymmetrical dual universe. The four force fields (gravity, the strong force, electromagnetism, and the weak force) are derived from the 4-step evolution of the cyclic dual un-

iverse. Baryonic matter and dark matter are in the positive-energy universe, while a part of dark energy is in the negative-energy universe.

The posited theory of everything based on the three postulates explains cosmology, the composition (baryonic matter, dark matter, and dark energy) in the universe, the periodic table of elementary particles (quarks, leptons, and bosons), the galaxy evolution, superconductivity, black hole, thermodynamic, and quantum mechanics.

References

[1] Woit, P. (2006) Not Even Wrong: The Failure of String Theory and the Search for Unity in Physical Law. Basic Books, New York.

[2] Chung, D. (2014) *Journal of Modern Physics*, **5**, 1234-1243. http://dx.doi.org/10.4236/jmp.2014.514123

[3] Chung, D. and Hefferlinm, R. (2013) *Journal of Modern Physics*, **4**, 21-26. http://dx.doi.org/10.4236/jmp.2013.44A004

[4] Chung, D. (2014) *Journal of Modern Physics*, **5**, 1467-1472. http://dx.doi.org/10.4236/jmp.2014.515148

[5] Steven Weinberg, S. (1989) *Review Modern Physics*, **61**, 1-23. http://dx.doi.org/10.1103/RevModPhys.61.1

[6] Chung, D. and Krasnoholovets, V. (2013) *Journal of Modern Physics*, **4**, 27-31. http://dx.doi.org/10.4236/jmp.2013.44A005

[7] Chung, D. (2015) *Journal of Modern Physics*, **6**, 1820-1832. http://dx.doi.org/10.4236/jmp.2015.613186

[8] Carroll, S. (2010) From Eternity to Here: The Quest for the Ultimate Theory of Time. Dutton, New York.

[9] Chung, D. (2015) *Journal of Modern Physics*, **6**, 1249-1260. http://dx.doi.org/10.4236/jmp.2015.69130

[10] Chung, D. (2015) *Journal of Modern Physics*, **6**, 1189-1194. http://dx.doi.org/10.4236/jmp.2015.69123

[11] Chung, D. and Krasnoholovets, V. (2013) *Journal of Modern Physics*, **4**, 77-84. http://dx.doi.org/10.4236/jmp.2013.47A1009

[12] Chung, D. (2014) *Journal of Modern Physics*, **5**, 464-472. http://dx.doi.org/10.4236/jmp.2014.56056

[13] Chung, D. (2014) *International Journal of Astronomy and Astrophysics*, **4**, 374-383. http://dx.doi.org/10.4236/ijaa.2014.42032

[14] Chung, D. (2015) *Global Journal of Science Frontier Research A*, **15**, 119-125.

[15] Chung, D. (2015) *Journal of Modern Physics*, **6**, 26-36. http://dx.doi.org/10.4236/jmp.2015.61005

[16] Chung, D. and Krasnoholovets, V. (2013) *Journal of Modern Physics*, **4**, 1-6. http://dx.doi.org/10.4236/jmp.2013.47A1001

[17] Chung, D. (2014) *Global Journal of Science Frontier Research A*, **14**, 1-8.

[18] Bounias, M. and Krasnoholovets, V. (2003) *The International Journal of Systems and Cybernetics*, **32**, 1005-1020.

[19] Jarosik, N., Bennett, C.L., Dunkley, J., Gold, B., Greason, M.R., Halpern, M., *et al.* (2011) *The Astrophysical Journal*: *Supplement Series*, **192**, 14. http://dx.doi.org/10.1088/0067-0049/192/2/14

[20] Greene, B. (2011) The Hidden Reality: Parallel Universes and the Deep Laws of the Cosmos. Alfred A. Knopf, New York.

[21] Randall, L. (2005) Warped Passages: Unraveling the Mysteries of the Universe's Hidden Dimensions. Harper Collins, New York.

[22] Randall, L. and Sundrum, R. (1999) *Physical Review Letters*, **83**, 3370-3373. http://dx.doi.org/10.1103/PhysRevLett.83.3370

[23] Maldacena, J. (1998) *Advances in Theoretical and Mathematical Physics*, **2**, 231-252. http://dx.doi.org/10.4310/ATMP.1998.v2.n2.a1

[24] Padmanabhan, T. (2003) *Physics Reports*, **380**, 235-320. http://dx.doi.org/10.1016/S0370-1573(03)00120-0

Journal of Modern Physics, 2016, 7, 656-661
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jmp
http://dx.doi.org/10.4236/jmp.2016.77065

Net Force $F = \gamma^3 ma$ at High Velocity

Olivier Serret

Cugnaux, France
Email: o.serret@free.fr

Received 24 February 2016; accepted 25 April 2016; published 28 April 2016

Abstract

Newton's theory of gravitation has been outdated by relativity theory explaining specific phenomena like perihelion precession of Mercury, light deflection and very recently the detection of gravitational waves. But the disappearance of the obvious gravitational force and the variation of time are arguable concepts difficult to directly prove. Present methodology is based on hypotheses as expressed in a previous article: a universal time and an inertial mass variable according to the Lorentz factor (which could not be envisioned at Newton's age). Because this methodology is mainly stood on Newtonian mechanics, it will be called neo-Newtonian mechanics. This theory is in coherence with the time of the Quantum Mechanics. In Newtonian mechanics, all forces, including gravitational force, are deducted from the linear momentum. Introducing the variable inertial mass, the result of the demonstration is an updated expression of the net force at high velocity: $F = \gamma^3 m_g a$. If such a factor in γ^3 can look a bit strange at first sight for a force, let us remind that the lost energy in a synchrotron is already measured in γ^4. Next article will be on the perihelion precession of Mercury within neo-Newtonian mechanics.

Keywords

Net Force, Strength, Neo-Newtonian, Lorentz Factor, General Relativity Theory, Circular Motion, Synchrotron Radiation, High Velocity

1. Introduction

1.1. Relativity Hypotheses

For a century, hypothesis of a variable time is laid down by the special theory of relativity. This hypothesis can explain many Nature observations, experiments and formulas, for example, the demonstration of the Lorentz factor. Because of such good explanations, the hypothesis of a variable time has been validated. Nevertheless, it remains some paradoxes and some predictions which are difficult to measure directly, as a reversible time, an "imaginary" time or even the time variation itself.

How to cite this paper: Serret, O. (2016) Net Force $F = \gamma^3 ma$ at High Velocity. *Journal of Modern Physics*, **7**, 656-661.
http://dx.doi.org/10.4236/jmp.2016.77065

And in developing his ideas about the consequences of the equivalence principle between gravitational mass and inertial mass, Einstein leads to a new vision of gravitation which is to replace that of Newton: the general theory of relativity. The most important aspect is the disappearance of gravitational force concept. For Einstein, the motion of a body is not determined by strength, but by the configuration of space-time [1]. For example, relativity theory explains the deflection of light and the perihelion precession of Mercury, and predicts the gravitational waves which have been very recently detected.

But the absence of gravitational force and a variable time according to the reference frame remain concepts difficult to directly prove.

1.2. The Purpose

The question is: is it possible to explain such phenomena within another theory, *i.e.* using gravitational forces and a universal time? It is what we will try to do in this article.

A universal time would give in coherence with the universal time of the Quantum Mechanics.

1.3. Neo-Newtonian Hypotheses

The basis has been laid down in a previous article [2]: Lorentz factor can be demonstrated without using a variable time! It is only necessary to consider a variable inertial mass, different of the gravitational mass, and the energy of the particle linked to the inertial mass. If Newton distinguished the concepts of gravitational mass from the inertial mass [3], he could not be envisioned a variation of the inertial mass only detected at very high velocity (let us remind in the 17th century, Huygens was only trying to estimate the light celerity [4]). We will call these hypotheses: the neo-Newtonian mechanics. We compare them in **Chart 1**.

Now in this article, we will check the consequence of these hypotheses on force expression in general (which includes resultant gravitational force).

2. Net Force Demonstration

The linear momentum p is by definition the product of the mass of a body by its velocity [5]:

$$p = m \cdot v \tag{1}$$

It is a general formula, the mass m is the inertial mass (it is not the gravitational mass).
So the linear momentum can be written more precisely

$$p = m_i \cdot v \tag{2}$$

With m_i the inertial mass

According to neo-Newtonian demonstration [2], the inertial mass is linked to the gravitational mass m_g by the Lorentz factor γ

$$m_i = \gamma \cdot m_g \tag{3}$$

with

$$\gamma = \frac{1}{\sqrt{1 - v^2/s^2}} \tag{4}$$

Chart 1. Comparison of hypotheses.

Hypotheses	Newtonian	Relativity	Neo-Newtonian
Space	Constant	Variable	Constant
Time	Universal	Variable	Universal
Gravitational mass	m_g = constant	$m_G = \gamma m_g$ = variable	m_g = constant
Inertial mass	$m_i = m_g$ = constant	$m_i = m_G$ = variable	$m_i = \gamma m_g$ = variable
Speed limit	None	Light celerity c	Asymptote s $(s \approx c)$

By property of the net force F according to the second Newton's law of motion:

$$F = \frac{\mathrm{d}p}{\mathrm{d}t} \tag{5}$$

Because gravitational mass is constant:

$$\frac{\mathrm{d}m_g}{\mathrm{d}t} = 0 \tag{6}$$

so

$$F = \gamma m_g \frac{\mathrm{d}v}{\mathrm{d}t} + \frac{\mathrm{d}\gamma}{\mathrm{d}t} m_g v \tag{7}$$

$$F = \gamma m_g \frac{\mathrm{d}v}{\mathrm{d}t}\left(1 + \frac{1}{\gamma}\frac{\mathrm{d}\gamma}{\frac{\mathrm{d}v}{\mathrm{d}t}} v\right) \tag{8}$$

$$F = \gamma m_g \frac{\mathrm{d}v}{\mathrm{d}t}\left(1 + \frac{1}{\gamma}\frac{\mathrm{d}\gamma}{\mathrm{d}v} v\right) \tag{9}$$

And due to Equation (4):

$$\gamma = \left(1 - v^2/s^2\right)^{\left(-\frac{1}{2}\right)} \tag{4bis}$$

$$\frac{\mathrm{d}\gamma}{\mathrm{d}v} = \frac{-1}{2}\left(\frac{-2v}{s^2}\right)\left(1 - v^2/s^2\right)^{\left(-\frac{3}{2}\right)} \tag{10}$$

$$\frac{\mathrm{d}\gamma}{\mathrm{d}v} = \left(\frac{v}{s^2}\right)(\gamma)^3 \tag{11}$$

So, with Equation (9):

$$F = \gamma m_g \frac{\mathrm{d}v}{\mathrm{d}t}\left(1 + \frac{1}{\gamma}\frac{v}{s^2}\gamma^3 v\right) \tag{12}$$

$$F = \gamma m_g \frac{\mathrm{d}v}{\mathrm{d}t}\left(1 + \frac{v^2}{s^2}\gamma^2\right) \tag{13}$$

And again due to Equation (4):

$$\gamma = \left(1 - v^2/s^2\right)^{\left(-\frac{1}{2}\right)} \tag{4ter}$$

$$\gamma^{-2} = \left(1 - v^2/s^2\right)^1 \tag{14}$$

$$1 = \gamma^2\left(1 - v^2/s^2\right) \tag{15}$$

$$1 + \frac{v^2}{s^2}\gamma^2 = \gamma^2 \tag{16}$$

And so, with Equation (13)

$$F = \gamma m_g \frac{\mathrm{d}v}{\mathrm{d}t}\left(\gamma^2\right) \tag{17}$$

$$F = \gamma^3 m_g \frac{\mathrm{d}v}{\mathrm{d}t} \tag{18}$$

By definition, the acceleration a is:

$$a = \frac{\mathrm{d}v}{\mathrm{d}t} \tag{19}$$

So the updated property of the net force is:

$$F = \gamma^3 m_g a \tag{20}$$

3. Comments

3.1. Comparison

This expression in γ^3 can look a bit strange at first sight.

Let us remind the synchrotron radiation. The cyclotron is used for particles, and the synchrotron is used for particles at velocities close to light celerity. The loss of energy per turn by synchrotron radiation can be measured as follows [6]-[8]:

$$W = \gamma^4 \left(\frac{4\pi K e^2}{3r}\right)\left(\frac{V}{c}\right)^3 \tag{21}$$

formula which can also be written:

$$W = \gamma^3 \gamma \frac{4\pi K e^2 V^2 V}{3rc^3} \tag{21bis}$$

or

$$W = \gamma^3 \left(\frac{\gamma V^2}{r}\right) V \left(\frac{4\pi K e^2}{3c^3}\right) \tag{21ter}$$

And let us remind a work is a force by a length, and a length is a velocity by a time. So

$$W' = F \cdot l \tag{22}$$

$$W' = \left(\gamma^3 ma\right) \cdot \left(Vt\right) \tag{23}$$

$$W' = \gamma^3 \left(\frac{\gamma V^2}{r}\right) V \left(mt\right) \tag{24}$$

Then, the work W' of a γ^3 force [Equation (24)] appears to be homogeneous with the measure in a synchrotron of the loss of energy W [Equation (24ter)] of particles at very high velocity. This synchrotron effect can be checked not only in a laboratory but also in pulsed emission gamma-ray radiation from pulsar [9].

3.2. Numerical Application

This γ^3 factor can be detected only at very high velocity. At very high velocity, it is of course easier to measure when the body stays close, *i.e.* on a constant periodic movement, as the circular motion. For example:

- Planet revolution (Mercury is the fastest planet of our solar system).
- Particle in a cyclotron.
- Particle in a synchrotron.

Let us check the value of γ^3 at various velocities. See **Chart 2** and/or **Figure 1**.

It confirms

- Variation of γ^3 could not envisioned at Newton's age when the higher motion known was Mercury velocity (with a $\gamma^3 = 1.00000004$).
- Expression of the net force with γ^3 can be checked with a synchrotron.

3.3. Meaning

That means that, at very high velocity,

- For a same variation of velocity (or acceleration), the net force will be slightly higher than traditionally expected.

Chart 2. Value of γ^3 according to the velocity.

	Velocity			γ^3
	[km/h]	[m/s]	[v/c]	
Planet (Mercury)	170,000	47,000	0.01%	1.00000004
Cyclotron	15,000,000	4,000,000	1%	1.0003
Synchrotron	700,000,000	200,000,000	67%	2.4

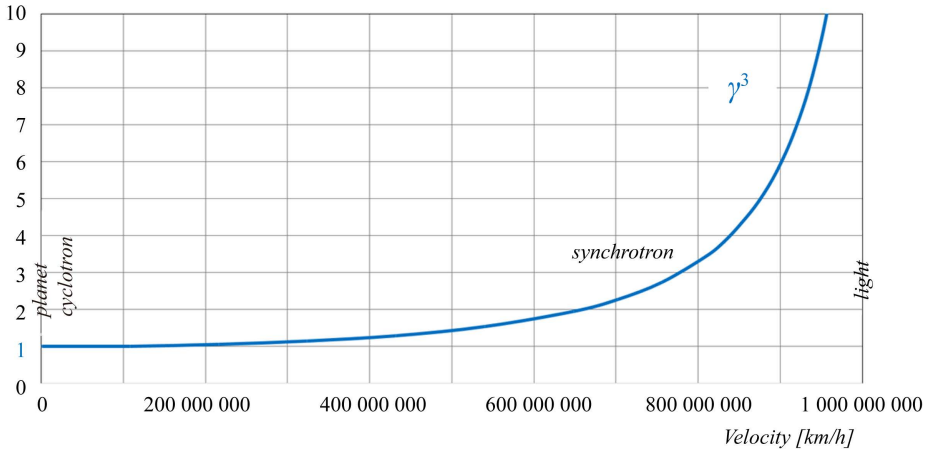


Figure 1. Value of γ^3 according to the velocity.

- For a same net force, the variation of velocity (or acceleration) will be slightly lower than traditionally expected and at usual velocity, $\gamma \approx 1$, and we find back the usual formula:

$$F = m_g a \tag{25}$$

4. Conclusions

First, we remind results of a previous article: Lorentz factor can be demonstrated without using a variable time, but using a variable inertial mass. Such a hypothesison time, called neo-Newtonian theory, is in coherence with the Quantum Mechanics.

Then in present article, consequence of this hypothesis has been checked on net force expression. Deducted and demonstrated from the linear momentum, net force is so expressed to: $F = \gamma^3 m_g a$.

This γ^3 factor can be detected only at very high velocity. At very high velocity, it is of course easier to measure on a constant periodic movement, as the circular motion. For example, the synchrotron radiation (in synchrotron laboratory or in pulsed emission gamma-ray radiation from pulsar): the electromagnetic energy emitted by electrons or protons at circular velocity close to light celerity is done with the factor γ^4.

Application of such a neo-Newtonian hypotheses on the perihelion precession of Mercury (the faster of the planets of our solar system), the deflection of light or the Doppler Effect will be done in next articles.

Acknowledgements

I would like to thank the reviewers for their advice about the looking of my article.

References

[1] Esslinger, O. (2015) Astronomie & Astrophysique. http://www.astronomes.com/la-fin-des-etoiles-massives/relativite-generale/

[2] Serret, O. (2015) *Journal of Modern Physics*, **6**, 252-259. http://dx.doi.org/10.4236/jmp.2015.63030

[3] Newton, I. (1686) Philosophiae Naturalis Principia Mathematica.

http://sites.trin.cam.ac.uk/manuscripts/NQ_16_200/manuscript.php?fullpage=1/

[4] Huygens, C. (1690) Traité de la lumière. https://books.google.fr/books?id=No8PAAAAQAAJ&pg=PA9&hl=fr#v=onepage&q&f=false

[5] Queyrel, J.-L. and Mesplede, J. (1993) Précis de physique—Mécanique, 73, Bréal.

[6] Nave, R. (2014) Synchrotron Radiation. Georgia State University, Atlanta. http://hyperphysics.phy-astr.gsu.edu/hbase/Particles/synchrotron.html

[7] Beckmann, V. (2006) Synchrotron Radiation. NASA. http://asd.gsfc.nasa.gov/Volker.Beckmann/school/download/Longair_Radiation2.pdf

[8] Barletta, W. (2016) Synchrotron Radiation, USPAS. http://uspas.fnal.gov/materials/09UNM/Unit_11_Lecture_18_Synchrotron_radiation.pdf

[9] Mazure, A. and Baza, S. (2007) L'Univers dans tous ses états, 77-91. Ed. Dunod-Quai des Sciences.

Journal of Modern Physics, 2016, 7, 662-679
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jmp
http://dx.doi.org/10.4236/jmp.2016.77066

On the Origin of Charge-Asymmetric Matter. II. Localized Dirac Waveforms

Alexander Makhlin

Rapid Research Inc., Southfield, MI, USA
Email: amakhlin@comcast.net

Received 25 February 2016; accepted 25 April 2016; published 28 April 2016

Abstract

This paper continues the author's work [1], where a new framework of the matter-induced physical geometry was built and an intrinsic nonlinearity of the Dirac equation was discovered. Here, the nonlinear Dirac equation is solved and the localized configurations are found analytically. Of the two possible types of the potentially stationary localized configurations of the Dirac field, only one is stable with respect to the action of an external field and it corresponds to a positive charge. A connection with the global charge asymmetry of matter in the Universe and with the recently observed excess of the cosmic positrons is discussed.

Keywords

Nonlinear Dirac Field, Localization, Cosmological Charge Asymmetry

1. Introduction

This paper continues the author's study of the long-standing question of how the physical Dirac field of a real matter becomes a finite-sized particle, and it is approached here as a practical problem. The problem is posed and solved in a new framework of the matter-induced affine geometry [1], which deduces the geometric relations in the space-time continuum from the dynamic properties of the Dirac field. The intuitive argument of a possible auto-localization of the Dirac field followed from an observation [1] that the local time flows slower at higher invariant density, and then from the wave nature of the Dirac equation. Its further consequence must be the (well-known but not clearly understood) charge asymmetry of the observed localized matter. In the present work, these qualitative expectations are confirmed by explicit calculations.

The earlier developed [1] mathematical background for the present work is based on the following ideas and results. It is observed that if at a point in spacetime continuum (the principal differentiable manifold $\mathbb{M}$) a physical Dirac field is defined, then the latter determines the tetrad of *Dirac currents*. These are linearly in-

dependent and Lorentz-orthogonal and can serve as local algebraic basis for any four-dimensional vector space, including the infinitesimal displacements in coordinate space.

The Dirac currents are employed as the Cartan's moving frame in spacetime which, in its turn, results in the technique of covariant derivatives for the vector and spinor fields. The physics is naturally brought into this mathematical picture by the equations of motion of the Dirac field, which made an artificial tangent (pseudo) Euclidean space unnecessary. Differential identities derived from equations of motion fully determine all the components of the matter-induced affine connection (the Ricci coefficients of rotation of the tetrad) in $\mathbb{M}$ and without resorting to a particular coordinate system. Thus determined connections completely define an *affine geometry* (endowed with the connection but with no metric). Thus defined connection depends on the Dirac field which makes the Dirac equation nonlinear.

With known connections, it became possible to find the coordinate lines and coordinate surfaces of the matter-induced affine geometry, which have a clear physical meaning and quite high degree of symmetry. The congruence of lines of the timelike vector current appeared to be normal, thus determining the family of the hypersurfaces of the constant world time τ. The lines of the spacelike axial current appeared to be straight and their congruence normal. They define the surfaces of the constant distance ρ. The two-dimensional surfaces of constant ρ at a given time τ were proved to be just spherical surfaces.

Below, the inevitable localization of the Dirac field into particles observed in real world, but not explained by any theory so far, is confirmed by the analytic solutions of the nonlinear Dirac equation in one-body approximation. One of the solutions has maximum near its center and is clearly associated with a stable localized positive charge. Another one has minimum and is sought to be an intrinsically unstable negative charge, which can be only weakly localized by an external field.

The content of the paper is organized as follows. In Section 2 we use the previously developed [1] tools of the matter-induced affine geometry to write down the Dirac equation in its most general coordinate-independent form. Then, in Section 3 we derive the formulae that connect the Dirac matrices in the principal manifold $\mathbb{M}$ and in arithmetic $\mathbb{R}^4$. In Sections 4 and 5, the Dirac equation is written down in a mixed representation, with derivatives in $\mathbb{M}$, and coordinates and Dirac matrices in $\mathbb{R}^4$. This representation is well suited for finding the analytic solution. These solutions are found in Section 6 and their stability is discussed in Section 7. The conceptual questions of the charge-asymmetric real world are briefly discussed in the Summary.

2. The Framework

In the first part of this work we explored differential identities for the four Dirac currents, vector current $\boldsymbol{j}$, axial current $\boldsymbol{J}$, and two " charged currents", $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$. Using them, we found all components of the affine connection ω_{ABC}, as well as connection Γ_B of the Dirac field in principal manifold $\mathbb{M}$,

$$\Gamma_B = ieA_B + (1/4)\,\omega_{ACB}\rho_1\alpha^A\rho_1\alpha^C. \tag{2.1}$$

The connection (2.1) determines the covariant derivative of the Dirac field and it enters the Dirac equation as $\alpha^B\Gamma_B$,

$$\alpha^B\left[\partial_B\psi - \Gamma_B\psi\right] = -im\rho_1\psi. \tag{2.2}$$

The nonzero elements of the ω_{ABC} in the tetrad basis of the normalized Dirac currents $\boldsymbol{e}_A$ are as follows,

$$\omega_{030} = -\omega_{131} = -\omega_{232} = Q,\quad \omega_{12D} = 2e\tilde{A}_{[D]},\quad (D = 0,1,2,3), \tag{2.3}$$

where $Q \equiv \partial_{[3]}\ln\mathcal{R} = -m\mathcal{P}/\mathcal{R} = -m\sin\mathcal{Y}$ is the derivative of the invariant density $\mathcal{R}$ in the direction of the axial current and it has an algebraic representation via the pseudoscalar density $\mathcal{P}$. These formulae assume that $\tilde{A}_{[D]} = +A_D$ for the right-handed spatial triad $\boldsymbol{e}_{[1]}, \boldsymbol{e}_{[2]}, \boldsymbol{e}_{[3]}$ with $\boldsymbol{\Theta} = \mathcal{R}\boldsymbol{e}_{[1]}$, $\boldsymbol{\Phi} = \mathcal{R}\boldsymbol{e}_{[2]}$ and the naturally outward directed axial current $\boldsymbol{J} = \mathcal{R}\boldsymbol{e}_3$, *i.e.* $\left[\boldsymbol{e}_{[1]}\times\boldsymbol{e}_{[2]}\right] = \boldsymbol{e}_{[3]}$ [c.f Equations (A.9), (A.10)]. When the latter is directed inward, but we still wish $\boldsymbol{e}_{[3]}$ to point outward, then we have to take $\boldsymbol{\Theta} = \mathcal{R}\boldsymbol{e}_{[2]}$, $\boldsymbol{\Phi} = \mathcal{R}\boldsymbol{e}_{[1]}$ and replace $\omega_{12D} \to \omega_{21D} = -\omega_{12D}$ (or $\tilde{A}_{[D]} = -A_D$) in Equation (2.3)[1].

[1]Throughout this paper, when uppercase index A of the basis $\boldsymbol{e}_A \equiv \boldsymbol{e}_{[A]}$, $(A = 0,1,2,3)$ takes a particular numeric value we put it in brackets $[0],[1],\cdots$. The lowercase indices a that are related to the tetrad $\boldsymbol{h}_a \equiv \boldsymbol{h}_{(a)}$ are put in parentheses, $(0),(1),\cdots$.

It is instructive to see how the operator $D_A = \partial_A - \Gamma_A$ carries out the parallel transport of the Dirac spinor ψ in different directions. Substituting the results (2.3) into connection (2.1), it is straightforward to obtain,

$$\begin{aligned}
&\alpha^{[0]}\Gamma_0 = (1/2)Q\alpha^{[3]} + 2ieA_{[0]}\alpha^{[0]}\Pi = (1/2)Q\alpha^{[3]} + ie\left[A_{[0]}\alpha^{[0]} - \tilde{A}_{[0]}\rho_{[3]}\alpha^{[3]}\right],\\
&\alpha^{[3]}\Gamma_3 = +2ieA_{[3]}\alpha^{[3]}\Pi = ie\left[A_{[3]}\alpha^{[3]} - \tilde{A}_{[3]}\rho_{[3]}\alpha^{[0]}\right],\\
&\alpha^{[1]}\Gamma_1 = (1/2)Q\alpha^{[3]} + 2ieA_{[1]}\alpha^{[1]}\Pi = (1/2)Q\alpha^{[3]} + ie\left[A_{[1]}\alpha^{[1]} + i\tilde{A}_{[1]}\alpha^{[2]}\right],\\
&\alpha^{[2]}\Gamma_2 = (1/2)Q\alpha^{[3]} + 2ieA_{[2]}\alpha^{[2]}\Pi = (1/2)Q\alpha^{[3]} + ie\left[A_{[2]}\alpha^{[2]} - i\tilde{A}_{[2]}\alpha^{[1]}\right],
\end{aligned} \tag{2.4}$$

where $\Pi = \left(1 \mp i\gamma^{[1]}\gamma^{[2]}\right)\Big/2 = S^{-1}\left(1 \pm \sigma^{(3)}\right)S\Big/2$. The upper and lower signs in the projector Π (accordingly, the sign in $\tilde{A}_{[D]} = \pm A_{[D]}$) correspond to the outward and inward directions of the axial current, respectively, which then determines the right- and left-oriented spatial triplets $\boldsymbol{e}_{[1]}, \boldsymbol{e}_{[2]}, \boldsymbol{e}_{[3]}$. It will be shown below, that, from the perspective of the localized solutions, this orientation is translated into the bump of the positive charge and to the dip of the negative one, respectively, *i.e.* $\pm = -\text{sign}\left(\partial_{[3]}\mathcal{R}\right)$. Therefore, depending on this sign, only the *locally inward or locally outward* components, (d_L, d_R) or (u_L, u_R), interact with the electromagnetic potential but with the doubled coupling constant $2e$. In a sense, the charge conjugation goes together with spatial reflection. The matrix ρ_3 differentiate between the right and left components.

With the connection (2.4) the Dirac equation becomes a nonlinear equation and its explicit form reads as,

$$\begin{aligned}
&\alpha^{[0]}\left[\partial_{[0]} - ieA_{[0]} + i\rho_{[3]}e\tilde{A}_{[3]}\right]\psi + \alpha^{[3]}\left[\partial_{[3]} - ieA_{[3]} + i\rho_{[3]}e\tilde{A}_{[0]} - (3/2)Q\right]\psi\\
&+\alpha^{[1]}\left[\partial_{[1]} - ieA_{[1]} - e\tilde{A}_{[2]}\right]\psi + \alpha^{[2]}\left[\partial_{[2]} - ieA_{[2]} + e\tilde{A}_{[1]}\right]\psi + im\rho_{[1]}\psi = 0,
\end{aligned} \tag{2.5}$$

where anomalous term $-3Q/2$ singles out the direction of the axial current among others even when $A_\mu = 0$.

This equation is valid in every connected domain where $R^2 > 0$ and the Dirac currents define a non-degenerate orthogonal tetrad $e_A^\mu(\psi)$. As anticipated, it is invariant in a most broad sense—it depends neither on choice of coordinates x^μ in $\mathbb{R}^4$ nor on a tetrad system h_a^μ (also in $\mathbb{R}^4$) not even on a particular choice of the γ-matrices. The latter is always taken for granted since one can introduce a new Dirac field $\psi' = S\psi$ leaving the gamma matrices unchanged. But this trick works only for re-parameterizations in $\mathbb{R}^4$, *i.e.* change of the Lorentz frame or transformations between orthogonal coordinates. It cannot be employed in the principal manifold $\mathbb{M}$ just because the Dirac field is a coordinate scalar.

Finally, Equation (2.5) is nonlinear because both the connection ω_{ACB} and the Dirac matrices $\alpha^A = V_a^A(\psi)\alpha^a$ in it depend on the Dirac field $\psi \in \mathbb{M}$. The dependence of ω_{ACB} on the Dirac field is due to (2.3). The dependence of the Dirac matrices on ψ, $\alpha^A = V_a^A(\psi)\alpha^a$, is not so explicit but not less important and it cannot be avoided. Indeed, in the basis $[A]$ each of the currents $\boldsymbol{J}_A$ has only one nonzero component, e.g.,

$$j^A = \psi^+\alpha^A\psi = V_a^A j^a = RV_a^A V_0^a = R\delta_{[0]}^A.$$

The latter cannot be achieved without an explicit dependence $\alpha^A(\psi)$. Indeed, with $\psi \in \mathbb{M}$ and numerical matrices α^a the current j^a will have all components. Obviously, this is a significant technical difficulty. However, only this dependence solves a conceptual problem of independence of the equation of motion for the physical Dirac field in $\mathbb{M}$ on a particular choice of the tetrad $\boldsymbol{h}_a$ and of the matrices α^a in tangent T_p. Therefore, we begin with the establishing rules of transformation of the 16 Dirac matrices between $\mathbb{M}$ and $\mathbb{R}^4$.

3. Dirac Matrices in Principal Manifold $\mathbb{M}$

Historically, the Dirac equation for the free field ψ was formulated as $i\alpha^a\partial_a\psi - m\beta\psi = 0$ with the aid of *Hermitian* Dirac matrices $\alpha^a = \left(\alpha^a\right)^+$ and $\beta = \beta^+$, which satisfy the commutation relations,

$$\alpha^a\beta\alpha^b + \alpha^b\beta\alpha^a = 2\beta\eta^{ab}, \quad \alpha^a\beta + \beta\alpha^a = 0, \quad \beta^2 = 1. \tag{3.1}$$

Usually one assumes that $\alpha^a = \left(1, \alpha^i\right)$; $a = 0,1,2,3$; $i = 1,2,3$ (so that $\alpha^0 = 1$ is a unit matrix) but this is not required. An apparently symmetric form of commutation relations (3.1) emerges (along with the equation, $i\gamma^a\partial_a\psi - m\psi = 0$) in terms of the matrices $\gamma^a = \left(\gamma^{(0)}, \gamma^i\right) = \left(\beta, \beta\alpha^i\right)$,

$$\gamma^a\gamma^b+\gamma^b\gamma^a=2\eta^{ab}. \tag{3.2}$$

Neither of these matrices is uniquely defined. However, if there exist two sets of the matrices, γ^a and $\gamma^{[A]}$, that satisfy (2) then, according to the Pauli's fundamental theorem, there exists such a nonsingular *S*, that

$$\gamma^{[\kappa]}=S^{-1}\gamma^{(\kappa)}S, \tag{3.3}$$

where $\kappa=0,1,2,3$, $[\kappa]$ is a number standing for superscript *A* and (κ) is the same number for superscript *a*. There are sixteen linearly independent 4×4 matrices $O_p=\left(1,\gamma^a,\gamma^a\gamma^b,\cdots\right)$, all of which are the products of 1, 2, 3 or 4 different gamma. Therefore, $O_{[p]}=S^{-1}O_pS=\left(1,\gamma^{[a]},\gamma^{[a]}\gamma^{[b]},\cdots\right)$, which is an indisputable technical advantage.

By their definition, the matrices γ^a are not Hermitian. However, since β and α^i are Hermitian and anti-commuting, the Hermit-conjugated γ-matrices are $\left(\gamma^a\right)^+=\gamma^{(0)}\gamma^a\gamma^{(0)}$. If, by the same token, $\gamma^{[A]}=\gamma^{[0]}\alpha^{[A]}$ (with Hermitian $\gamma^{[0]}$ and $\alpha^{[A]}$), then $\left(\gamma^{[A]}\right)^+=\gamma^{[0]}\gamma^{[A]}\gamma^{[0]}$, which yields,

$$S^{-1}\gamma^{(A)}S=\gamma^{[A]}=\gamma^{[0]}\left(\gamma^{[A]}\right)^+\gamma^{[0]}=\gamma^{[0]}\left(S^{-1}\gamma^AS\right)^+\gamma^{[0]}=\left(\gamma^{[0]}S^+\gamma^{(0)}\right)\gamma^A\left(\gamma^{[0]}S^+\gamma^{(0)}\right)^{-1}.$$

Multiplying this by *S* from the left and by $\gamma^{[0]}S^+\gamma^{(0)}$ from the right, we find,

$$\gamma^A\left(S\gamma^{[0]}S^+\gamma^{(0)}\right)=\left(S\gamma^{[0]}S^+\gamma^{(0)}\right)\gamma^A. \tag{3.4}$$

The matrix $\left(S\gamma^{[0]}S^+\gamma^{(0)}\right)$ commutes with all the matrices γ^A and must be the unit matrix, *viz.*,

$$\gamma^{[0]}S^{-1}=S^+\gamma^{(0)}. \tag{3.5}$$

On the one hand, we can continue as

$$\alpha^{[A]}=\gamma^{[0]}\gamma^{[A]}=\gamma^{[0]}S^{-1}\gamma^AS=S^+\gamma^{(0)}\gamma^AS=S^+\alpha^AS. \tag{3.6}$$

On the other hand, condition (3.5) means that $\gamma^{[0]}=S^+\gamma^{(0)}S\neq S^{-1}\gamma^{(0)}S$, which conflicts with Equation (3.3), because matrix *S* is not unitary. This conflict can be avoided by adopting a slightly different agreement (that does not affect any of the common usages of the gamma-matrices). Namely, let us denote $\beta=\rho_1$ and define γ-matrices as $\gamma^a=\rho_1\alpha^a$ and $\gamma^{[A]}=\rho_1\alpha^{[A]}$. Now we must replace both $\gamma^{[0]}$ and $\gamma^{(0)}$ in Equation (4) by ρ_1, so that $S^+\rho_1=\rho_1S^{-1}$ and $\gamma^{[a]}=\rho_1\alpha^{[a]}=\rho_1S^+\alpha^{(a)}S=S^{-1}\rho_1\alpha^{(a)}S=S^{-1}\gamma^{(a)}S$, $a=0,1,2,3$, in compliance with (3.3). Choosing $\alpha^{(0)}=1$, we have $\gamma^{(0)}=\rho_1$, $\alpha^{[0]}=S^+S$, $\gamma^{[0]}=\rho_1S^+S=S^+S\rho_1=\left(\gamma^{[0]}\right)^+$.

Throughout this paper, we are only interested in a special case of the transformations (3.3) and (3.6),

$$\gamma^{[A]}=V_a^A\gamma^a,\ \ \alpha^{[A]}=V_a^A\alpha^a, \tag{3.7}$$

where the transformation matrix V_a^A is real and has the properties,

$$V_A^aV_a^B=\delta_A^B,\ \ V_A^aV_b^A=\delta_b^a. \tag{3.8}$$

Then the commutation relations (3.1) are the same for γ^a and γ^A and *S* must be a solution of the matrix equation,

$$\alpha^{[A]}=S^+\alpha^AS=V_a^A(\psi)\alpha^a. \tag{3.9}$$

Though V_a^A has a character of a Lorentz transformation, it has no infinitesimal prototype. Since $S^+=\rho_1S^{-1}\rho_1$, we also have a habitual $\gamma^{[A]}=\rho_1\alpha^{[A]}=S^{-1}\gamma^AS=V_a^A(\psi)\gamma^a$. However, in the basis of matrices $\gamma^{[A]}$, the Pauli-conjugated Dirac spinor must be defined as $\bar\psi=\psi^+\rho_1$ and *not* as $\bar\psi=\psi^+\gamma^{[0]}$.

The set O_p of 16 linearly independent elements of Clifford algebra comprised of various products of the γ^a - (or the γ^A-) matrices is in one-to-one correspondence with 16 Hermitian matrices, $\left(1,\rho_i,\sigma^i,\rho_i\sigma^k=\sigma^k\rho_i\right)$, $i,k=1,2,3$, where $\rho_1=\gamma^0$, $\rho_2=\gamma^1\gamma^2\gamma^3$, $\rho_3=i\gamma^0\gamma^1\gamma^2\gamma^3=i\rho_1\rho_2$ and $\sigma^i=i\rho_2\gamma^i=i\gamma^1\gamma^2\gamma^3\gamma^i=\rho_3\alpha^i$. The Dirac matrices, ρ_i and σ^i, satisfy the same commutation relations as the Pauli matrices, $\sigma^i\sigma^k=\delta_{ik}+i\epsilon_{ikl}\sigma^l$,

and $\rho_a\rho_b = \delta_{ab} + i\epsilon_{abc}\rho_c$. Finally, it is straightforward to check that the matrix $\rho_3 = i\gamma^0\gamma^1\gamma^2\gamma^3$ (commonly known as $-\gamma^5$) is an invariant of transformations (3.3),

$$\rho_{[3]} = \frac{i}{4!}\epsilon_{ABCD}\gamma^A\gamma^B\gamma^C\gamma^D = \frac{i}{4!}\epsilon_{ABCD}V_a^AV_b^BV_c^CV_d^D\gamma^a\gamma^b\gamma^c\gamma^d = \frac{i}{4!}\epsilon_{abcd}\gamma^a\gamma^b\gamma^c\gamma^d = \rho_3 \tag{3.10}$$

Then the matrix $\rho_2 = i\rho_1\rho_3$ is transformed like ρ_1, so that

$$\rho_{[3]} = S^{-1}\rho_3 S = \rho_3,\ \rho_{[i]} = S^+\rho_i S = \rho_i,\ \ i = 1,2. \tag{3.11}$$

As long as $S^+\rho_3 S = S^+S\rho_3 = \rho_3 S^+S = \rho_3\alpha^{[0]}$, the matrices σ on the $\mathbb{M}$, being defined as $\sigma^{[I]} = \rho_3\alpha^{[I]}$, are transformed as

$$\sigma^{[I]} = S^+\sigma^I S = \rho_3\left(V^I_{(0)}\alpha^{(0)} + V^I_j\alpha^j\right) = V^I_{(0)}\rho_3 + V^I_j\sigma^j \tag{3.12}$$

(as it should be for the spatial components of the axial current J^a)[2].

4. The Nonlinear Dirac Equation, Explicitly

So far, we have been studying the general geometric properties of the Dirac field in the scope of the affine geometry and carefully avoiding any assumptions about what *a solution of the Dirac equation* that has these properties can be. All the previously established [1] properties of the Dirac currents belong (along with the Dirac field itself) to the principal differentiable manifold $\mathbb{M}$. Without resorting to any particular coordinate manifold $\mathbb{R}^4$ we have established in [1] the following facts:

(i) The congruence of lines of the vector field $e^\mu_{[0]}$ is normal. The family $S_{(123)}$ of hypersurfaces, $\tau(x) = \text{const}$, of the constant world time τ is extrinsically flat; τ is a holonomic coordinate and it can be taken for x^0 in $\mathbb{R}^4$.

(ii) The congruence of lines of the vector field $e^\mu_{[3]}$ is normal and geodesic. The hypersurfaces $S_{(012)}$ of the constant radius ρ have constant extrinsic curvature and the holonomic coordinate ρ can serve as x^3 in $\mathbb{R}^4$.

(iii) The two-dimensional surfaces $S_{(12)}$ of constant τ and ρ are just spheres, *i.e.* umbilical (with two equal Gauss' curvatures) surfaces with constant mean (extrinsic) curvature $H = m\mathcal{P}/\mathcal{R} = -m\partial_{[3]}\ln\mathcal{R}$. The latter is determined by the Dirac field within principal manifold $\mathbb{M}$ and depends only on the radius ρ. The intrinsic (sectional) curvature, $R^t_{1212} = 2e\left(\partial_{[1]}A_{[2]} - \partial_{[2]}A_{[1]}\right) - 4e^2\left(A^2_{[1]} + A^2_{[2]}\right) = 2eF_{12} = 2eB_{[3]}$, is due to the *external* electromagnetic field. It coincides with projection of the magnetic field onto the direction of the axial current.

(iv) The two-dimensional surfaces $S_{(03)}$ are covered by a well-defined coordinate net formed by the streamlines of the vector and axial currents. This net can be identically mapped between the principal manifold $\mathbb{M}$ and the arithmetic $\mathbb{R}^4$.

These general observations can be summarized as follows. For any solution of the Dirac equation, which is not homogeneous in spatial directions, *the spherical symmetry is the property of a solution, thus being a dynamic symmetry*.

In order to find a solution of the Dirac equation, one has to specify a coordinate basis in $\mathbb{R}^4$ and a basis of the Dirac matrices. Here, we shall employ the numerical matrices α^a in spinor representation (A.7) and associate them with a tetrad $\boldsymbol{h}^\mu_{(a)}$. Then, $\alpha^A = V^A_a\alpha^a$, while the derivatives $\mathcal{D}_{[A]}$ will stay in the basis $\boldsymbol{e}_A$, which is associated with coordinate surfaces determined in the principal manifold $\mathbb{M}$. In this mixed representation, Dirac equation reads as

$$\begin{bmatrix} V^A_{(0)}+V^A_{(3)} & V^A_{(1)}-iV^A_{(2)} & 0 & 0 \\ V^A_{(1)}+iV^A_{(2)} & V^A_{(0)}-V^A_{(3)} & 0 & 0 \\ 0 & 0 & V^A_{(0)}-V^A_{(3)} & -V^A_{(1)}+iV^A_{(2)} \\ 0 & 0 & -V^A_{(1)}-iV^A_{(2)} & V^A_{(0)}+V^A_{(3)} \end{bmatrix} \begin{bmatrix} \mathcal{D}_A\left(u_R e^{i\phi^u_R}\right) \\ \mathcal{D}_A\left(d_R e^{i\phi^d_R}\right) \\ \mathcal{D}_A\left(u_L e^{i\phi^u_L}\right) \\ \mathcal{D}_A\left(d_L e^{i\phi^d_L}\right) \end{bmatrix} = -im \begin{bmatrix} u_L e^{i\phi^u_L} \\ d_L e^{i\phi^d_L} \\ u_R e^{i\phi^u_R} \\ d_R e^{i\phi^d_R} \end{bmatrix} \tag{4.1}$$

[2]Then the charge-conjugated spinor $\psi_c = \boldsymbol{C}\psi^* = \rho_2\sigma^2\psi^*$ becomes $\psi_c = \rho_2\sigma^{[2]}\psi^*$. In particular, $\Lambda^a_{(-)} = \psi^+\alpha^a\psi_c \to \Lambda^{[a]}_{(-)} = \psi^+\alpha^{[a]}S^{-1}\rho_2\sigma^2 S\psi^* = \psi^+S^+\alpha^a\rho_2\sigma^2 S\psi^*$. At the same time, $\gamma^{[0]}\gamma^{[t]} = S^{-1}\alpha^t S$ and $i\gamma^{[1]}\gamma^{[2]} = S^{-1}\sigma^3 S$, ...

The operators $\mathcal{D}_A$, which are copied from Equation (2.5), are as follows,

$$\begin{aligned}&\mathcal{D}_{[0]}=\partial_{[0]}-ieA_{[0]}+i\rho_3 e\tilde{A}_{[3]}, \qquad \mathcal{D}_{[1]}=\partial_{[1]}-ieA_{[1]}-e\tilde{A}_{[2]},\\ &\mathcal{D}_{[3]}=\partial_{[3]}-ieA_{[3]}+i\rho_3 e\tilde{A}_{[0]}-3Q/2, \quad \mathcal{D}_{[2]}=\partial_{[2]}-ieA_{[2]}+e\tilde{A}_{[1]},\end{aligned} \tag{4.2}$$

where ρ_3 differentiate between the right and left components and it stands for $+1$ for u_R, d_R and for -1 for u_L, d_L. The coordinate net formed by the integral lines of the tetrad vectors $\boldsymbol{e}_{[0]}$ and $\boldsymbol{e}_{[3]}$ that covers the two-dimensional surface $S_{(03)}$ in $\mathbb{M}$ is holonomic and the vectors $\boldsymbol{h}_{(0)}, \boldsymbol{h}_{(3)}$ in $\mathbb{R}^4$ can be chosen tangent to this surface. In order for the other two tetrad vectors, $\boldsymbol{h}_{(1)}$ and $\boldsymbol{h}_{(2)}$, to be normal to this surface, it is necessary that the components $V^{(1)}_{[0]}=V^{(2)}_{[0]}=V^{(1)}_{[3]}=V^{(2)}_{[3]}=0$. Just by inspection of Equations (A.9), we see that this is possible only when either $d_R=d_L=0$ or $u_R=u_L=0$. In both cases, as seen from Equations (A.10), we have $V^{(0)}_{[1]}=V^{(3)}_{[1]}=V^{(0)}_{[2]}=V^{(3)}_{[2]}=0$. In other words, the spacetime with the matter-induced anholonomic basis can be viewed as a direct product of the two-dimensional subspaces, $S_{(03)}\otimes S_{(12)}$. This is sufficient to treat the up- and down-polarizations separately,

$$\psi_u=\begin{bmatrix} u_R\exp\left(i\phi^u_R\right)\\ 0\\ u_L\exp\left(i\phi^u_L\right)\\ 0\end{bmatrix}, \quad \psi_d=\begin{bmatrix} 0\\ d_R\exp\left(i\phi^d_R\right)\\ 0\\ d_L\exp\left(i\phi^d_L\right)\end{bmatrix}. \tag{4.3}$$

Having only u_R, u_L or d_R, d_L components, the states ψ_u and ψ_d cannot bear quantum numbers of an angular momentum. For the up-polarized ψ_u, we have $J^{(3)}=+\left|J^{(3)}\right|$, $Q\equiv\partial_{[3]}\ln\mathcal{R}=-m\sin\mathcal{Y}<0$. In this case [C.f. (A.9)-(A.11)], $\mathcal{R}=\mathcal{R}_u=2u_Ru_L$ and the matrix $\alpha^{(a)}V^{[A]}_{(a)}$ in the l.h.s. of Equation (4.1) simplifies to

$$\begin{aligned}&V^{[0]}_{(0)}+V^{[0]}_{(3)}=V^{[3]}_{(0)}+V^{[3]}_{(3)}=u_R/u_L\,, V^{[0]}_{(0)}-V^{[0]}_{(3)}=V^{[3]}_{(3)}-V^{[3]}_{(0)}=u_L/u_R\,,\\ &V^{[1]}_{(1)}\pm iV^{[1]}_{(2)}=\mp i\left(V^{[2]}_{(1)}\pm iV^{[2]}_{(2)}\right)=\mathrm{e}^{\mp i\left(\phi^u_L+\phi^u_R\right)}.\end{aligned}$$

Accordingly, system (4.1) for ψ_u becomes

$$\begin{aligned}&u_R\left[\mathcal{D}_{[0]}+\mathcal{D}_{[3]}\right]u_R\mathrm{e}^{i\phi^u_R}=-imu^2_L\mathrm{e}^{i\phi^u_L}, \ \mathrm{e}^{-i\left(\phi^u_R+\phi^u_L\right)}\left[\mathcal{D}_{[1]}+i\mathcal{D}_{[2]}\right]u_R\mathrm{e}^{i\phi^u_R}=0,\\ &u_L\left[\mathcal{D}_{[0]}-\mathcal{D}_{[3]}\right]u_L\mathrm{e}^{i\phi^u_L}=-imu^2_R\mathrm{e}^{i\phi^u_R}, \ \mathrm{e}^{-i\left(\phi^u_R+\phi^u_L\right)}\left[\mathcal{D}_{[1]}+i\mathcal{D}_{[2]}\right]u_L\mathrm{e}^{i\phi^u_L}=0.\end{aligned} \tag{4.4}$$

For the down-polarized ψ_d, we have $J^{(3)}=-\left|J^{(3)}\right|$, $Q=\partial_{[3]}\ln\mathcal{R}=+m\sin\mathcal{Y}>0$. Here, $\mathcal{R}=\mathcal{R}_d=2d_Rd_L$ and the elements of the matrix in the l.h.s. of Equation (4.1) become,

$$\begin{aligned}&V^{[0]}_{(0)}-V^{[0]}_{(3)}=V^{[3]}_{(0)}-V^{[3]}_{(3)}=d_R/d_L\,, V^{[0]}_{(0)}+V^{[0]}_{(3)}=-\left(V^{[3]}_{(0)}+V^{[3]}_{(3)}\right)=d_L/d_R\,,\\ &-\left(V^{[1]}_{(1)}\pm iV^{[1]}_{(2)}\right)=\mp i\left(V^{[2]}_{(1)}\pm iV^{[2]}_{(2)}\right)=\mathrm{e}^{\mp i\left(\phi^d_L+\phi^d_R\right)}.\end{aligned}$$

Now, the system (4.1) reads as

$$\begin{aligned}&d_R\left[\mathcal{D}_{[0]}+\mathcal{D}_{[3]}\right]d_R\mathrm{e}^{i\phi^d_R}=-imd^2_L\mathrm{e}^{i\phi^d_L}, \ \mathrm{e}^{-i\left(\phi^d_R+\phi^d_L\right)}\left[\mathcal{D}_{[1]}+i\mathcal{D}_{[2]}\right]d_R\mathrm{e}^{i\phi^d_R}=0,\\ &d_L\left[\mathcal{D}_{[0]}-\mathcal{D}_{[3]}\right]d_L\mathrm{e}^{i\phi^d_L}=-imd^2_R\mathrm{e}^{i\phi^d_R}, \ \mathrm{e}^{-i\left(\phi^d_R+\phi^d_L\right)}\left[\mathcal{D}_{[1]}+i\mathcal{D}_{[2]}\right]d_L\mathrm{e}^{i\phi^d_L}=0.\end{aligned} \tag{4.5}$$

Remembering about the sign due to ρ_3, we obtain the following formulae for all the differential operators involved,

$$\begin{aligned}
&\mathcal{D}_{[0]}+\mathcal{D}_{[3]}=\partial_{[0]}+\left(\partial_{[3]}-\frac{3}{2}Q\right)-ie\left[\left(A_{[0]}-\tilde{A}_{[0]}\right)+\left(A_{[3]}-\tilde{A}_{[3]}\right)\right],\\
&\mathcal{D}_{[0]}-\mathcal{D}_{[3]}=\partial_{[0]}-\left(\partial_{[3]}-\frac{3}{2}Q\right)-ie\left[\left(A_{[0]}-\tilde{A}_{[0]}\right)-\left(A_{[3]}-\tilde{A}_{[3]}\right)\right],\\
&\mathcal{D}_{[1]}+i\mathcal{D}_{[2]}=\partial_{[1]}+i\partial_{[2]}-ie\left[\left(A_{[1]}-\tilde{A}_{[1]}\right)+i\left(A_{[2]}-\tilde{A}_{[2]}\right)\right],\\
&\mathcal{D}_{[1]}-i\mathcal{D}_{[2]}=\partial_{[1]}-i\partial_{[2]}-ie\left[\left(A_{[1]}+\tilde{A}_{[1]}\right)-i\left(A_{[2]}+\tilde{A}_{[2]}\right)\right].
\end{aligned}\tag{4.6}$$

In Equations (4.4) and (4.5), the operator $\mathcal{D}_{[0]}+\mathcal{D}_{[3]}$ acts only on u_R and d_R while $\mathcal{D}_{[0]}-\mathcal{D}_{[3]}$ only on u_L and d_L.

5. Solutions of the Nonlinear Equations

So far we were expanding the vector of spacetime displacement dx^μ in terms of the basis $\boldsymbol{e}_A$ of the tetrad determined by the Dirac currents $\mathrm{d}x^\mu=e^\mu_A\mathrm{d}S^A$. But the true physical variables are the world time τ and the distance ρ. They are holonomic coordinates, because $\mathrm{d}\tau=\mathcal{R}\mathrm{d}S^{[0]}$ and $\mathrm{d}\rho=\mathcal{R}\mathrm{d}S^{[3]}$ are the total differentials of the independent coordinates $\mathrm{d}x^\mu\in\mathbb{R}^4$,

$$\tau_2-\tau_1=\int_{x(\tau_1)}^{x(\tau_2)}j_\mu(x)\mathrm{d}x^\mu=\int\mathcal{R}\mathrm{d}S^{[0]},\ \rho_2-\rho_1=\int_{x(\rho_1)}^{x(\rho_2)}J_\mu(x)\mathrm{d}x^\mu=\int\pm\mathcal{R}\mathrm{d}S^{[3]}.\tag{5.1}$$

Here, the upper sign is for the ψ_u, where $\mathcal{J}_3=u_R^2+u_L^2>0$. The lower sign is for ψ_d, where $\mathcal{J}_3=-d_R^2-d_L^2<0$ and the axial current is directed inward. The world time τ and the radial variable ρ, being defined as invariants in $\mathbb{M}$, can immediately be used in arithmetic $\mathbb{R}^4$.

5.1. Reduction to the Physical Variables

At the points where $j^{(3)}=V^{(3)}_{[0]}=0$ and $\mathcal{J}^{(0)}=V^{(0)}_{[3]}=0$ (in general, a 2-d surface) the relation between spatial components, $[\boldsymbol{\Theta}\times\boldsymbol{\Phi}]/\mathcal{R}^2=+\boldsymbol{\mathcal{J}}/\mathcal{R}$, is an algebraic identity. For the axial current directed outward, *i.e.* $\mathcal{J}_3>0$, we take $\mathcal{J}^\mu=+\mathcal{R}e^\mu_{[3]}$, $\Theta^\mu=\mathcal{R}e^\mu_{[1]}$ and $\Phi^\mu=\mathcal{R}e^\mu_{[2]}$, so that $\boldsymbol{e}_{[3]}=\left[\boldsymbol{e}_{[1]}\times\boldsymbol{e}_{[2]}\right]$. In this case, we change the variables in Equation (4.4) as follows,

$$\begin{aligned}
&\partial_{[0]}\to m\mathcal{R}\partial_\tau,\partial_{[3]}\to m\mathcal{R}\partial_\rho,eA_{[0]}\to\frac{e}{m}\mathcal{R}A_\tau,eA_{[3]}\to\frac{e}{m}\mathcal{R}A_\rho,\\
&\frac{\omega_{120}}{2}=e\tilde{A}_{[0]}\to\frac{e}{m}\mathcal{R}\tilde{A}_\tau,\frac{\omega_{123}}{2}=e\tilde{A}_{[3]}\to\frac{e}{m}\mathcal{R}\tilde{A}_\rho,e\tilde{A}_{[1]}\to\frac{e}{m}\tilde{A}_{[1]},\ e\tilde{A}_{[2]}\to\frac{e}{m}\tilde{A}_{[2]}
\end{aligned}\tag{5.2}$$

Adopting the physical variables (5.2) in Equations (4.4) we obtain the equations that eventually must be solved. In these equations, according to (4.6), there is an operator $\left(\partial_{[3]}-\frac{3}{2}\partial_{[3]}\ln\mathcal{R}\right)f=\mathcal{R}^{3/2}\partial_{[3]}\left(\mathcal{R}^{-3/2}f\right)$ $=\mathcal{R}\cdot\mathcal{R}^{3/2}\partial_\rho\left(\mathcal{R}^{-3/2}f\right)$. Since $\partial_A\mathcal{R}=\partial_A\mathcal{S}=\partial_A\mathcal{P}=\partial_A\mathcal{Y}=0$ for $A=0,1,2$, a simple calculation with $\partial_\tau\mathcal{R}=\partial_\tau\mathcal{Y}=0$ yields the system,

$$\begin{aligned}
&\frac{\mathcal{R}}{2}\left(\partial_\rho+\partial_\tau\right)\left(\frac{u_R^2}{\mathcal{R}^3}\right)+i\mathcal{R}\left(\frac{u_R^2}{\mathcal{R}^3}\right)\left(\partial_\rho+\partial_\tau\right)\phi^u_R=-i\left(\frac{u_L^2}{\mathcal{R}^3}\right)\mathrm{e}^{+i\mathcal{Y}_u},&\text{(a)}\\
&-\frac{\mathcal{R}}{2}\left(\partial_\rho-\partial_\tau\right)\left(\frac{u_L^2}{\mathcal{R}^3}\right)-i\mathcal{R}\left(\frac{u_L^2}{\mathcal{R}^3}\right)\left(\partial_\rho-\partial_\tau\right)\phi^u_L=-i\left(\frac{u_R^2}{\mathcal{R}^3}\right)\mathrm{e}^{-i\mathcal{Y}_u},&\text{(b)}\\
&\mathrm{e}^{-i\left(\phi^u_L+\phi^u_R\right)}\left[\partial_{[1]}+i\partial_{[2]}\right]u_R\mathrm{e}^{i\phi^u_R}=0,&\text{(c)}\\
&\mathrm{e}^{-i\left(\phi^u_L+\phi^u_R\right)}\left[\partial_{[1]}+i\partial_{[2]}\right]u_L\mathrm{e}^{i\phi^u_L}=0,&\text{(d)}
\end{aligned}\tag{5.3}$$

where $\mathcal{Y}_u=\phi^u_L-\phi^u_R$. For the axial current directed inward, in order to preserve an intuitive physical understanding of a distance *from* an object, we want $\boldsymbol{e}_{[3]}$ be directed outward. Then the triplet $\left(\boldsymbol{e}_{[1]},\boldsymbol{e}_{[2]},\boldsymbol{e}_{[3]}\right)$ will be

left-handed. We have to take $\mathcal{J}^{\mu}=-\mathcal{R}e^{\mu}_{[3]}$, $\Theta^{\mu}=\mathcal{R}e^{\mu}_{[2]}$, and $\Phi^{\mu}=\mathcal{R}e^{\mu}_{[1]}$ in order for the vector product $\left[\boldsymbol{e}_{[1]}\times\boldsymbol{e}_{[2]}\right]=\boldsymbol{e}_{[3]}$ to represent the external normal and the triplet $\left(\boldsymbol{e}_{[1]},\boldsymbol{e}_{[2]},\boldsymbol{e}_{[3]}\right)$ to be right-handed. This results in the interchange of the tetrad indices $1\leftrightarrow 2$ in Equations (2.3), or, equivalently, in the change of the sign of the tetrad components of the vector potential, $e\tilde{A}_{B}\to -e\tilde{A}_{B}$. Thus, the string of the change of variables becomes

$$\begin{gathered}\partial_{[0]}\to m\mathcal{R}\partial_{\tau},\ \partial_{[3]}\to -m\mathcal{R}\partial_{\rho},\ eA_{[0]}\to\frac{e}{m}\mathcal{R}A_{\tau},\ eA_{[3]}\to -\frac{e}{m}\mathcal{R}A_{\rho},\\ e\tilde{A}_{[0]}\to -\frac{e}{m}\mathcal{R}\tilde{A}_{\tau},\ e\tilde{A}_{[3]}\to +\frac{e}{m}\mathcal{R}\tilde{A}_{\rho},\ e\tilde{A}_{[1]}\to -\frac{e}{m}\tilde{A}_{[1]},\ e\tilde{A}_{[2]}\to -\frac{e}{m}\tilde{A}_{[2]}.\end{gathered}\tag{5.4}$$

Note, that in the course of the change of variables outlined above, the sign of the $e\tilde{A}_{[3]}$ has been changed twice. Now, using the physical variables (5.4) in Equations (4.5) we arrive at a similar system,

$$\begin{aligned}&-\frac{\mathcal{R}}{2}\left(\partial_{\rho}-\partial_{\tau}\right)\left(\frac{d_R^2}{\mathcal{R}^3}\right)+i\mathcal{R}\left(\frac{d_R^2}{\mathcal{R}^3}\right)\left[\left(\partial_{\tau}-\partial_{\rho}\right)\phi_R^d-\frac{2e}{m}\left(A_{\tau}-A_{\rho}\right)\right]=-i\left(\frac{d_L^2}{\mathcal{R}^3}\right)\mathrm{e}^{i\mathcal{Y}_d}, && \text{(a)}\\ &\frac{\mathcal{R}}{2}\left(\partial_{\rho}+\partial_{\tau}\right)\left(\frac{d_L^2}{\mathcal{R}^3}\right)+i\mathcal{R}\left(\frac{d_L^2}{\mathcal{R}^3}\right)\left[\left(\partial_{\tau}+\partial_{\rho}\right)\phi_L^d-\frac{2e}{m}\left(A_{\tau}+A_{\rho}\right)\right]=-i\left(\frac{d_R^2}{\mathcal{R}^3}\right)\mathrm{e}^{-i\mathcal{Y}_d}, && \text{(b)}\\ &\mathrm{e}^{-i\left(\phi_L^d+\phi_R^d\right)}\left[\partial_{[1]}+i\partial_{[2]}-\frac{2ie}{m}\left(A_{[1]}+iA_{[2]}\right)\right]d_R\mathrm{e}^{i\phi_R^d}=0, && \text{(c)}\\ &\mathrm{e}^{-i\left(\phi_L^d+\phi_R^d\right)}\left[\partial_{[1]}+i\partial_{[2]}-\frac{2ie}{m}\left(A_{[1]}+iA_{[2]}\right)\right]d_L\mathrm{e}^{i\phi_L^d}=0, && \text{(d)}\end{aligned}\tag{5.5}$$

where $\mathcal{Y}_d=\phi_L^d-\phi_R^d$. The difference between ψ_u and ψ_d is seen right in the equations of motion. The tetrad components of an external field along holonomic coordinates, $A_{\tau},A_{\rho}\in S_{(03)}$, affect only ψ_d-mode. The associated with the non-holonomic coordinates angular components $A_{[1]},A_{[2]}\in S_{(12)}$ are assembled as the ladder operators and affect ψ_d pushing it up to the state ψ_u. This difference between the last two equations of systems (5.3) and (5.5) points to a generic instability of the ψ_d-mode[3]. It is discussed in Section 7.

5.2. Reduction to the Real-Valued Functions

As the last step before solving systems (5.3) and (5.5) we split real and imaginary parts of the first two equations of these systems and reduce equations to a form convenient for finding the solutions. For the mode ψ_u the result reads as

$$\begin{aligned}&\frac{\mathcal{R}}{2}\left(\frac{\partial}{\partial\rho}+\frac{\partial}{\partial\tau}\right)\left(\frac{u_R^2}{\mathcal{R}^3}\right)=\left(\frac{u_L^2}{\mathcal{R}^3}\right)\sin\mathcal{Y}_u, && \text{(a)}\\ &\mathcal{R}\left(\frac{\partial}{\partial\rho}+\frac{\partial}{\partial\tau}\right)\phi_R^u=-\frac{u_L^2}{u_R^2}\cos\mathcal{Y}_u && \text{(a′)}\\ &\frac{\mathcal{R}}{2}\left(\frac{\partial}{\partial\rho}-\frac{\partial}{\partial\tau}\right)\left(\frac{u_L^2}{\mathcal{R}^3}\right)=\left(\frac{u_R^2}{\mathcal{R}^3}\right)\sin\mathcal{Y}_u, && \text{(b)}\\ &\mathcal{R}\left(\frac{\partial}{\partial\rho}-\frac{\partial}{\partial\tau}\right)\phi_L^u=\frac{u_R^2}{u_L^2}\cos\mathcal{Y}_u. && \text{(b′)}\end{aligned}\tag{5.6}$$

For the mode ψ_d the result is somewhat different,

[3]Since $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are the "angular" directions, it is instructive to recall that the operators $L_{+}=L_1+iL_2$ are ladder operators for the angular momentum that moves eigenstate of the L_z up. Both systems (5.3) and (5.5) contain only L_{+}. While ψ_u cannot be pushed further up (and is stable), the ψ_d is readily pushed up to the ψ_u. One can view these transitions as a manifestation of the ψ_d-waveform's "motion". In fact, it is a flow of surrounding Dirac matter with $\mathcal{R}\geq 1$ that looks like a motion of the ψ_d-dip (or void).

$$\begin{aligned}
&\frac{\mathcal{R}}{2}\left(\frac{\partial}{\partial\rho}-\frac{\partial}{\partial\tau}\right)\left(\frac{d_R^2}{\mathcal{R}^3}\right)=-\left(\frac{d_L^2}{\mathcal{R}^3}\right)\sin\mathcal{Y}_d, && \text{(a)}\\
&\mathcal{R}\left(\frac{\partial}{\partial\rho}-\frac{\partial}{\partial\tau}\right)\phi_R^d=\frac{d_L^2}{d_R^2}\cos\mathcal{Y}_d+\frac{2e}{m}\mathcal{R}\left(A_\rho-A_\tau\right), && \text{(a}'\text{)}\\
&\frac{\mathcal{R}}{2}\left(\frac{\partial}{\partial\rho}+\frac{\partial}{\partial\tau}\right)\left(\frac{d_L^2}{\mathcal{R}^3}\right)=-\left(\frac{d_R^2}{\mathcal{R}^3}\right)\sin\mathcal{Y}_d, && \text{(b)}\\
&\mathcal{R}\left(\frac{\partial}{\partial\rho}+\frac{\partial}{\partial\tau}\right)\phi_L^d=-\frac{d_R^2}{d_L^2}\cos\mathcal{Y}_d+\frac{2e}{m}\mathcal{R}\left(A_\rho+A_\tau\right). && \text{(b}'\text{)}
\end{aligned} \tag{5.7}$$

The phases ϕ_R^u and ϕ_L^u are affected in ψ_u by the right and left lightlike components of the vector potential, respectively, but with the coupling constant $2e$. Conversely, the phases ϕ_L^d and ϕ_R^d of the ψ_d are not affected at all.

Next, adding and subtracting Equations (5.6.a′) and (5.6.b′) and recalling that $\phi_L^u-\phi_R^u=\mathcal{Y}_u$ we find that

$$\begin{aligned}
&\mathcal{R}\frac{\partial\mathcal{Y}_u}{\partial\rho}=\mathcal{R}\frac{\partial\mathcal{Z}_u}{\partial\tau}+\left(\mathcal{X}_u^2+\frac{1}{\mathcal{X}_u^2}\right)\cos\mathcal{Y}_u, && \text{(a)}\\
&\frac{\mathrm{d}\mathcal{R}}{\mathrm{d}\rho}=-\sin\mathcal{Y}_u, && \text{(b)}\\
&\mathcal{R}\frac{\partial\mathcal{Z}_u}{\partial\rho}=-\left(\mathcal{X}_u^2-\frac{1}{\mathcal{X}_u^2}\right)\cos\mathcal{Y}_u. && \text{(c)}\\
&\left[\partial_{[1]}+i\partial_{[2]}\right]\mathcal{Z}_u=0, && \text{(d)}
\end{aligned} \tag{5.8}$$

where $\mathcal{Z}_u=\phi_L^u+\phi_R^u$ and $u_L/u_R=\mathcal{X}_u$. Repeating the same for the mode ψ_d we obtain,

$$\begin{aligned}
&-\mathcal{R}\frac{\partial\mathcal{Y}_d}{\partial\rho}=\mathcal{R}\frac{\partial\mathcal{Z}_d}{\partial\tau}+\left(\mathcal{X}_d^2+\frac{1}{\mathcal{X}_d^2}\right)\cos\mathcal{Y}_d-\frac{4e}{m}\mathcal{R}A_\tau, && \text{(a)}\\
&\frac{\mathrm{d}\mathcal{R}}{\mathrm{d}\rho}=+\sin\mathcal{Y}_d, && \text{(b)}\\
&\mathcal{R}\frac{\partial\mathcal{Z}_d}{\partial\rho}=-\left(\mathcal{X}_d^2-\frac{1}{\mathcal{X}_d^2}\right)\cos\mathcal{Y}_d+\frac{4e}{m}\mathcal{R}A_\rho. && \text{(c)}\\
&\left[\partial_{[1]}+i\partial_{[2]}\right]\mathcal{Z}_d=\frac{4e}{m}\left(A_{[1]}+iA_{[2]}\right), && \text{(d)}
\end{aligned} \tag{5.9}$$

where $\mathcal{Z}_u=\phi_L^d+\phi_R^d$ and $d_L/d_R=\mathcal{X}_d$. Equations (5.8.d) and (5.9.d) are easily obtained from Equations (5.3.c,d) and (5.5.c,d) because none of the amplitudes u_R,u_L and d_R,d_L and of the phase differences $\mathcal{Y}_u,\mathcal{Y}_d$ depend on the angular variables $S^{[1]}$ and $S^{[2]}$. We postpone discussion of the Equations (5.3.c,d) and (5.5.c,d), which are responsible for the stability or instability of the solutions, till Section 7.

Before looking for the stationary modes of the nonlinear Dirac equation we are going to learn whether they can emerge as asymptotic configurations at $\tau\to\infty$ of a transient process that can begin from an arbitrary perturbation or are they *ad hoc* constructed isolated solutions. By adding and subtracting Equations (5.6.a,b), with the l.h.s. reduced to the logarithmic derivatives, and some simple algebra we obtain

$$\frac{\partial\mathcal{X}_u}{\partial\tau}=\mathcal{X}_u\left(\mathcal{X}_u-\frac{1}{\mathcal{X}_u}\right)^2\frac{\partial\ln\mathcal{R}}{\partial\rho},\quad \frac{\partial\mathcal{X}_u}{\partial\rho}=\mathcal{X}_u\left(\mathcal{X}_u^2-\frac{1}{\mathcal{X}_u^2}\right)\frac{\partial\ln\mathcal{R}}{\partial\rho}, \tag{5.10}$$

where $\partial_\rho\ln\mathcal{R}=-\sin\mathcal{Y}_u/\mathcal{R}$. Excluding from these two equations the $\partial_\rho\ln\mathcal{R}$, one finds a first-order wave equation, $\partial_\tau\mathcal{X}+c(\mathcal{X})\partial_\rho\mathcal{X}=0$, with the wave velocity $c(\mathcal{X})=\left(1-\mathcal{X}^2\right)/\left(1+\mathcal{X}^2\right)$. Because $c(1)=0$, the "propagation" of $\mathcal{X}$ stops at $\mathcal{X}=1$. Since $\mathcal{R}$ depends only on ρ, both Equations (5.10) are easily integrated,

$$\mathcal{X}_u^2(\tau,\rho)=1-\frac{1}{2\partial_\rho \ln\mathcal{R}\cdot\tau+C_2(\rho)},\quad \mathcal{X}_u^2(\tau,\rho)=\frac{\mathcal{R}^4-C_1(\tau)}{\mathcal{R}^4+C_1(\tau)},\tag{5.11}$$

where the constants of integration $C_1(\tau)$ and $C_2(\rho)$ are arbitrary functions of only one argument. Since $\mathcal{X}_u(\infty,\rho)=1$ (and then $C_1(\infty)=0$), we find that at the asymptotic world time τ the coefficients in front of $\cos\mathcal{Y}_u(\rho)$ in Equations (5.8.a) and (5.8.c) become 2 and 0, respectively. Assuming further that $e=0$ (no external field), we find that $\partial_\rho\mathcal{Z}_u=0$ and thus $\mathcal{Z}_u=\mathcal{Z}_u(\tau)$. Now, $\partial_\tau\mathcal{Z}_u$ is the only potentially τ-dependent term in Equation (5.8.a); then it cannot depend on τ. Therefore, the only option is $\partial_\tau\mathcal{Z}_u=-2E=\text{const}$, $\mathcal{Z}_u=-2E\tau$, and it immediately follows that $u_L^2=u_R^2=u^2=\mathcal{R}/2$ (which is an evidence that the particle is at rest!). Equations (5.11) are compatible only in the limit of $\tau\to\infty$ since they imply $\partial_\tau\mathcal{R}=0$; a transient process naturally requires that $\partial_\tau\mathcal{R}\neq 0$. Similar results are true for the mode ψ_d.

6. Stationary Solutions

Being interested here only in stationary states we assume a trivial dependence of the phases of Dirac field components on τ, $\psi\propto e^{-iE\tau}$, and replace, $\phi_R\to\phi_R(\rho)-E\tau$, $\phi_L\to\phi_L(\rho)-E\tau$, throughout this section. Then, $u_L^2=u_R^2=u^2=\mathcal{R}/2$ and $d_L^2=d_R^2=d^2=\mathcal{R}/2$. Taking further the coupling constant $e=0$, which is, in fact, equivalent to a one-body approximation, we end up with an autonomous system of two ODEs for two unknown functions (the amplitude $\mathcal{R}(\rho)$ and the phase difference $\mathcal{Y}(\rho)$) of the natural parameter ρ (and not the affine parameter *s*!) along the radial geodesic lines.

6.1. Localized Solution for the ψ_u-Mode of the Dirac Field

In the stationary case, Equations (5.8) for the ψ_u-mode with the axial current directed outward, read as

$$\begin{aligned}\mathcal{R}(\rho)\frac{\mathrm{d}\mathcal{Y}_u(\rho)}{\mathrm{d}\rho}&=-2\epsilon\mathcal{R}(\rho)+2\cos\mathcal{Y}_u(\rho),\qquad\text{(a)}\\ \mathcal{R}(\rho)\frac{\mathrm{d}\mathcal{R}(\rho)}{\mathrm{d}\rho}&=-\mathcal{R}(\rho)\sin\mathcal{Y}_u(\rho),\qquad\text{(b)}\end{aligned}\tag{6.1}$$

where $\epsilon=E/m$. The characteristic equation for this system,

$$\frac{\mathrm{d}\mathcal{Y}_u}{-2\epsilon\mathcal{R}+2\cos\mathcal{Y}_u}=-\frac{\mathrm{d}\mathcal{R}}{\mathcal{R}\sin\mathcal{Y}_u},\tag{6.2}$$

is easily solved in terms of $w(\mathcal{R})=\cos\mathcal{Y}_u$. Then, $\mathcal{R}w'_{\mathcal{R}}-2w+2\epsilon\mathcal{R}=0$, and

$$\cos Y_u=C\mathcal{R}^2+2\epsilon\mathcal{R},\tag{6.3}$$

is the first integral of system (1) depending on one, yet undetermined, constant *C*.

1. General (periodic) solution. Solving Equation (6.3) for $\mathcal{R}$, and taking into account two possible signs of *C*, one can rewrite Equation (1a) as

$$\frac{\mathrm{d}\mathcal{Y}}{\mathrm{d}\rho}=\mp2\sqrt{C}\cdot\sqrt{\frac{\epsilon^2}{C}+\cos\mathcal{Y}},\ C>0\ \text{ and }\ \frac{\mathrm{d}\mathcal{Y}}{\mathrm{d}\rho}=\mp2\sqrt{|C|}\cdot\sqrt{\frac{\epsilon^2}{|C|}-\cos\mathcal{Y}}.\ C<0.\tag{6.4}$$

Thus, the dependence $\rho(\mathcal{Y})$ in the cases $C>0$ and $C<0$ is given by the quadratures [2],

$$\rho(\mathcal{Y})=\frac{\mp1}{\sqrt{C(1+b^2)}}\int_0^{\mathcal{Y}/2}\frac{\mathrm{d}\phi}{\sqrt{1-\frac{2}{1+b^2}\sin^2\phi}}=\frac{\mp1}{\sqrt{C(1+b^2)}}F\left(\frac{\mathcal{Y}}{2}\middle|\frac{2}{1+b^2}\right)],\ C>0,\tag{6.5}$$

$$\rho(\mathcal{Y})=\frac{\mp1}{\sqrt{|C|(b^2-1)}}\int_0^{\mathcal{Y}/2}\frac{\mathrm{d}\phi}{\sqrt{1-\frac{2}{1-b^2}\sin^2\phi}}=\frac{\mp1}{\sqrt{C(b^2-1)}}F\left(\frac{\mathcal{Y}}{2}\middle|\frac{2}{1-b^2}\right)0,\ C<0,\tag{6.6}$$

where $b^2=\epsilon^2/|C|>0$ and $w=F(\Phi\,|\,k^2)=\mathrm{sn}^{-1}(\sin\Phi\,|\,k^2)$ is the incomplete elliptic integral of the first

kind[4],

$$F\left(\Phi\,|\,k^2\right)=\int_0^{\Phi}\left(1-k^2\sin^2\phi\right)^{-1/2}\mathrm{d}\phi=\int_0^{X}\left[\left(1-x^2\right)\left(1-k^2x^2\right)\right]^{-1/2}\mathrm{d}x,\quad X=\sin\Phi. \tag{6.7}$$

Its inverse is a well-known Jacobi's amplitude function, $\Phi=\mathrm{am}\left(w\,|\,k^2\right)$. Leaving aside for a while the case of $C<0$, we readily find that

$$\begin{aligned}&\sin\frac{\mathcal{Y}}{2}=\mathrm{sn}\left(u\,|\,k^2\right),\ \cos\frac{\mathcal{Y}}{2}=\mathrm{cn}\left(u\,|\,k^2\right),\\&\sin\mathcal{Y}=2\mathrm{sn}\left(u\,|\,k^2\right)\mathrm{cn}\left(u\,|\,k^2\right),\ \cos\mathcal{Y}=\mathrm{cn}^2\left(u\,|\,k^2\right)-\mathrm{sn}^2\left(u\,|\,k^2\right),\end{aligned} \tag{6.8}$$

where $u=\sqrt{\epsilon^2+C}\ \rho=F\left(\mathcal{Y}/2\,|\,2/\left(1+b^2\right)\right)$, $k^2=2/\left(1+b^2\right)$. Now Equation (6.1b) becomes,

$$\frac{\mathrm{d}\mathcal{R}(\rho)}{\mathrm{d}\rho}=-\sin\mathcal{Y}(\rho)=-2\mathrm{sn}\left(u\,|\,k^2\right)\mathrm{cn}\left(u\,|\,k^2\right), \tag{6.9}$$

and, since $\int\mathrm{sn}\left(u\,|\,k^2\right)\mathrm{cn}\left(u\,|\,k^2\right)\mathrm{d}u=-\mathrm{dn}\left(u\,|\,k^2\right)/k^2$ [2], the latter equation is readily integrated,

$$\mathcal{R}(\rho)=\frac{\sqrt{\epsilon^2+C}}{C}\,\mathrm{dn}\left(\sqrt{\epsilon^2+C}\,\rho\,\middle|\,\frac{2}{1+b^2}\right),\ C>0. \tag{6.10}$$

In the second case of $C<0$ we would have

$$\mathcal{R}(\rho)=\frac{\sqrt{\epsilon^2-|C|}}{|C|}\,\mathrm{dn}\left(\sqrt{\epsilon^2-|C|}\,\rho\,\middle|\,\frac{2}{1-b^2}\right),\ C<0. \tag{6.11}$$

The Jacobi's elliptic functions $\mathrm{sn}\left(u\,|\,k^2\right)$, $\mathrm{cn}\left(u\,|\,k^2\right)$ and $\mathrm{dn}\left(u\,|\,k^2\right)$ are known to be double-periodic functions of their argument. While periodic behavior of the phase $\mathcal{Y}(\rho)$ cannot *a priori* be excluded, *periodicity in radial direction is impossible for the invariant density* $\mathcal{R}(\rho)$, simply because it would conflict with the physical localization.

2. Localized (aperiodic) solution. There is, however, a special case when the module of the elliptic function $k=1$ and the periodicity disappears (the period becomes infinite). For the Equation (6.10), this means that $b^2=\epsilon^2/|C|=1$ so that $\mathrm{dn}\left(u\,|\,1\right)=1/\cosh u$ (as well as $\mathrm{cn}\left(u\,|\,1\right)=1/\cosh u$ and $\mathrm{sn}\left(u\,|\,1\right)=\tanh u$). For the Equation (6.11) the same would mean $b^2=-1$, which is impossible, since $b^2>0$, by definition. Hence, the case of $C<0$ must be dropped from further consideration.

The constant C of integration in the Equation (6.3) is now uniquely determined as $C=\epsilon^2=\left(E/m\right)^2$, and the equation of characteristics of system (6.1) becomes

$$\cos\mathcal{Y}+1=2\cos^2\left(\mathcal{Y}/2\right)=\left(\epsilon\mathcal{R}+1\right)^2. \tag{6.12}$$

Since the Jacobi's elliptic functions with module $k=1$ are elementary functions, it is much easier to return to the original system (6.1) and the characteristic equation (6.12) with $C=\epsilon^2$, using the latter as a constraint. After using the constraint (with the signs to be determined later), $\epsilon\mathcal{R}+1=\pm\sqrt{2}\cos\left(\mathcal{Y}_u/2\right)$, the system (6.1) simplifies to

$$\begin{aligned}&\frac{\mathrm{d}\mathcal{Y}_u}{\mathrm{d}\rho}=-2^{3/2}\epsilon\cos\frac{\mathcal{Y}_u}{2}, &&\text{(a)}\\&\frac{\mathrm{d}\mathcal{R}}{\mathrm{d}\rho}=-\sin\mathcal{Y}_u=-2\sin\frac{\mathcal{Y}_u}{2}\cos\frac{\mathcal{Y}_u}{2}, &&\text{(b)}\end{aligned} \tag{6.13}$$

and its first equation is readily integrated to $\rho(\mathcal{Y})$ first, and then yields $\mathcal{Y}(\rho)$

[4]These expressions have no practical value and will be used below for a sole purpose of proving that the modules of the elliptic integrals must equal +1 by the physics of the problem. Then, and only then is $\mathcal{R}(\rho)$ not oscillating in radial direction. This uniquely fixes the constant as $|C|=\epsilon^2$ and guarantee that elliptic integrals become smooth elementary functions. The limits of integration in (6.5) are tentative.

$$\sqrt{2}\epsilon\rho = \tanh^{-1}\left(\sin\frac{\mathcal{Y}_u}{2}\right),\ \sin\frac{\mathcal{Y}_u}{2} = -\tanh\left(\sqrt{2}\,\epsilon\rho\right),\ \cos\frac{\mathcal{Y}_u}{2} = \frac{1}{\cosh\left(\sqrt{2}\,\epsilon\rho\right)}. \tag{6.14}$$

When $\rho\to\infty$, we have $\epsilon\mathcal{R}+1\to 0$, which is possible only when $\epsilon = E/m < 0$. We also obtain the anticipated $\sin\mathcal{Y}(\infty)=0$ and $\cos\mathcal{Y}_u(\infty)=-1$, *i.e.* $\mathcal{Y}_u(\infty)=\pi$. Returning the result of integration into Equations (6.12) and (6.13b), we simplify the latter to

$$\epsilon\mathcal{R}+1 = \frac{-\sqrt{2}}{\cosh\left(\sqrt{2}\,\epsilon\rho\right)},\ \frac{\mathrm{d}\mathcal{R}}{\mathrm{d}\rho} = -\sin\mathcal{Y}_u(\rho) = -2\frac{\sinh\left(\sqrt{2}|\epsilon|\rho\right)}{\cosh^2\left(\sqrt{2}\epsilon\rho\right)}. \tag{6.15}$$

In order for this solution to be interpreted as an isolated particle at rest, we must require that $E=-m$. Thus the solution

$$\mathcal{R}(\rho) = \frac{\sqrt{2}}{\cosh\left(\sqrt{2}\rho\right)}+1, \tag{6.16}$$

is the mode with the negative energy with respect to the vacuum level zero attributed to $\mathcal{R}=1$. Finally, in natural units,

$$\sin\frac{\mathcal{Y}_u}{2} = \tanh\left(\sqrt{2}\,m\rho\right),\ \mathcal{R}(\rho) = \frac{\sqrt{2}}{\cosh\left(\sqrt{2}m\rho\right)}+1. \tag{6.17}$$

This result also follows from Equation (6.9), since $\mathrm{dn}(u\,|\,1) = 1/\cosh u$. We can take the radius ρ_0 of the spherical surface, where $\mathrm{d}\mathcal{R}/\mathrm{d}\rho$ reaches its maximum (the inflection point) for the size of the particle. Here, $\sin\mathcal{Y}_u(\rho_0)=1$, and, consequently, $\sinh\left(\sqrt{2}\,m\rho_0\right)=1$, $\cosh\left(\sqrt{2}\,m\rho_0\right)=\sqrt{2}$. Therefore (in natural units),

$$\rho_0 = \frac{\sinh^{-1}(1)}{\sqrt{2}m} = \frac{0.623}{m}\ \text{and}\ s_0 = \frac{\rho_0}{\mathcal{R}(\rho_0)} = \frac{1}{m},$$

as it was previously contemplated. At the radius ρ_0, also as expected, the phase is $\mathcal{Y}_u(\rho_0)=\pi/2$. Indeed, $\cos\left(\mathcal{Y}_u(\rho_0)/2\right) = 1/\sqrt{2} = \cos(\pi/4)$ and $\sin\mathcal{Y}_u(\rho_0)=1$, $\mathcal{R}_u(\rho_0)=2$. The peak amplitude $\mathcal{R}_u(0)=1+\sqrt{2}$.

6.2. Dirac Field in ψ_d-Mode

We expect that in real world the mode ψ_d with the axial current looking inward will be unstable and not similar, even qualitatively, to the mode ψ_u. However, it is instructive to repeat the previous steps and consider only Equations (5.7) leaving aside Equations (5.5.c,d). Then most of the analysis remains the same and only Equations (6.1) and (6.12)-(6.16) are modified. Equations (6.1) now read as

$$\begin{aligned}\mathcal{R}(\rho)\frac{\mathrm{d}\mathcal{Y}_d(\rho)}{\mathrm{d}\rho} &= +2\epsilon\mathcal{R}(\rho)-2\cos\mathcal{Y}_d(\rho), &&\text{(a)}\\ \mathcal{R}(\rho)\frac{\mathrm{d}\mathcal{R}(\rho)}{\mathrm{d}\rho} &= +\mathcal{R}(\rho)\sin\mathcal{Y}(\rho), &&\text{(b)}\end{aligned} \tag{6.18}$$

and the change of the sign of ϵ and of the slope does not affect the characteristic equation (6.3) except that we must replace $\epsilon\to-\epsilon, \cos\mathcal{Y}_u\to-\cos\mathcal{Y}_d$ in it. Then the cases $C > 0$ and $C < 0$ must be swapped in Equations (6.4)-(6.11) with the conclusion that constant C must be determined as $C=-\epsilon^2=-(E/m)^2$, and Equation (6.3) of characteristics of system (6.18) reads as

$$1-\cos\mathcal{Y} = 2\sin^2(\mathcal{Y}/2) = (1-\epsilon\mathcal{R})^2. \tag{6.19}$$

After using the constraint, $1-\epsilon\mathcal{R} = -\sqrt{2}\sin(\mathcal{Y}_d/2)$, the system (6.18) becomes,

$$\begin{aligned}\frac{\mathrm{d}\mathcal{Y}_d}{\mathrm{d}\rho} &= -2^{3/2}|\epsilon|\sin\frac{\mathcal{Y}_d}{2}, &&\text{(a)}\\ \frac{\mathrm{d}\mathcal{R}}{\mathrm{d}\rho} &= \sin\mathcal{Y}_d = 2\sin\frac{\mathcal{Y}_d}{2}\cos\frac{\mathcal{Y}_d}{2}, &&\text{(b)}\end{aligned} \tag{6.20}$$

and its first equation is readily integrated as

$$\sqrt{2}m\rho = \tanh^{-1}\left(\cos\frac{\mathcal{Y}_d}{2}\right),\ \cos\frac{\mathcal{Y}_d}{2} = \tanh\left(\sqrt{2}\epsilon\rho\right),\ \sin\frac{\mathcal{Y}_d}{2} = \frac{1}{\cosh\left(\sqrt{2}\,\epsilon\rho\right)}. \tag{6.21}$$

Acting as previously, we simplify the constrain and Equation (6.20.b) to

$$1-\epsilon\mathcal{R} = \frac{\sqrt{2}}{\cosh\left(\sqrt{2}\,\epsilon\rho\right)},\ \ \sin\mathcal{Y}_d(\rho) = +\frac{2\sinh\left(\sqrt{2}\,\epsilon\rho\right)}{\cosh^2\left(\sqrt{2}\,\epsilon\rho\right)} = \frac{\mathrm{d}\mathcal{R}}{\mathrm{d}\rho}, \tag{6.22}$$

where the second equation is identical to (6.18.b) and is a consequence of the first one. When $\rho\to\infty$, we have $1-\epsilon\mathcal{R}\to 0$, which is possible only when $\epsilon = E/m > 0$. Here, the condition of a particle at rest requires that $\epsilon = E/m = +1$. We also obtain the anticipated $\sin\mathcal{Y}(\infty) = 0$ and $\cos\mathcal{Y}_d(\infty) = 1$, *i.e.* $\mathcal{Y}_d(\infty) = \pi$. Thus the solution (in natural units)

$$\mathcal{R}(\rho) = 1-\frac{\sqrt{2}}{\cosh\left(\sqrt{2}m\rho\right)}, \tag{6.23}$$

can be interpreted as an isolated particle at rest with the positive energy $E = +m$, which is 2*m* higher than that for the similar localized static ψ_u-mode. Here, once again, $\mathcal{R}(\infty) = 1$. If the auto-localization is a real process it must favor localization not of ψ_d that has a dip, but the bump of ψ_u. This is also a hint that an *ad hoc* created ψ_d can be unstable (as it is in Nature). We elaborate on it in the last section.

Finally, for the mode with a dip of the invariant density in its interior, the invariant density reaches its theoretical minimum, $\mathcal{R}(\rho_0) = 0$, at the inflection point $s_0 = 1/m$ [5]. At this point we have $\sin\mathcal{Y}_d(\rho_0) = 1$, *i.e.* $\mathcal{Y}_d(\rho_0) = \pi/2$. Inside this radius the density *R*, as formally defined by (6.23), becomes negative, which is impossible. This can be a yet another indication that *an isolated localized* negative charge is unstable (at least in the absence of external field or of stable third bodies nearby). In other words, even being localized, it most likely is "an agile shallow deepening on a hill". Indeed, in real world of a stable matter, all electrons are light and only weakly localized around atomic nuclei, so that normal matter is charge-neutral. The heavy inward-polarized particles (e.g., antiprotons) are found only rarely and they would not be detected without abundant normal matter nearby. These probably are "deep holes on a high hill". Verification of this hypothesis is not a one-body problem and is beyond the scope of this work.

7. Stability and an Effective Lagrangian

The two exact solutions of the Dirac equation in one-body approximation, given by Equations (6.13)-(6.16) for the modes ψ_u, and by Equations (6.21)-(6.23) for the mode ψ_d, seem to be very similar to each other except that ψ_u has a bump and ψ_d has a dip of the invariant density near the center. According to the initial hypothesis, they correspond to positive and negative charges, respectively. The primary guess was [1, 2] that the former must be localized better and (if being unstable) live longer than the later, solely because the proper time in their interior flows the slower, the higher the invariant density is. Beyond the one-body approximation, the difference between these solutions is encoded mainly in the last two equations of the system (5.3) for ψ_u and (5.5) for ψ_d. In the case of ψ_u they do not depend on the external field A_μ, while in the case of ψ_d they do. Furthermore, the tetrad components $A_{[1]} + iA_{[2]}$ in Equations (5.5.c,d) oscillate with time as $\mathrm{e}^{-2im\tau}$ and can cause a transition from ψ_d to ψ_u.

The field A_μ in the Dirac equation is an external field. Remarkably, whatever this field is, the Dirac field determines world time across every auto-localized object. In a sense, all solutions of Equations (5.3) and (5.5) with the energy $\epsilon = E/m$ are the static solutions. But it is well-known that not all static solutions are stable. Solutions (6.16) and (6.23) obtained in absence of an external field are both truly static since there is nothing in Equations (6.1) and (6.18) that could have trigger instability. To investigate the effects of instability one must return to Equations (5.5.c,d) and also to Equations (5.8) and (5.9), which also account for the external field A_D

[5] In general, none of the Dirac currents vanishes at $\mathcal{R} = 0$; they all become proportional to one lightlike vector that must have both up- and down-components. Then nothing can identify the surface $S_{(12)}$ of constant τ and ρ as a two-dimensional sphere.

and dynamics of the sums of phases, $\mathcal{Z} = \phi_L + \phi_R$. The problem has two different aspects, *viz.*, formation of a perturbation and its decay.

Below, we try to specify both aspects and speculate regarding possible approaches/tools. The following terminology seems most appropriate for the discussion. Let us consider the components of ψ_u and ψ_d as the wave functions of the initial state and denote them as $|u\rangle_i, |d\rangle_i$. Next, let us contract Dirac equation with the Hermit conjugated wave function of a "final state", ${}_f\langle u|, {}_f\langle d|$ and consider ${}_f\langle ...\rangle_i$ as " transition amplitudes".

7.1. Creation of Perturbations in Dirac Vacuum

The problem of what may trigger the initial (and almost necessarily unstable) configuration is the most subtle one. Classically, one has to start with arbitrary initial field ψ and a plausible external field $A_\tau \pm A_\rho$ (i.g., of the cosmic microwave background). In quasi-static regime, the interaction of reasonably well defined initial states $|u\rangle_i$ with the lightlike components $A_\tau \pm A_\rho$ of the vector potential is not distractive, since Equations (5.3.a,b) can contribute only to diagonal (with respect to the spin) matrix elements,

$$-4ie_f\langle d_R|\left(A_\tau + A_\rho\right)|d_R\rangle_i, \quad -4ie_f\langle d_L|\left(A_\tau - A_\rho\right)|d_L\rangle_i. \tag{7.1}$$

These are *not* the transitions between up- and down-states. Regardless how weak this interaction is, it takes place in enormous space and for astronomical times. It can collapse to a solitary excitation just because such excitations exist. This mechanism can be considered as a potential source of the cosmic positron excess (for an extensive review see Ref. [3]). Furthermore, in Equations (5.3.c,d), which could have trigger transition from up- to down-states, there is no interaction terms at all. Thus, solution (6.17) of Equations (5.3), which is associated with a positive charge, is expected to be stable.

7.2. Decay of an Initial Perturbation

If an initial finite waveform is given, a reasonable theory must predict its decay into stable solitary configurations. Equations (5.5.c,d) (unlike (5.3.c,d)) prompt the interaction

$$-4ie_f\langle u_R|\mathrm{e}^{i\left(\phi_L^d+\phi_R^d\right)}\left(A_{(1)}+iA_{(2)}\right)|d_R\rangle_i \quad \text{and} \quad -4ie_f\langle u_L|\mathrm{e}^{i\left(\phi_L^d+\phi_R^d\right)}\left(A_{(1)}+iA_{(2)}\right)|d_L\rangle_i, \tag{7.2}$$

that affects stability of the localized inward-polarized state. In these formulae, $A_{(1)}$ and $A_{(2)}$ are the components of vector potential with respect to a judiciously chosen basis $(\boldsymbol{h}_1, \boldsymbol{h}_2)$ on the surface $S_{(12)} \in \mathbb{M}$ mapped onto $\mathbb{R}^4$. The transition from unstable mode to the stable one is due to the charged Dirac currents that naturally oscillate as $\mathrm{e}^{-2im\tau}$, and this transition can be triggered by almost any external electromagnetic field. The latter can be random or regular and originate, e.g., from the cosmic background. Possibly, they can even stabilize the ψ_d mode for a long time. This could explain the difference between an apparently stable particle in a storage ring and a visibly unstable particle in the natural world.

7.3. Similarity to Magnetic Resonance?

The matrix elements (7.2) are intimately connected with the dynamics of the spin 1/2 in magnetic field, where quantum and classical equations of motion coincide. Indeed, the sectional curvature[6] of the spherical surface $S_{(12)}$ (the curvature of the lines of the charged currents $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$),

$$R^t_{1212} = 2e\left(\partial_{[1]}A_{[2]} - \partial_{[2]}A_{[1]}\right) - 4e^2\left(A^2_{[1]} + A^2_{[2]}\right) = 2eF_{12} = 2eB_{[3]}, \tag{7.3}$$

is totally due to the projection of the *external magnetic field* onto radial direction of the axial current. If such a projection is not zero, it will cause flip of the spin polarization into the outward direction of the stable ψ_u-mode.

7.4. An Effective Lagrangian

More accurate approach that would allow one to go beyond the lowest order approximation can probably be

[6]The sectional curvature of a surface spanned by a net of the lines of the vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ equals to the angle by which the basis $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ is rotated after moving along an infinitesimal loop within this surface.

based on the so-called effective Lagrangian, $\mathcal{L}=\psi^{+}\left[i\alpha^{A}D_{A}\psi-m\rho_{1}\right]\psi$, with the operator of Equation (2) in brackets. The terms depending on A_{μ} in it can be viewed as the interaction with the outside sources. Retaining the interaction term ($e\neq 0$), actually, leads beyond the one-body approximation. Below, solely for the purpose of stability analysis, we add the alien up- and/or down-components as a perturbation. The state is supposed to be stable if the alien components dissipate due to the interaction. It will be genuinely unstable if the interaction enforces dissipation of the native components. We continue to dub the configurations with $u_{L}^{2}+u_{R}^{2}>d_{L}^{2}+d_{R}^{2}$ as ψ_{u} (with native *u* and an admixture of alien *d*). Those with $u_{L}^{2}+u_{R}^{2}<d_{L}^{2}+d_{R}^{2}$ are dubbed as ψ_{d} (with native *d* and alien *u*).

Let us look at the terms associated with the charged currents Θ^{a} and Φ^{a} and consider the matrix element,

$$T_{ab}=\langle b|T|a\rangle=\psi_{b}^{+}\left[\alpha^{[1]}\left(-ieA_{[1]}-e\tilde{A}_{[2]}\right)\psi+\alpha^{[2]}\left(-ieA_{[2]}+e\tilde{A}_{[1]}\right)\right]\psi_{a}, \tag{7.4}$$

between the configurations ψ_{a} and ψ_{b}. Here, $\tilde{A}_{D}$ stands for A_{D} when the triplet $\left(\boldsymbol{e}_{[1]},\boldsymbol{e}_{[2]},\boldsymbol{e}_{[3]}\right)$ forms the right-handed system, and for $-A_{D}$ when this triplet is left-handed. As an illustration, consider a particular term assuming native u_{L},u_{R} and alien d_{L},d_{R}; then T_{ab} is

$$\begin{aligned}\psi_{b}^{+}T_{+}\psi_{a}&=\psi_{b}^{+}\left[eA_{[1]}\left(-i\alpha^{[1]}+\alpha^{[2]}\right)+eA_{[2]}\left(-i\alpha^{[2]}-\alpha^{[1]}\right)\right]\psi_{a}\\&=-2ie\left(A_{[1]}-iA_{[2]}\right)\cdot\psi_{b}^{+}\rho_{3}\sigma^{+}\psi_{a}\\&=-2eA_{\mu}\left(e_{[1]}^{\mu}-ie_{[2]}^{\mu}\right)\psi_{b}^{+}\mathcal{O}_{u}\rho_{3}\sigma^{+}\mathcal{O}_{d}\psi_{a}.\end{aligned} \tag{7.5}$$

Here, $\sigma^{+}=\left(\sigma^{[1]}+i\sigma^{[2]}\right)/2$ is the ladder (spin-flip) operator for the projection of spin 1/2 onto the positive direction $\boldsymbol{e}_{[3]}$ of the right-hand oriented triplet. Let us recall that $\mathcal{O}_{u/d}=\left(1\pm\sigma^{3}\right)/2$ are the projection operators onto the up-/down-components of the Dirac spinor. In detail, the action of the operator T_{+} is as follows. The ladder operator $\rho_{3}\sigma^{+}$ eliminates the native components u_{R} and u_{L} (acting on ψ_{a} as $\mathcal{O}_{d}$) and replaces them with the alien d_{R} and d_{L}, producing $\psi'=\left(d_{R},0,-d_{L},0\right)$. Since $\sigma^{+}\mathcal{O}_{d}=\sigma^{+}$, this can be viewed as a two-step action. Namely, the $\mathcal{O}_{d}$ (inherited from connection (2.4)) filters out the d_{R} and d_{L} in their alien position, and then σ^{+} moves them "up", thus filtering out the positive helicity of the native "up"-final state $\psi_{b}^{+}\mathcal{O}_{u}$. In other words, $\psi_{b}^{+}T_{+}\psi_{a}\propto\left(u_{bR}^{*}d_{aR}-u_{bL}^{*}d_{aL}\right)e^{-2im\tau}$. If the state ψ_{a} was a pure up-state ψ_{u} and had no components $\left(d_{R},d_{L}\right)$ at all, then $\psi_{b}^{+}T_{+}\psi_{u}=0$; this is the case of Equations (3)—the ψ_{u} does not interact with the external $A_{[1]},A_{[2]}$. Conversely, *a solitary localized state* ψ_{d} that has only $\left(d_{R},d_{L}\right)$ is unstable under this interaction and the charged currents will smoothen it, or even cause its decay. This reproduces the primitive analysis of Equations (7.1) and (7.2).

Since the effective Lagrangian is nonlinear, there are many open questions, which cannot be addressed comprehensively within the scope of the present work. For example, it is not clear *a priori*, which of states, initial or final, should determine the nonlinear terms. These issues will be discussed separately. Of highest priority are the questions about time scales of the processes that contribute to the transition amplitudes (2) as well as about stability of the uniform distribution of the invariant density.

8. Summary

1. The method. The most intriguing discovery of this work is that Dirac field endows spacetime with a matter-induced affine geometry (MIAG), which is *fully determined by a real matter*. This is possible solely because the Dirac field satisfies equations of motion. Then, and only then, the geometry is independent of a particular coordinate background. Possibly, this result can look strange for mathematicians. But it should not surprise physicists, who know very well that nothing in spacetime can be measured without localized material objects. So far, the method of MIAG determined the shape of a solitary localized object as spherical dynamically and with no conjectures. The problem of signals still has to be worked out.

2. The results. The author's conjecture [4] that there exists a generic mechanism of the Dirac field auto-localization into finite-sized positively charged Dirac particles is rigorously confirmed. The explicit solution representing such a particle is found. It possesses the following properties,

(i) A solitary Dirac field waveform in free space can be stable with respect to the interaction with an external electromagnetic field A_{μ} only if this waveform is formed solely by outward polarized components. The solution that represents such a waveform has *negative energy* $E=-m$.

(ii) An apparently complementary inward-polarized solution with negative charge has *positive energy* $E = +m$. It cannot be stable as a strongly localized object; its instability is due to the indispensable "charged currents" Θ and Φ. They oscillate twice faster than stationary Dirac field, $\Theta \pm i\Phi \propto e^{\pm 2iE\tau}$. The corresponding tetrad components $A_{[1]}, A_{[2]}$ of the vector potential affect only the inward polarized waveform, thus making it unstable. This "motion" is confined to within the spheres of a constant radius within a localized object[7]. Similar oscillations also show up in the theory of the Compton scattering as the *t*-channel transitions of electron into the negative energy states. These transitions are responsible for the classical part of the Compton cross-section (Thompson scattering).

(iii) The difference in degree and the time duration of the localization obviously makes the localized charges of opposite sign unequivocally different particles. The correlation between the signs of electric charge, shape and polarization explains the interdependence between the discrete *C*- and *P*-transformations as a natural property of the simplest localized waveforms. While *C* qualitatively stands for the charge conjugation, *P* is not an abstract reflection symmetry in a flat space; it stands for the interchange of *inward* and *outward*. In a sense, these two discrete transformations do not exist separately; in this sense, *CP* is a physical symmetry between the corresponding processes[8].

3. The prospects. Our major perception of vacuum is absence of localized matter. This means that in the vacuum $\mathcal{R}$ is constant, e.g., $\mathcal{R} = 1$. Since Dirac equation is a hyperbolic system, the Dirac field must experience refraction towards domains where $\mathcal{R} > 1$, amplifying $\mathcal{R}$ even more, which resembles a well-known nonlinear effect of self-focusing. The opposite trend must be observed in domains where $\mathcal{R} < 1$; the Dirac waves tend to escape them. This idea can be phrased more precisely as: *Identification of the sign of* $\log \mathcal{R}$ *with the sign of electric charge leads to a dynamic picture of an empirically known charge-asymmetric world in which stable positively charged elementary Dirac objects are highly localized* (*and presumably heavy*), *while negatively charged objects tend to be poorly localized* (*and presumably light*). This mechanism of localization is generic and points to the picture that stunningly resembles the today's world. It must be worked out in greater details with the prospect that the issue of cosmological charge asymmetry, first addressed long ago by A.D. Sakharov [5], as well as the currently observed positron excess [3], could be better understood.

Meanwhile, to validate our approach in cosmological context, two major questions must be answered,

(i) What (if anything) can trigger a spontaneous creation of a proton alone (without an antiproton)? This is the most formidable problem.

(ii) Let a $p\bar{p}$ pair be created in an energetic process and the antiproton be thoroughly isolated from a normal matter (except for the cosmic background radiation). Will it live infinitely long? If not, then how will it decay? This question does not seem unbearable[9] and can be solved by methods developed in this one and previous author's papers (work in progress).

Acknowledgements

I am indebted M.E. Osinovsky for his advice on subtle issues of spinor analysis and differential geometry and for critically reading the manuscript. This work is supported by the Rapid Research, Inc.

References

[1] Makhlin, A. (2016) *Journal of Modern Physics*, **7**, 587-610. http://dx.doi.org/10.4236/jmp.2016.77061

[2] Byrd, P.F. and Friedman, M.D. (1971) Handbook of Elliptic Integrals for Engineers and Scientists. 2nd Edition, Springer-Verlag, Berlin. http://dx.doi.org/10.1007/978-3-642-65138-0

[3] Serpico, P.D. (2012) *Astroparticle Physics*, **39-40**, 2-11. http://dx.doi.org/10.1016/j.astropartphys.2011.08.007

[4] Makhlin, A. (2010) Localization, CP-Symmetry and Neutrino Signals of the Dirac Matter. arxiv:1005.2693 [math-ph]

[5] Sakharov, A.D. (1967) *JETP Letters*, **5**, 24-27.

[7] This motion cannot be interpreted as an oscillation of a *mean coordinate* -- the famous Schrödinger *Zitterbewegung*.

[8] This is in contrast with the view of Dirac field as the representation of the Lorentz group. In that framework, the Poincaré invariance is presumed, and all states can be obtained from a single state by a sequence of the Lorentz transformations.

[9] Some theories of Grand Unification predict proton's decay with a lifetime greater than the currently estimated age of Universe. From our perspective, only antiproton can be unstable.

Appendix: Notation and Algebraic Conventions

All observables associated with the Dirac field are bilinear forms built with the aid of *Hermitian* Dirac matrices $\alpha^i = \left(\alpha^i\right)^+$ and $\beta = \beta^+$, which satisfy the commutation relations

$$\alpha^a \beta \alpha^b + \alpha^b \beta \alpha^a = 2\beta\eta^{ab}, \quad \alpha^a \beta + \beta \alpha^a = 0, \quad \beta^2 = 1, \tag{A.1}$$

Throughout this paper, the Dirac matrices associated with a tetrad $h_a^\mu \in \mathbb{R}^4$ are numeric and are chosen in the spinor representation,

$$\alpha^0 = \begin{pmatrix} \mathbf{1} & 0 \\ 0 & \mathbf{1} \end{pmatrix}, \quad \alpha^i = \begin{pmatrix} \tau_i & 0 \\ 0 & -\tau_i \end{pmatrix}, \quad \sigma^i = \begin{pmatrix} \tau_i & 0 \\ 0 & \tau_i \end{pmatrix}$$

$$\rho_1 = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}, \ \rho_2 = \begin{pmatrix} 0 & -i\cdot\mathbf{1} \\ i\cdot\mathbf{1} & 0 \end{pmatrix}, \ \rho_3 = \begin{pmatrix} \mathbf{1} & 0 \\ 0 & -\mathbf{1} \end{pmatrix}. \tag{A.2}$$

where τ_i are the 2×2 Pauli matrices.

If the Dirac spinor is written down in terms of modules and phases of its components,

$$\psi = \begin{bmatrix} u_R \exp\left(i\phi_R^u\right) \\ d_R \exp\left(i\phi_R^d\right) \\ u_L \exp\left(i\phi_L^u\right) \\ d_L \exp\left(i\phi_L^d\right) \end{bmatrix}, \tag{A.3}$$

then, with the Dirac matrices (A.7), the scalars and the four Dirac currents have the following components,

$$j^a = \begin{bmatrix} u_L^2 + d_L^2 + u_R^2 + d_R^2 \\ 2u_L d_L \cos\left(\phi_L^u - \phi_L^d\right) - 2u_R d_R \cos\left(\phi_R^u - \phi_R^d\right) \\ -2u_L d_L \sin\left(\phi_L^u - \phi_L^d\right) + 2u_R d_R \sin\left(\phi_R^u - \phi_R^d\right) \\ u_L^2 - d_L^2 - u_R^2 + d_R^2 \end{bmatrix},$$

$$\mathcal{J}^a = \begin{bmatrix} u_L^2 + d_L^2 - u_R^2 - d_R^2 \\ 2u_L d_L \cos\left(\phi_L^u - \phi_L^d\right) + 2u_R d_R \cos\left(\phi_R^u - \phi_R^d\right) \\ -2u_L d_L \sin\left(\phi_L^u - \phi_L^d\right) - 2u_R d_R \sin\left(\phi_R^u - \phi_R^d\right) \\ u_L^2 - d_L^2 + u_R^2 - d_R^2 \end{bmatrix}, \tag{A.4}$$

$$\Theta^a = \begin{bmatrix} -2u_L d_R \cos\left(\phi_L^u + \phi_R^d\right) + 2d_L u_R \cos\left(\phi_R^u + \phi_L^d\right) \\ 2u_L u_R \cos\left(\phi_L^u + \phi_R^u\right) - 2d_L d_R \cos\left(\phi_R^d + \phi_L^d\right) \\ -2u_L u_R \sin\left(\phi_L^u + \phi_R^u\right) - 2d_L d_R \sin\left(\phi_R^d + \phi_L^d\right) \\ -2u_L d_R \cos\left(\phi_L^u + \phi_R^d\right) - 2d_L u_R \cos\left(\phi_R^u + \phi_L^d\right) \end{bmatrix},$$

$$\Phi^a = \begin{bmatrix} -2u_L d_R \sin\left(\phi_L^u + \phi_R^d\right) + 2d_L u_R \sin\left(\phi_R^u + \phi_L^d\right) \\ 2u_L u_R \sin\left(\phi_L^u + \phi_R^u\right) - 2d_L d_R \sin\left(\phi_R^d + \phi_L^d\right) \\ 2u_L u_R \cos\left(\phi_L^u + \phi_R^u\right) + 2d_L d_R \cos\left(\phi_R^d + \phi_L^d\right) \\ -2u_L d_R \sin\left(\phi_L^u + \phi_R^d\right) - 2d_L u_R \sin\left(\phi_R^u + \phi_L^d\right) \end{bmatrix}, \tag{A.5}$$

$$\mathcal{S}+i\mathcal{P}=2\left(u_R u_L \mathrm{e}^{i\left(\phi_L^u-\phi_R^u\right)}+d_R d_L \mathrm{e}^{i\left(\phi_L^d-\phi_R^d\right)}\right)=\mathcal{R}\,\mathrm{e}^{i\mathcal{Y}},$$

$$\mathcal{R}^2=4\left[u_R^2u_L^2+d_R^2d_L^2+2u_R u_L d_R d_L\cos\left(\phi_L^u-\phi_R^u-\phi_L^d+\phi_R^d\right)\right]. \quad \text{(A.6)}$$

Journal of Modern Physics, 2016, 7, 680-698
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jmp
http://dx.doi.org/10.4236/jmp.2016.77067

The Case against Dark Matter and Modified Gravity: Flat Rotation Curves Are a Rigorous Requirement in Rotating Self-Gravitating Newtonian Gaseous Discs

Dimitris M. Christodoulou[1], Demosthenes Kazanas[2]

[1]Department of Mathematical Sciences and Lowell Center for Space Science and Technology, University of Massachusetts Lowell, Lowell, MA, USA
[2]Laboratory for High-Energy Astrophysics, NASA Goddard Space Flight Center, Greenbelt, MD, USA
Email: dimitris_christodoulou@uml.edu, demos.kazanas@nasa.gov

Received 19 February 2016; accepted 25 April 2016; published 28 April 2016

Abstract

By solving analytically the various types of Lane-Emden equations with rotation, we have discovered two new coupled fundamental properties of rotating, self-gravitating, gaseous discs in equilibrium: isothermal discs must, on average, exhibit strict power-law density profiles in radius x on their equatorial planes of the form Ax^{k-1}, where A and $k-1$ are the integration constants; and "flat" rotation curves precisely such as those observed in spiral galaxy discs. Polytropic discs must, on average, exhibit strict density profiles of the form $\left[\ln\left(Ax^{k}\right)\right]^{n}$, where n is the polytropic index; and "flat" rotation curves described by square roots of upper incomplete gamma functions. By "on average", we mean that, irrespective of the chosen boundary conditions, the actual profiles must oscillate around and remain close to the strict mean profiles of the analytic singular equilibrium solutions. We call such singular solutions the "intrinsic" solutions of the differential equations because they are demanded by the second-order equations themselves with no regard to the Cauchy problem. The results are directly applicable to gaseous galaxy discs that have long been known to be isothermal and to protoplanetary discs during the extended isothermal and adiabatic phases of their evolution. In galactic gas dynamics, they have the potential to resolve the dark matter—modified gravity controversy in a sweeping manner, as they render both of these hypotheses unnecessary. In protoplanetary disc research, they provide observers with a powerful new probing tool, as they predict a clear and simple connection between the radial density profiles and the rotation curves of self-gravitating discs in their very early (pre-Class 0) phases of evolution.

How to cite this paper: Christodoulou, D.M. and Kazanas, D. (2016) The Case against Dark Matter and Modified Gravity: Flat Rotation Curves Are a Rigorous Requirement in Rotating Self-Gravitating Newtonian Gaseous Discs. *Journal of Modern Physics*, **7**, 680-698. http://dx.doi.org/10.4236/jmp.2016.77067

Keywords

Dark Matter, Gravitation, Galaxies, Protoplanetary Discs

1. Introduction

A large number of observations, mostly in the 21 cm emission line of neutral hydrogen, have firmly established that the rotation curves of spiral galaxy discs do not exhibit a Keplerian falloff; in fact, most of them remain flat or slightly increasing as far away from the centers as they can be observed [1]-[22]. The radial scales over which the neutral hydrogen discs can be observed reach out to ~100 kpc in the largest spiral galaxies and since the rotation curves remain flat, it was postulated by many researchers that some unseen extended mass distribution ought to exist all the way out to hundreds of kiloparsecs from the galaxy centers. Thus the Dark Matter Hypothesis was born, and soon the news leaked out to the rest of the physics community and intensive and extensive searches for dark matter particles and fields boomed into existence. On the other hand, some researchers who certainly felt uncomfortable with this new "aetherial" hypothesis proposed that Newtonian gravity should instead be modified at galactic scales and beyond, in order to solve the problem of the fast rotation of HI galaxy discs [23]-[33].

The gas in spiral galaxies is distributed in centrally concentrated, vertically thin discs. For this reason, it was expected that the rotation curves had to turn over at some intermediate radius and begin a decline that would be indicative of the absence of substantial amounts of matter at large radii. This view about the luminous matter is nowadays considered so settled and clear that it has made its way into introductory Astronomy textbooks that compare and contrast the kinematics of spiral galaxies to the kinematics observed in our solar system (the Keplerian falloff mentioned above). In this work, we show that this elementary perception is quite naive and totally wrong because it ignores the influence (in fact, the dominance) of pressure and enthalpy gradients in self-gravitating gaseous discs. Personally, we believe that it amounts to a blunder because, before the results described below and even back in the 1980s when all this was unfolding, we had important clues that pointed out the importance of gas pressure in determining the equilibrium structures of gaseous discs. For instance, the sound-crossing time at 10 kpc in a 10^4 K cold galaxy disc is 1 Gyr, a value that lies to within 1.5 - 3 of the rotation timescales at 10 kpc in all galaxies with rotation speeds of 100 - 200 km·s^{-1}. Gallagher *et al*. [34] noted that the star formation histories of a sample of irregular and spiral galaxies did not indicate the presence of dark matter in the low-mass galaxies of the sample. But the most obvious clue was the existence of the Mestel disc [35], a centrally concentrated potential-density pair with a flat rotation curve (see also [36]). The Mestel disc has always been considered just a toy model [37] and the situation did not improve when Schulz [38] showed that the finite Mestel disc requires significantly less mass to produce the flat rotation curve. The main argument against the universal adoption of the Mestel disc has been the absence of a physical law or reason that would make galaxy discs assume this specific surface density profile and rotation law.

The above argument is not based on solid reasoning. When nature shows us that she has widely adopted a specific property (the flat rotation curves in galaxy discs), Aristotelian Logic dictates that we should search for a new law or reason, in order to understand the universality of this property and establish its physical meaning; not to create ghosts (particles and fields), aethers, and new forces that effectively facilitate our aversion to confronting the facts.

We do not claim that the Mestel disc [35] is the answer to establishing the universality of flat rotation curves in galaxy discs; only that it has always been a telling clue that gravity does not pull the strings and is not in control in gaseous self-gravitating discs. Furthermore, we have solved the full Newtonian problem and we now know precisely how such universal rotation curves emerge in spiral galaxy discs. The resolution of this ubiquitous problem is the subject of this paper. Before we can delve into the physics of the problem, we need to correct some common misconceptions that appear in the theory of second-order differential equations and which also have made their way into the textbooks. We do so in Section 2. Then, in Section 3, we revisit the theory of rotating Newtonian isothermal gaseous-disc equilibrium models and we calculate analytically the mean shapes of their density profiles and their rotation curves. The results match precisely the shapes of the rotation curves of spiral galaxy discs with no additional assumptions of any kind. So these results make a strong case against both

dark matter and modified gravity and their implications have far-reaching consequences all the way to cosmology. For completeness, we describe in Section 4 polytropic models that also demand monotonically increasing rotation curves because they are subject to the same physical principles. These models are also applicable to very young protoplanetary discs (certainly to pre-Class 0 discs and possibly to the youngest Class 0 non-Keplerian discs). Finally, we conclude with a discussion of all the pertinent issues and our results in Section 5.

2. Second-Order Differential Equations and the Cauchy Problem

In mathematical physics, the trivial solutions of the various second-order differential equations are commonly ignored as being uninteresting; and too much attention is paid to the Cauchy problem in determining arbitrary constants as opposed to the internal properties of the equations themselves that have no regard for externally imposed conditions of any type. Both of these practices are damaging as they work to hinder our efforts toward solving the physical problems described by the differential equations in the first place. Such practices are relevant to all linear and nonlinear second-order equations of physics, so we can discuss and clarify the various issues involved by using any well-known equation. We choose to make use of the Bessel differential equation in this section.

The Bessel equation [39]

$$y'' + \frac{1}{x}y' + \left(1 - \frac{m^2}{x^2}\right)y = 0, \ \left(m = \text{const.}\right), \tag{1}$$

has regular solutions that are called Bessel functions of order m and a trivial solution $y = 0$. The trivial solution is not singular as it can be obtained from the regular solutions by an appropriate choice of the arbitrary constants. Nevertheless, its name indicates that $y = 0$ is of no interest at all. It is also well-known that the Bessel functions all oscillate about the x-axis [39], but this statement is grossly inaccurate and obscures the truth: the regular solutions oscillate about the trivial solution which just happens to coincide with the x-axis in this case. We demonstrate this important point by solving numerically an inhomogeneous $m = 0$ Bessel equation of the form

$$y'' + \frac{1}{x}y' + y = K, \ \left(K = \text{const.}\right), \tag{2}$$

along with nonsingular boundary conditions that attempt to initially push the regular solutions away from the new trivial solutions $y = K$: $y(1) = 1$, $y'(1) = -10$ for $K = 5$; and $y(2) = 1$, $y'(2) = 10$ for $K = -5$. The results are shown in **Figure 1**. Both regular solutions have nothing to do with the x-axis; instead, they turn around and settle into oscillations that clearly occur about the new trivial solutions $y = K$. This behaviour can be demonstrated for all linear second-order equations of mathematical physics with oscillatory regular solutions [40] and for (non)linear equations of the Lane-Emden type [41]-[45]. The lesson to be learned is that the so-called trivial solutions of second-order equations are not at all trivial. They are in fact favoured by the differential equations themselves which have no regard for the externally imposed boundary conditions. Thus, we will heretofore call these solutions the *intrinsic solutions* that are preferred and demanded by the equations themselves, irrespective of the externally imposed Cauchy problem.

When the Cauchy problem is solved, as in **Figure 1**, the externally imposed boundary conditions are usually at odds with the underlying equation and the regular solutions cannot match the favoured intrinsic solution. As a result, the regular solutions are forced by the equation itself to oscillate about the intrinsic solution as soon as they intersect this favoured solution the first time. Thus, the intrinsic solutions act as attractors of the regular solutions which, in turn, are forced to always stay near and around the more dominant intrinsic solutions. We view this behaviour as a triumph of the differential equation (and its intrinsically favoured solution) over the Cauchy problem (and the particular solution it strives to produce).

This striking behaviour remains intact in at least some nonlinear second-order equations. Very clear examples in which rotation is involved can be found in [43] and [44]. Here we provide two additional nonlinear examples of the dominance of intrinsic solutions over regular solutions drawn from Lane-Emden equations [46] [47] in the absence of rotation. The isothermal Lane-Emden equations, take the form

$$y'' + \frac{D-1}{x}y' + \mathrm{e}^{y} = 0, \tag{3}$$

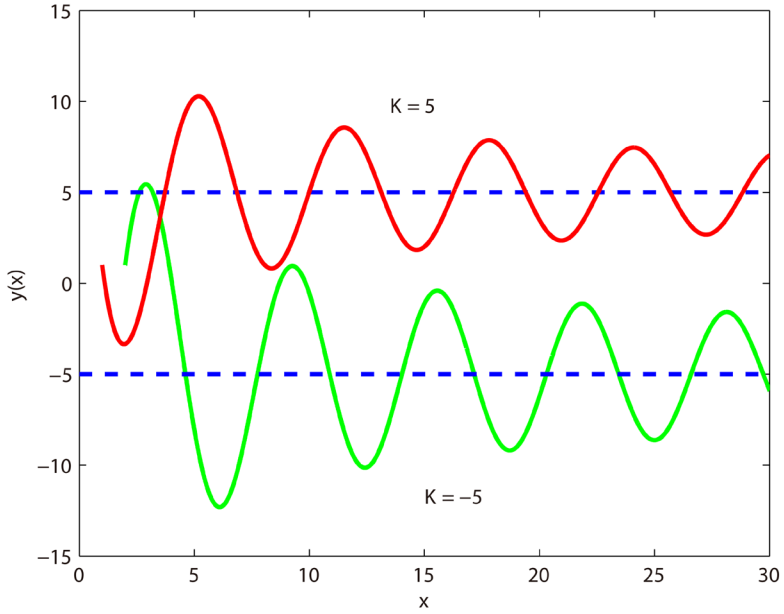


Figure 1. Numerical solutions of the $m = 0$ inhomogeneous Bessel differential Equation (2) subject to the following boundary conditions: in the $K = 5$ case, we use $y(1) = 1$ and $y'(1) = -10$; in the $K = -5$ case, we use $y(2) = 1$ and $y'(2) = 10$. In both cases, the regular solutions are forced to oscillate and stay near the dominant intrinsic solutions $y = K$ (dashed lines).

where D is the dimensionality of space ($D = 3$ in spherical coordinates and $D = 2$ in cylindrical coordinates). Singular and regular solutions have been obtained in many applications [48]-[50] [37] [42] [44] and they are all nonoscillatory.[1] The reason for this is quite obvious: Equation (3) does not have an intrinsic solution because $e^y \neq 0$. Why the latter condition precludes an intrinsic solution will become clear in the next section, where we describe a procedure for obtaining intrinsic solutions.

In stark contrast, the polytropic Lane-Emden equations, take the form

$$y'' + \frac{D-1}{x} y' + y^n = 0, \tag{4}$$

where $n > 0$ is the polytropic index, and it possesses the intrinsic solution $y = 0$. Although few analytic solutions are known [50] [52], numerical integrations show that, depending on n, this equation has both oscillatory and nonoscillatory solutions. For Equation (4), we have derived a precise criterion for the existence of oscillatory solutions [45]. This criterion predicts that for $D = 2$ (cylindrical form), all solutions with odd integer n-values are oscillatory; while for $D = 3$ (spherical form), only the $n = 1$ and $n = 3$ integer-n solutions are oscillatory. Numerical integrations (using the physical boundary conditions $y(0) = 1$, $y'(0) = 0$) easily confirm these results. The reason for the existence of nonoscillatory solutions is that for the corresponding choices of n, the differential equation is not a harmonic oscillator [40]. This is also true for the modified Bessel equation [39] that is known to possess only nonoscillatory solutions. Its real solutions cannot be oscillatory[2] and they are then prohibited from intersecting the intrinsic solution $y = 0$ more than once [40].

3. Isothermal Self-Gravitating Newtonian Gaseous Discs

In what follows, we use the arbitrary scaling constants R_o and ρ_o to normalize the disc radius R and density $\rho(R)$, respectively. We thus define the dimensionless radius $x \equiv R/R_o$ and density $\tau(x) \equiv \rho(R)/\rho_o$. Velocities $V(R)$ are also normalized consistently by the constant $V_o = R_o\sqrt{4\pi G\rho_o}$, where G is the Newtonian

[1]The periodic solution found in [51] describes a Cartesian slab and not a disc or cylinder.

[2]We note in passing that, as a result of the substitution $x \to ix$ that produces the modified Bessel equation [39], its solutions are oscillatory about the imaginary axis in the complex plane.

gravitational constant, in which case we also define the dimensionless rotation velocity $v(x) \equiv V(R)/V_o$. The same scaling also applies to the sound speed C_o of the gas which in this section is a constant, *i.e.*, the dimensionless sound speed is $c_o \equiv C_o/V_o$.

The cylindrical isothermal Lane-Emden equation [46] [47] with rotation can then be written in dimensionless form as

$$c_o^2 \cdot \frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x} x \frac{\mathrm{d}}{\mathrm{d}x} \ln \tau + \tau = \frac{1}{x}\frac{\mathrm{d}v^2}{\mathrm{d}x}. \tag{5}$$

This equation describes the radial (x) equilibrium of a rotating, self-gravitating, gaseous disc or cylinder in which the gas obeys the isothermal equation of state $p(x) = c_o^2 \tau(x)$, where p is the dimensionless pressure of the gas. Equation (5) is valid exactly for infinite cylinders and to a high degree of approximation in the equatorial (symmetry) planes of discs (see the **Appendix**). This latter point has been demonstrated convincingly by the calculations in [42] [53] [54]. In particular, the latter two investigations of finite discs uncovered equatorial density profiles that were strictly oscillatory under proper boundary conditions, just as was predicted by the analysis of Section 2 above.

Hayashi *et al*. [53] and Schmitz [55] studied also the stability of such equilibria and found that, except for the very flattened discs and the nearly spherical configurations, the intermediate models are stable. The very flattened discs in [53], in particular, were unstable to ring formation that causes their equatorial power-law density profiles to become oscillatory, in agreement with the numerical solutions of Equation (5) obtained in [44].

Despite the exact analytic results of the researchers quoted above, an objection has been raised over the years concerning the validity of using cylindrical coordinates to study axisymmetric, vertically thin discs rather than just infinite cylinders; and this is also the major sticking point for the present work, so it needs to be addressed in detail. We defer the analysis of the Lane-Emden equations for vertically thin discs to an Appendix because it is much easier to follow the derivations after the intrinsic analytic solution has been obtained for cylinders as follows.

Christodoulou & Kazanas [44] described a procedure for obtaining the intrinsic solution of Equation (5): If we equate the last two terms:

$$\tau(x) = \frac{1}{x}\frac{\mathrm{d}v^2(x)}{\mathrm{d}x}, \tag{6}$$

then this is an intrinsic solution provided that the rest of the equation (the radial variation of the logarithmic gradient of the enthalpy) vanishes:

$$\frac{\mathrm{d}}{\mathrm{d}x} x \frac{\mathrm{d}}{\mathrm{d}x} \ln \tau(x) = 0. \tag{7}$$

Equations (6) and (7) form a system in which $v(x)$ is totally dependent on $\tau(x)$. First we solve Equation (7) to obtain the radial density profile:

$$\tau(x) = Ax^{k-1}, \ (A, k = \text{const.}), \tag{8}$$

and then we solve Equation (6) to determine the rotation curve of the intrinsic solution:

$$v(x) = \sqrt{Ag(x) + B}, \ (B = \text{const.}), \tag{9}$$

where

$$g(x) \equiv \begin{cases} x^{k+1}/(k+1), & \text{if } k \neq -1 \\ \ln x, & \text{if } k = -1 \end{cases}. \tag{10}$$

The solution contains 3 free parameters, the integration constants A, B, and k. Parameter B sets the vertical scale of the rotation curve $v(x)$; here we choose $B = 1$. This choice can be made for $k \leq -1$ and it implies that the power-law density profile is cut off at an inner boundary $x = x_1$, where $v(x_1)$ drops to zero.

Figures 2-4 show the shapes of the rotation curves obtained from Equations (9) and (10) for various choices of the constants A and k. The results are scale-invariant (a property of power-law density profiles), so the radial

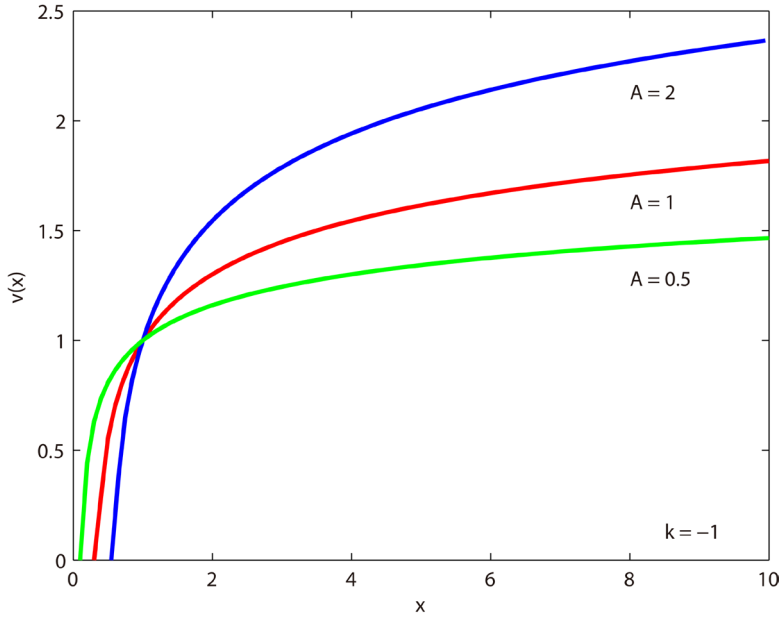


Figure 2. Rotation curves of the intrinsic solution of the isothermal Lane-Emden equation for $k=-1$, $B=1$, and various values of A. The choice $B=1$ implies the existence of an inner edge at which the velocity drops to zero.

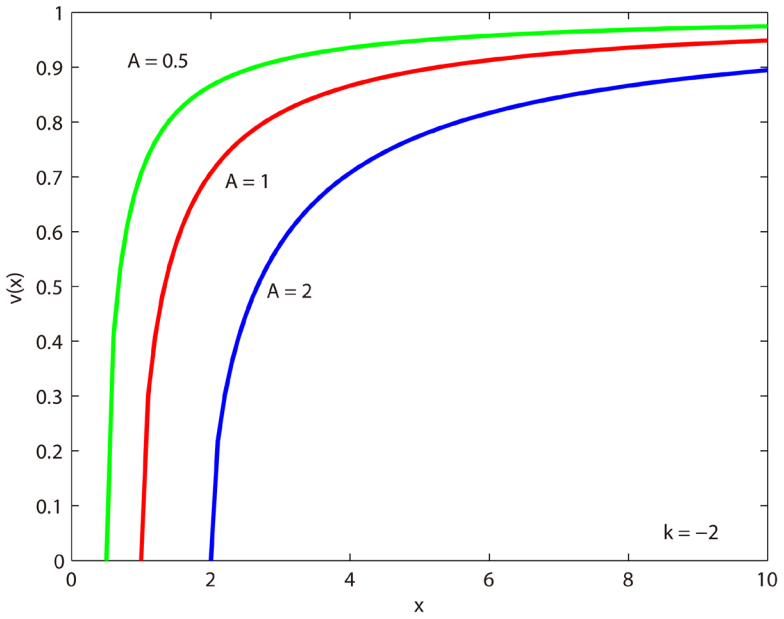


Figure 3. As in **Figure 2**, but for $k=-2$.

scale is arbitrary. It is not surprising (see Section 3.1 below) that, in these Newtonian models, one sees most of the shapes of the "flat" rotation curves observed in spiral galaxies.

The only shapes missing from the figures are those of the few falling rotation curves shown in [18]. Apparently, in some compact galaxies, the above equilibrium profiles did not endure (the sound-crossing time at 30 kpc for $C_o = 10\ \mathrm{km\cdot s^{-1}}$ is 3 Gyr, so it seems that there has been enough time to achieve equilibrium); perhaps because of interactions with nearby galaxies; or because the "external" gravity of the massive bulges eliminates the intrinsic solution (see Section 4.1 and footnote 4 below). But even in these objects, the falloff is slower than Keplerian which implies that gas pressure is still fighting to establish its own preferred profiles. This must be the

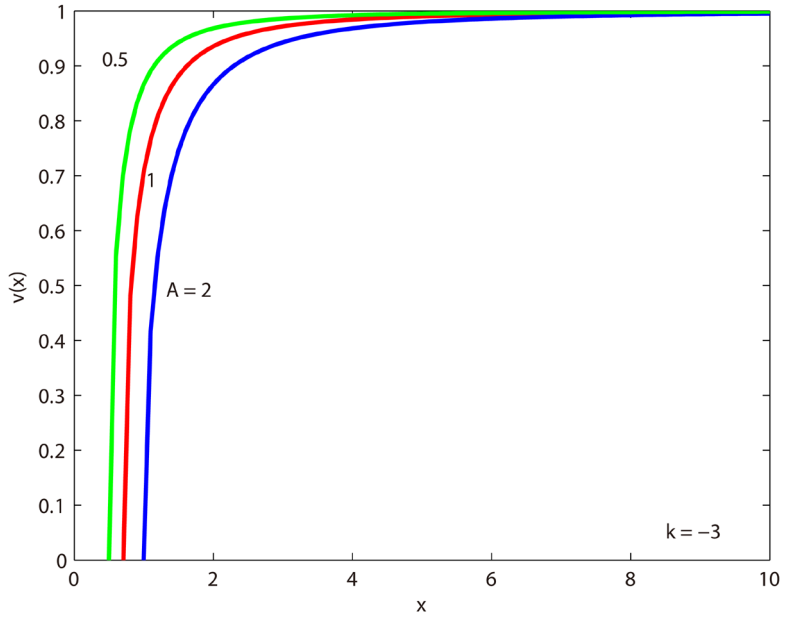


Figure 4. As in **Figure 2**, but for $k=-3$. **Figures 2-4** show that the curves become flatter for steeper values of the power-law index k.

case since the intrinsic solutions are favoured by the equilibrium differential equation itself (see Section 2) in the absence of other external forces. In the two galaxies observed by Casertano & van Gorkom [18], certain segments of the rotation curves are flat and that indicates to us that the radial self-gravitational equilibrium has fallen apart at some locations but not (yet) everywhere in these discs.

The above results can be summarized as follows: The derived density profiles are simple power laws in radius and the rotation curves are flat or slightly increasing at large radii (Equations (9) and (10)) irrespective of the value of the index k. Spiral galaxies have always been fitted with exponential density profiles, thus we do not know which values of k occur in nature. Galaxy profiles will have to be fitted again, but the payoff this time will be substantial: when k is determined from observations, the large-scale rotation curve (away from the center) will also be obtained independently from the velocity measurements. Thus, the observational results will be tested for consistency within the same data set in each case. This also holds true for protoplanetary discs in their early isothermal or adiabatic (see Section 4 below) phases; but first we need to find such purely self-gravitating discs in a pre-Class 0 YSO (Young Stellar Object) stage, and the 21 cm HI line in emission or absorption may give us a chance [56].

3.1. Physical Interpretation

But how can such power-law density profiles produce and support flat or slowly increasing rotation curves? The problem has always been that the centrifugal force remains too high in the outer regions of the disc where Newtonian gravity weakens substantially. How do these equilibrium models get around this discrepancy? The answer to these questions was given in [44] and we repeat it here: The Lane-Emden Equation (5) is a second-order differential equation. As such, it respects but does not rely solely on force balance. (As will be readily seen in Equation (12) below, force balance is guaranteed by the way that the specific enthalpy of the gas is operating.) This second-order equation describes locally the detailed radial $(\mathrm{d}/\mathrm{d}x)$ variation of the *logarithmic gradients* $(\mathrm{d}/\mathrm{d}\ln x)$ of the potentials involved in the struggle to reach equilibrium. So, initially, it is the log-gradient of the enthalpy (with help from rotation) that sets out to oppose the log-gradient of the gravitational potential. This competition can be seen by rewriting Equation (5) in the form

$$\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{\mathrm{d}}{\mathrm{d}\ln x}\left(h+\psi\right)\right]=\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\left[v^2\right], \tag{11}$$

where $h(x)\equiv\int \mathrm{d}p/\tau$ and $\psi(x)$ are the dimensionless enthalpy per unit mass and the Newtonian self-gra-

vitational potential, respectively. This equation is exact for cylinders and approximately valid for discs at large radii ($x \gg 1, \quad \tau \ll 1$), where the vertical (z) gradient of the vertical gravitational force $\partial^2\psi/\partial z^2$ becomes small and can be neglected.

In that case, how and why does the average intrinsic solution (Equations (8) and (9)) come into existence? The answer to these questions is even more illuminating: The intrinsic solution was derived above by imposing two separate conditions, that the centrifugal force should match the gravitational force (see Equation (6) and the last two terms in Equation (11)):

$$\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}\psi}{\mathrm{d}\ln x}\right)=\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\left(v^2\right)\Rightarrow\frac{v^2}{x}=\frac{\mathrm{d}\psi}{\mathrm{d}x}; \tag{12}$$

while, at the same time, the log-gradient of the enthalpy should retire from the competition by assuming a constant profile (see Equation (7) and the leading term in Equation (11)):

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}h}{\mathrm{d}\ln x}\right)=0. \tag{13}$$

This occurs only for a power-law density profile (Equation (8)) and for a specific rotation profile (Equations (9) and (10)), and these profiles become internal properties characteristic of the equilibrium disc or cylinder. Stated more simply, up until now people believed that rotation was not related to the structure of the equilibrium disc and that they could adopt any arbitrary rotation profile for gaseous self-gravitating discs. We see now that this is not true, the radial density and rotation profiles are strongly coupled and uniquely determined through the intrinsic solution discussed above.

The only surprise in this narrative is the unique way that the differential equation finds to promote and establish the above intrinsic solution: rather than trying to simultaneously balance the variations of the three potentials involved at every single radius (not possible because the corresponding timescales vary widely at different radii), the disc assumes gradually a logarithmic specific-enthalpy profile determined from the solution of Equation (13):

$$\frac{\mathrm{d}h}{\mathrm{d}x}\propto\frac{1}{x}\Rightarrow h\left(x\right)\propto\ln x. \tag{14}$$

So it is the thermodynamic potential $h(x)$ of the gas that becomes logarithmic, and not the gravitational potential that people have been trying to make it so for nearly 50 years! Naturally, the disc establishes such a logarithmic profile because this law guarantees precise force balance (Equation (12)) at all of its equatorial radii. The action of $h(x)$ unfolds in the physical disc from inside-out over timescales of the order of the local sound-crossing time R/C_o. So the global equilibrium becomes complete after a time $\sim R_{\max}/C_o$, where $R_{\max}$ is the outer radius of the disc.

We understand physically the preceding results in the following manner: Self-gravity is a long-range force (by Gauss's law, the gravitational potential at radius x depends on the entire mass interior to x) and it cannot adjust its potential in the disc to effect local changes to the density distribution. In other words, the density is a source term in the Poisson equation that determines the gravitational potential, but not the other way around. In stark contrast, enthalpy is a local potential whose action depends only on the local behaviour of the pressure and the density of the disc. When $h(x)$ assumes its logarithmic radial profile (Equation (14)), it dictates that the local density adjust according to the local log-gradient of the pressure

$$\tau\left(x\right)\propto\frac{\mathrm{d}p\left(x\right)}{\mathrm{d}\ln x}. \tag{15}$$

By doing that, the enthalpy uses the density in order to modify the sourcing of both the gravitational potential (via the Poisson equation $\nabla^2\psi=\tau$) and the centrifugal potential (via Equation (6)). The result of this tactic is a rotation law that is entirely dependent on the distribution of the density/enthalpy (see Equation (6)) that does not feed back (v does not enter in Equation (7)). The rotation so produced is capable of balancing gravity at all radii all by itself (Equation (12)), and the enthalpy retires from the struggle for equilibrium (Equation (7)) having implicitly won the competition at every radius.

It is important to keep in mind that the enthalpy wins the struggle because the two equations of the intrinsic equilibrium solution (Equations (6), (7) or Equations (12), (13)) and the Poisson equation ($\nabla^2\psi=\tau$) do not

allow for feedback loops and counter-sourcing by self-gravity or rotation. Thus, the state of rotation, the density profile, and the local self-gravity have all been manipulated unilaterally by the action of the enthalpy.

3.2. Concluding Remarks

The above equations give us a new probe into gaseous astrophysical disc systems (Equations (6) and (7)). For spiral galaxy discs, this probe ought to routinely confirm the above-described density-rotation coupling, as the observations already exist. At the same time, we are full of anticipation about what we can learn from the very early phases of purely self-gravitating protoplanetary discs (before the central protostars form and dominate the dynamics) for which the rotation curves have not been measured with accuracy yet, but the radial density profiles in Class 0 YSOs have been obtained [57] [58]. Our prediction, of course, is that, if found, such early discs (preferably earlier than Class 0) will be observed to have "flat" rotation curves.[3]

4. Polytropic Self-Gravitating Newtonian Gaseous Discs

The cylindrical polytropic Lane-Emden equation [46] [47] with rotation can be written in dimensionless form as

$$nc_o^2 \cdot \frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x} x \frac{\mathrm{d}}{\mathrm{d}x}\tau^{1/n} + \tau = \frac{1}{x}\frac{\mathrm{d}v^2}{\mathrm{d}x}, \tag{16}$$

where $n>0$ is the polytropic index and the dimensionless constant sound speed c_o was defined for $\rho=\rho_o$. (In general, the square of the sound speed $c^2(x)\equiv \mathrm{d}p/\mathrm{d}\tau$ varies as $\tau^{1/n}$ across the medium.) This equation describes the radial (x) equilibrium of a rotating, self-gravitating, gaseous disc or cylinder in which the gas obeys a polytropic equation of state $p\propto\tau^{1+1/n}$. As in Section 3, Equation (16) is valid exactly for infinite cylinders and to a high degree of approximation in the equatorial (symmetry) planes of discs (see the **Appendix**). This latter point is supported by the calculations in [43] [63] [64] that studied also the stability of thin-disc and cylindrical equilibria and found large regions of the parameter space with stable models for all values of $n>1$; and a sizeable region in which flattened discs with power-law density profiles were unstable to ring formation that causes their profiles to become oscillatory, just as was predicted by the analysis of Section 2 above.

We repeat the procedure outlined in Section 3 in order to obtain the intrinsic solution of Equation (16): If we equate again the last two terms:

$$\tau(x) = \frac{1}{x}\frac{\mathrm{d}v^2}{\mathrm{d}x}, \tag{17}$$

then this is an intrinsic solution provided that the rest of the equation (the radial variation of the logarithmic gradient of the enthalpy) vanishes:

$$\frac{\mathrm{d}}{\mathrm{d}x} x \frac{\mathrm{d}}{\mathrm{d}x}\tau^{1/n} = 0. \tag{18}$$

Equations (17) and (18) form a system in which $v(x)$ is totally dependent on $\tau(x)$. First we solve Equation (18) to obtain the radial density profile:

$$\tau(x) = \left[\ln\left(Ax^k\right)\right]^n, \ (A,k=\text{const.}), \tag{19}$$

and then we solve Equation (17) to determine the rotation curve of the intrinsic solution:

$$v(x) = \sqrt{\frac{A^{-2/k}}{2}\left(\frac{-k}{2}\right)^n \cdot \Gamma\left(n+1, \ln\frac{A^{-2/k}}{x^2}\right) + B}, \tag{20}$$

where B is the integration constant, $A>0$, $n>0$, $k<0$, $Ax^k\geq 1$ (*i.e.*, $x\leq A^{-1/k}$), and the upper incomplete gamma function is defined as

$$\Gamma(\alpha,\xi) \equiv \int_\xi^\infty \mathrm{e}^{-t}t^{\alpha-1}\mathrm{d}t, \ (\xi\geq 0). \tag{21}$$

[3]In the case of the Class 0 young system L1527 [59] [60], the rotation curve is not flat, but we did not catch this 0.3 Myr old system early enough (the sound-crossing time at 100 AU for a 10 K cold gas with C_o = 0.3 km·s^{-1} is 1500 yr). On the other hand, Yen *et al*. [61] [62] report several other Class 0 YSOs whose rotation is not Keplerian outside of the inner few AU.

The solution contains 4 free parameters, the integration constants A, B, k, and the polytropic index n. Parameter B can adjust the vertical scale of the rotation curve $v(x)$, but here we opted to use $B=0$ in what follows. This choice is equivalent to the boundary condition that $v(0)=0$.

Figures 5-8 show the shapes of the rotation curves obtained from Equation (20) for two polytropes with $n=1.5$ and $n=3$ and for various choices of the constants A and k. As in the isothermal case of Section 3, the rotation profiles are slowly increasing or flat with radius x. In this case however, we need to obey the condition $x \le A^{-1/k}$ ($\xi \ge 0$ in Equation (21) and $\tau \ge 0$ in Equation (19)) in the calculation of the gamma function and so the rotation curves terminate when x reaches its maximum value where $\tau = 0$ as well. Two basic trends are

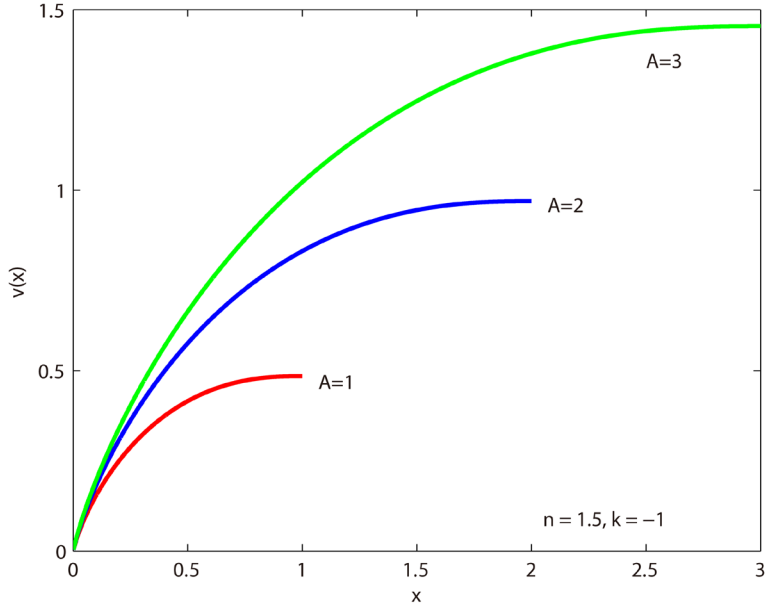


Figure 5. Rotation curves of the intrinsic solution of the $n=1.5$ polytropic Lane-Emden equation for $k=-1$, $B=0$, and various values of A.

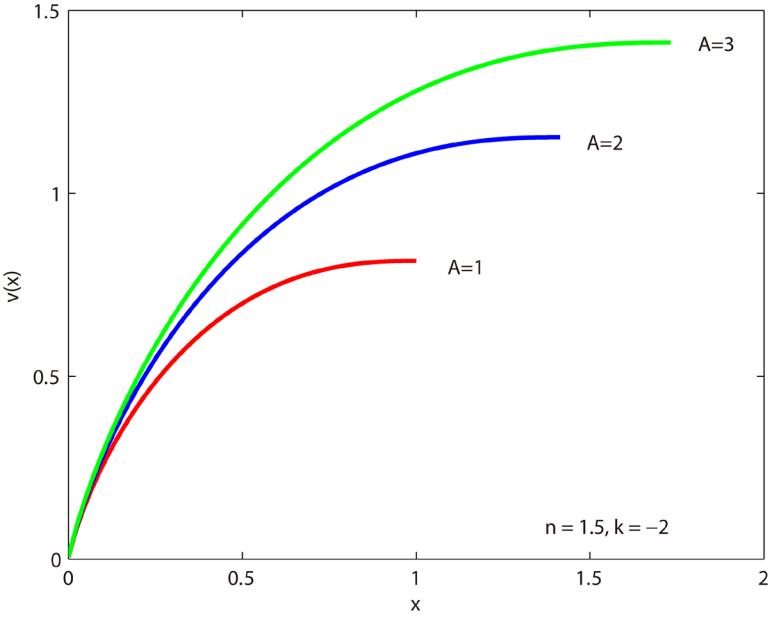


Figure 6. As in **Figure 5**, but for $k=-2$. **Figure 5** and **Figure 6** show that, for fixed n, the curves rise more steeply for steeper values of the index k.

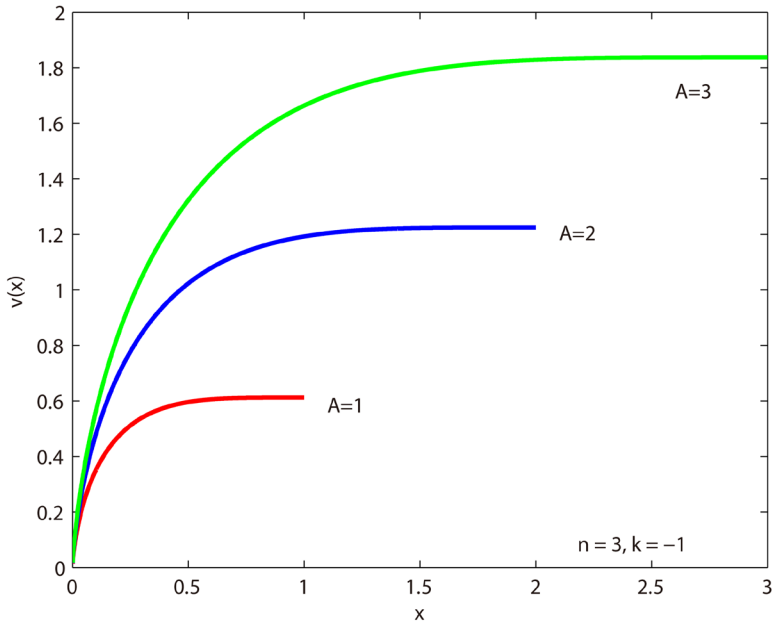


Figure 7. Rotation curves of the intrinsic solution of the $n = 3$ polytropic Lane-Emden equation for $k = -1$, $B = 0$, and various values of A.

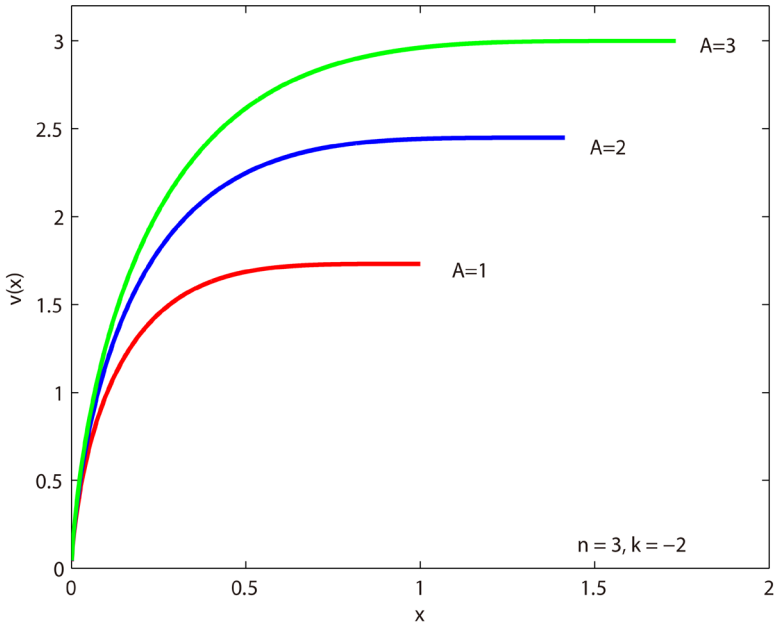


Figure 8. As in **Figure 7**, but for $k = -2$. **Figure 6** and **Figure 8** show that, for fixed k, the curves become flatter for higher values of the polytropic index n.

noted in the figure captions as well: 1) for fixed n, the curves rise more steeply for steeper values of the index $k < 0$; and 2) for fixed k, the curves become flatter for higher values of the polytropic index $n > 0$.

4.1. Physical Interpretation

The polytropic Lane-Emden equation with rotation and its intrinsic solution assume the exact same forms as in the isothermal case (Section 3.1) when the polytropic equation of state $p \propto \tau^{1+1/n}$ is used to introduce the specific enthalpy $h(x) \equiv \int \mathrm{d}p/\tau$. Therefore, the fundamental equations discussed in Section 3.1 also apply to

polytropic models with finite radial extent and the enthalpy plays the exact same role in manipulating the source terms of the gravitational potential and the rotational potential.

The fact that the thermodynamic potential $h(x)$ operates locally in the exact same manner (*i.e.*, $h \propto \ln x$) in polytropic and isothermal equilibria helps us correct another common misconception: Since the time that the results of Hayashi *et al*. [53] came to light (they studied isothermal self-gravitating gaseous discs that oddly exhibited power-law radial density profiles and "flat" rotation curves in equilibrium), it has been often stated that isothermal disc models tend to exhibit flat rotation curves because they happen to have a "special" mass distribution (*i.e.*, their specific angular momentum is proportional to the mass interior to radius *x*, or their mass grows linearly with *x*). This is not true. The above results show without a doubt that there are no special equilibrium models; and that the isothermal and the polytropic equilibria are both subject to the same fundamental physics at the local level, where the thermodynamic potential $h(x)$ operates and dominates when the only gravity it faces is the self-gravity of the disc. As for the validity of using the cylindrical coordinate system with its "special" ∇^2 operator for discs, this issue is addressed in detail in the **Appendix**.

Furthermore, as we have seen above, a flat rotation curve does not imply and does not need a linearly increasing mass distribution. Even if the underlying density profile is decreasing in radius (as the light does in spiral galaxy discs), a rising rotation curve is a requirement in the equilibrium solutions derived in this section and in Section 3 so long as more mass is added incrementally with increasing radius (this is simply a restatement of Gauss's law). Therefore, one can assume a constant mass-to-light ratio and build a Newtonian self-gravitating galaxy disc model with a declining mass distribution that will still be required to have a rising rotation curve provided that the disc may be thin, but not razor-thin; that $h(x) \propto \ln x$ on the equatorial plane; and that its volumetric density profile is a power law in radius *x* (see the **Appendix**).

4.2. Concluding Remarks

The polytropic intrinsic solution (Equations (17) and (18)) gives us a new probe into nonisothermal astrophysical disc systems. In particular, protoplanetary discs undergo various early phases of adiabatic evolution (figure 2 in [65]). We predict that, if found, such discs will be observed to have the same fundamental characteristics (Equations (19) and (20)) irrespective of the polytropic index *n* appropriate for each adiabatic phase. In fact, observations of the density profiles, fitted with Equation (19), may be able to determine, not only the profile constants *k* and *A*, but also the value of *n*, thereby deriving the equation of state of the gas independently from theoretical models. The only problem is that we need to find such systems very early in their development (see footnote 3 above), and this is a difficult task [56] [57] [66].

5. Discussion

In this paper, we have investigated the Lane-Emden equations with rotation that describe the equilibrium structures of rotating, self-gravitating, gaseous discs and cylinders (Sections 3 and 4; see also the **Appendix**). We have obtained new analytic singular solutions that we call intrinsic solutions because they are dictated and favoured by the differential equations themselves with no regard to physical boundary conditions that are externally imposed and that shape up the regular solutions of the equations (the Cauchy problem). In Sections 2-4, we have effectively shown that second-order differential equations (both linear and nonlinear) are at odds with the Cauchy problem because the equations show a strong preference for their own intrinsic solutions that, for inhomogeneous equations of the Lane-Emden type, cannot be obtained by solving the boundary-value problem (hence they are singular). The regular Cauchy-type solutions are then attracted to and forced to oscillate about the intrinsic solutions, which means that they do their best to match those dominant solutions. This results in "regular" oscillatory density and rotation profiles whose averages are precisely the underlying intrinsic solutions [40] [44] [45]. In this sense, the differential equations succeed in imposing their preferences to the Cauchy problem. This, by itself, is an important conclusion that has ramifications beyond astrophysics for the theory of second-order differential equations of mathematical physics.

The intrinsic solutions are very much related to the so-called trivial solutions of differential equations [44]. We now understand that there are no trivial solutions; in fact such solutions of second-order equations are quite dominant (Section 2): in many cases of interest, knowing the trivial solution of an equation implies that we know the average behaviour of all the regular solutions that depend on various types of boundary conditions but, nevertheless, end up oscillating about the intrinsic solution, provided that the differential equation is a harmonic

oscillator (as the ordinary Bessel equations in Section 2; the nonisothermal Lane-Emden equations without rotation in Section 2; and the Lane-Emden equations with rotation in Sections 3 and 4).

The mean density and rotation profiles that we have derived analytically in Sections 3 and 4 are dominated by natural logarithms and power laws. This is the implicit reason that Marr [67] has recently suceeded in matching the shapes of the rotation curves of a sample of 37 spiral galaxies by using the log-normal probability distribution (and no dark matter or modified gravity) to describe the density profiles in the equatorial planes of the discs. This surface density distribution (Equation (2) in [67]) is equivalent to a variable power law of the form $x^{k(x)}$ in normalized radius x in which the index $k(x)$ varies slowly across the disc as

$$k(x) = -1 - \frac{1}{2\sigma^2} \ln x, \tag{22}$$

where σ^2 is the variance of the distribution, a free parameter to be fitted for each spiral galaxy model. In principle, a slowly varying $k(x)$ is permitted by our analytic solution because the specific enthalpy $h(x)$ is a strictly local potential function (Equation (13) in Section 3.1); and if the power-law index k has to vary radially in order for the equilibrium disc to obey some other fundamental law (e.g., as Marr states, the total entropy of the overall configuration should be maximized), then such adjustment may occur on timescales determined by the local sound speed.

Returning to the astrophysical context, the intrinsic solutions of the various types of the Lane-Emden equation with rotation have, for the first time, succeeded in explaining the flat rotation curves of spiral galaxy discs without the need of invoking dark matter or modified gravity. Flat rotation curves are a rigorous requirement of Newtonian gravity in gaseous self-gravitating astrophysical discs (Sections 3.1 and 4.1). The only fair way to describe the results of Sections 3 and 4 is that Sir Isaac Newton [68] is vindicated once again, and our searches for dark matter in the universe and our attempts to modify Newtonian gravity (references are listed in Section 1) have sadly amounted to just a "wild-goose chase". In fact, since the flat rotation curves of spiral galaxies can now be explained at such a fundamental level, the massive observational results collected over the years must be considered as yet another test that Newtonian gravity has successfully passed on scales of ~10-100 kpc and at nonrelativistic velocities. This is an impressive achievement when compared to previous tests conducted on and limited to scales no larger than that of our solar system.

Our results render the Dark Matter Hypothesis unnecessary on galaxy scales. This removes the largest pillar of this hypothesis (flat rotation curves have remained to this day the "strongest piece of evidence" in favour of dark matter, but not any longer). But this "aetherial" hypothesis is not about to roll over and die without a fight. The next areas of confrontation will be larger than galaxy scales and cosmology. We are very much encouraged from a recent report of the absence of dark matter on larger than galaxy-disc scales: Magain & Chantry [69] analyzed 25 gravitational lenses and found that their mass determinations indicate the absence of extended dark matter haloes all the way out to distances comparable to the Einstein ring (the separation between lensed quasar images).

On the other hand, we do not anticipate any serious problems materializing in cosmology because we do not believe that there currently is any credible observational evidence in favour of dark matter or modified gravity on those largest scales. In fact, some results that argue against the necessity for dark matter on various scales have timidly begun to appear [70]-[76]. For these reasons, here is how we approach the issue of dark matter now: Christiaan Huygens presented his theory of elastic longitudinal light waves propagating in "aether" to the Paris Academy of Sciences in 1678 and published his views a few years later [77]. That year, the physics world entered a Dark Age that lasted nearly 200 years, until the genious of Michelson & Morley [78] finally showed that the universe is not filled with aether. It seems that we also live in another Dark Age, the "Dark Matter Dark Age", that commenced in the 1970s when K. C. Freeman [1] and others [2]-[22] reported that the rotation curves of spiral galaxies were not falling with radius. By all accounts, the current Dark Age has lasted for nearly 50 years; and the sooner we get out of it, the better for our understanding of the large scales of the universe around us.

The analytic solutions that we have derived in this work should also find applications in the field of protoplanetary disc research (Sections 3.2 and 4.2), especially if very young, purely self-gravitating discs could be found in the future [56]. At present, only observations of Class 0 YSOs are widely available [57] [58], but these systems are not young enough and do not have flat rotation curves; their protostars have formed and they are changing the dynamics and kinematics of the discs.[4] We are encouraged however by the report of Tsitali *et al*.

[66] who discovered a rising rotation curve in the inner 2000-8000 AU of the "first hydrostatic core" candidate Cha-MMS1.

Acknowledgements

We thank Joel Tohline for feedback and guidance over many years; and John Marr, Hsi-Wei Yen, and Earl Schulz for many fruitful discussions, especially those concerning the observations of astrophysical self-gravitating discs.

References

[1] Freeman, K.C. (1970) *Astrophysical Journal*, **160**, 811. http://dx.doi.org/10.1086/150474

[2] Rogstad, D.H. and Shostak, G.S. (1972) *Astrophysical Journal*, **176**, 315. http://dx.doi.org/10.1086/151636

[3] Roberts, M.S. and Rots, A.H. (1973) *Astronomy & Astrophysics*, **26**, 483-485.

[4] Bosma, A. (1978) The Distribution and Kinematics of Neutral Hydrogen in Spiral Galaxies of Various Morphological Types. Ph.D. Thesis, University of Groningen, Groningen.

[5] Rubin, V.C. (1980) *Astrophysical Journal*, **238**, 808-817. http://dx.doi.org/10.1086/158041

[6] Rubin, V.C., Ford Jr., W.K. and Thonnard, N. (1980) *Astrophysical Journal*, **238**, 471-487. http://dx.doi.org/10.1086/158003

[7] Bosma, A. (1981a) *Astronomical Journal*, **86**, 1791. http://dx.doi.org/10.1086/113062

[8] Bosma, A. (1981b) *Astronomical Journal*, **86**, 1825. http://dx.doi.org/10.1086/113063

[9] Rubin, V.C., Ford Jr., W.K., Thonnard, N. and Burstein, D. (1982) *Astrophysical Journal*, **261**, 439-456. http://dx.doi.org/10.1086/160355

[10] Bahcall, J.N. and Casertano, S. (1985) *Astrophysical Journal*, **293**, L7-L10. http://dx.doi.org/10.1086/184480

[11] van Albada, T.S., Sancisi, R., Petrou, M. and Tayler, R.J. (1986) *Philosophical Transactions of the Royal Society A*, **320**, 447-464. http://dx.doi.org/10.1098/rsta.1986.0128

[12] Kent, S.M. (1987) *The Astronomical Journal*, **93**, 816-832. http://dx.doi.org/10.1086/114366

[13] Begeman, K.G. (1987) HI Rotation Curves of Spiral Galaxies. PhD Thesis, University of Groningen, Groningen.

[14] Persic, M. and Salucci, P. (1988) *Monthly Notices of the Royal Astronomical Society*, **234**, 131-154. http://dx.doi.org/10.1093/mnras/234.1.131

[15] Begeman, K.G. (1989) *Astronomy & Astrophysics*, **223**, 47-60.

[16] Persic, M. and Salucci, P. (1990) *Monthly Notices of the Royal Astronomical Society*, **245**, 577-581.

[17] Carignan, C., Charbonneau, P., Boulanger, F. and Viallefond, F. (1990) *Astronomy & Astrophysics*, **234**, 43-52.

[18] Casertano, S. and van Gorkom, J.H. (1991) *The Astronomical Journal*, **101**, 1231-1241. http://dx.doi.org/10.1086/115759

[19] Broeils, A.H. (1992) Dark and Visible Matter in Spiral Galaxies. PhD Thesis, University of Groningen, Groningen.

[20] Persic, M. and Salucci, P. (1995) *Astrophysical Journal Supplement*, **99**, 501. http://dx.doi.org/10.1086/192195

[21] Persic, M., Salucci, P. and Stel, F. (1996) *Monthly Notices of the Royal Astronomical Society*, **281**, 27-47. http://dx.doi.org/10.1093/mnras/278.1.27

[22] Salucci, P. and Persic, M. (1997) Dark Halos Around Galaxies. In: Salucci, P. and Persic, M., Eds., *Dark and Visible Matter in Galaxies*, ASP Conference Series, Vol. 117, Astronomical Society of the Pacific, San Francisco, 1.

[23] Milgrom, M. (1983) *Astrophysical Journal*, **270**, 365-370. http://dx.doi.org/10.1086/161130

[24] Milgrom, M. (1983) *Astrophysical Journal*, **270**, 371-389. http://dx.doi.org/10.1086/161131

[25] Milgrom, M. (1983) *Astrophysical Journal*, **270**, 384. http://dx.doi.org/10.1086/161132

[26] Tohline, J.E. (1983) Stabilizing a Cold Disk with a 1/r Force Law. In: Athanassoula, E., Ed., *Internal Kinematics and Dynamics of Galaxies*, IAU Symposium 100, Reidel, Dordrecht, 205.

[27] Felten, J.E. (1984) *Astrophysical Journal*, **286**, 3-6. http://dx.doi.org/10.1086/162569

[28] Sanders, R.H. (1984) *Astronomy & Astrophysics*, **136**, L21-L23.

[4]When a protostar grows at the center of a protoplanetary disc, the enthalpy is defeated by gravity because the enthalpy cannot source and manipulate this component of the gravitational field via the Poisson equation (see Section 3.1); and the intrinsic solutions described in this work are no longer valid for such gaseous discs subject to "external" gravitational fields.

[29] Sanders, R.H. (1986) *Monthly Notices of the Royal Astronomical Society*, **223**, 539-555. http://dx.doi.org/10.1093/mnras/223.3.539

[30] Mannheim, P.D. and Kazanas, D. (1989) *Astrophysical Journal*, **342**, 635-638. http://dx.doi.org/10.1086/167623

[31] Christodoulou, D.M. (1991) *Astrophysical Journal*, **372**, 471-477. http://dx.doi.org/10.1086/169992

[32] Mannheim, P.D. and O'Brien, J.G. (2011) *Physical Review Letters*, **106**, Article ID: 121101. http://dx.doi.org/10.1103/PhysRevLett.106.121101

[33] Mannheim, P.D. and O'Brien, J.G. (2012) *Physical Review D*, **85**, Article ID: 124020. http://dx.doi.org/10.1103/PhysRevD.85.124020

[34] Gallagher III, J.S., Hunter, D.A. and Tutukov, A.V. (1984) *Astrophysical Journal*, **284**, 544-556. http://dx.doi.org/10.1086/162437

[35] Mestel, L. (1963) *Monthly Notices of the Royal Astronomical Society*, **126**, 553-575. http://dx.doi.org/10.1093/mnras/126.6.553

[36] Jalali, M.A. and Abolghasemi, M. (2002) *Astrophysical Journal*, **580**, 718-724. http://dx.doi.org/10.1086/343849

[37] Binney, J. and Tremaine, S. (1987) Galactic Dynamics. Princeton University Press, Princeton.

[38] Schulz, E. (2012) *Astrophysical Journal*, **747**, 106. http://dx.doi.org/10.1088/0004-637X/747/2/106

[39] Watson, G.N. (1922) A Treatise on the Theory of Bessel Functions. Cambridge University Press, Cambridge.

[40] Christodoulou, D.M., Graham-Eagle, J. and Katatbeh, Q.D. (2016) *Advances in Difference Equations*, **2016**, 48. http://dx.doi.org/10.1186/s13662-016-0774-x

[41] Robe, H. (1968) *Annales d'Astrophysique*, **31**, 549-558.

[42] Schmitz, F. and Ebert, R. (1986) *Astronomy & Astrophysics*, **154**, 214-218.

[43] Schneider, M. and Schmitz, F. (1995) *Astronomy & Astrophysics*, **301**, 933-940.

[44] Christodoulou, D.M. and Kazanas, D. (2007) Exact Solutions of the Isothermal Lane-Emden Equation with Rotation and Implications for the Formation of Planets and Satellites. arXiv:0706.3205v2.

[45] Christodoulou, D.M., Katatbeh, Q.D. and Graham-Eagle, J. (2016) *Journal of Inequalities and Applications*. (In Press)

[46] Lane, J.H. (1870) *American Journal of Science and Arts*, Second Series, **50**, 57-74.

[47] Emden, R. (1907) Gaskugeln. B.G. Teubner, Leipzig.

[48] Stodólkiewicz, J.S. (1963) *Acta Astronautica*, **13**, 30-54.

[49] Ostriker, J. (1964) *Astrophysical Journal*, **140**, 1056. http://dx.doi.org/10.1086/148005

[50] Horedt, G.P. (2004) Polytropes. Kluwer Academic Publishers, Dordrecht.

[51] Schmid-Burgk, J. (1967) *Astrophysical Journal*, **149**, 727. http://dx.doi.org/10.1086/149305

[52] Chandrasekhar, S. (1939) An Introduction to the Study of Stellar Structure. University of Chicago Press, Chicago.

[53] Hayashi, C., Narita, S. and Miyama, S.M. (1982) *Progress of Theoretical Physics*, **68**, 1949-1966. http://dx.doi.org/10.1143/PTP.68.1949

[54] Narita, S., Kiguchi, M., Miyama, S.M. and Hayashi, C. (1990) *Monthly Notices of the Royal Astronomical Society*, **244**, 349-356.

[55] Schmitz, F. (1986) *Astronomy & Astrophysics*, **169**, 171-177.

[56] Kamp, I., Freudling, W., Robberto, M., Chengalur, J. and Keto, E. (2008) *Physica Scripta*, **2008**, Article ID: 014013. http://dx.doi.org/10.1088/0031-8949/2008/T130/014013

[57] Williams, J.P. and Cieza, L.A. (2011) *Annual Review of Astronomy and Astrophysics*, **49**, 67-117. http://dx.doi.org/10.1146/annurev-astro-081710-102548

[58] Belloche, A. (2013) *EAS Publications Series*, **62**, 25-66. http://dx.doi.org/10.1051/eas/1362002

[59] Tobin, J.J., Hartmann, L., Chiang, H.-F., Wilner, D.J., Looney, L.W., Loinard, L., Calvet, N. and D'Alessio, P. (2012) *Nature*, **492**, 83-85. http://dx.doi.org/10.1038/nature11610

[60] Tobin, J.J., Hartmann, L., Chiang, H.-F., Wilner, D.J., Looney, L.W., Loinard, L., Calvet, N. and D'Alessio, P. (2013) *Astrophysical Journal*, **771**, 48. http://dx.doi.org/10.1088/0004-637X/771/1/48

[61] Yen, H.-W., Koch, P.M., Takakuwa, S., Ho, P.T.P., Ohashi, N. and Tang, Y.-W. (2015) *Astrophysical Journal*, **799**, 193. http://dx.doi.org/10.1088/0004-637X/799/2/193

[62] Yen, H.-W., Takakuwa, S., Koch, P.M., Aso, Y., Koyamatsu, S., Krasnopolsky, R. and Ohashi, N. (2015) *Astrophysical Journal*, **812**, 129.

[63] Schmitz, F. and Ebert, R. (1987) *Astronomy & Astrophysics*, **181**, 41-49.

[64] Schmitz, F. (1988) *Astronomy & Astrophysics*, **200**, 120-126.

[65] Tohline, J.E. (2002) *Annual Review of Astronomy and Astrophysics*, **40**, 349-385. http://dx.doi.org/10.1146/annurev.astro.40.060401.093810

[66] Tsitali, A.E., Belloche, A., Commerçon, B. and Menten, K.M. (2013) *Astronomy & Astrophysics*, **557**, A98. http://dx.doi.org/10.1051/0004-6361/201321204

[67] Marr, J.H. (2015) *Monthly Notices of the Royal Astronomical Society*, **448**, 3229-3241. http://dx.doi.org/10.1093/mnras/stv216

[68] Newton, I. (1687) Philosophæ Naturalis Principia Mathematica. S. Pepys, Reg. Soc. Praesses, London.

[69] Magain, P. and Chantry, V. (2013) Gravitational Lensing Evidence against Extended Dark Matter Halos. arXiv: 1303.6896.

[70] Jordi, K., Grebel, E.K., Hilker, M., Baumgardt, H., Frank, M., Kroupa, P., Haghi, H., Côté, P. and Djorgovski, S.G. (2009) *The Astronomical Journal*, **137**, 4586-4596. http://dx.doi.org/10.1088/0004-6256/137/6/4586

[71] Lane, R.R., Kiss, L.L., Lewis, G.F., Ibata, R.A., Siebert, A., Bedding, T.R. and Székely, P. (2009) *Monthly Notices of the Royal Astronomical Society*, **400**, 917-923. http://dx.doi.org/10.1111/j.1365-2966.2009.15505.x

[72] Lane, R.R., Salinas, R. and Richtler, T. (2015) *Astronomy & Astrophysics*, **574**, A93. http://dx.doi.org/10.1051/0004-6361/201424074

[73] Moni Bidin, C., Carraro, G., Mendez, R.A. and Smith, R. (2012) *Astrophysical Journal*, **751**, 30. http://dx.doi.org/10.1088/0004-637X/751/1/30

[74] Moni Bidin, C., Smith, R., Carraro, G., Mendez, R.A. and Moyano, M. (2015) *Astronomy & Astrophysics*, **573**, A91. http://dx.doi.org/10.1051/0004-6361/201424675

[75] Lelli, F. (2014) *Galaxies*, **2**, 292-299. http://dx.doi.org/10.3390/galaxies2030292

[76] López-Corona, O. (2015) *Journal of Physics*: *Conference Series*, **600**, 012046.

[77] Huygens, C. (1690) Treatise of Light. Pierre van der Aa, Leyden.

[78] Michelson, A.A. and Morley, E.W. (1887) *American Journal of Science*, **34**, 333-345. http://dx.doi.org/10.2475/ajs.s3-34.203.333

[79] Huré, J.-M., Hersant, F., Carreau, C. and Busset, J.-P. (2008) *Astronomy & Astrophysics*, **490**, 477-486. http://dx.doi.org/10.1051/0004-6361:200809682

[80] Myers, P.C. (2009) *Astrophysical Journal*, **700**, 1609-1625. http://dx.doi.org/10.1088/0004-637X/700/2/1609

[81] Fischera, J. and Martin, P.G. (2012) *Astronomy & Astrophysics*, **542**, A77. http://dx.doi.org/10.1051/0004-6361/201218961

Appendix: Vertically Thin Discs Versus Infinite Cylinders

When applied to the equatorial planes ($Z=0$) of thin discs, Equations (5) and (16) do not include the specific enthalpy term

$$\left.\frac{\mathrm{d}^2 h(x,z)}{\mathrm{d}z^2}\right|_{z=0}, \tag{23}$$

where $z \equiv Z/R_o$. This term is small for cold gases since it scales as the sound speed squared c_o^2 . Nevertheless, strong objections have been raised about the validity of this approximation. Here we address such objections as follows.

The analysis presented in Sections 3 and 4 shows that the specific enthalpy in cylinders assumes a logarithmic radial profile (Equation (14)). Now, pressure is an isotropic force and its nature is to push spherically out in self-gravitating gases. Therefore, Equation (14) suggests that, in axisymmetry, the specific enthalpy of the gas could assume the spherical form

$$h(r) \propto \ln r = \frac{1}{2}\ln\left(x^2+z^2\right), \tag{24}$$

where r is the spherical radius normalized by R_o . Such behaviour is suppressed in cylinders by dropping the dependence on z from the equations. But, in principle, it should not be suppressed off-hand in discs, so the influence of the terms (23) and (24) should at least be quantified in the equatorial planes of discs:

1) Applying the radial (x) component of the cylindrical Laplacian to Equation (24) and then setting $z=0$, we confirm Equation (14); that is, this term of the Lane-Emden equations vanishes on the equatorial plane, as was also found in the intrinsic solutions of Sections 3 and 4.

2) Taking the derivative with respect to z in Equation (24) and then setting $z=0$, we obtain the correct symmetry condition that $\mathrm{d}h/\mathrm{d}z=0$ on the equatorial plane of the disc. Then, using the equation for vertical hydrostatic balance, $\mathrm{d}h/\mathrm{d}z+\mathrm{d}\psi/\mathrm{d}z=0$, we also obtain the correct boundary condition for the self-gravitational potential, that is $\mathrm{d}\psi/\mathrm{d}z=0$ at $z=0$.

3) Combining Equations (23) and (24), we find that

$$\left.\frac{\mathrm{d}^2 h(x,z)}{\mathrm{d}z^2}\right|_{z=0} \propto \frac{1}{x^2}. \tag{25}$$

This is the term that was ignored for discs in the radial force balance described by Equation (12). But it makes only a minor contribution to the force balance over all radii where $c_o \ll v(x)$. Specifically, it modifies Equations (6) and (17) to

$$\tau(x) = \frac{1}{x}\frac{\mathrm{d}v^2}{\mathrm{d}x} - \frac{\ell}{x^2}, \tag{26}$$

where $\ell<0$ is a proportionality constant of order c_o^2 . For our models, $\ell=(k-1)c_o^2$ for isothermal discs and $\ell=knc_o^2$ for polytropic discs. Since $\ell<0$ (because $k<0$), the last term in Equation (26) opposes self-gravity. By integration of this equation, we find that its contribution to the rotation profile $v^2(x)$ is logarithmic:

$$v^2(x) = \int \tau(x)\,x\mathrm{d}x + \ell\ln x + B, \tag{27}$$

where B is the integration constant. We see now that the dimensionless mass per unit length

$$m(x) \equiv \int \tau(x)\,x\mathrm{d}x, \tag{28}$$

makes the largest contribution to the rotation curve. Neglecting the term $\ell\ln x$, we can write

$$v^2(x) \approx B + m(x). \tag{29}$$

For $B=0$ (*i.e.*, $v(0)=0$) and $v^2 = x(\mathrm{d}\psi/\mathrm{d}x)$, this equation reduces to the cylindrical Gauss's law applied onto the equatorial plane of the disc in the limit of $z\to 0$. This result differs conclusively from the conventional thinking that is responsible for the acceptance of dark matter in galaxies; that $v^2/r = Gm/r^2$, thus a

constant v requires $m \propto r$. Odd as it may seem, Equation (29) makes sense for discs because their equatorial planes do possess cylindrical symmetry in the limit of $z \to 0$, no matter which coordinate system is used. For this reason, Equation (27) can also be derived by using the Laplacian in spherical coordinates in the limit of $z \to 0$, $r \to x$, and for $x \gg 1$, where the inertial terms of the two coordinate systems become unimportant.

Similar results can also be obtained for Equations (3) and (4) in which the inertial terms $(D-1)y'/x \to 0$ for $x \gg 1$ in both coordinate systems with $D = 2$ and $D = 3$; and from some of the finite razor-thin disc/ring models of Huré *et al.* [79] with power-law indices > -1 (their Figure 4) in which the rotation curves can be calculated and turn out to be slowly increasing functions of radius. For power-law indices < -1, edge effects become important and blur the picture. But in the absence of radial boundaries, the infinitely extended discs show a self-similar power-law potential (their Equation (66)) capable of generating flat and slowly increasing rotation curves for shallow surface density profiles. These models support our results because the radial distribution of their equatorial surface densities is a shallow power law that places sufficient mass at all radii to keep the rotation curves rising.

On the other hand, the models in [79] with steep power-law surface densities ($k < -1$) exhibit falling rotation curves. This observation helps us understand a problem that plagues such two-dimensional models and that has been summarized in [37]. In all of these models, the surface densiy profiles were obtained by collapsing spheroidal homoeoids down to an equatorial plane. Then, when the surface densities decline steeply with radius x, there is not enough mass interior to x at large radii to cause the rotation curves to continue rising with radius. This is because some of the interior mass comes from homoeoids exterior to x and this mass does not attract at x. The Mestel disc [35] [79] [38] is the marginal collapsed model among all classical potential-surface density pairs in which all mass interior to x attracts, while the exterior mass does not contribute to radius x. So this model has barely enough mass at all radii to maintain a flat rotation curve. This explains also the peculiar property of the Mestel disc [35] to be the only two-dimensional model that obeys the cylindrical Gauss's law at $Z = 0$ for every radius R [37]:

$$V^2(R) = R\frac{\mathrm{d}\Phi}{\mathrm{d}R} = \frac{GM(R)}{R}, \tag{30}$$

where $M(R)$ is the mass enclosed within radius R and $\Phi(R)$ is the self-gravitational potential.

We argue that all other razor-thin models that do not satisfy Gauss's law at $Z = 0$ are physically questionable. This is because the homoeoids from which they were constructed do satisfy Gauss's law over similar Gaussian surfaces. Our models do not suffer from such questionable behaviour because they follow the three-dimensional density distribution $\rho(R,0)$ in the equatorial plane of a thin (or thick) disc; and these models obey Gauss's law at every radius R in the limit of $Z \to 0$ for a Gaussian surface that matches their symmetry (an infinitesimally tall cylindrical surface of radius R).

Rising rotation curves were also found by Marr [67] who studied models in which $1/x$ surface density profiles were truncated at a radius near the last observed point. By not being there to pull outward, the exterior mass allowed the inward force due to the interior mass to amplify, and this caused the rotation curves to turn up near the peripheries of the discs, despite the fact that these were essentially Mestel discs. In any case, our analytic solutions (Sections 3 and 4) avoid the above pitfalls (mixing of homoeoids of different sizes, truncation creating sharp edges); and they show that the volumetric density at all equatorial radii is always sufficient to continue pulling the rotation curves higher.

For $B \geq 0$ in Equation (29), the term $B + m$ determines the rotation curves of the models. Hidden in this term are the two different types of Newtonian models that we have discovered and that require monotonically rising (not flat) rotation curves in discs and cylinders:

1) In all isothermal models with $k > -1$ and in all polytropic models (**Figures 5-8**), the indefinite integral in Equation (27) is positive definite. Then $B = 0$ implying that $v(0) = 0$, $v(x) = \sqrt{m(x)}$, and the rotation curves rise at all radii as more interior mass is added incrementally by the radial integration with increasing x.

2) On the other hand, in the isothermal models with $k < -1$ (**Figure 3** and **Figure 4**), the indefinite intergal in Equation (27) is negative definite (although the integral for the mass is still positive). Then, in effect, $v^2 = B - |m(x)|$ (where now it is necessary that $B > 0$), and the rotation curves approach asymptotically the constant value $v = \sqrt{B}$ as $|m(x)|$ decreases toward zero with increasing x. Because $\tau(x)$ is a very steeply decreasing function of x in such cases, then $m(x) \to 0$ from below quite fast, and that makes the rotation

curves appear quite flat over most radii in **Figure 3** and **Figure 4**. Here, $B > 0$ implies the inner boundary condition that $v(x_1) = 0$ at a cutoff radius x_1.

Returning to the pressure term ($\ell \ln x$) in Equation (27), if it is included in $v^2(x)$ in future models, this term has the potential to drive some rotation curves down at large radii because it is negative for $x > 1$ (since $\ell < 0$). This behaviour can occur only in discs since this term is not present in infinite cylinders. So, unlike discs, cylindrical star-forming filaments [80] [81] can *never* exhibit falling rotation curves in equilibrium; except for the unlikely case that their age would be smaller than the sound-crossing time, a sign of extreme youth indicating that a global equilibrium has not been established yet.

Our analysis in Sections 3 and 4 was carried out without the $\ell \ln x$ term of Equation (27) in order to bring out the physics of such Newtonian systems. Nevertheless, the pressure term in Equation (27) or some similar approximation may be of interest to researchers planning to remodel the rotation curves of spiral galaxies. But it does not appear to be necessary to modeling the rotation of pre-Class 0 protoplanetary discs because such starless discs are subject to extended infall of matter [58] that ought to create cylindrical, vertically thick quasi-equilibria to a good approximation.

Journal of Modern Physics, 2016, 7, 699-728
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jmp
http://dx.doi.org/10.4236/jmp.2016.77068

Airy, Beltrami, Maxwell, Einstein and Lanczos Potentials Revisited

J.-F. Pommaret

CERMICS, Ecole des Ponts ParisTech, Marne-la-Vallée, France
Email: jean-francois.pommaret@wanadoo.fr, pommaret@cermics.enpc.fr

Received 19 February 2016; accepted 25 April 2016; published 28 April 2016

Abstract

The purpose of this paper is to revisit the well known *potentials*, also called *stress functions*, needed in order to study the parametrizations of the stress equations, respectively provided by G.B. Airy (1863) for 2-dimensional elasticity, then by E. Beltrami (1892), J.C. Maxwell (1870) for 3-dimensional elasticity, finally by A. Einstein (1915) for 4-dimensional elasticity, both with a variational procedure introduced by C. Lanczos (1949, 1962) in order to relate potentials to Lagrange multipliers. Using the methods of *Algebraic Analysis*, namely mixing differential geometry with homological algebra and combining the *double duality test* involved with the *Spencer cohomology*, we shall be able to extend these results to an arbitrary situation with an arbitrary dimension *n*. We shall also explain why double duality is perfectly adapted to variational calculus with differential constraints as a way to eliminate the corresponding Lagrange multipliers. For example, the *canonical parametrization* of the stress equations is just described by the formal adjoint of the $n^2(n^2-1)/12$ components of the linearized Riemann tensor considered as a linear second order differential operator but the minimum number of potentials needed is equal to $n(n-1)/2$ for any *minimal parametrization*, the Einstein parametrization being "*in between*" with $n(n+1)/2$ potentials. We provide all the above results without even using indices for writing down explicit formulas in the way it is done in any textbook today, but it could be *strictly impossible* to obtain them without using the above methods. We also revisit the *possibility* (Maxwell equations of electromagnetism) or the *impossibility* (Einstein equations of gravitation) to obtain canonical or minimal parametrizations for various equations of physics. It is nevertheless important to notice that, when *n* and the algorithms presented are known, most of the calculations can be achieved by using computers for the corresponding symbolic computations. Finally, though the paper is mathematically oriented as it aims providing new insights towards the mathematical foundations of general relativity, it is written in a rather self-contained way.

Keywords

Stress Equations, Stress Functions, Elasticity Theory, Lagrange Multipliers, Formal Adjoint, Control

How to cite this paper: Pommaret, J.-F. (2016) Airy, Beltrami, Maxwell, Einstein and Lanczos Potentials Revisited. *Journal of Modern Physics*, **7**, 699-728. http://dx.doi.org/10.4236/jmp.2016.77068

Theory, General Relativity, Einstein Equations, Lanczos Potentials, Algebraic Analysis, Riemann Tensor, Weyl Tensor

1. Introduction

The language of *differential modules* has been recently introduced in control theory as a way to understand in an intrinsic way the *structural properties* of systems of ordinary differential (OD) or partial differential (PD) equations (controllability, observability, identifiability, ...) [1]-[10]. A similar comment can be done for optimal control that is for variational calculus with differential constraints, and the author thanks Prof. Lars Andersson (Einstein Institute, Potsdam) for having suggested him to study the Lanczos potential within this new framework.

We start providing a few explicit examples in order to convince the reader that the corresponding computations are often becoming so tricky that nobody could achieve them or even imagine any underlying general algorithm, for example in the study of the mathematical foundations of control theory, elasticity theory or general relativity.

EXAMPLE 1.1: *OD Control Theory*

With one independent variable *x*, for example the time *t* in control theory or the curvilinear abcissa *s* in the study of a beam, and three unknowns $\left(y^1, y^2, y^3\right)$. Setting formally $d_x y^k = y_x^k$ for $k = 1, 2, 3$ and so on, let us consider the system made by the two *first order* OD equations depending on a variable coefficient $a(x)$:

$$y_x^3 - a(x)\, y^2 - y_x^1 = 0, \quad y^3 - y_x^2 + y_x^1 = 0$$

In control theory, if $a = cst$ is a constant parameter, one could bring the system to *any first order Kalman form* and check that the corresponding control system is *controllable* if and only if $a(a-1) \neq 0$, that is $a \neq 0$ and $a \neq 1$ (exercise), independently of the choice of 1 *input* and 2 *outputs* among the 3 control variables [11]. In addition to that, using the second OD equation in the form $y^3 = y_x^2 - y_x^1$ and substituting in the first, we get the only *second order* OD equation:

$$y_{xx}^1 + y_x^1 - y_{xx}^2 + a(x)\, y^2 = 0$$

a result leading to a kind of "vicious circle" because the only way to test controllability is ... to bring this second order equation back to a first order system and there are a lot of possibilities. Again, *in any case*, the only critical values are $a = 0$ and $a = 1$. Of course, one could dream about a direct approach providing the same result in an intrinsic way. Introducing the operator $d = d_x$ as the (formal) derivative with respect to *x*, we may rewrite the last equation in the form:

$$d(d+1)\, y^1 = \left(d^2 - a\right) y^2$$

Replacing the operators $d(d+1)$ and $d^2 - a$ by the polynomials $\chi(\chi+1)$ and $\chi^2 - a$, the two polynomials have a common root $\chi = 0 \Rightarrow a = 0$ or $\chi = -1 \Rightarrow a = 1$ and we find back the desired critical values but such a result is not intrinsic at all. However, we notice that, for example $a = 0 \Rightarrow d\left((d+1)\, y^1 - d y^2\right) = 0$. Introducing $z' = y_x^1 + y^1 - y_x^2$, we get $z_x' = 0$ while $a = 1 \Rightarrow (d+1)\left(d y^1 - (d-1)\, y^2\right)$ that is, setting $z'' = y_x^1 - y^2$, we get now $z_x'' + z'' = 0$. Calling "*torsion element*" any scalar quantity made from the unknowns and their derivatives but satisfying at least one OD equation, we discover that such quantities do exist ... if and only if $a = 0$ or $a = 1$ (exercise). Of course, the existence of any torsion element breaks at once the controllability of the system but the converse is not evident at all, a result leading nevertheless to the feeling that *a control system is controllable if and only if no torsion element can be found* and such an idea can be extended "*mutatis mutandis*" to any system of PD equations [6]. However, this result could be useful if and only if there is a test for checking such a property of the system.

Now, using a variable parameter $a(x)$, *not a word of the preceding approach is left* but the concept of a torsion element still exists. We shall prove, at the end of the paper, that the condition $a(a-1) \neq 0$ becomes $\partial_x a + a^2 - a \neq 0$ and that the computations needed are quite far from the previous ones. We ask the reader familiar with classical control theory to make his mind a few minutes (or hours!) to agree with us by trying to re-

cover himself such a differential condition.

EXAMPLE 1.2: *OD Optimal Control Theory*

OD optimal control is the study of OD variational calculus with OD constraints described by OD control systems. However, while studying optimal control, the author of this paper has been surprised to discover that, *in all cases*, the OD constraints were defined by means of controllable control systems. It is only at the end of this paper that the importance of such an assumption will be explained. For the moment, we shall provide an example allowing to exhibit all the difficulties involved. For this, let $y^1 = f^1(x), y^2 = f^2(x)$ be a solution of the following single input/single output (SISO) OD control system where *a* is a constant parameter:

$$y^1_x + y^1 - y^2_x - ay^2 = 0$$

Proceeding as before, the two polynomials replacing the respective operators are $\chi+1$, $\chi+a$ and can only have the common root $a=1$. Accordingly, the system is controllable if and only if $a\neq 1$ for any choice of input and output. Now, let us introduce the so-called "*cost function*" and let us look at the extremum of the integral $\int \frac{1}{2}\left(\left(y^1\right)^2-\left(y^2\right)^2\right)\mathrm{d}x$ under the previous OD constraint. It is well known that the proper way to study such a problem is to introduce a *Lagrange multiplier* λ and to vary the new integral:

$$\int\left[\frac{1}{2}\left(\left(y^1\right)^2-\left(y^2\right)^2\right)+\lambda\left(y^1_x+y^1-y^2_x-ay^2\right)\right]\mathrm{d}x$$

The corresponding *Euler-Lagrange* (EL) equations are:

$$\begin{cases} y^1 \rightarrow -\lambda_x+\lambda+y^1=0 \\ y^2 \rightarrow \lambda_x-a\lambda-y^2=0 \end{cases}$$

to which we must add the OD constraint when varying λ. Summing the two EL equations, we get $(a-1)\lambda = y^1-y^2$ and *two possibilities*:

1) $a=1\Rightarrow y^1-y^2=0$ compatible with the constraint.

2) $a\neq 1\Rightarrow \lambda=\left(y^1-y^2\right)/(a-1)$.

Substituting, we get:

$$\begin{cases} y^1_x-y^2_x-ay^1+y^2=0 \\ y^1_x-y^2_x+y^1-ay^2=0 \end{cases}$$

This system may not be formally integrable. Indeed, by substraction, we get $(a+1)\left(y^1-y^2\right)=0$ and must consider the following *two possibilities*:

$$\begin{cases} \bullet\ a=-1\Rightarrow y^1_x+y^1-y^2_x+y^2=0 \\ \bullet\ a\neq -1\Rightarrow y^1=y^2=0 \end{cases}$$

Summarising the results so far obtained, we discover that *the Lagrange multiplier is known if and only if the system is controllable*. Also, if $a=-1$, we may exhibit the parametrization $\xi_x-\xi=y^1, \xi_x+\xi=y^2$ and the cost function becomes $2\xi\xi_x=d_x\left(\xi^2\right)$. Equivalently, *when the system is controllable it can be parametrized and the variational problem with constraint becomes a variational problem without any constraint* which, *sometimes*, does not provide EL equations. We finally understand that extending such a situation to PD variational calculus with PD constraints needs new techniques.

EXAMPLE 1.3: *Elasticity Theory*

In classical elasticity, the *stress* tensor density $\sigma=\left(\sigma^{ij}=\sigma^{ji}\right)$ existing inside an elastic body is a symmetric 2-tensor density introduced by A. Cauchy in 1822. The corresponding Cauchy *stress equations* can be written as $\partial_r\sigma^{ir}=f^i$ where the right member describes the local density of forces applied to the body, for example gravitation. With zero second member, we study the possibility to "*parametrize*" the system of PD equations $\partial_r\sigma^{ir}=0$, namely to express its general solution by means of a certain number of arbitrary functions or *potentials*, called *stress functions*. Of course, the problem is to know about the number of such functions and the order of the parametrizing operator. In what follows, the space has *n* local coordinates $x=\left(x^i\right)=\left(x^1,\cdots,x^n\right)$. For

$n=1,2,3$ one may introduce the Euclidean metric $\omega=\left(\omega_{ij}=\omega_{ji}\right)$ while, for $n=4$, one may consider the Minkowski metric. A few definitions used thereafter will be provided later on.

- $n=1$: There is no possible parametrization of $\partial_x\sigma=0$.
- $n=2$: The stress equations become $\partial_1\sigma^{11}+\partial_2\sigma^{12}=0,\partial_1\sigma^{21}+\partial_2\sigma^{22}=0$. Their second order parametrization $\sigma^{11}=\partial_{22}\phi,\sigma^{12}=\sigma^{21}=-\partial_{12}\phi,\sigma^{22}=\partial_{11}\phi$ has been provided by George Biddell Airy (1801-1892) in 1863 [12]. It can be simply recovered in the following manner:

$$\partial_1\sigma^{11}=\partial_2\left(-\sigma^{12}\right)\Rightarrow\exists\varphi,\sigma^{11}=\partial_2\varphi,\sigma^{12}=-\partial_1\varphi,$$

$$\partial_2\sigma^{22}=\partial_1\left(-\sigma^{21}\right)\Rightarrow\exists\psi,\sigma^{22}=\partial_1\psi,\sigma^{21}=-\partial_2\psi$$

$$\sigma^{12}=\sigma^{21}\Rightarrow\partial_1\varphi=\partial_2\psi\Rightarrow\exists\phi,\varphi=\partial_2\phi,\psi=\partial_1\phi$$

We get the second order system:

$$\begin{cases}\sigma^{11}\equiv\partial_{22}\phi=0 & \boxed{\begin{matrix}1 & 2\\ 1 & \bullet\\ 1 & \bullet\end{matrix}}\\ -\sigma^{12}\equiv\partial_{12}\phi=0 & \\ \sigma^{22}\equiv\partial_{11}\phi=0 & \end{cases}$$

which is involutive with one equation of class 2, 2 equations of class 1 and it is easy to check that the 2 corresponding first order CC are just the stress equations. As we have a system with constant coefficients, we may use localization [6] [13] in order to transform the 2 PD equations into the 2 *linear* equations $\chi_1\sigma^{11}+\chi_2\sigma^{12}=0,\ \chi_1\sigma^{21}+\chi_2\sigma^{22}=0$ and get

$$\sigma^{11}=-\frac{\chi_2}{\chi_1}\sigma^{12}=-\frac{\left(\chi_2\right)^2}{\chi_1\chi_2}\sigma^{12},\quad\sigma^{22}=-\frac{\chi_1}{\chi_2}\sigma^{12}=-\frac{\left(\chi_1\right)^2}{\chi_1\chi_2}\sigma^{12}$$

Setting $\sigma^{12}=-\chi_1\chi_2\phi$, we finally get $\sigma^{11}=\left(\chi_2\right)^2\phi,\sigma^{22}=\left(\chi_1\right)^2\phi$ and obtain the previous parametrization by delocalizing, that is replacing now χ_i by ∂_i.

- $n=3$: Things become quite more delicate when we try to parametrize the 3 PD equations:

$$\partial_1\sigma^{11}+\partial_2\sigma^{12}+\partial_3\sigma^{13}=0,\quad\partial_1\sigma^{21}+\partial_2\sigma^{22}+\partial_3\sigma^{23}=0,\quad\partial_1\sigma^{31}+\partial_2\sigma^{32}+\partial_3\sigma^{33}=0$$

Of course, localization could be used similarly by dealing with the 3 *linear* equations:

$$\chi_1\sigma^{11}+\chi_2\sigma^{12}+\chi_3\sigma^{13}=0,\chi_1\sigma^{21}+\chi_2\sigma^{22}+\chi_3\sigma^{23}=0,\chi_1\sigma^{31}+\chi_2\sigma^{32}+\chi_3\sigma^{33}=0$$

having rank 3 for 6 unknowns but, even if we succeed bringing all the fractions to the same denominator as before after easy but painful calculus, there is an additional difficulty which is well hidden. Indeed, coming back to the previous Example when $a=cst$, say $a=1$, we should get $\left(d^2+d\right)y^1=\left(d^2-1\right)y^2\Rightarrow\chi\left(\chi+1\right)y^1=\left(\chi+1\right)\left(\chi-1\right)y^2\Rightarrow\chi y^1=\left(\chi-1\right)y^2\Rightarrow y^1=\frac{\chi-1}{\chi}y^2$. Hence, setting $y^2=\chi y,y^1=\left(\chi-1\right)y$, we only get a parametrization of the *first order* OD equation $z\equiv\dot{y}^1-\dot{y}^2+y^2=0$ leading to $\dot{z}+z=0$. Accordingly, localization does indeed provide a parametrization, ... if we already know there exists a possibility to parametrize the given system or if we are able to check that we have obtained such a parametrization by using involution, a way to supersede the use of Janet or Gröbner bases as was proved for the case $n=2$ [14]. Also, if we proceed along such a way, we should surely loose any geometric argument that could exist.

A direct computational approach has been provided by Eugenio Beltrami (1835-1900) in 1892 [15], James Clerk Maxwell (1831-1879) in 1870 [16] and Giacinto Morera (1856-1909) in 1892 [17] by introducing the 6 stress functions $\phi_{ij}=\phi_{ji}$ through the parametrization obtained by considering:

$$\sigma^{11}=\partial_{33}\phi_{22}+\partial_{22}\phi_{33}-2\partial_{23}\phi_{23}$$

$$\sigma^{12}=\sigma^{21}=\partial_{13}\phi_{23}+\partial_{23}\phi_{13}-\partial_{33}\phi_{12}-\partial_{12}\phi_{33}$$

and the additional 4 relations obtained by using a cyclic permutation of $(1,2,3)$. The system:

$$\left\{\begin{array}{ll}\sigma^{11}\equiv\partial_{33}\phi_{22}+\partial_{22}\phi_{33}-2\partial_{23}\phi_{23}=0 & 1\ 2\ 3\\ -\sigma^{12}\equiv\partial_{33}\phi_{12}+\partial_{12}\phi_{33}-\partial_{13}\phi_{23}-\partial_{23}\phi_{13}=0 & 1\ 2\ 3\\ \sigma^{22}\equiv\partial_{33}\phi_{11}+\partial_{11}\phi_{33}-2\partial_{13}\phi_{13}=0 & 1\ 2\ 3\\ \sigma^{13}\equiv\partial_{23}\phi_{12}+\partial_{12}\phi_{23}-\partial_{22}\phi_{13}-\partial_{13}\phi_{22}=0 & 1\ 2\ \bullet\\ -\sigma^{23}\equiv\partial_{23}\phi_{11}+\partial_{11}\phi_{23}-\partial_{12}\phi_{13}-\partial_{13}\phi_{12}=0 & 1\ 2\ \bullet\\ \sigma^{33}\equiv\partial_{22}\phi_{11}+\partial_{11}\phi_{22}-2\partial_{12}\phi_{12}=0 & 1\ 2\ \bullet\end{array}\right.$$

is involutive with 3 equations of class 3, 3 equations of class 2 and no equation of class 1. The 3 CC are describing the stress equations which admit therefore a parametrization ... justifying the localization approach "*a posteriori*" but without any geometric framework [18].

Surprisingly, the Maxwell parametrization is obtained by keeping $\phi_{11}=A,\phi_{22}=B,\phi_{33}=C$ while setting $\phi_{12}=\phi_{23}=\phi_{31}=0$ in order to obtain the system:

$$\left\{\begin{array}{ll}\sigma^{11}\equiv\partial_{33}B+\partial_{22}C=0 & 1\ 2\ 3\\ \sigma^{22}\equiv\partial_{33}A+\partial_{11}C=0 & 1\ 2\ 3\\ -\sigma^{23}\equiv\partial_{23}A=0 & 1\ 2\ \bullet\\ \sigma^{33}\equiv\partial_{22}A+\partial_{11}B=0 & 1\ 2\ \bullet\\ -\sigma^{13}\equiv\partial_{13}B=0 & 1\ \bullet\ \bullet\\ -\sigma^{12}\equiv\partial_{12}C=0 & 1\ \bullet\ \bullet\end{array}\right.$$

However, *this system may not be involutive* and no CC can be found "*a priori*" because the coordinate system is surely not δ-regular. Indeed, effecting the linear change of coordinates $x^1\to x^1+x^3, x^2\to x^2+x^3, x^3\to x^3$, we obtain the involutive system:

$$\left\{\begin{array}{ll}\partial_{33}C+\partial_{13}C+\partial_{23}C+\partial_{12}C=0 & 1\ 2\ 3\\ \partial_{33}B+\partial_{13}B=0 & 1\ 2\ 3\\ \partial_{33}A+\partial_{23}A=0 & 1\ 2\ 3\\ \partial_{23}C-\partial_{13}B-\partial_{13}C-\partial_{12}C+\partial_{22}C=0 & 1\ 2\ \bullet\\ \partial_{23}A-\partial_{22}C+\partial_{13}B+2\partial_{12}C-\partial_{11}C=0 & 1\ 2\ \bullet\\ \partial_{22}A+\partial_{22}C-2\partial_{12}C+\partial_{11}B+\partial_{11}C=0 & 1\ 2\ \bullet\end{array}\right.$$

and it is easy to check that the 3 CC obtained just amount to the desired 3 stress equations when coming back to the original system of coordinates. Again, *if there is a geometrical background, this change of local coordinates is hidding it totally*. Moreover, we notice that the stress functions kept in the procedure are just the ones on which ∂_{33} is acting. The reason for such an apparently technical choice is related to very general deep arguments in the theory of differential modules that will only be explained at the end of the paper. The Morera parametrization is obtained similarly by keeping now $\phi_{23}=L,\phi_{13}=M,\phi_{12}=N$ while setting $\phi_{11}=\phi_{22}=\phi_{33}=0$.

- $n\geq 4$: As already explained, localization *cannot* be applied directly as we don't know if a parametrization may exist and in any case no analogy with the previous situations $n=1,2,3$ could be used. Moreover, no known differential geometric background could be used at first sight in order to provide a hint towards the solution. Now, if ω is the Minkowski metric and $\phi=GM/r$ is the gravitational potential, then $\phi/c^2\ll 1$ and a perturbation $\Omega\in S_2T^*$ of ω may satisfy in vacuum the 10 second order *Einstein equations* for the 10 Ω:

$$E_{ij}\equiv\omega^{rs}\left(d_{ij}\Omega_{rs}+d_{rs}\Omega_{ij}-d_{ri}\Omega_{sj}-d_{sj}\Omega_{ri}\right)-\omega_{ij}\left(\omega^{rs}\omega^{uv}d_{rs}\Omega_{uv}-\omega^{ru}\omega^{sv}d_{rs}\Omega_{uv}\right)=0$$

by introducing the corresponding second order *Einstein operator* $S_2T^*\xrightarrow{\text{Einstein}}S_2T^*:\Omega\to E$ when $n=4$ [19]. Though it is well known that the corresponding second order *Einstein operator* is parametrizing the stress equations, the challenge of parametrizing Einstein equations has been proposed in 1970 by J. Wheeler for 1000 $ and solved *negatively* in 1995 by the author who only received 1 $. We shall see that, *exactly as before and*

though it is quite striking, the key ingredient will be to use the linearized Riemann tensor considered as a second order operator [6] [20]. As an *even more striking fact*, we shall discover that the condition $n \geq 4$ has only to do with Spencer cohomology for the symbol of the *conformal Killing equations*.

EXAMPLE 1.4: *PD Control Theory*

The aim of this last example is to prove that the possibility to exhibit two different parametrizations of the stress equations which has been presented in the previous example has surely nothing to do with the proper mathematical background of elasticity theory!

For this, let us consider the (trivially involutive) inhomogeneous PD equations with two independent variables (x^1, x^2), two unknown functions (η^1, η^2) and a second member ζ:

$$\partial_2 \eta^1 - \partial_1 \eta^2 + x^2 \eta^2 = \zeta$$

Multiplying on the left by a test function λ and integrating by parts, the corresponding inhomogeneous adjoint system of PD equations is:

$$\begin{cases} \eta^1 \rightarrow -\partial_2 \lambda = \mu^1 \\ \eta^2 \rightarrow \partial_1 \lambda + x^2 \lambda = \mu^2 \end{cases}$$

Using crossed derivatives, we get $\lambda = \partial_2 \mu^2 + \partial_1 \mu^1 + x^2 \mu^1$ and substituting, we get the two CC:

$$\begin{cases} -\partial_{22} \mu^2 - \partial_{12} \mu^1 - x^2 \partial_2 \mu^1 - 2\mu^1 = \nu^1 \\ \partial_{12} \mu^2 + \partial_{11} \mu^1 + 2x^2 \partial_1 \mu^1 + x^2 \partial_2 \mu^2 + (x^2)^2 \mu^1 - \mu^2 = \nu^2 \end{cases}$$

The corresponding generating CC for the second member (ν^1, ν^2) is:

$$\partial_2 \nu^2 + \partial_1 \nu^1 + x^2 \nu^1 = 0$$

Therefore ν^2 is differentially dependent on ν^1 but ν^1 is also differentially dependent on ν^2.

Multiplying the first equation by the test function ξ^1, the second equation by the test function ξ^2, adding and integrating by parts, we get the canonical parametrization $(\xi^1, \xi^2) \rightarrow (\eta^1, \eta^2)$:

$$\begin{cases} \mu^2 \rightarrow -\partial_{22} \xi^1 + \partial_{12} \xi^2 - x^2 \partial_2 \xi^2 - 2\xi^2 = \eta^2 \\ \mu^1 \rightarrow -\partial_{12} \xi^1 + x^2 \partial_2 \xi^1 - \xi^1 + \partial_{11} \xi^2 - 2x^2 \partial_1 \xi^2 + (x^2)^2 \xi^2 = \eta^1 \end{cases} \quad \begin{array}{|cc|} \hline 1 & 2 \\ 1 & \bullet \\ \hline \end{array}$$

of the initial system with zero second member. The system (up to sign) is involutive and the kernel of this parametrization has differential rank equal to 1.

Keeping $\xi^1 = \xi$ while setting $\xi^2 = 0$, we get the first minimal parametrization $\xi \rightarrow (\eta^1, \eta^2)$:

$$\begin{cases} -\partial_{22} \xi = \eta^2 \\ -\partial_{12} \xi + x^2 \partial_2 \xi - \xi = \eta^1 \end{cases} \quad \begin{array}{|cc|} \hline 1 & 2 \\ 1 & \bullet \\ \hline \end{array}$$

The system is again involutive (up to sign) and the parametrization is minimal because the kernel of this parametrization has differential rank equal to 0. With a similar comment, setting now $\xi^1 = 0$ while keeping $\xi^2 = \xi'$, we get the second minimal parametrization $\xi' \rightarrow (\eta^1, \eta^2)$:

$$\begin{cases} \partial_{11} \xi' - 2x^2 \partial_1 \xi' + (x^2)^2 \xi' = \eta^1 \\ \partial_{12} \xi' - x^2 \partial_2 \xi' - 2\xi' = \eta^2 \end{cases}$$

EXAMPLE 1.5: *PD Optimal Control Theory*

Let us revisit briefly the foundation of n-dimensional elasticity theory as it can be found today in any textbook, restricting our study to $n = 2$ for simplicity. If $x = (x^1, x^2)$ is a point in the plane and $\xi = (\xi^1(x), \xi^2(x))$ is the displacement vector, lowering the indices by means of the Euclidean metric, we may introduce the "small" deformation tensor $\epsilon = (\epsilon_{ij} = \epsilon_{ji} = (1/2)(\partial_i \xi_j + \partial_j \xi_i))$ with $n(n+1)/2 = 3$ (independent) *components* $(\epsilon_{11}, \epsilon_{12} = \epsilon_{21}, \epsilon_{22})$. If we study a part of a deformed body, for example a thin elastic plane sheet, by means of a variational principle, we may introduce the local density of free energy $\varphi(\epsilon) = \varphi(\epsilon_{ij} \,|\, i \leq j) = \varphi(\epsilon_{11}, \epsilon_{12}, \epsilon_{22})$ and vary the total free energy $\Phi = \int \varphi(\epsilon) \mathrm{d}x$ with $\mathrm{d}x = \mathrm{d}x^1 \wedge \mathrm{d}x^2$ by introducing $\sigma^{ij} = \partial \varphi / \partial \epsilon_{ij}$ for $i \leq j$ in order

to obtain $\delta\Phi=\int\left(\sigma^{11}\delta\epsilon_{11}+\sigma^{12}\delta\epsilon_{12}+\sigma^{22}\delta\epsilon_{22}\right)\mathrm{d}x$. Accordingly, *the "decision" to define the stress tensor* σ *by a symmetric matrix with* $\sigma^{12}=\sigma^{21}$ *is purely artificial within such a variational principle*. Indeed, the usual Cauchy device (1828) assumes that each element of a boundary surface is acted on by a surface density of force $\boldsymbol{\sigma}$ with a linear dependence $\boldsymbol{\sigma}=\left(\sigma^{ir}(x)n_r\right)$ on the outward normal unit vector $\boldsymbol{n}=(n_r)$ and does not make any assumption on the stress tensor. It is only by an equilibrium of forces and couples, namely the well known *phenomenological static torsor equilibrium*, that one can "*prove*" the symmetry of σ. However, even if we assume this symmetry, *we now need the different summation* $\sigma^{ij}\delta\epsilon_{ij}=\sigma^{11}\delta\epsilon_{11}+2\sigma^{12}\delta\epsilon_{12}+\sigma^{22}\delta\epsilon_{22}=\sigma^{ir}\partial_r\delta\xi_i$. An integration by parts and a change of sign produce the integral $\int\left(\partial_r\sigma^{ir}\right)\delta\xi_i\mathrm{d}x$ leading to the stress equations $\partial_r\sigma^{ir}=0$ already considered. This classical approach to elasticity theory, based on invariant theory with respect to the group of rigid motions, cannot therefore describe equilibrium of torsors by means of a variational principle where the proper torsor concept is totally lacking. It is however widely used through the technique of "*finite elements*" where it can also be applied to electromagnetism (EM) with similar quadratic (piezoelectricity) or cubic (photoelasticity) Lagrangian integrals. In this situation, the 4-potential *A* of EM is used in place of ξ while the EM field $\mathrm{d}A=F=(\boldsymbol{B},\boldsymbol{E})$ is used in place of ϵ and the Maxwell equations $\mathrm{d}F=0$ are used in place of the Riemann CC for ϵ.

However, there exists another equivalent procedure dealing with a *variational calculus with constraint*. Indeed, as we shall see later on, the deformation tensor is not any symmetric tensor as it must satisfy $n^2\left(n^2-1\right)/12$ compatibility conditions (CC), that is only $\partial_{22}\epsilon_{11}+\partial_{11}\epsilon_{22}-2\partial_{12}\epsilon_{12}=0$ when $n=2$. In this case, introducing the *Lagrange multiplier* λ, *we have to vary the new integral* $\int\left[\varphi(\epsilon)+\lambda\left(\partial_{22}\epsilon_{11}+\partial_{11}\epsilon_{22}-2\partial_{12}\epsilon_{12}\right)\right]\mathrm{d}x$ *for an arbitrary* ϵ. Setting $\lambda=-\phi$, a double integration by parts now provides the parametrization $\sigma^{11}=\partial_{22}\phi$, $\sigma^{12}=\sigma^{21}=-\partial_{12}\phi,\sigma^{22}=\partial_{11}\phi$ of the stress equations by means of the Airy function ϕ and the *formal adjoint* of the Riemann CC, *on the condition to observe that we have in fact* $2\sigma^{12}=-2\partial_{12}\phi$ as another way to understand the deep meaning of the factor "2" in the summation. The same variational calculus with constraint may thus also be used in order to "shortcut" the introduction of the EM potential.

Finally, using the *constitutive relations* of the material establishing an isomorphism $\sigma\leftrightarrow\epsilon$, one can also introduce a local free energy $\psi(\sigma)$ in a variational problem having now for constraint the stress equations, with the same comment as above (see [6], p. 915, for more details). The well known *Minkowski constitutive relations* $(\boldsymbol{B},\boldsymbol{E})\leftrightarrow(\boldsymbol{H},\boldsymbol{D})$ can be similarly used for EM.

In arbitrary dimension, the above compatibility conditions are nothing else but the linearized Riemann tensor in Riemannian geometry, a crucial mathematical tool in the theory of general relativity and a good reason for studying the work of Cornelius Lanczos (1893-1974) as it can be found in [21]-[23] or in a few modern references [24]-[31]. The starting point of Lanczos has been to take EM as a model in order to introduce a Lagrangian that should be quadratic in the Riemann tensor $\left(\rho^k_{l,ij}\Rightarrow\rho_{ij}=\rho^r_{i,rj}=\rho_{ji}\Rightarrow\rho=\omega^{ij}\rho_{ij}\right)$ while considering it independently of its expression through the second order derivatives of a metric (ω_{ij}) with inverse (ω^{ij}) or the first order derivatives of the corresponding Christoffel symbols (γ^k_{ij}). According to the previous paragraph, the corresponding variational calculus *must* involve PD constraints made by the Bianchi identities and the new Lagrangian to vary must therefore contain as many Lagrange multipliers as the number of Bianchi identities (*care*!) that can be written under the form:

$$\nabla_r\rho^k_{l,ij}+\nabla_i\rho^k_{l,jr}+\nabla_j\rho^k_{l,ri}=0\Rightarrow\nabla_r\rho^r_{l,ij}=\nabla_i\rho_{lj}-\nabla_j\rho_{li}$$

Meanwhile, Lanczos and followers have been looking for a kind of parametrization of the Bianchi identities, exactly like the Lagrange multiplier has been used as an Airy potential for the stress equations. However, we shall prove that the definition of a *Riemann candidate* and the answer to this question cannot be done without the knowledge of the Spencer cohomology. Moreover, we have pointed out the existence of well known couplings between elasticity and electromagnetism, namely piezoelectricity and photoelasticity, which are showing that, in the respective Lagrangians, the EM field is on equal footing with the deformation tensor and *not* with the Riemann tensor. This fact is showing the *shift by one step* that must be used in the physical inter-pretation of the differential sequences involved and cannot be avoided. Meanwhile, the *ordinary derivatives* ∂_i can be used in place of the *covariant derivatives* ∇_i when dealing with the linearized framework as the Christoffel symbols vanish when Euclidean or Minkowskian metrics are used.

The next tentative of Lanczos has been to extend his approach to the Weyl tensor:

$$\tau^k_{l,ij} = \rho^k_{l,ij} - \frac{1}{(n-2)}\left(\delta^k_i\rho_{lj} - \delta^k_j\rho_{li} + \omega_{lj}\omega^{ks}\rho_{si} - \omega_{li}\omega^{ks}\rho_{sj}\right) + \frac{1}{(n-1)(n-2)}\left(\delta^k_i\omega_{lj} - \delta^k_j\omega_{li}\right)\rho$$

The main problem is now that the Spencer cohomology of the symbols of the conformal Killing equations, in particular the 2-acyclicity, will be *absolutely needed* in order to study the Vessiot structure equations providing the Weyl tensor and its relation with the Riemann tensor. It will follow that *the CC for the Weyl tensor are not first order contrary to the CC for the Riemann tensor made by the Bianchi identities*, another reason for justifying the above shift by one step.

Finally, comparing the various parametrizations already obtained in the previous examples, it seems that the procedures are similar, even when dealing with systems having variable coefficients. The purpose of the paper is to prove that, in order to obtain a general algorithm, we shall need a lot of new tools involving at the same time *commutative algebra*, *homological algebra*, *differential algebra* and *differential geometry* that will be recalled in the next sections. Finally, like in any good crime story, it is only at the real end of the paper that we shall be able to revisit and compare all these examples in a unique framework.

2) MODULE THEORY

Before entering the heart of the next section dealing with extension modules, we need a few technical definitions and results from commutative algebra [13] [32].

DEFINITION 2.1: A *ring A* is said to be *unitary* if it has a (unique) element $1 \in A$ such that $1a = a1 = a$, $\forall a \in A$ and *commutative* if $ab = ba, \forall a, b \in A$. A non-zero element $a \in A$ is called a *zero-divisor* if one can find a non-zero $b \in A$ such that $ab = 0$ and a ring is called an *integral domain* if it has no zero-divisor. From now on, all rings considered will be unitary integral domains as we shall deal mainly with rings of partial differential operators.

DEFINITION 2.2: A ring *K* is called a *field* if every non-zero element $a \in K$ is a *unit*, that is one can find an element $b \in K$ such that $ab = 1 \in K$.

DEFINITION 2.3: A *module M* over a ring *A* or simply an *A-module* is a set of elements $x, y, z, \cdots$ which is an abelian group for an addition $(x, y) \to x + y$ with an action $A \times M \to M : (a, x) \to ax$ satisfying:

- $a(x + y) = ax + ay, \forall a \in A, \forall x, y \in M$
- $a(bx) = (ab)x, \forall a, b \in A, \forall x \in M$
- $(a + b)x = ax + bx, \forall a, b \in A, \forall x \in M$
- $1x = x, \forall x \in M$

The set of modules over a ring *A* will be denoted by $mod(A)$. A module over a field is called a *vector space*.

DEFINITION 2.4: A map $f : M \to N$ between two *A*-modules is called a *homomorphism* over *A* if $f(x + y) = f(x) + f(y), \forall x, y \in M$ and $f(ax) = af(x), \forall a \in A, \forall x \in M$. We successively define:

- $ker(f) = \{x \in M \mid f(x) = 0\}$
- $coim(f) = M/ker(f)$
- $im(f) = \{y \in N \mid \exists x \in M, f(x) = y\}$
- $coker(f) = N/im(f)$

with an isomorphism $coim(f) \simeq im(f)$ induced by *f*.

DEFINITION 2.5: We say that a chain of modules and homomorphisms is a *sequence* if the composition of two successive such homomorphisms is zero. A sequence is said to be *exact* if the kernel of each map is equal to the image of the map preceding it. An injective homomorphism is called a *monomorphism*, a surjective homomorphism is called an *epimorphism* and a bijective homomorphism is called an *isomorphism*. A short exact sequence is an exact sequence made by a monomorphism followed by an epimorphism.

PROPOSITION 2.6: If one has a short exact sequence:

$$0 \to M' \xrightarrow{f} M \xrightarrow{g} M'' \to 0$$

then the following conditions are equivalent:

- There exists an epimorphism $u : M \to M'$ such that $u \circ f = id_{M'}$ (*left inverse* of *f*).
- There exists a monomorphism $v : M'' \to M$ such that $g \circ v = id_{M''}$ (*right inverse* of *g*).

DEFINITION 2.7: In the above situation, we say that the short exact sequence *splits*. The relation $f \circ u + v \circ g = id_M$ provides an isomorphism $(u, g) : M \to M' \oplus M''$ with inverse $f + v : M' \oplus M'' \to M$.

The short exact sequence $0 \to \mathbb{Z} \to \mathbb{Q} \to \mathbb{Q}/\mathbb{Z} \to 0$ cannot split over $\mathbb{Z}$.

For the sake of clarity, as a few results will also be valid for modules over non-commutative rings, we shall denote by ${}_AM_B$ a *bimodule M* which is a left module for *A* with operation $(a,x) \to ax$ and a right module for *B* with operation $(x,b) \to xb$. In the commutative case, lower indices are not needed. If $M = {}_AM$ and $N = {}_AN$ are two left *A*-modules, the set of *A*-linear maps $f : M \to N$ will be denoted by $hom_A(M,N)$ or simply $hom(M,N)$ when there will be no confusion and there is a canonical isomorphism $hom(A,M) \simeq M : f \to f(1)$ with inverse $x \to (a \to ax)$. When *A* is commutative, $hom(M,N)$ is again an *A*-module for the law $(bf)(x) = f(bx)$. In the non-commutative case, things are much more complicate and we have:

LEMMA 2.8: Given ${}_AM$ and ${}_AN_B$, then $hom_A(M,N)$ becomes a right module over *B* for the law $(fb)(x) = f(x)b$. A similar result can be obtained with ${}_AM_B$ and ${}_AN$, where $hom_A(M,N)$ now becomes a left module over *B* for the law $(bf)(x) = f(xb)$.

THEOREM 2.9: If M, M', M'' are *A*-modules, the sequence:

$$M' \xrightarrow{f} M \xrightarrow{g} M'' \to 0$$

is exact if and only if the sequence:

$$0 \to hom(M'',N) \to hom(M,N) \to hom(M',N)$$

is exact for any *A*-module *N*.

COROLLARY 2.10: The short exact sequence:

$$0 \to M' \to M \to M'' \to 0$$

splits if and only if the short exact sequence:

$$0 \to hom(M'',N) \to hom(M,N) \to hom(M',N) \to 0$$

is exact for any module *N*.

DEFINITION 2.11: If *M* is a module over a ring *A*, a *system of generators* of *M* over *A* is a family $\{x_i\}_{i \in I}$ of elements of *M* such that any element of *M* can be written $x = \sum_{i \in I} a_i x_i$ with only a finite number of nonzero a_i. An *A*-module is called *noetherian* if every submodule of *M* (and thus *M* itself) is finitely generated.

One has the following standard technical result:

PROPOSITION 2.12: In a short exact sequence of modules, the central module is noetherian if and only if the two other modules are noetherian. As a byproduct, if *A* is a noetherian ring and *M* is a finitely generated module over *A*, then *M* is noetherian.

Accordingly, if *M* is generated by $\{x_1, \cdots, x_r\}$, there is an epimorphism $A^r \to M : (1,0,\cdots,0) \to x_1, \cdots, (0,\cdots,1) \to x_r$. The kernel of this epimorphism is thus also finitely generated, say by $\{y_1, \cdots, y_s\}$ and we therefore obtain the exact sequence $A^s \to A^r \to M \to 0$ that can be extended inductively to the left. *Such a property will always be assumed in the sequel.*

DEFINITION 2.13: In this case, we say that *M* is *finitely presented*.

We now turn to the definition and brief study of tensor products of modules over rings that will not be necessarily commutative unless stated explicitly.

Let $M = M_A$ be a right *A*-module and $N = {}_AN$ be a left *A*-module. We may introduce the free $\mathbb{Z}$-module made by finite formal linear combinations of elements of $M \times N$ with coefficients in $\mathbb{Z}$.

DEFINITION 2.14: The tensor product of *M* and *N* over *A* is the $\mathbb{Z}$-module $M \otimes {}_AN$ obtained by quotienting the above $\mathbb{Z}$-module by the submodule generated by the elements of the form:

$$(x+x',y)-(x,y)-(x',y),(x,y+y')-(x,y)-(x,y'),(xa,y)-(x,ay)$$

and the image of (x,y) will be denoted by $x \otimes y$.

It follows from the definition that we have the relations:

$$(x+x') \otimes y = x \otimes y + x' \otimes y, x \otimes (y+y') = x \otimes y + x \otimes y', xa \otimes y = x \otimes ay$$

and there is a canonical isomorphism $M \otimes_A A \simeq M, A \otimes_A N \simeq N$. When *A* is commutative, we may use left modules only and $M \otimes {}_AN$ becomes a left *A*-module.

EXAMPLE 2.15: If $A = \mathbb{Z}, M = \mathbb{Z}/2\mathbb{Z}$ and $N = \mathbb{Z}/3\mathbb{Z}$, we have $(\mathbb{Z}/2\mathbb{Z}) \otimes_{\mathbb{Z}} (\mathbb{Z}/3\mathbb{Z}) = 0$ because $x \otimes y = 3(x \otimes y) - 2(x \otimes y) = x \otimes 3y - 2x \otimes y = 0 - 0 = 0$.

We present the technique of *localization* in order to introduce rings and modules of fractions.

Definition 2.16: A subset S of a ring A is said to be *multiplicatively closed* if $\forall s,t \in S \Rightarrow st \in S$ and $1 \in S$. By a *left ring of fractions* or *left localization* of a noncommutative ring A with respect to a multiplicatively closed subset S of A, we mean a *ring* denoted by $S^{-1}A$ with a monomorphism $A \to S^{-1}A : a \to 1^{-1}a$ or simply a such that:

1) s is invertible in $S^{-1}A$, with inverse $s^{-1}1$ or simply $s^{-1}, \forall s \in S$.

2) Each element of $S^{-1}A$ or *fraction* has the form $s^{-1}a$ for some $s \in S, a \in A$.

We have to distinguish carefully $s^{-1}a$ from as^{-1} and we recover the standard notation a/s of the commutative case when two fractions a/s and b/t can be reduced to the same denominator $st = ts$. The following proposition is essential for constructing localizations.

Proposition 2.17: If there exists a left localization of A with respect to S, then we must have $Sa \cap As \neq \varnothing$, $\forall a \in A, \forall s \in S$. A set S satisfying this condition is called a *left Ore set*.

Proof: As $S^{-1}A$ *must* be a ring, the element as^{-1} in $S^{-1}A$ must be of the form $t^{-1}b$ for some $t \in S$, $b \in A$. Accordingly, $as^{-1} = t^{-1}b \Rightarrow ta = bs$ with $t \in S, b \in A$.

Q.E.D.

Lemma 2.18: If S is a left Ore set in a ring A, then $As \cap At \cap S \neq \varnothing, \forall s,t \in S$ and two fractions can be brought to the same denominator.

Proof: From the left Ore condition, we can find $u \in S$ and $a \in A$ such that $us = at \in S$. More generally, we can find $u, v \in A$ such that $us = vt \in S$ and we successively get:

$$(us)^{-1}(ua) = s^{-1}u^{-1}ua = s^{-1}a, \quad (vt)^{-1}(vb) = t^{-1}v^{-1}vb = t^{-1}b$$

so that the two fractions $s^{-1}a$ and $t^{-1}b$ can be brought to the same denominator $us = vt$.

Q.E.D.

Let us now define an equivalence relation on $S \times A$ by saying that $(s,a) \sim (t,b)$ if one can find $u, v \in S$ such that $us = vt \in S$ and $ua = vb$. Such a relation is clearly reflexive and symmetric, thus we only need to prove that it is transitive. So let $(s_1, a_1) \sim (s_2, a_2)$ and $(s_2, a_2) \sim (s_3, a_3)$. Then we can find $u_1, u_2 \in A$ such that $u_1 s_1 = u_2 s_2 \in S$ and $u_1 a_1 = u_2 a_2$. Also we can find $v_2, v_3 \in A$ such that $v_2 s_2 = v_3 s_3 \in S$ and $v_2 a_? = v_3 a_3$. Now, from the Ore condition, one can find $w_1, w_3 \in A$ such that $w_1 u_1 s_1 = w_3 v_3 s_3 \in S$ and thus $w_1 u_2 s_2 = w_3 v_2 s_2 \in S$, that is to say $(w_1 u_2 - w_3 v_2) s_2 = 0$. As A is an integral domain, we have $w_1 u_2 - w_3 v_2 = 0 \Rightarrow w_1 u_2 = w_3 v_2 \Rightarrow w_1 u_1 a_1 = w_1 u_2 a_2 = w_3 v_2 a_2 = w_3 v_3 a_3$ as wished. We finally define $S^{-1}A$ to be the quotient of $S \times A$ by the above equivalence relation with $\theta : A \to S^{-1}A : a \to 1^{-1}a$. The sum $(s,a) + (t,b)$ will be defined to be $(us = vt, ua + vb)$ and the product $(s,a) \times (t,b)$ will be defined to be $(s^{-1}a)(t^{-1}b) = s^{-1}(at^{-1})b = s^{-1}u^{-1}cb = (us)^{-1}(cb)$ whenever $ua = ct$.

A similar approach can be used in order to define and construct modules of fractions whenever S satisfies the two conditions of the last proposition. For this we need a preliminary lemma:

LEMMA 2.19: If S is a left Ore set in a ring A and M is a left module over A, the set:

$$t_S(M) = \{x \in M \mid \exists s \in S, sx = 0\}$$

is a submodule of M called the *S-torsion submodule* of M.

Proof: If $x, y \in t_S(M)$, we may find $s,t \in S$ such that $sx = 0, ty = 0$. Now, we can find $u, v \in A$ such that $us = vt \in S$ and we successively get $us(x+y) = usx + vty = 0 \Rightarrow x + y \in t_S(M)$. Also, $\forall a \in A$, using the Ore condition for S, we can find $b \in A, t \in S$ such that $ta = bs$ and we get $tax = bsx = 0 \Rightarrow ax \in t_S(M)$.

Q.E.D.

DEFINITION 2.20: By a *left module of fractions* or *left localization* of M with respect to S, we mean a left module $S^{-1}M$ over $S^{-1}A$ both with a homomorphism $\theta = \theta_S : M \to S^{-1}M : x \to 1^{-1}x = s^{-1}sx$ such that:

1) Each element of $S^{-1}M$ has the form $s^{-1}x$ for $s \in S, x \in M$.

2) $ker(\theta_S) = t_S(M)$.

In order to construct $S^{-1}M$, we shall define an equivalence relation on $S \times M$ by saying that $(s,x) \sim (t,y)$ if there exists $u, v \in A$ such that $us = vt \in S$ and $ux = vy$. The main property of localization is expressed by the following theorem:

Theorem 2.21: If one has an exact sequence:

$$M' \xrightarrow{f} M \xrightarrow{g} M''$$

then one also has the exact sequence:

$$S^{-1}M' \xrightarrow{S^{-1}f} S^{-1}M \xrightarrow{S^{-1}g} S^{-1}M''$$

where $S^{-1}f\left(s^{-1}x\right)=s^{-1}f\left(x\right)$.

As a link between tensor product and localization, we notice that the multiplication map $S^{-1}A\times M\to S^{-1}M$ given by $\left(s^{-1}a,x\right)\to s^{-1}ax$ induces an isomorphism $S^{-1}A\otimes_A M\to S^{-1}M$ of modules over $S^{-1}A$ when $S^{-1}A$ is considered as a right module over *A* with $\left(s^{-1}a\right)b=s^{-1}ab$ and *M* as a left module over *A*. In particular, when *A* is a commutative integral domain and $S=A-\{0\}$, the field $K=Q\left(A\right)=S^{-1}A$ is called the field of fractions of *A* and we have the short exact sequence:

$$0\to A\to K\to K/A\to 0$$

If now *M* is a left *A*-module, we may tensor this sequence by *M* on the right with $A\otimes M=M$ but we do not get in general an exact sequence. The defect of exactness *on the left* is nothing else but the *torsion submodule* $t\left(M\right)=\left\{x\in M\mid \exists 0\neq s\in A, sx=0\right\}\subseteq M$ and we have the long exact sequence:

$$0\to t\left(M\right)\to M\to K\otimes_A M\to K/A\otimes_A M\to 0$$

as we may describe the central map as follows:

$$x\to 1\otimes x=s^{-1}s\otimes x=s^{-1}\otimes sx,\quad \forall 0\neq s\in A$$

As we saw in the Introduction, such a result allows to understand why controllability has to do with localization which is introduced implicitly through the *transfer matrix* in control theory. In particular, a module *M* is said to be a *torsion module* if $t\left(M\right)=M$ and a *torsion-free module* if $t\left(M\right)=0$.

DEFINITION 2.22: A module in $mod\left(A\right)$ is called a *free module* if it has a *basis*, that is a system of generators linearly independent over *A*. When a module *F* is free, the number of generators in a basis, and thus in any basis, is called the *rank* of *F* over *A* and is denoted by $rank_A\left(F\right)$. In particular, if *F* is free of finite rank *r*, then $F\simeq A^r$. More generally, a module *P* is said to be *projective* if there exists another (projective) module *Q* such that $P\oplus Q\simeq F$ and any short exact sequence splits if it ends with a projective module (see [6], p. 638-644) for a formal test).

If *M* is any module over a ring *A* and *F* is a maximum free submodule of *M*, then $M/F=T$ is a torsion module. Indeed, if $x\in M, x\notin F$, then one can find $a\in A$ such that $ax\in F$ because, otherwise, $F\subset\left\{F,x\right\}$ should be free submodules of *M* with a strict inclusion. In that case, the *rank* of *M* is by definition the rank of *F* over *A*. When *A* is commutative, one has:

LEMMA 2.23: $rk_A\left(M\right)=dim_K\left(K\otimes_A M\right)$.

Proof: Taking the tensor product by *K* over *A* of the short exact sequence $0\to F\to M\to T\to 0$, we get an isomorphism $K\otimes_A F\simeq K\otimes_A M$ because $K\otimes_A T=0$ (exercise) and the lemma follows from the definition of the rank.

Q.E.D.

PROPOSITION 2.24: (*additivity property of the rank*) If $0\to M'\xrightarrow{f}M\xrightarrow{g}M''\to 0$ is a short exact sequence of modules over a ring *A*, then we have $rk_A\left(M\right)=rk_A\left(M'\right)+rk_A\left(M''\right)$.

Proof : Let us consider the following diagram with exact left/right columns and central row:

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & 0 & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0\to & & F' & \to & F'\oplus F'' & \to & F'' & \to 0 & \\
 & & \downarrow i' & & \downarrow i & & \downarrow i'' & & \\
0\to & & M' & \xrightarrow{f} & M & \xrightarrow{g} & M'' & \to 0 & \\
 & & \downarrow p' & & \downarrow p & & \downarrow p'' & & \\
0\to & & T' & \to & T & \to & T'' & \to 0 & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & 0 & & 0 & & 0 & &
\end{array}$$

where $F'(F'')$ is a maximum free submodule of $M'(M'')$ and $T' = M'/F'(T'' = M''/F'')$ is a torsion module. Pulling back by g the image under i'' of a basis of F'', we may obtain by linearity a map $\sigma : F'' \to M$ and we define $i = f \circ i' \circ \pi' + \sigma \circ \pi''$ where $\pi' : F' \oplus F'' \to F'$ and $\pi'' : F' \oplus F'' \to F''$ are the canonical projections on each factor of the direct sum. We have $i|_{F'} = f \circ i'$ and $g \circ i = g \circ \sigma \circ \pi'' = i'' \circ \pi''$. Hence, the diagram is commutative and thus exact with $rk_A(F' \oplus F'') = rk_A(F') + rk_A(F'')$ trivially. Finally, if T' and T'' are torsion modules, it is easy to check that T is a torsion module too and $F' \oplus F''$ is thus a maximum free submodule of M.

Q.E.D.

DEFINITION 2.25: If $f : M \to N$ is any morphism, the *rank* of f will be defined to be $rk_A(f) = rk_A(im(f))$.

We provide a few additional properties of the rank that will be used in the sequel. For this we shall set $M^* = hom_A(M, A)$ and, for any morphism $f : M \to N$ we shall denote by $f^* : N^* \to M^*$ the corresponding morphism which is such that $f^*(h) = h \circ f, \forall h \in hom_A(N, A)$.

PROPOSITION 2.26: When A is a commutative integral domain and M is a finitely presented module over A, then $rk_A(M) = rk_A(M^*)$.

Proof: Applying $hom_A(\bullet, A)$ to the short exact sequence in the proof of the preceding lemma while taking into account $T^* = 0$, we get a monomorphism $0 \to M^* \to F^*$ and obtain therefore $rk_A(M^*) \leq rk_A(F^*)$. However, as $F \simeq A^r$ with $r < \infty$ because M is finitely generated, we get $F^* \simeq A^r$ too because $A^* \simeq A$. It follows that $rk_A(M^*) \leq rk_A(F^*) = rk_A(F) = rk_A(M)$ and thus $rk_A(M^*) \leq rk_A(M)$.

Now, if $F_1 \xrightarrow{d_1} F_0 \to M \to 0$ is a finite presentation of M, applying $hom_A(\bullet, A)$ to this presentation, we get the ker/coker exact sequence:

$$0 \leftarrow N \leftarrow F_1^* \xleftarrow{d_1^*} F_0^* \leftarrow M^* \leftarrow 0$$

Applying $hom_A(\bullet, A)$ to this sequence while taking into account the isomorphisms $F_0^{**} \simeq F_0, F_1^{**} \simeq F_1$, we get the ker/coker exact sequence:

$$0 \to N^* \to F_1 \xrightarrow{d_1} F_0 \to M \to 0$$

Counting the ranks, we obtain:

$$rk_A(N) - rk_A(M^*) = rk_A(F_1^*) - rk_A(F_0^*) = rk_A(F_1) - rk_A(F_0) = rk_A(N^*) - rk_A(M)$$

and thus:

$$\left(rk_A(M) - rk_A(M^*)\right) + \left(rk_A(N) - rk_A(N^*)\right) = 0$$

As both two numbers in this sum are non-negative, they must be zero and we finally get $rk_A(M) = rk_A(M^*), rk_A(N) = rk_A(N^*)$.

Q.E.D.

COROLLARY 2.27: Under the condition of the proposition, we have $rk_A(f) = rk_A(f^*)$.

Proof: Introducing the *ker* / *coker* exact sequence:

$$0 \to K \to M \xrightarrow{f} N \to Q \to 0$$

we have: $rk_A(f) + rk_A(Q) = rk_A(N)$. Applying $hom_A(\bullet, A)$ and taking into account Theorem 2.9, we have the exact sequence:

$$0 \to Q^* \to N^* \xrightarrow{f^*} M^*$$

and thus: $rk_A(f^*) + rk_A(Q^*) = rk_A(N^*)$. Using the preceding proposition, we get $rk_A(Q) = rk_A(Q^*)$ and $rk_A(N) = rk_A(N^*)$, that is to say $rk_A(f) = rk_A(f^*)$.

Q.E.D.

3) HOMOLOGICAL ALGEBRA

We need a few definitions and results from homological algebra [13] [33] [34] and start recalling the well known Cramer's rule for linear systems through the exactness of the ker/coker sequence for modules when $f : M \to N$ is a linear map (homomorphism):

$$0 \to ker(f) \to M \xrightarrow{f} N \to coker(f) \to 0$$

In the case of vector spaces over a field K, we successively have $rk(f) = dim(im(f))$, $dim(ker(f)) = dim(M) - rk(f)$, $dim(coker(f)) = dim(N) - rk(f) = nb$ of compatibility conditions, and obtain by substraction:

$$dim(ker(f)) - dim(M) + dim(N) - dim(coker(f)) = 0$$

In the case of modules, we may replace the dimension by the rank and obtain the same relations because of the additive property of the rank. We may also define *cohomology theory* as follows:

DEFINITION 3.1: If one has a sequence $L \xrightarrow{f} M \xrightarrow{g} N$, that is if $g \circ f = 0$, then one may introduce the submodules $coboundary = B = im(f) \subseteq ker(g) = cocycle = Z \subseteq M$ and define the cohomology at M to be the quotient $H = Z/B$.

We now introduce the *extension modules* in an elementary manner, using the standard notation $hom_A(M, A) = M^*$. Using a *free resolution* of an A-module M, that is to say a long exact sequence:

$$\cdots \xrightarrow{d_2} F_1 \xrightarrow{d_1} F_0 \to M \to 0$$

where $F_0, F_1, \cdots$ are free modules, namely modules isomorphic to powers of A and $M = coker(d_1) = F_0/im(d_1)$. We may *take out* M and obtain the *deleted sequence*:

$$\cdots \xrightarrow{d_2} F_1 \xrightarrow{d_1} F_0 \to 0$$

which is of course no longer exact. We may apply the functor $hom_A(\bullet, A)$ and obtain the sequence:

$$\cdots \xleftarrow{d_2^*} F_1^* \xleftarrow{d_1^*} F_0^* \leftarrow 0$$

in order to state:

DEFINITION 3.2: We set:

$$ext^0(M) = ext_A^0(M, A) = ker(d_1^*) = M^*,$$
$$ext^i(M) = ext_A^i(M, A) = ker(d_{i+1}^*)/im(d_i^*), \forall i \geq 1$$

The extension modules have the following three main properties [6] [13] [33] [34]:

PROPOSITION 3.3: The extension modules do not depend on the resolution of M chosen.

PROPOSITION 3.4: If $0 \to M' \to M \to M'' \to 0$ is a short exact sequence of A-modules, then we have the following *connecting long exact sequence*:

$$0 \to M''^* \to M^* \to M'^* \to ext^1(M'') \to ext^1(M) \to ext^1(M') \to ext^2(M'') \to ext^2(M) \to \cdots$$

of extension modules. Moreover $ext^i(P) = 0, \forall i \geq 1$ whenever P is a projective module.

PROPOSITION 3.5: $ext^i(M)$ is a torsion module, $\forall i \geq 1$.

Proof: Having in mind that $B_i = im(d_i^*)$ and $Z_i = ker(d_{i+1}^*)$, we obtain $rk(B_i) = rk(d_i^*) = rk(d_i)$ and $rk(Z_i) = rk(F_i^*) - rk(d_{i+1}^*) = rk(F_i) - rk(d_{i+1})$. However, we started from a resolution, that is an exact sequence in which $rk(d_i) + rk(d_{i+1}) = rk(F_i)$. It follows that $rk(B_i) = rk(Z_i)$ and thus $rk(H_i) = rk(Z_i) - rk(B_i) = 0$, that is to say $ext^i(M)$ is a torsion module for $i \geq 1$, $\forall M \in mod(A)$.

Q.E.D.

The next theorem and its corollary constitute the main results that will be used for applications through a classification of modules [1] [2] [6]-[8] [34] [35]:

THEOREM 3.6: The following long exact sequence:

$$0 \to ext^1(N) \to M \xrightarrow{\epsilon} M^{**} \to ext^2(N) \to 0$$

is isomorphic to the *ker/coker* long exact sequence for the central morphism ϵ which is defined by $\epsilon(x)(f) = f(x), \forall x \in M, \forall f \in M^*$.

Proof: Introducing $K = im(d_1^*)$, we may obtain two short exact sequences, a left one starting with K and a right one finishing with K as follows:

$$\begin{array}{ccccccccc} & & & & \overset{d_1^*}{\leftarrow} & & & & \\ 0 \leftarrow & N & \leftarrow & F_1^* & & F_0^* & \leftarrow & M^* & \leftarrow 0 \\ & & & \nwarrow & & \swarrow & & & \\ & & & & K & & & & \\ & & & \swarrow & & \nwarrow & & & \\ & & & 0 & & 0 & & & \end{array}$$

Using the two corresponding long exact connecting sequences, we get $ext^1(K) \simeq ext^2(N)$ from the one starting with N^* which is also providing the left exact column of the next diagram and the exact central row of this diagram from the one starting with K^*. The Theorem is finally obtained by a chase proving that the full diagram is commutative and exact:

$$\begin{array}{ccccccccc} & & F_1 & = & F_1 & & & & \\ & & \downarrow & & \downarrow d_1 & & & & \\ 0 \rightarrow & K^* & \overset{d_1}{\rightarrow} & F_0 & \rightarrow & M^{**} & \rightarrow & ext^2(N) & \rightarrow 0 \\ & \downarrow & & \downarrow & \nearrow & & & & \\ 0 \rightarrow & ext^1(N) & \rightarrow & M & & & & & \\ & \downarrow & & \downarrow & & & & & \\ & 0 & & 0 & & & & & \end{array}$$

Q.E.D.

COROLLARY 3.7: $t(M) = ext^1(N) = ker(\epsilon)$.

Proof: As $ext^1(N) \subseteq M$ is a torsion module, we have therefore $ext^1(N) \subseteq t(M)$. Now, if $x \in t(M)$, we may find $0 \neq a \in A$ such that $ax = 0$ and $\epsilon(x)(f) = f(x) \Rightarrow f(ax) = af(x) = 0 \Rightarrow f(x) = 0$ because A is an integral domain, that is $t(M) \subseteq ker(\epsilon) = ext^1(N)$ and thus $t(M) = ext^1(N) = ker(\epsilon)$.

Q.E.D.

DEFINITION 3.8: A module M will be called *torsion-free* if $ext^1(N) = t(M) = 0$ and *reflexive* if $ext^1(N) = 0, ext^2(N) = 0$.

Despite all these results, a major difficulty still remains. Indeed, we have $M = coker(d_1) = {}_AM$ as a left module over A but, using the bimodule structure of $A = {}_AA_A$ and Lemma 2.13, it follows that $M^* = ker(d_1^*) = M_A^*$ is a right module over A and thus $N = coker(d_1^*) = N_A$ is also a right module over A. However, as we shall see, all the differential modules used through applications will be left modules over the ring of differential operators and it will therefore not be possible to use dual sequences as we did without being able to "*pass from left to right and vice-versa*". For this purpose we now need many delicate results from differential geometry, in particular a way to deal with the *formal adjoint* of an operator as we did many times in the Introduction.

4) SYSTEM THEORY

If E is a vector bundle over the base manifold X with projection π and local coordinates $(x, y) = (x^i, y^k)$ projecting onto $x = (x^i)$ for $i = 1, \cdots, n$ and $k = 1, \cdots, m$, identifying a map with its graph, a (local) section $f : U \subset X \rightarrow E$ is such that $\pi \circ f = id$ on U and we write $y^k = f^k(x)$ or simply $y = f(x)$. For any change of local coordinates $(x, y) \rightarrow (\overline{x} = \varphi(x), \overline{y} = A(x)y)$ on E, the change of section is $y = f(x) \rightarrow \overline{y} = \overline{f}(\overline{x})$ such that $\overline{f}^l(\varphi(x)) \equiv A_k^l(x) f^k(x)$. The new vector bundle E^* obtained by changing the *transition matrix* A to its inverse A^{-1} is called the *dual vector bundle* of E. Differentiating with respect to x^i and using new coordinates y_i^k in place of $\partial_i f^k(x)$, we obtain $\overline{y}_r^l \partial_i \varphi^r(x) = A_k^l(x) y_i^k + \partial_i A_k^l(x) y^k$. Introducing a multi-index $\mu = (\mu_1, \cdots, \mu_n)$ with length $|\mu| = \mu_1 + \cdots + \mu_n$ and prolonging the procedure up to order q, we may construct in this way, by patching coordinates, a vector bundle $J_q(E)$ over X, called the *jet bundle of order* q with local coordinates $(x, y_q) = (x^i, y_\mu^k)$ with $0 \leq |\mu| \leq q$ and $y_0^k = y^k$. We have therefore epimorphisms $\pi_q^{q+r} : J_{q+r}(E) \rightarrow J_q(E), \forall q, r \geq 0$. For a later use, we shall set $\mu + 1_i = (\mu_1, \cdots, \mu_{i-1}, \mu_i + 1, \mu_{i+1}, \cdots, \mu_n)$ and define the operator $j_q : E \rightarrow J_q(E) : f \rightarrow j_q(f)$ on sections by the local formula $j_q(f) : (x) \rightarrow (\partial_\mu f^k(x) | 0 \leq |\mu| \leq q, k = 1, \cdots, m)$. Moreover, a jet coordinate y_μ^k is said to be of *class* i if $\mu_1 = \cdots = \mu_{i-1} = 0$, $\mu_i \neq 0$. We finally introduce the *Spencer operator* $D : J_{q+1}(E) \rightarrow T^* \otimes J_q(E) : f_{q+1} \rightarrow j_1(f_q) - f_{q+1}$ with

$\left(Df_{q+1}\right)^k_{\mu,i} = \partial_i f^k_\mu - f^k_{\mu+1_i}$.

DEFINITION 4.1: A *system* of PD equations of order q on E is a vector subbundle $R_q \subset J_q(E)$ locally defined by a constant rank system of linear equations for the jets of order q of the form $a^{\tau\mu}_k(x)\, y^k_\mu = 0$. Its *first prolongation* $R_{q+1} \subset J_{q+1}(E)$ will be defined by the equations $a^{\tau\mu}_k(x)\, y^k_\mu = 0, a^{\tau\mu}_k(x)\, y^k_{\mu+1_i} + \partial_i a^{\tau\mu}_k(x)\, y^k_\mu = 0$ which may not provide a system of constant rank as can easily be seen for $xy_x - y = 0 \Rightarrow xy_{xx} = 0$ where the rank drops at $x = 0$.

The next definition of *formal integrability* (FI) will be crucial for our purpose.

DEFINITION 4.2: A system R_q is said to be *formally integrable* if the R_{q+r} are vector bundles $\forall r \geq 0$ (regularity condition) and no new equation of order $q+r$ can be obtained by prolonging the given PD equations more than r times, $\forall r \geq 0$ or, equivalently, we have induced epimorphisms $\pi^{q+r+1}_{q+r} : R_{q+r+1} \to R_{q+r}$, $\forall r \geq 0$ allowing to compute "*step by step*" formal power series solutions.

A formal test has been first sketched by C. Riquier in 1910 [36], then improved by M. Janet in 1920 [20] [37] and by E. Cartan in 1945 [38], finally rediscovered in 1965, totally independently, by B. Buchberger who introduced Gröbner bases, using the name of his thesis advisor. However all these tentatives have been largely superseded and achieved in an intrinsic way, again totally independently of the previous approaches, by D.C. Spencer in 1965 [20] [39] [40].

DEFINITION 4.3: The family g_{q+r} of vector spaces over X defined by the purely linear equations $a^{\tau\mu}_k(x)\, v^k_{\mu+\nu} = 0$ for $|\mu| = q, |\nu| = r$ is called the *symbol* at order $q+r$ and only depends on g_q.

The following procedure, *where one may have to change linearly the independent variables if necessary*, is the key towards the next definition which is intrinsic even though it must be checked in a particular coordinate system called *δ-regular* (see [6] [20] and [39] for more details):

- *Equations of class n*: Solve the maximum number β^n_q of equations with respect to the jets of order q and class n. Then call $\left(x^1, \cdots, x^n\right)$ *multiplicative variables*.

– – – – – – – – – – – – – – – –

- *Equations of class i*: Solve the maximum number of *remaining* equations with respect to the jets of order q and class i. Then call $\left(x^1, \cdots, x^i\right)$ *multiplicative variables* and $\left(x^{i+1}, \cdots, x^n\right)$ *non-multiplicative variables*.

– – – – – – – – – – – – – – – –

- *Remaining equations of order* $\leq q-1$: Call $\left(x^1, \cdots, x^n\right)$ *non-multiplicative variables*.

DEFINITION 4.4: The above multiplicative and non-multiplicative variables can be visualized respectively by integers and dots in the corresponding *Janet board*. A system of PD equations is said to be *involutive* if its first prolongation can be achieved by prolonging its equations only with respect to the corresponding multiplicative variables. The following numbers are called *characters*:

$$\alpha^i_q = m(q+n-i-1)!/\left((q-1)!(n-i)!\right) - \beta^i_q, \quad \forall 1 \leq i \leq n \quad \Rightarrow \quad \alpha^1_q \geq \cdots \geq \alpha^n_q$$

For an involutive system, $\left(y^{\beta^n_q+1}, \cdots, y^m\right)$ can be given arbitrarily.

For an involutive system of order q in the above *solved form*, we shall use to denote by y_{pri} the *principal jet coordinates*, namely the leading terms of the solved equations in the sense of involution. Accordingly, any formal derivative of a principal jet coordinate is again a principal jet coordinate. The remaining jet coordinates will be called *parametric jet coordinates* and denoted by y_{par}. Now, the symbol of j_q is the zero symbol and is thus trivially involutive at any order q. Accordingly, if we introduce the *multiplicative variables* $x^1, \cdots, x^i$ for the parametric jets of order q and class i, the formal derivative or a parametric jet of strict order q and class i by one of its multiplicative variables is uniquely obtained and cannot be a principal jet of order $q+1$ which is coming from a uniquely defined principal jet of order q and class i.

PROPOSITION 4.5: Using the Janet board and the definition of involutivity, we get:

$$dim\left(g_{q+r}\right) = \sum_{i=1}^n \frac{(r+i-1)!}{r!(i-1)!}\alpha^i_q \Rightarrow dim\left(R_{q+r}\right) = dim\left(R_{q-1}\right) + \sum_{i=1}^n \frac{(r+i)!}{r!i!}\alpha^i_q$$

Let T be the tangent vector bundle of vector fields on X, T^* be the cotangent vector bundle of 1-forms on X and $\wedge^s T^*$ be the vector bundle of s-forms on X with usual bases $\left\{dx^I = dx^{i_1} \wedge \cdots \wedge dx^{i_s}\right\}$ where we have set

$I=(i_1<\cdots<i_s)$. Also, let S_qT^* be the vector bundle of symmetric q-covariant tensors. Moreover, if $\xi,\eta\in T$ are two vector fields on *X*, we may define their *bracket* $[\xi,\eta]\in T$ by the local formula $([\xi,\eta])^i(x)=\xi^r(x)\partial_r\eta^i(x)-\eta^s(x)\partial_s\xi^i(x)$ leading to the *Jacobi identity* $[\xi,[\eta,\zeta]]+[\eta,[\zeta,\xi]]+[\zeta,[\xi,\eta]]=0,\forall\xi,\eta,\zeta\in T$. We may finally introduce the *exterior derivative* $d:\wedge^rT^*\to\wedge^{r+1}T^*:\omega=\omega_I dx^I\to d\omega=\partial_i\omega_I dx^i\wedge dx^I$ with $d^2=d\circ d\equiv 0$ in the *Poincaré sequence*:

$$\wedge^0T^*\xrightarrow{d}\wedge^1T^*\xrightarrow{d}\wedge^2T^*\xrightarrow{d}\cdots\xrightarrow{d}\wedge^nT^*\to 0$$

In a purely algebraic setting, one has [20] [39] [40]:

PROPOSITION 4.6: There exists a map $\delta:\wedge^sT^*\otimes S_{q+1}T^*\otimes E\to\wedge^{s+1}T^*\otimes S_qT^*\otimes E$ which restricts to $\delta:\wedge^sT^*\otimes g_{q+1}\to\wedge^{s+1}T^*\otimes g_q$ and $\delta^2=\delta\circ\delta=0$.

Proof: Let us introduce the family of s-forms $\omega=\{\omega^k_\mu=v^k_{\mu,I}dx^I\}$ and set $(\delta\omega)^k_\mu=dx^i\wedge\omega^k_{\mu+1_i}$. We obtain at once $(\delta^2\omega)^k_\mu=dx^i\wedge dx^j\wedge\omega^k_{\mu+1_i+1_j}=0$.

Q.E.D.

The kernel of each δ in the first case is equal to the image of the preceding δ but this may no longer be true in the restricted case and we set (see [39], p. 85-88 for more details):

DEFINITION 4.7: We denote by $B^s_{q+r}(g_q)\subseteq Z^s_{q+r}(g_q)$ and $H^s_{q+r}(g_q)=Z^s_{q+r}(g_q)/B^s_{q+r}(g_q)$ respectively the coboundary space, cocycle space and cohomology space at $\wedge^sT^*\otimes g_{q+r}$ of the restricted δ-sequence which only depend on g_q and may not be vector bundles. The symbol g_q is said to be *s-acyclic* if $H^1_{q+r}=\cdots=H^s_{q+r}=0,\forall r\geq 0$, *involutive* if it is n-acyclic and *finite type* if $g_{q+r}=0$ becomes trivially involutive for r large enough. For a later use, we notice that a symbol g_q is involutive *and* of finite type if and only if $g_q=0$. Finally, $S_qT^*\otimes E$ is involutive $\forall q\geq 0$ if we set $S_0T^*\otimes E=E$.

FI CRITERION 4.8: If $\pi^{q+1}_q:R_{q+1}\to R_q$ is an epimorphism of vector bundles and g_q is 2-acyclic (involutive), then R_q is formally integrable (involutive).

EXAMPLE 4.9: The system R_2 defined by the three PD equations

$$y_{33}=0,\quad y_{23}-y_{11}=0,\quad y_{22}=0$$

is homogeneous and thus automatically formally integrable but g_2 is not involutive though finite type because $g_4=0$. Elementary computations of ranks of matrices show that the δ-map:

$$0\to\wedge^2T^*\otimes g_3\xrightarrow{\delta}\wedge^3T^*\otimes g_2\to 0$$

is a 3×3 isomorphism and thus g_3 is 2-acyclic with $dim(g_3)=1$, a *crucial intrinsic* property totally absent from any "old" work and quite more easy to handle than its Koszul dual.

The main use of involution is to construct differential sequences that are made up by successive *compatibility conditions* (CC) of order one. In particular, when R_q is involutive, the differential operator $\mathcal{D}:E\xrightarrow{j_q}J_q(E)\xrightarrow{\Phi}J_q(E)/R_q=F_0$ of order *q* with space of solutions $\Theta\subset E$ is said to be *involutive* and one has the canonical *linear Janet sequence* ([31], p. 144):

$$0\to\Theta\to E\xrightarrow{\mathcal{D}}F_0\xrightarrow{\mathcal{D}_1}F_1\xrightarrow{\mathcal{D}_2}\cdots\xrightarrow{\mathcal{D}_n}F_n\to 0$$

where each other operator is first order involutive and generates the CC of the preceding one with the *Janet bundles* $F_r=\wedge^rT^*\otimes J_q(E)/\left(\wedge^rT^*\otimes R_q+\delta\left(\wedge^{r-1}T^*\otimes S_{q+1}T^*\otimes E\right)\right)$. As the Janet sequence can be "cut at any place", that is can also be constructed anew from any intermediate operator, *the numbering of the Janet bundles has nothing to do with that of the Poincaré sequence for the exterior derivative*, contrary to what many physicists still believe ($n=3$ with $\mathcal{D}=div$ provides the simplest example). Moreover, the fiber dimension of the Janet bundles can be computed at once inductively from the board of multiplicative and non-multiplicative variables that can be exhibited for $\mathcal{D}$ by working out the board for $\mathcal{D}_1$ and so on. For this, the number of rows of this new board is the number of dots appearing in the initial board while the number $nb(i)$ of dots in the column *i* just indicates the number of CC of class *i* for $i=1,\cdots,n$ with $nb(i)<nb(j),\forall i<j$. When R_q is *not* involutive but formally integrable and the *r*-prolongation of its symbol g_q becomes 2-acyclic, it is known that the generating CC are of order $r+1$ (see [39], Example 6, p. 120 and previous Example).

DEFINITION 4.10: More generally, a differential sequence is said to be *formally exact* if each operator generates the CC of the operator preceding it.

EXAMPLE 4.11: ([41], §38, p 40, is providing the first intuition of formal integrability) The second order

system $y_{11}=0, y_{13}-y_2=0$ is neither formally integrable nor involutive. Indeed, we get $d_3y_{11}-d_1(y_{13}-y_2)=y_{12}$ and $d_3y_{12}-d_2(y_{13}-y_2)=y_{22}$, that is to say *each first and second* prolongation does bring a new second order PD equation. Considering the new system $y_{22}=0, y_{12}=0, y_{13}-y_2=0, y_{11}=0$, the question is to decide whether this system is involutive or not. In such a simple situation, as there is no PD equation of class 3, the evident permutation of coordinates $(1,2,3)\to(3,2,1)$ provides the following involutive second order system with one equation of class 3, 2 equations of class 2 and 1 equation of class 1:

$$\left\{\begin{array}{ll}\Phi^4\equiv y_{33}=0 & \boxed{\begin{array}{ccc}1&2&3\\1&2&\bullet\\1&2&\bullet\\1&\bullet&\bullet\end{array}}\\ \Phi^3\equiv y_{23}=0 & \\ \Phi^2\equiv y_{22}=0 & \\ \Phi^1\equiv y_{13}-y_2=0 & \end{array}\right.$$

We have $\alpha_2^3=0,\alpha_2^2=0,\alpha_2^1=2$ and the corresponding CC system is easily seen to be the following involutive first order system:

$$\left\{\begin{array}{ll}\Psi^4\equiv d_3\Phi^3-d_2\Phi^4=0 & \boxed{\begin{array}{ccc}1&2&3\\1&2&3\\1&2&3\\1&2&\bullet\end{array}}\\ \Psi^3\equiv d_3\Phi^2-d_2\Phi^3=0 & \\ \Psi^2\equiv d_3\Phi^1-d_1\Phi^4+\Phi^3=0 & \\ \Psi^1\equiv d_2\Phi^1-d_1\Phi^3+\Phi^2=0 & \end{array}\right.$$

The final CC system is the involutive first order system:

$$\left\{\Omega\equiv d_3\Psi^1-d_2\Psi^2+d_1\Psi^4-\Psi^3=0\ \boxed{1\ \ 2\ \ 3}\right.$$

We get therefore the (formally exact) Janet sequence:

$$0\to\Theta\to 1\to 4\to 4\to 1\to 0$$

However, keeping only Φ^1 and Φ^4 while using the fact that d_{33} commutes with $d_{13}-d_2$, we get the formally exact sequence $0\to\Theta\to 1\to 2\to 1\to 0$ which is *not* a Janet sequence. We finally check that each Φ^1,Φ^2,Φ^3 is separately differentially dependent on Φ^4 because we have successively $d_3\Phi^3-d_2\Phi^4=0$, $d_{33}\Phi^2-d_{22}\Phi^4=0, d_{33}\Phi^1-d_{13}\Phi^4+d_2\Phi^4=0$.

Finally, we may extend the restriction $D:R_{q+1}\to T^*\otimes R_q$ of the Spencer operator to:

$$D:\wedge^rT^*\otimes R_{q+1}\to\wedge^{r+1}T^*\otimes R_q:\alpha\otimes f_{q+1}\to d\alpha\otimes f_q+(-1)^r\,\alpha\wedge Df_{q+1}\Rightarrow D^2=D\circ D\equiv 0$$

in order to construct the *first Spencer sequence* which is another resolution of Θ because the kernel of the first D is such that $f_{q+1}\in R_{q+1}, Df_{q+1}=0\Leftrightarrow f_{q+1}=j_{q+1}(f), f\in\Theta$ when q is large enough.

5) DIFFERENTIAL MODULES

Let K be a *differential field*, that is a field containing $\mathbb{Q}$ with n commuting *derivations* $\{\partial_1,\cdots,\partial_n\}$ with $\partial_i\partial_j=\partial_j\partial_i=\partial_{ij},\forall i,j=1,\cdots,n$ such that $\partial_i(a+b)=\partial_ia+\partial_ib$, $\partial_i(ab)=(\partial_ia)b+a\partial_ib,\forall a,b\in K$ and $\partial_i(1/a)=-(1/a^2)\partial_ia,\forall a\in K$. Using an implicit summation on multi-indices, we may introduce the (noncommutative) *ring of differential operators* $D=K[d_1,\cdots,d_n]=K[d]$ with elements $P=a^\mu d_\mu$ such that $|\mu|<\infty$ and $d_ia=ad_i+\partial_ia$. The highest value of $|\mu|$ with $a^\mu\neq 0$ is called the *order* of the *operator P* and the ring D with multiplication $(P,Q)\to P\circ Q=PQ$ is filtred by the order q of the operators. We have the *filtration* $0\subset K=D_0\subset D_1\subset\cdots\subset D_q\subset\cdots\subset D_\infty=D$. Moreover, it is clear that D, as an algebra, is generated by $K=D_0$ and $T=D_1/D_0$ with $D_1=K\oplus T$ if we identify an element $\xi=\xi^id_i\in T$ with the vector field $\xi=\xi^i(x)\partial_i$ of differential geometry, but with $\xi^i\in K$ now. It follows that $D={}_DD_D$ is a *bimodule* over itself, being at the same time a left D-module ${}_DD$ by the composition $P\to QP$ and a right D-module D_D by the composition $P\to PQ$ with $D_rD_s=D_{r+s},\forall r,s\geq 0$.

If we introduce *differential indeterminates* $y=(y^1,\cdots,y^m)$, we may extend $d_iy_\mu^k=y_{\mu+1_i}^k$ to $\Phi^\tau\equiv a_k^{\tau\mu}y_\mu^k\xrightarrow{d_i}d_i\Phi^\tau\equiv a_k^{\tau\mu}y_{\mu+1_i}^k+\partial_ia_k^{\tau\mu}y_\mu^k$ for $\tau=1,\cdots,p$. Therefore, setting $Dy^1+\cdots+Dy^m=Dy\simeq D^m$ and calling $I=D\Phi\subset Dy$ the *differential module of equations*, we obtain by residue the *differential module* or *D-module* $M=Dy/D\Phi$, denoting the residue of y_μ^k by $\overline{y}_\mu^k$ when there can be a confusion. Introducing the two free differential modules $F_0\simeq D^{m_0}, F_1\simeq D^{m_1}$, we obtain equivalently the *free presentation* $F_1\xrightarrow{d_1}F_0\to M\to 0$ of order q when $m_0=m, m_1=p$ and $d_1=D$. We shall moreover assume that $\mathcal{D}$ provides a

strict morphism, namely that the corresponding system R_q is formally integrable. It follows that *M* can be endowed with a *quotient filtration* obtained from that of D^m which is defined by the order of the jet coordinates y_q in $D_q y$. We have therefore the *inductive limit* $0 = M_{-1} \subseteq M_0 \subseteq M_1 \subseteq \cdots \subseteq M_q \subseteq \cdots \subseteq M_\infty = M$ with $d_i M_q \subseteq M_{q+1}$ but it is important to notice that $D_r D_q = D_{q+r} \Rightarrow D_r M_q = M_{q+r}, \forall q, r \geq 0 \Rightarrow M = DM_q, \quad \forall q \geq 0$ *in this particular case*. It also follows from Noetherian arguments and involution that $D_r I_q = I_{q+r}, \quad \forall r \geq 0$ though we have in general only $D_r I_s \subseteq I_{r+s}, \forall r \geq 0, \forall s < q$. As $K \subset D$, we may introduce the *forgetful functor* $for : mod(D) \to mod(K) : {}_D M \to {}_K M$.

More generally, introducing the successive CC as in the preceding section while changing slightly the numbering of the respective operators, we may finally obtain the *free resolution* of *M*, namely the exact sequence $\cdots \xrightarrow{d_3} F_2 \xrightarrow{d_2} F_1 \xrightarrow{d_1} F_0 \to M \to 0$. In actual practice, *one must never forget that* $\mathcal{D} = \Phi \circ j_q$ *acts on the left on column vectors in the operator case and on the right on row vectors in the module case*. Also, with a slight abuse of language, when $\mathcal{D} = \Phi \circ j_q$ is involutive as in Section 2 and thus $R_q = ker(\Phi)$ is involutive, one should say that *M* has an *involutive presentation* of order *q* or that M_q is *involutive*.

DEFINITION 5.1: Setting $P = a^\mu d_\mu \in D \xleftrightarrow{ad} ad(P) = (-1)^{|\mu|} d_\mu a^\mu \in D$, we have $ad(ad(P)) = P$ and $ad(PQ) = ad(Q)\, ad(P), \forall P, Q \in D$. Such a definition can be extended to any matrix of operators by using the transposed matrix of adjoint operators and we get:

$$\langle \lambda, \mathcal{D}\xi \rangle = \langle ad(\mathcal{D})\lambda, \xi \rangle + div(...)$$

from integration by part, where λ is a row vector of test functions and $\langle \, \rangle$ the usual contraction. We quote the useful formulas $[ad(\xi), ad(\eta)] = ad(\xi)\, ad(\eta) - ad(\eta)\, ad(\xi) = -ad([\xi, \eta]), \forall \xi, \eta \in T$ (*care about the minus sign*) and $rk_D(\mathcal{D}) = rk_D(ad(\mathcal{D}))$ as in ([32], p. 610-612).

REMARK 5.2: As can be seen from the examples of the Introduction, when $\mathcal{D}$ is involutive, then $ad(\mathcal{D})$ may not be involutive. In the differential framework, we may set $rk_D(\mathcal{D}) = m - \alpha_q^n = \beta_q^n$. Comparing to similar concepts used in *differential algebra*, this number is just the maximum number of differentially independent equations to be found in the differential module *I* of equations. Indeed, pointing out that differential indeter- minates in differential algebra are nothing else than jet coordinates in differential geometry and using standard notations, we have $K\{y\} = \lim_{q \to \infty} K[y_q]$. In that case, the differential ideal *I automatically* generates a prime differential ideal $\mathfrak{p} \subset K\{y\}$ providing a *differential extension* L/K with $L = Q(K\{y\}/\mathfrak{p})$ and *differential transcendence degree* $difftrd(L/K) = \alpha_q^n$, a result explaining the notations [39]. Now, from the dimension formulas of R_{q+r}, we obtain at once $rk_D(M) = \alpha_q^n$ and thus $rk_D(\mathcal{D}) = m - rk_D(M)$ in a coherent way with any free presentation of *M* starting with $\mathcal{D}$. However, $\mathcal{D}$ acts on the left in differential geometry but on the right in the theory of differential modules. For an operator of order zero, we recognize the fact that the rank of a matrix is equal to the rank of the transposed matrix.

PROPOSITION 5.3: If $f \in aut(X)$ is a local diffeomorphisms on *X*, we may set $x = f^{-1}(y) = g(y)$ and we have the *identity*:

$$\frac{\partial}{\partial y^k}\left(\frac{1}{\Delta(g(y))}\partial_i f^k(g(y))\right) \equiv 0 \Rightarrow \frac{\partial}{\partial y^k}\left(\frac{1}{\Delta}\frac{\partial f^k}{\partial x^i}\mathcal{A}^i\right) = \frac{1}{\Delta}\frac{\partial f^k}{\partial x^i}\frac{\partial A^i}{\partial y^k} = \frac{1}{\Delta}\partial_i \mathcal{A}^i$$

and the adjoint of the well defined intrinsic operator $\wedge^0 T^* \xrightarrow{d} \wedge^1 T^* = T^* : A \to \partial_i A$ is (*minus*) the well defined intrinsic operator $\wedge^n T^* \xleftarrow{d} \wedge^n T^* \otimes T \simeq \wedge^{n-1} T^* : \partial_i \mathcal{A}^i \leftarrow \mathcal{A}^i$. Accordingly, if we have an operator $E \xrightarrow{\mathcal{D}} F$, we obtain the *formal adjoint* operator $\wedge^n T^* \otimes E^* \xleftarrow{ad(\mathcal{D})} \wedge^n T^* \otimes F^*$.

Now, with operational notations, let us consider the two differential sequences:

$$\xi \xrightarrow{\mathcal{D}} \eta \xrightarrow{\mathcal{D}_1} \zeta$$

$$\nu \xleftarrow{ad(\mathcal{D})} \mu \xleftarrow{ad(\mathcal{D}_1)} \lambda$$

where $\mathcal{D}_1$ generates all the CC of $\mathcal{D}$. Then $\mathcal{D}_1 \circ \mathcal{D} \equiv 0 \Leftrightarrow ad(\mathcal{D}) \circ ad(\mathcal{D}_1) \equiv 0$ but $ad(\mathcal{D})$ may not generate all the CC of $ad(\mathcal{D}_1)$. Passing to the module framework, we just recognize the definition of $ext_D^1(M)$. Now, exactly like we defined the differential module *M* from $\mathcal{D}$, let us define the differential module *N* from $ad(\mathcal{D})$. Then $ext_D^1(N) = t(M)$ does not depend on the presentation of *M*.

Having in mind that *D* is a *K*-algebra, that *K* is a left *D*-module with the standard action $(D, K) \to K : (P, a)$

$\to P(a):(d_i,a)\to \partial_i a$ and that D is a bimodule over itself, *we have only two possible constructions leading to the following two definitions*:

DEFINITION 5.4: We may define the *inverse system* $R=hom_K(M,K)$ of M and introduce $R_q=hom_K(M_q,K)$ as the *inverse system of order q*.

DEFINITION 5.5: We may define the right differential module $M^*=hom_D(M,D)$.

The first definition is leading to the *inverse systems* introduced by Macaulay in [41] (see [43] for details). As for the second, we have (see [1], p. 21, [6], p. 483-495, [42], [43] for details).

THEOREM 5.6: When M and N are left D-modules, then $hom_K(M,N)$ and $M\otimes_K N$ are left D-modules. In particular $R=hom_K(M,K)$ is also a left D-module for the *Spencer operator*. Moreover, the structures of left D-modules existing therefore on $M\otimes_K N$ and $hom_K(N,L)$ are now coherent with the *adjoint isomorphism* for $mod(D)$:

$$\varphi: hom_D(M\otimes_K N,L)\xrightarrow{\simeq} hom_D(M,hom_K(N,L)),\ \forall L,M,N\in mod(D)$$

Proof: For any $f\in hom_K(M,N)$, let us define:

$$(af)(m)=af(m)=f(am)\quad \forall a\in K,\forall m\in M$$

$$(\xi f)(m)=\xi f(m)-f(\xi m)\quad \forall \xi=\xi^i d_i\in T,\forall m\in M$$

It is easy to check that $\xi a=a\xi+\xi(a)$ in the operator sense and that $\xi\eta-\eta\xi=[\xi,\eta]$ is the standard bracket of vector fields. We have in particular with d in place of any d_i:

$$\begin{aligned}((da)f)(m)&=(d(af))(m)=d(af(m))-af(dm)\\&=(\partial a)f(m)+ad(f(m))-af(dm)\\&=(a(df))(m)+(\partial a)f(m)\\&=((ad+\partial a)f)(m)\end{aligned}$$

For any $m\otimes n\in M\otimes_K N$ with arbitrary $m\in M$ and $n\in N$, we may then define:

$$a(m\otimes n)=am\otimes n=m\otimes an\in M\otimes_K N$$

$$\xi(m\otimes n)=\xi m\otimes n+m\otimes \xi n\in M\otimes_K N$$

and conclude similarly with:

$$\begin{aligned}(da)(m\otimes n)&=d(a(m\otimes n))=d(am\otimes n)\\&=d(am)\otimes n+am\otimes dn\\&=(\partial a)m\otimes n+a(dm)\otimes n+am\otimes dn\\&=(ad+\partial a)(m\otimes n)\end{aligned}$$

Using K in place of N, we finally get $(d_i f)^k_\mu=(d_i f)(y^k_\mu)=\partial_i f^k_\mu-f^k_{\mu+1_i}$ that is *we recognize exactly the Spencer operator* with *now* $Df=dx^i\otimes d_i f$ and thus:

$$\begin{aligned}&(d_i(d_j f))^k_\mu=\partial_{ij}f^k_\mu-(\partial_i f^k_{\mu+1_j}+\partial_j f^k_{\mu+1_i})+f^k_{\mu+1_i+1_j}\\&\Rightarrow d_i(d_j f)=d_j(d_i f)=d_{ij}f\end{aligned}$$

In fact, R is the *projective limit* of $\pi^{q+r}_q: R_{q+r}\to R_q$ in a coherent way with jet theory [18] [19].

The next result is entrelacing the two left structures that we have just provided through the formula $(g(m))(n)=f(m\otimes n)\in N$ defining the map φ whenever $f\in hom_D(M\otimes_K N,L)$ is given and $\varphi(f)=g$. Using any $\xi\in T$, we get successively in L:

$$\begin{aligned}\big(\xi(g(m))\big)(n) &= \xi\big((g(m))(n)\big) - (g(m))(\xi n)\\ &= \xi\big(f(m\otimes n)\big) - f(m\otimes \xi n)\\ &= f\big(\xi(m\otimes n)\big) - f(m\otimes \xi n)\\ &= f(\xi m\otimes n + m\otimes \xi n) - f(m\otimes \xi n)\\ &= f(\xi m\otimes n) = \big(g(\xi m)\big)(n)\end{aligned}$$

and thus $\xi(g(m)) = g(\xi m), \forall m \in M$ or simply $\xi \circ g = g \circ \xi$.

Q.E.D.

COROLLARY 5.7: If *M* and *N* are right *D*-modules, then $hom_K(M,N)$ is a left *D*-module. Moreover, if *M* is a left *D*-module and *N* is a right *D*-module, then $M \otimes_K N$ is a right *D*-module.

Proof: If *M* and *N* are right *D*-modules, we just need to set $(\xi f)(m) = f(m\xi) - f(m)\xi, \forall \xi \in T, \forall m \in M$ and conclude as before. Similarly, if *M* is a left *D*-module and *N* is a right *D*-module, we just need to set $(m\otimes n)\xi = m\otimes n\xi - \xi m\otimes n$.

Q.E.D.

REMARK 5.8: When $M = {}_DM \in mod(D)$ and $N = N_D$, , then $hom_K(N,M)$ cannot be endowed with any left or right differential structure. When $M = M_D$ and $N = N_D$, then $M \otimes_K N$ cannot be endowed with any left or right differential structure (see [1], p. 24 for more details).

As $D = {}_DD_D$ is a bimodule, then $M^* = hom_D(M,D)$ is a right *D*-module according to Lemma 2.13 and the module *N* defined by the ker/coker sequence $0 \leftarrow N \leftarrow F_1^* \xleftarrow{\mathcal{D}^*} F_0^* \leftarrow M^* \leftarrow 0$ is thus a right module N_D.

COROLLARY 5.9: We have the *side changing* procedure $N_D \to N = {}_DN = hom_K(\wedge^n T^*, N_D)$ with inverse $M = M \to M_D = \wedge^n T^* \otimes_K M$ whenever $M, N \in mod(D)$.

Proof: According to the above Theorem, we just need to prove that $\wedge^n T^*$ has a natural right module structure over *D*. For this, if $\alpha = a dx^1 \wedge \cdots \wedge dx^n \in T^*$ is a volume form with coefficient $a \in K$, we may set $\alpha \cdot P = ad(P)(a)\, dx^1 \wedge \cdots \wedge dx^n$ when $P \in D$. As *D* is generated by *K* and *T*, we just need to check that the above formula has an intrinsic meaning for any $\xi = \xi^i d_i \in T$. In that case, we check at once:

$$\alpha \cdot \xi = -\partial_i(a\xi^i)\, dx^1 \wedge \cdots \wedge dx^n = -\mathcal{L}(\xi)\alpha$$

by introducing the Lie derivative of α with respect to ξ, along the intrinsic formula $\mathcal{L}(\xi) = i(\xi)d + di(\xi)$ where $i(\,)$ is the interior multiplication and *d* is the exterior derivative of exterior forms. According to well known properties of the Lie derivative, we get:

$$\begin{aligned}&\alpha\cdot(a\xi) = (\alpha\cdot\xi)\cdot a - \alpha\cdot\xi(a),\\ &\alpha\cdot(\xi\eta - \eta\xi) = -\big[\mathcal{L}(\xi), \mathcal{L}(\eta)\big]\alpha = -\mathcal{L}\big([\xi,\eta]\big)\alpha = \alpha\cdot[\xi,\eta].\end{aligned}$$

Q.E.D.

Collecting all the results so far obtained, if a differential operator $\mathcal{D}$ is given in the framework of differential geometry, we may keep the same notation $\mathcal{D}$ in the framework of differential modules which are *left* modules over the ring *D* of linear differential operators and apply duality, provided we use the notation $\mathcal{D}^*$ and deal with *right* differential modules or use the notation $ad(\mathcal{D})$ and deal again with *left* differential modules by using the *left* $\leftrightarrow$ *right* *conversion* procedure.

DEFINITION 5.10: If a differential operator $\xi \xrightarrow{\mathcal{D}} \eta$ is given, a *direct problem* is to find (generating) *compatibility conditions* (CC) as an operator $\eta \xrightarrow{\mathcal{D}_1} \zeta$ such that $\mathcal{D}\xi = \eta \Rightarrow \mathcal{D}_1\eta = 0$. Conversely, given $\eta \xrightarrow{\mathcal{D}_1} \zeta$, the *inverse problem* will be to look for $\xi \xrightarrow{\mathcal{D}} \eta$ such that $\mathcal{D}_1$ generates the CC of $\mathcal{D}$ and we shall say that $\mathcal{D}_1$ *is parametrized by* $\mathcal{D}$ *if such an operator* $\mathcal{D}$ *is existing*.

REMARK 5.11: Of course, solving the direct problem (Janet, Spencer) is *necessary* for solving the inverse problem. However, though the direct problem always has a solution, the inverse problem may not have a solution at all and the case of the Einstein operator is one of the best non-trivial PD counterexamples (compare [10] to [34]). It is rather striking to discover that, in the case of OD operators, it took almost 50 years to understand that the possibility to solve the inverse problem was equivalent to the controllability of the corresponding control system (compare [11] to [34]).

As $ad\left(ad\left(P\right)\right)=P,\forall P\in D$, any operator is the adjoint of a certain operator and we get:

FORMAL TEST 5.12: The *double duality test* needed in order to check whether $t\left(M\right)=0$ or not and to find out a parametrization if $t\left(M\right)=0$ has 5 steps which are drawn in the following diagram where $ad\left(\mathcal{D}\right)$ generates the CC of $ad\left(\mathcal{D}_1\right)$ and $\mathcal{D}_1'$ generates the CC of $\mathcal{D}$:

$$\begin{array}{ccccccc} & & & & & \zeta' & 5 \\ & & & & \nearrow^{\mathcal{D}_1'} & & \\ 4 & \xi & \xrightarrow{\mathcal{D}} & \eta & \xrightarrow{\mathcal{D}_1} & \zeta & 1 \\ & & & & & & \\ 3 & \nu & \xleftarrow{ad(\mathcal{D})} & \mu & \xleftarrow{ad(\mathcal{D}_1)} & \lambda & 2 \end{array}$$

THEOREM 5.13: $\mathcal{D}_1$ parametrized by $\mathcal{D}\Leftrightarrow\mathcal{D}_1=\mathcal{D}_1'\Leftrightarrow t\left(M\right)=0\Leftrightarrow ext^1\left(N\right)=0$.

REMARK 5.14: When an operator $\mathcal{D}_1$ can be parametrized by an operator $\mathcal{D}$, we may ask whether or not $\mathcal{D}$ can be parametrized again by an operator $\mathcal{D}_{-1}$ and so on. A good comparison can be made with hunting rifles as a few among them, called double rifles, are equipped with a double trigger mechanism, allowing to shoot again once one has already shot. In a mathematical manner, the question is to know whether the diffe- rential module defined by $\mathcal{D}_1$ is torsion-free or reflexive. The main difficulty is that these intrinsic properties highly depend on the choice of the parametrizing operator. The simplest example is provided by the Poincaré sequence for *n* = 3 made by the successive $grad,curl,div$ operators. Indeed, any student knows that *curl* is parametrizing *div* and that *grad* is parametrizing *curl*. However, we may parametrize $\partial_1\eta^1+\partial_2\eta^2+\partial_3\eta^3=0$ by choosing $\partial_3\xi^1=\eta^1,\partial_3\xi^2=\eta^2,-\partial_1\xi^1-\partial_2\xi^2=\eta^3$ with 2 potentials $\left(\xi^1,\xi^2\right)$ only instead of the usual 3 potentials $\left(\xi^1,\xi^2,\xi^3\right)$ and cannot proceed ahead as before. Other important examples will be provided in the next section dealing with applications, in particular the one involving Einstein equations when $n=4$. This comment points out the reason for using the extension modules.

It remains to study a delicate question on which all the examples of the Introduction were focussing. Indeed, if a parametrization of a given system of OD or PD equations is possible, that is, equivalently, if the corresponding differential module is torsion-free, it appears that different parametrizations may exist with quite different numbers of potentials needed. Accordingly, it should be useful to know about the possibility to have upper and lower bounds for these numbers when $n>1$, particularly in elasticity theory, because when $n=1$, an OD module *M* with $t\left(M\right)=0$ being *automatically* isomorphic to a free module *E*, the number of potentials needed is equal to $rk_D\left(M\right)=rk_D\left(E\right)$. We shall use the language of differential modules in order to improve and apply a few results already presented in ([7], Theorem 7+ Appendix).

THEOREM 5.15: Let $F_1\xrightarrow{\mathcal{D}_1}F_0\rightarrow M\rightarrow 0$ be a finite free presentation of the differential module $M=coker\left(\mathcal{D}_1\right)$ and assume we already know that $t\left(M\right)=0$ by using the formal test. Accordingly, we have obtained the exact sequence $F_1\xrightarrow{\mathcal{D}_1}F_0\xrightarrow{\mathcal{D}}E$ of free differential modules where $\mathcal{D}$ is the parametrizing operator. Then, there exists other parametrizations $F_1\xrightarrow{\mathcal{D}_1}F_0\xrightarrow{\mathcal{D}'}E'$ called *minimal parametrizations* and such that $coker\left(\mathcal{D}'\right)$ is a torsion module.

Proof: We first explain the reason for using the word "*minimal*". Indeed, we have $rk_D\left(M\right)=rk_D\left(F_0\right)-rk_D\left(im\left(\mathcal{D}_1\right)\right)=rk_D\left(\mathcal{D}\right)\leq rk_D\left(E\right)$ but also $rk_D\left(M\right)=rk_D\left(\mathcal{D}'\right)=rk_D\left(E'\right)$ and thus $rk_D\left(E'\right)=rk_D\left(M\right)\leq rk_D\left(E\right)$ as a way to get a lower bound for the number of potentials but not to get a differential geometric framework.

While applying the formal test in the operator language, $ad\left(\mathcal{D}\right)$ is describing the (generating) CC of $ad\left(\mathcal{D}_1\right)$ and we shall denote by $ad\left(\mathcal{D}_{-1}\right)$ the (generating) CC of $ad\left(\mathcal{D}\right)$ as we did in Example 1.3. In the module framework, going on with left differential modules, when *F* is a free left module, we shall denote by $\tilde{F}$ the corresponding *converted* left differential module of the right differential module F^*. The reader not familiar with duality may look at the situations met in electromagnetism and elasticity in ([6], p. 492-495). If $L=coker\left(ad\left(\mathcal{D}_{-1}\right)\right)\simeq im\left(ad\left(\mathcal{D}\right)\right)\subset\tilde{F}_0$ and $\tilde{E}'$ is the largest free differential submodule of *L* (D^3 in Example 1.3, *D* in Example 1.4), then $T=L/\tilde{E}'$ is a torsion module and we have the following commutative and exact diagram:

$$\begin{array}{ccccccccc}
 & & & & 0 & & 0 & & \\
 & & & & \downarrow & & \downarrow & & \\
0 & & \rightarrow & & \tilde{E}' & = & \tilde{E}' & \rightarrow & 0 \\
\downarrow & & & & \downarrow & \swarrow & \downarrow & & \\
\tilde{E}_{-1} & & \xrightarrow{ad(\mathcal{D}_{-1})} & & \tilde{E} & \rightarrow & L & \rightarrow & 0 \\
\| & & & & \downarrow & & \downarrow & & \\
\tilde{E}_{-1} & & \rightarrow & & \tilde{E}'' & \rightarrow & T & \rightarrow & 0 \\
\downarrow & & & & \downarrow & & \downarrow & & \\
0 & & & & 0 & & 0 & &
\end{array}$$

where the central vertical monomorphism $\tilde{E}' \rightarrow \tilde{E}$ is obtained by pulling a basis of $\tilde{E}'$ back to $\tilde{E}$ as we did in the diagram of Proposition 2.24. Coming back to the operators $ad(\mathcal{D})$ and $ad(\mathcal{D}_1)$, we get the following commutative and exact diagram allowing to define $ad(\mathcal{D}')$ by composition:

$$\begin{array}{ccccccc}
\tilde{F}_1 & \xleftarrow{ad(\mathcal{D}_1)} & \tilde{F}_0 & \xleftarrow{ad(\mathcal{D})} & \tilde{E} & & \\
 & & \nwarrow & \swarrow & & & \\
 & & \| & L & \uparrow & & \\
 & & \swarrow & \nwarrow & & & \\
 & & \tilde{F}_0 & \xleftarrow{ad(\mathcal{D}')} & \tilde{E}' & \leftarrow & 0 \\
 & & & & \uparrow & \nwarrow & \\
 & & & & 0 & & 0
\end{array}$$

We have:

$$ad(\mathcal{D}_1) \circ ad(\mathcal{D}') \equiv 0 \Rightarrow \mathcal{D}' \circ \mathcal{D}_1 \equiv 0 \Rightarrow ker(D) = im(\mathcal{D}_1) \subseteq ker(\mathcal{D}') \subset F_0$$

and obtain by duality the following commutative and exact diagram where $M' = coim(\mathcal{D}')$:

$$\begin{array}{ccccccccc}
 & & & & & & & & 0 \\
 & & & & & & & \nearrow & \\
 & & 0 & & & & M & & \\
 & & \downarrow & & & \nearrow & \downarrow & \searrow & \\
0 \rightarrow & ker(\mathcal{D}) & \rightarrow & F_0 & & \xrightarrow{\mathcal{D}} & & E & \\
 & & & & & \downarrow & & 0 & \\
 & & & & & \downarrow & \nearrow & & \\
 & & \downarrow & & \| & & M' & & \downarrow \\
 & & & & & \nearrow & \downarrow & \searrow & \\
0 \rightarrow & ker(\mathcal{D}') & \rightarrow & F_0 & & \xrightarrow{\mathcal{D}'} & & E' & \\
 & & & & & \downarrow & & & \\
 & & & & & 0 & & &
\end{array}$$

However, though the upper sequence $F_1 \xrightarrow{\mathcal{D}_1} F_0 \rightarrow M \rightarrow 0$ is exact by definition because $M = coim(\mathcal{D}) = coker(\mathcal{D}_1)$, the lower induced sequence $F_1 \xrightarrow{\mathcal{D}_1} F_0 \rightarrow M' \rightarrow 0$ may not be exact. With $rk_D = rk$ for simplicity, $t(M) = 0$ and the induced epimorphism $M \rightarrow M' \rightarrow 0$, we obtain:

$$rk(M) = rk(\mathcal{D}) = rk(ad(\mathcal{D})) = rk(\tilde{F}_0) - rk(ad(\mathcal{D}_1)) = rk(F_0) - rk(\mathcal{D}_1)$$

$$rk(M') = rk(\mathcal{D}') = rk(ad(\mathcal{D}')) = rk(\tilde{E}') = rk(L) = rk(\tilde{F}_0) - rk(ad(\mathcal{D}_1)) = rk(F_0) - rk(\mathcal{D}_1)$$

$$\Rightarrow rk(M) = rk(M') \Rightarrow rk(ker(M \to M')) = 0 \Rightarrow ker(M \to M') \subseteq t(M) \subseteq (M)$$

$$\Rightarrow ker(M \to M') = 0 \Rightarrow M \simeq M'$$

Accordingly, $\mathcal{D}'$ is a minimal parametrization of $\mathcal{D}_1$ contrary to $\mathcal{D}$ in general and we invite the reader to repeat the proof by using operators and their adjoints as in the formal test.

Q.E.D.

6) APPLICATIONS

EXAMPLE 6.1: *OD Control Theory Revisited*

The following result is well known and can be found in any textbook of algebra [6] [13] [32]:

PROPOSITION 6.2: If *A* is a principal ideal domain, that is if any ideal in *A* is generated by a single element, then any torsion-free module over *A* is free and thus projective.

As this is the case of the ring $D = K[d_x] = K[d]$ when $n = 1$, we obtain the following corollary of the preceding parametrizing Theorem, allowing to extend the Kalman test of controllability to PD systems with variable coefficients as we did in the Introduction (see [6]-[9] [11] for details).

COROLLARY 6.3: If $\mathcal{D}_1$ is *surjective*, then $ad(\mathcal{D}_1)$ is injective if and only if *M* is projective.

Proof: As $\mathcal{D}_1$ is surjective, replacing *M* by *P*, we have the following short exact sequence:

$$0 \to F_1 \xrightarrow{\mathcal{D}_1} F_0 \to P \to 0$$

As *P* is projective, this short exact sequence splits with $F_0 \simeq P \oplus F_1$ [6] [13] [32]. Using Proposition 2.6, we can construct a *right inverse* operator $\mathcal{P}_1$ of $\mathcal{D}_1$ with *now* $\mathcal{D}_1 \circ \mathcal{P}_1 = id_{F_1}$ for the corresponding *morphisms*. Applying duality and Corollary 2.10, we get the short exact sequence:

$$0 \leftarrow F_1^* \xleftarrow{\mathcal{D}_1^*} F_0^* \leftarrow P^* \leftarrow 0$$

It follows that $\mathcal{D}_1^*$ is surjective and the adjoint operator $ad(\mathcal{D}_1)$ is injective because $N = 0$.

Conversely, if $ad(\mathcal{D}_1)$ is injective, there exists a *left inverse* $ad(\mathcal{P}_1)$ of $ad(\mathcal{D}_1)$ providing a *right inverse* $\mathcal{P}_1$ of $\mathcal{D}_1$ (*care*). We may thus use again Corollary 2.10 because $F_0^{**} \simeq F_0$ and $F_1^{**} \simeq F_1$. Meanwhile, we have proved that, if $n = 1$ and $t(M) = 0$, it is always possible to find an injective parametrization but Example 1.4 is showing that this result is no longer true when $n > 1$.

Q.E.D.

Multiplying the control system of Example 1.1 by a test function λ and integrating by parts, the kernel of the operator thus obtained is defined by the OD equations:

$$\lambda_{xx} - \lambda_x = 0, \quad \lambda_{xx} - a\lambda = 0 \quad \Rightarrow \lambda_x - a\lambda = 0 \quad \Rightarrow (\partial_x a + a^2 - a)\lambda = 0$$

The formal adjoint of the operator defining the control system is thus injective if and only if we have $\partial_x a + a^2 - a\lambda \neq 0$, a result *absolutely not evident at first sight* but explaining why *we used the same notation for a test function and for a Lagrange multiplier*.

EXAMPLE 6.4: *Elasticity Theory Revisited*

The *Killing operator* $\mathcal{D}: T \to S_2T^*$ is a defined by $\xi \in T \to \mathcal{D}\xi = \mathcal{L}(\xi)\omega = \Omega = 2\epsilon \in S_2T^*$ with $\omega_{rj}\partial_i\xi^r + \omega_{ir}\partial_j\xi^r + \xi^r\partial_r\omega_{ij} = \Omega_{ij} = 2\epsilon_{ij}$ where ξ is the displacement vector, $\mathcal{L}(\xi)\omega$ is the Lie derivative of ω with respect to ξ and ϵ is the infinitesimal deformation tensor of textbooks. It is a *Lie operator* because its solutions $\Theta \subset T$ satisfy $[\Theta, \Theta] \subset \Theta$. The corresponding first order *Killing system* $R_1 \subset J_1(T)$ is *not* involutive because its symbol $g_1 \subset T^* \otimes T$ is finite type with first prolongation $g_2 = 0$ and thus $rk(\mathcal{D}) = n$. Accordingly, as ω is a flat constant metric, the second order CC are described by an operator $\mathcal{D}_1$ coming from the linearization of the Riemann tensor obtained in a standard way by setting $\omega \to \omega + t\Omega$ with a small parameter $t \ll 1$, dividing by *t* and taking the limit when $t \to 0$. Finally, raising the index by means of the metric, the adjoint operator $\wedge^n T^* \otimes T^* \xleftarrow{ad(\mathcal{D})} \wedge^n T^* \otimes S_2T$ is defined by the *intrinsic stress equations* $\nabla_r\sigma^{ir} \equiv \partial_r\sigma^{ir} + \gamma^i_{rs}\sigma^{rs} = f^i$ where ∇ is the covariant derivative and γ the Christoffel symbols ([6], p. 494, [44], p. 236).

- Airy parametrization of the stress equations when $n = 2$ gives $rk(E') = rk(E) = 1$ and we have thus 1 potential only. By duality, working out the corresponding adjoint operators, we obtain the two formally exact

sequences:

$$\begin{array}{ccccccccc} & & & \overset{Killing}{} & & \overset{Riemann}{} & & & \\ & & 2 & \rightarrow & 3 & \rightarrow & 1 & \rightarrow 0 \\ & \overset{Cauchy}{} & & & & \overset{Airy}{} & & & \\ 0 \leftarrow & 2 & \leftarrow & 3 & \leftarrow & 1 & & & \end{array}$$

Accordingly, the canonical and the minimal parametrizations coincide when $n=2$. We discover that the Airy parametrization is nothing else than the formal adjoint of the Riemann CC for the deformation tensor:

$$\partial_i \xi_j + \partial_j \xi_i = \Omega_{ij} \Rightarrow \partial_{22}\Omega_{11} + \partial_{11}\Omega_{22} - 2\partial_{12}\Omega_{12} = 0$$

where the indices of the *displacement vector* $\left(\xi^1, \xi^2\right)$ are lowered by means of the euclidean metric of $\mathbb{R}^2$. We do not believe this result is known in such a general framework.

- Beltrami parametrization of the stress equations when $n=3$ gives $rk\left(E\right)=6$ and we have thus 6 potentials. However, Maxwell/Morera parametrizations of the stress equations when $n=3$ both give $rk\left(E'\right)=3$ and we have thus 3 potentials only.

$$\begin{array}{ccccccccccc} & & & & \overset{Killing}{} & & \overset{Riemann}{} & & \overset{Bianchi}{} & & \\ & & & 3 & \rightarrow & 6 & \rightarrow & 6 & \rightarrow & 3 \rightarrow 0 & \\ & & & & & & & & & & \\ & & \overset{Cauchy}{} & & & & \overset{Beltrami}{} & & & & \\ 0 \leftarrow & 3 & \leftarrow & 6 & \leftarrow & 6 & \leftarrow & & 3 & & \\ & & & \| & & & \uparrow & & & & \\ & & & \underset{Maxwell}{} & & & & & & & \\ & & & 6 & \leftarrow & 3 & & & & & \end{array}$$

Accordingly, the canonical parametrization has 6 potentials while *any* minimal parametrization has 3 potentials. We finally notice that the Cauchy operator is parametrized by the Beltrami operator which is *again* parametrized by the adjoint of the Bianchi operator obtained by linearizing the Bianchi identities existing for the Riemann tensor, *a property not held by any minimal parametrization* as we already noticed.

- For $n=4$, we shall prove below that the *Einstein parametrization* of the stress equations is neither canonical nor minimal in the following diagram:

$$\begin{array}{ccccccccccc} & & \overset{Killing}{} & & \overset{Riemann}{} & & \overset{Bianchi}{} & & & & \\ & 4 & \rightarrow & 10 & \rightarrow & 20 & \rightarrow & 20 & \rightarrow & 6 & \rightarrow 0 \\ & & & \| & & \downarrow & & \downarrow & & & \\ & & & \underset{Einstein}{} & & \overset{div}{} & & & & & \\ & & & 10 & \rightarrow & 10 & \rightarrow & 4 & \rightarrow & 0 & \\ & \overset{Cauchy}{} & & & & & & & & & \\ 0 \leftarrow & 4 & \leftarrow & 10 & \leftarrow & 20 & \leftarrow & 20 & & & \\ & & & \| & & \uparrow & & & & & \\ & & & \underset{Einstein}{} & & & & & & & \\ & & & 10 & \leftarrow & 10 & & & & & \end{array}$$

obtained by using the fact that the Einstein operator, linearization of the Einstein tensor at the Minkowski metric, is self-adjoint, the 6 terms being exchanged between themselves [10] [45]. The upper *div* induced by *Bianchi* has *nothing to do* with the lower *Cauchy* stress equations, contrary to what is still believed today. It also follows that the Einstein equations in vacuum cannot be parametrized as we have the following diagram of operators (see [6] and [34] for more details or [10] for a computer algebra exhibition of this result):

$$\begin{array}{ccccc} & & & & 20 \\ & & & \overset{Riemann}{\nearrow} & \\ & \overset{Killing}{} & & \overset{Einstein}{} & \\ 4 & \rightarrow & 10 & \rightarrow & 10 \\ & & & & \\ & \overset{Cauchy}{} & & \overset{Einstein}{} & \\ 4 & \leftarrow & 10 & \leftarrow & 10 \end{array}$$

- It remains therefore to compute all these numbers for an arbitrary dimension $n \geq 2$. For this, we notice that

the successive prolongations $\rho_r(\Phi): J_{q+r}E \to J_r(F_0)$ defined by $d_\nu \Phi^\tau = z^\tau_\nu$ for $0 \le |\nu| \le r$ have kernel R_{q+r}. The *symbol morphism* $\sigma_r(\Phi): S_{q+r}T^* \otimes E \to S_rT^* \otimes F_0$ with kernel g_{q+r} is induced by the projection of $\rho_r(\Phi)$ onto $\rho_{r-1}(\Phi)$ (see [39], p. 256 or [41], p. 233 for details). If we use such a procedure for a first order system with no zero or first order CC, we have $q=1, E=T, F_0 = J_1(T)/R_1$. The Killing system R_1 is formally integrable (R_2 involutive) if and only if ω has *constant Riemannian curvature*:

$$\rho^k_{l,ij} = c\left(\delta^k_i \omega_{lj} - \delta^k_j \omega_{li}\right)$$

with $c=0$ when ω is the flat Minkowski metric [12] [20] [39]. In general, we may apply the Spencer δ-map to the top row obtained with $r=2$ in order to get the commutative diagram:

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & 0 & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 \to & & g_3 & \to & S_3T^* \otimes T & \to & S_2T^* \otimes F_0 & \to F_1 & \to 0 \\
 & & \downarrow \delta & & \downarrow \delta & & \downarrow \delta & & \\
0 \to & & T^* \otimes g_2 & \to & T^* \otimes S_2T^* \otimes T & \to & T^* \otimes T^* \otimes F_0 & \to 0 & \\
 & & \downarrow \delta & & \downarrow \delta & & \downarrow \delta & & \\
0 \to & & \wedge^2 T^* \otimes g_1 & \to & \underline{\wedge^2 T^* \otimes T^* \otimes T} & \to & \wedge^2 T^* \otimes F_0 & \to 0 & \\
 & & \downarrow \delta & & \downarrow \delta & & \downarrow & & \\
0 \to & & \wedge^3 T^* \otimes T & = & \wedge^3 T^* \otimes T & \to & 0 & & \\
 & & \downarrow & & \downarrow & & & & \\
 & & 0 & & 0 & & & &
\end{array}$$

with exact rows and exact columns but the first that may not be exact at $\wedge^2 T^* \otimes g_1$. We shall denote by $B^2(g_1)$ the *coboundary* as the image of the central δ, by $Z^2(g_1)$ the *cocycle* as the kernel of the lower δ and by $H^2(g_1) = Z^2(g_1)/B^2(g_1)$ the *Spencer δ- cohomology* at $\wedge^2 T^* \otimes g_1$.

In the classical Killing system, $g_1 \subset T^* \otimes T$ is defined by $\omega_{rj}(x)\xi^r_i + \omega_{ir}(x)\xi^r_j = 0 \Rightarrow \xi^r_r = 0, g_2 = 0, g_3 = 0$. Applying the previous diagram, we discover that the *Riemann tensor* $\left(\rho^k_{l,ij}\right) \subset \wedge^2 T^* \otimes T^* \otimes T$ is a section of the bundle $Riemann = F_1 = H^2(g_1) = Z^2(g_1)$ with $dim(Riemann) = \left(n^2(n+1)^2/4\right) - \left(n^2(n+1)(n+2)/6\right)$ $= \left(n^2(n-1)^2/4\right) - \left(n^2(n-1)(n-2)/6\right) = n^2\left(n^2-1\right)/12$ by using the top row or the left column. We obtain at once the two properties of the (linearized) Riemann tensor through the chase involved, namely $\left(R^k_{l,ij}\right) \in \wedge^2 T^* \otimes T^* \otimes T$ is killed by both δ and $\sigma_0(\Phi)$. However, we have no indices for F_1 and cannot therefore exhibit the *Ricci tensor* or the *Einstein tensor* of general relativity by means of the usual *contraction* or *trace*. We recall briefly their standard definitions by stating $R_{ij} = R_{ji} = R^r_{i,rj} \Rightarrow R = \omega^{ij}R_{ij} \Rightarrow E_{ij} = R_{ij} - \frac{1}{2}\omega_{ij}R$.

Similarly, going one step further, we get the (linearized) Bianchi identities with

$Bianchi = F_2 = H^3(g_1) = Z^3(g_1) \Rightarrow dim(Bianchi) = dim\left(\wedge^4 T^* \otimes T\right) - dim\left(\wedge^3 T^* \otimes g_1\right) = n^2\left(n^2-1\right)(n-2)/24$

as in ([46], p. 168-171). This approach is relating for the first time the concept of *Riemann tensor candidate*, introduced by Lanczos and others, to the Spencer δ-cohomology of the Killing symbols.

Counting the differential ranks is now easy because R_1 is formally integrable with finite type symbol and thus R_2 is involutive with symbol $g_2 = 0$. We get:

$$rk(Killing) = rk(Cauchy) = n$$

$$\Rightarrow rk(Riemann) = dim\left(S_2T^*\right) - n = \left(n(n+1)/2\right) - n = n(n-1)/2$$

$$rk(Bianchi) = \left(n^2\left(n^2-1\right)/12\right) - \left(n(n-1)/2\right) = n(n-1)(n-2)(n+3)/12$$

that is $rk(Bianchi) = 3$ when $n=3$ and $rk(Bianchi) = 14 = 20 - 6$ when $n=4$. Collecting all the results, we obtain that the canonical parametrization needs $n^2\left(n^2-1\right)/12$ potentials while any minimal parametrization only needs $n(n-1)/2$ potentials. The Einstein parametrization is thus "*in between*" because

$n(n-1)/2 < n(n+1)/2 < n^2(n^2-1)/12, \forall n \geq 4$.

The *conformal Killing system* $\hat{R}_1 \subset J_1(T)$ is defined by eliminating the function $A(x)$ in the system $\mathcal{L}(\xi)\omega = A(x)\omega$. It is also a *Lie operator* $\hat{\mathcal{D}}$ with solutions $\hat{\Theta} \subset T$ satisfying $\left[\hat{\Theta}, \hat{\Theta}\right] \subset \hat{\Theta}$. Its symbol $\hat{g}_1$ is defined by the linear equations $\omega_{rj}\xi_i^r + \omega_{ir}\xi_j^r - \frac{2}{n}\omega_{ij}\xi_r^r = 0$ which do not depend on any conformal factor and is finite type when $n \geq 3$ because $g_3 = 0$ but $\hat{g}_2$ is *now* 2-acyclic *only when* $n \geq 4$ and 3-acyclic *only when* $n \geq 5$ [20] [46]-[48]. It is known that $\hat{R}_2$ and thus $\hat{R}_1$ too (by a chase) are formally integrable if and only if ω has zero *Weyl tensor*:

$$\tau_{l,ij}^k \equiv \rho_{l,ij}^k - \frac{1}{(n-2)}\left(\delta_i^k \rho_{lj} - \delta_j^k \rho_{li} + \omega_{lj}\omega^{ks}\rho_{si} - \omega_{li}\omega^{ks}\rho_{sj}\right) + \frac{1}{(n-1)(n-2)}\left(\delta_i^k \omega_{lj} - \delta_j^k \omega_{li}\right)\rho = 0$$

We may use the formula $id_M - f \circ u = v \circ g$ of Proposition 2.6 in the *split short exact sequence* induced by the inclusions $R_1 \subset \hat{R}_1 \Rightarrow g_1 \subset \hat{g}_1$:

$$0 \to Ricci \to Riemann \to Weyl \to 0$$

according to the Vessiot structure equations, in particular if ω has constant Riemannian curvature and thus $\rho_{ij} = \rho_{i,rj}^r = c(n-1)\omega_{ij} \Rightarrow \rho = \omega^{ij}\rho_{ij} = cn(n-1)$ [20] [39] [45]-[47]. Using the same diagrams as before, we get $Weyl = \hat{F}_1 = H^2(\hat{g}_1) \neq Z^2(\hat{g}_1)$ for defining any *Weyl tensor candidate*. As a byproduct, the linearized Weyl operator is of order 2 with a symbol $\hat{h}_2 \subset S_2T^* \otimes \hat{F}_0$ which is *not* 2-acyclic by applying the δ-map to the short exact sequence:

$$0 \to \hat{g}_{3+r} \to S_{3+r}T^* \otimes T \xrightarrow{\sigma_{2+r}(\Phi)} \hat{h}_{2+r} \to 0$$

and chasing through the commutative diagram thus obtained with $r = 0,1,2$. As $\hat{h}_3$ becomes 2-acyclic after one prolongation of $\hat{h}_2$ only, it follows that *the generating CC for the Weyl operator are of order* 2, a result showing that the so-called Bianchi identities for the Weyl tensor are *not* CC in the strict sense of the definition as they do not involve only the Weyl tensor. Of course, these results could not have been discovered by Lanczos and followers because the formal theory of Lie pseudogroups and the Vessiot structure equations are still not acknowledged today.

For this reason, we provide a few hints in order to explain why the Vessiot structure equations *sometimes contain a few constants*, *sometimes none at all* as we just saw (see [39] [49] and [50] for more details). Indeed, isometries preserve the metric $\omega = (\omega_{ij} = \omega_{ji}) \in S_2T^*$ while conformal isometries preserve the symmetric tensor density $\hat{\omega} = \left(\hat{\omega}_{ij} = \omega_{ij} \Big/ \left(|det(\omega)|^{\frac{1}{n}}\right)\right)$. The respective variations are related by the similitude formula $\hat{\Omega}_{ij} \sim \Omega_{ij} - \frac{1}{n}\omega_{ij}\omega^{rs}\Omega_{rs}$ which only depends on ω and not on a conformal factor. It follows that $F_0 = S_2T^*$ and that $\hat{F}_0$ may be identified with the sub-bundle $\left\{\hat{\Omega} \in S_2T^* \mid \omega^{ij}\hat{\Omega}_{ij} = 0\right\}$ with the above well defined epimorphism $F_0 \to \hat{F}_0$ induced by the inclusion $R_1 \subset \hat{R}_1$. We set [39] [49] [50]:

DEFINITION 6.5: We say that a vector bundle F is associated with a Lie operator $\mathcal{D}$ if, for any solution $\xi \in \Theta \subset T$ of $\mathcal{D}$, there exists a first order operator $\mathcal{L}(\xi): F \to F$ called *Lie derivative* with respect to ξ and such that:

1) $\mathcal{L}(\xi + \eta) = \mathcal{L}(\xi) + \mathcal{L}(\eta) \quad \forall \xi, \eta \in \Theta$
2) $\left[\mathcal{L}(\xi), \mathcal{L}(\eta)\right] = \mathcal{L}([\xi, \eta]) \quad \forall \xi, \eta \in \Theta$
3) $\mathcal{L}(\xi)(f\eta) = f\mathcal{L}(\xi)\eta + (\xi \cdot f)\eta \quad \forall \xi \in \Theta, \forall f \in C^\infty(X), \forall \eta \in F$
4) If E and F are two such associated vector bundles, then:

$$\mathcal{L}(\xi)(\eta \otimes \zeta) = \mathcal{L}(\xi)\eta \otimes \zeta + \eta \otimes \mathcal{L}(\xi)\zeta, \quad \forall \xi \in \Theta, \forall \eta \in E, \forall \zeta \in F$$

In such a case, we may introduce $\Upsilon = \Upsilon(F) = \left\{\eta \in F \mid \mathcal{L}(\xi)\eta = 0, \forall \xi \in \Theta \subset T\right\}$.

PROPOSITION 6.6: Using capital letters for linearized objects, we have:

1) $\Upsilon(T) = C(\Theta) = \left\{\eta \in T \mid [\xi, \eta] = 0, \forall \xi \in \Theta\right\} = centralizer$ of Θ in T.

2) $\Upsilon_0 = \Upsilon(F_0) = \Upsilon(S_2T^*) = \{\Omega = A\omega \in S_2T^* \mid A = cst\}$.

3) $\Upsilon_1 = \Upsilon(F)_1 = \{R^k_{l,ij} = C(\delta^k_i\omega_{lj} - \delta^k_j\omega_{li}) \in F_1 \mid C = cst\}$.

4) $\hat{\Upsilon}_1 = \Upsilon(\hat{F}_1) = 0 \Rightarrow \Upsilon(Ricci) \simeq \Upsilon(Riemann)$.

5) The Lie derivative commutes with the Janet operators $\mathcal{D}, \mathcal{D}_1, \cdots, \mathcal{D}_n$.

We have in particular $\mathcal{D}_1 : \Upsilon_0 \to \Upsilon_1 : A \to C = -cA$ (*care to sign*).

Proof: Two (nondegenerate) metrics $\omega, \bar{\omega} \in S_2T^*$ give the same Killing system R_1 if and only if $\bar{\omega} = a\omega$ with the multiplicative group parameter $a = cst$. Therefore, if R_1 is FI, then the two metrics have respective constant curvatures c and $\bar{c} = c/a$. Setting $a = 1 + tA + \cdots \Rightarrow \bar{c} = c + tC + \cdots$ while linearizing these finite transformations with $t \ll 1$ gives $C = -cA$ when $t \to 0$.

Q.E.D.

However, we have yet not proved the most difficult result that could not be obtained without homological algebra and the next example will explain this additional difficulty.

EXAMPLE 6.7: With $\partial_{22}\xi = \eta^2, \partial_{12}\xi = \eta^1$ for $\mathcal{D}$, we get $\partial_1\eta^2 - \partial_2\eta^1 = \zeta$ for $\mathcal{D}_1$. Then $ad(\mathcal{D}_1)$ is defined by $\mu^2 = -\partial_1\lambda, \mu^1 = \partial_2\lambda$ while $ad(\mathcal{D})$ is defined by $\nu = \partial_{12}\mu^1 + \partial_{22}\mu^2$ but the CC of $ad(\mathcal{D}_1)$ are generated by $\nu' = \partial_1\mu^1 + \partial_2\mu^2$. Using operators, we have the differential sequences:

$$\begin{array}{ccccc} \xi & \overset{\mathcal{D}}{\to} & \eta & \overset{\mathcal{D}_1}{\to} & \zeta \\ & \overset{ad(\mathcal{D})}{\leftarrow} & & \overset{ad(\mathcal{D}_1)}{\leftarrow} & \\ \nu & & \mu & & \lambda \end{array}$$

where the upper sequence is formally exact at η but the lower sequence is not formally exact at μ.

Passing to the module framework, we obtain the sequences:

$$\begin{array}{ccccccc} D & \overset{\mathcal{D}_1}{\to} & D^2 & \overset{\mathcal{D}}{\to} & D & \to M \to 0 & \\ D & \overset{ad(\mathcal{D}_1)}{\leftarrow} & D^2 & \overset{ad(\mathcal{D})}{\leftarrow} & D & & \end{array}$$

where the lower sequence is not exact at D^2.

Therefore, we have to prove that the extension modules vanish, that is $ad(\mathcal{D})$ generates the CC of $ad(\mathcal{D}_1)$ and, conversely, that $\mathcal{D}_1$ generates the CC of $\mathcal{D}$. We also remind the reader that it has not been easy to exhibit the CC of the Maxwell or Morera parametrizations when $n = 3$ and that a direct checking for $n = 4$ should be strictly impossible. It has been proved by L. P. Eisenhart in 1926 [49] that the solution space Θ of the Killing system has $n(n+1)/2$ *infinitesimal generators* $\{\theta_\tau\}$ linearly independent over the constants if and only if ω had constant Riemannian curvature, namely zero in our case. As we have a Lie group of transformations preserving the metric, the three theorems of Sophus Lie assert than $[\theta_\rho, \theta_\sigma] = c^\tau_{\rho\sigma}\theta_\tau$ where the *structure constants c* define a Lie algebra $\mathcal{G}$. We have therefore $\xi \in \Theta \Leftrightarrow \xi = \lambda^\tau\theta_\tau$ with $\lambda^\tau = cst$. Hence, we may replace locally the Killing system by the system $\partial_i\lambda^\tau = 0$, getting therefore the differential sequence:

$$0 \to \Theta \to \wedge^0T^* \otimes \mathcal{G} \xrightarrow{d} \wedge^1T^* \otimes \mathcal{G} \xrightarrow{d} \cdots \xrightarrow{d} \wedge^nT^* \otimes \mathcal{G} \to 0$$

which is the tensor product of the Poincaré sequence by $\mathcal{G}$. Finally, it follows from Proposition 3.3 that the extension modules do not depend on the resolution used and thus vanish because the Poincaré sequence is self adjoint (up to sign), that is $ad(d)$ generates the CC of $ad(d)$ at any position, exactly like d generates the CC of d at any position. This (difficult) result explains why the differential modules we have met were torsion-free or even reflexive. We invite the reader to compare with the situation of the Maxwell equations in electro-magnetisme (see [6], p. 492-494 for more details). However, we have explained in [6] [45]-[47] [51] why neither the Janet sequence nor the Poincaré sequence can be used in physics and must be replaced by the *Spencer sequence* which is another resolution of Θ [39] [40] [46].

EXAMPLE 6.8: *PD Control Theory Revisited*

Comparing with the Theorem allowing to construct a minimal parametrization, we started with $\mathcal{D}_1\eta = \zeta$ and computed $ad(\mathcal{D}_1)\lambda = \mu$ with generating CC $ad(\mathcal{D})\mu = \nu$, obtaining therefore finally the generating CC $ad(D_{-1})\nu = 0$, that is $\partial_2\nu^2 + \partial_1\nu^1 + x^2\nu^1 = 0$. In that case, $rk(L) = 1$ in the diagram providing the minimal parametrization. This result explains why we had two potentials (ξ^1, ξ^2) in the canonical parametrization and

only one, namely $\left(\xi^1=\xi,0\right)$ or $\left(0,\xi^2=\xi'\right)$, in the minimal parametrizations but it is not possible to imagine the underlying procedure.

EXAMPLE 6.9: *OD/PD Optimal Control Revisited*

Using the notations of the Formal Test 5.12, let us assume that the two differential sequences:

$$\begin{array}{ccccc} \xi & \overset{\mathcal{D}}{\rightarrow} & \eta & \overset{\mathcal{D}_1}{\rightarrow} & \zeta \\ & \overset{ad(\mathcal{D})}{\leftarrow} & & \overset{ad(\mathcal{D}_1)}{\leftarrow} & \\ \nu & & \mu & & \lambda \end{array}$$

are *formally exact*, that is $\mathcal{D}_1$ generates the CC of $\mathcal{D}$ *and* $ad\left(\mathcal{D}\right)$ generates the CC of $ad\left(\mathcal{D}_1\right)$, namely ξ is a potential for $\mathcal{D}_1$ *and* λ is a potential for $ad\left(\mathcal{D}\right)$. We may consider a variational problem for a cost function $\varphi\left(\eta\right)$ under the linear OD or PD constraint described by $\mathcal{D}_1\eta=0$.

- Introducing convenient Lagrange multipliers λ while setting $dx=dx^1\wedge\cdots\wedge dx^n$ for simplicity, we must vary the integral:

$$\Phi=\int\left[\varphi\left(\eta\right)+\lambda\mathcal{D}_1\eta\right]\mathrm{d}x\Rightarrow\delta\Phi=\int\left[\left(\partial\varphi\left(\eta\right)/\partial\eta\right)\delta\eta+\lambda\mathcal{D}_1\delta\eta\right]\mathrm{d}x$$

Integrating by parts, we obtain the EL equations:

$$\partial\varphi\left(\eta\right)/\partial\eta+ad\left(\mathcal{D}_1\right)\lambda=0$$

to which we have to add the constraint $\mathcal{D}_1\eta=0$ obtained by varying λ. If $ad\left(\mathcal{D}_1\right)$ is an injective operator, in particular if $\mathcal{D}_1$ is formally surjective (no CC) while $n=1$ and M is torsion-free or $n\geq 1$ and M is projective, then one can obtain λ *explicitly* and eliminate it by substitution ([7]). Otherwise, using the CC $ad\left(\mathcal{D}\right)$ of $ad\left(\mathcal{D}_1\right)$ in order to eliminate λ, we have to study the formal integrability of the combined system:

$$ad\left(\mathcal{D}\right)\partial\varphi\left(\eta\right)/\partial\eta=0,\quad\mathcal{D}_1\eta=0$$

which may be a difficult task as we already saw through the examples of the Introduction.

- *We may also transform the given variational problem with constraint into a variational problem without any constraint if and only if the differential constraint can be parametrized*. Using the parametrization of $\mathcal{D}_1$ by $\mathcal{D}$, we may vary the integral:

$$\Phi=\int\varphi\left(\mathcal{D}\xi\right)\mathrm{d}x\Rightarrow\delta\Phi=\int\left(\partial\varphi\left(\eta\right)/\partial\eta\right)\mathcal{D}\delta\xi\mathrm{d}x$$

whenever $\eta=\mathcal{D}\xi$ and integrate by parts for arbitrary $\delta\xi$ in order to obtain the EL equations:

$$ad\left(\mathcal{D}\right)\partial\varphi\left(\eta\right)/\partial\eta=0,\quad\eta=\mathcal{D}\xi$$

in a coherent way with the previous approach and the Poincaré duality *geometry* $\leftrightarrow$ *physics*.

As a byproduct, if the *field equations* $\mathcal{D}_1\eta=0$ can be parametrized by a *potential* ξ through the formula $\mathcal{D}\xi=\eta$, then the *induction equations* $ad\left(\mathcal{D}\right)\mu=\nu$ can be obtained by duality in a coherent way with the double duality test, ... *on the condition to know what sequence must be used*.

However, we have already proved in [45]-[47] [51] [52] that the *Cauchy stress equations* must be replaced by the *Cosserat couple-stress equations* and that the *Janet sequence* (only used in this paper) must be thus replaced by the *Spencer sequence*. Accordingly, it becomes clear that the work of Lanczos and followers has been based on a *double confusion* between fields and inductions on one side, but also between the Janet sequence and the Spencer sequence on the other side.

FUNDAMENTAL RESULT 6.10: The Janet and Spencer sequences for *any* Lie operator of finite type are formally exact by construction, both with their corresponding adjoint sequences. Lanczos has been trying to parametrize $ad\left(\mathcal{D}_1\right)$ by $ad\left(\mathcal{D}_2\right)$ when $\mathcal{D}_1$ parametrizes $\mathcal{D}_2$. On the contrary, we have proved that one must parametrize $ad\left(\mathcal{D}\right)$ by $ad\left(\mathcal{D}_1\right)$ when $\mathcal{D}$ parametrizes $\mathcal{D}_1$ as in the famous *infinitesimal equivalence problem* ([20], p. 332-336), *with a shift by one step*. This is also the *only* way which is coherent with the corresponding non-linear sequences and the *finite equivalence problem* [39] [46] [47] [50] [52] [53].

2. Conclusion

The effective usefulness of the double duality test seems absolutely magical in actual practice but the reader may

not forget about the amount of mathematics needed from different domains. Unhappily, in our opinion based on a long experience in dealing with applications, the most difficult part is concerned with formal integrability and involution needed in order to compute the various differential ranks involved. However, the above methods, though largely superseding the pioneering approaches of Janet and Cartan, are still not known in mechanics and in mathematical physics, particularly in general relativity or even in control theory despite many tentatives done twenty years ago. We hope that this paper will help to improve this situation in a near future, in particular when dealing with *partial differential optimal control*, which is with variational calculus with OD or PD constraints along the way that has been initiated by Lanczos for eliminating the corresponding Lagrange multipliers or using them as potentials while studying the mathematical foundations of general relativity.

References

[1] Bjork, J.E. (1993) Analytic D-Modules and Applications. Kluwer Academic Publishers, Dordrecht. http://dx.doi.org/10.1007/978-94-017-0717-6

[2] Kashiwara, M. (1995) Algebraic Study of Systems of Partial Differential Equations, Mémoires de la Société Mathématique de France, 63 (Transl. from Japanese of His 1970 Master's Thesis).

[3] Oberst, U. (1990) *Acta Applicandae Mathematicae*, **20**, 1-175. http://dx.doi.org/10.1007/BF00046908

[4] Oberst, U. (2013) M*echanical Systems and Signal Processing*, **26**, 389-404.

[5] Palamodov, V.P. (1970) Linear Differential Operators with Constant Coefficients, Grundlehren der Mathematischen Wissenschaften 168. Springer, Berlin. http://dx.doi.org/10.1007/BF00046908

[6] Pommaret, J.-F. (2001) Partial Differential Control Theory. Kluwer, Dordrecht, 957 p. http://dx.doi.org/10.1007/978-94-010-0854-9

[7] Pommaret, J.-F. and Quadrat, A. (1999) *Systems & Control Letters*, **37**, 247-260. http://dx.doi.org/10.1016/S0167-6911(99)00030-4

[8] Pommaret, J.-F. and Quadrat, A. (1999) *IMA Journal of Mathematical Control and Informations*, **16**, 275-297. http://dx.doi.org/10.1093/imamci/16.3.275

[9] Quadrat, A. and Robertz, R. (2014) *Acta Applicandae Mathematicae*, **133**, 187-234. http://dx.doi.org/10.1007/s10440-013-9864-x

[10] Zerz, E. (2000) Topics in Multidimensional Linear Systems Theory. Lecture Notes in Control and Information Sciences (LNCIS) 256. Springer, Berlin.

[11] Kalman, E.R., Yo, Y.C. and Narenda, K.S. (1963) *Contrib. Diff. Equations*, **1**, 189-213.

[12] Airy, G.B. (1863) *Philosophical Transactions of the Royal Society London*, **153**, 49-80. http://dx.doi.org/10.1098/rstl.1863.0004

[13] Rotman, J.J. (1979) An Introduction to Homological Algebra, Pure and Applied Mathematics. Academic Press, New York..

[14] Pommaret, J.-F. (2007) *Computational & Applied Mathematics*, **2**, 1-21.

[15] Beltrami, E. (1892) *Atti della Accademia Nazionale dei Lincei Rendiconti*, **5**, 141-142.

[16] Maxwell, J.C. (1870) *Transactions of the Royal Society of Edinburgh*, **26**, 1-40. http://dx.doi.org/10.1017/S0080456800026351

[17] Morera, G. (1892) *Atti della Reale Accademia dei Lincei*, **1**, 137-141+233.

[18] Teodorescu, P.P. (1972) *Acta Mechanica*, **14**, 103-118. http://dx.doi.org/10.1007/BF01184852

[19] Einstein, A. (1915) Die Feldgleichungen der Gravitation, Sitz. Preus. Akademie der Wissenschaften zu Berlin, Berlin, 844-847.

[20] Pommaret, J.-F. (1978) Systems of Partial Differential Equations and Lie Pseudogroups. Gordon and Breach, New York. (Russian Translation by MIR, Moscow, 1983)

[21] Lanczos, C. (1949) *Reviews of Modern Physics*, **21**, 497-502. http://dx.doi.org/10.1103/RevModPhys.21.497

[22] Lanczos, C. (1962) *Reviews of Modern Physics*, **34**, 379-389. http://dx.doi.org/10.1103/RevModPhys.34.379

[23] Lanczos, C. (1949) The Variation Principles of Mechanics. 4th Edition, Dover, New York,.

[24] Bampi, F. and Caviglia, G. (1983) *General Relativity and Gravitation*, **15**, 375-386. http://dx.doi.org/10.1007/BF00759166

[25] Edgar, S.B. (2003) *Journal of Mathematical Physics*, **44**, 5375-5385. http://dx.doi.org/10.1063/1.1619203

[26] Edgar, S.B. and Höglund, A. (1997) *Proceedings of the Royal Society of London A*, **453**, 835-851. http://dx.doi.org/10.1098/rspa.1997.0046

[27] Edgar, S.B. and Höglund, A. (2000) *General Relativity and Gravitation*, **32**, 2307. http://dx.doi.org/10.1023/A:1001951609641

[28] Edgar, S.B. and Senovilla, J.M.M. (2004) *Classical and Quantum Gravity*, **21**, L133. http://dx.doi.org/10.1088/0264-9381/21/22/L01

[29] Massa, E. and Pagani, E. (1984) *General Relativity and Gravitation*, **16**, 805-816. http://dx.doi.org/10.1007/BF00762934

[30] O'donnell, P. and Pye, H. (2010) *Electronic Journal of Theoretical Physics*, **24**, 327-350.

[31] Roberts, M.D. (1996) *Il Nuovo Cimento*, **B110**, 1165-1176. http://dx.doi.org/10.1007/BF02724607

[32] Kunz, E. (1985) Introduction to Commutative Algebra and Algebraic Geometry. Birkhäuser, Boston.

[33] Bourbaki, N. (1980) Eléments de Mathématiques, Algèbre, Ch. 10. Algèbre Homologique. Masson, Paris.

[34] Pommaret, J.-F. (2005) Algebraic Analysis of Control Systems Defined by Partial Differential Equations, in Advanced Topics in Control Systems Theory. Lecture Notes in Control and Information Sciences (LNCIS) 311, Chapter 5, Springer, Berlin, 155-223. http://dx.doi.org/10.1007/11334774_5

[35] Pommaret, J.-F. (2013) *Multidimensional Systems and Signal Processing*, **26**, 405-437. http://dx.doi.org/10.1007/s11045-013-0265-0

[36] Riquier, C. (1910) Les Systèmes d'Equations aux Dérivées Partielles. Gauthiers-Villars, Paris.

[37] Janet, M. (1920) *Journal de Mathématiques*, **8**, 65-151.

[38] Cartan, E. (1945) Les Systèmes Différentiels Extérieurs et Leurs Applications Géométriques. Hermann, Paris.

[39] Pommaret, J.-F. (1994) Partial Differential Equations and Group Theory, New Perspectives for Applications. Mathematics and Its Applications 293, Kluwer, Dordrecht.

[40] Spencer, D.C. (1965) Overdetermined Systems of Partial Differential Equations. *Bulletin of the American Mathematical Society*, **75**, 1-114.

[41] Macaulay, F.S. (1916) The Algebraic Theory of Modular Systems, Cambridge Tracts 19. Cambridge University Press, London.

[42] Schneiders, J.-P. (1994) *Bulletin de la Société Royale des Sciences de Liège*, **63**, 223-295.

[43] Pommaret, J.-F. (2011) *Journal of Symbolic Computation*, **46**, 1049-1069. http://dx.doi.org/10.1016/j.jsc.2011.05.007

[44] Weyl, H. (1918) Space, Time. Matter, Berlin.

[45] Pommaret, J.-F. (2013) *Journal of Modern Physics*, **4,** 223-239. http://dx.doi.org/10.4236/jmp.2013.48A022

[46] Eisenhart, L.P. (1926) Riemannian Geometry. Princeton University Press, Princeton.

[47] Pommaret, J.-F. (1988) Lie Pseudogroups and Mechanics. Gordon and Breach, New York.

[48] Pommaret, J.-F. (2012) Spencer Operator and Applications: From Continuum Mechanics to Mathematical Physics. In: Gan, Y., Ed., *Continuum Mechanics-Progress in Fundamentals and Engineering Applications*, InTech, Chapter 1. http://www.intechopen.com/books/continuum-mechanics-progress-in-fundamentals-and-engineering-applications http://dx.doi.org/10.5772/35607

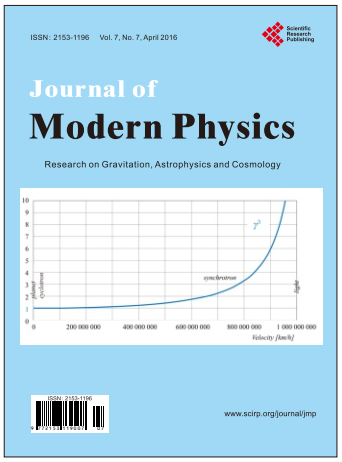
[49] Pommaret, J.-F. (2015) From Thermodynamics to Gauge Theory: The Virial Theorem Revisited. In: *Gauge Theories and Differential Geometry*, NOVA Science Publishers, Chapter 1, 1-44. http://arxiv.org/abs/1504.04118

[50] Pommaret, J.-F. (2014) *Journal of Modern Physics*, **5**, 157-170. http://dx.doi.org/10.4236/jmp.2014.55026

[51] Pommaret, J.-F. (2012) Deformation Cohomology of Algebraic and Geometric Structures. arXiv:1207.1964 [math.AP]

[52] Pommaret, J.-F. (2010) *Acta Mechanica*, **215**, 43-55. http://dx.doi.org/10.1007/s00707-010-0292-y

[53] Kumpera, A. and Spencer, D.C. (1972) Lie Equations, Ann. Math. Studies 73. Princeton University Press, Princeton.



Call for Papers

Journal of Modern Physics

ISSN: 2153-1196 (Print) ISSN: 2153-120X (Online)
http://www.scirp.org/journal/jmp

Journal of Modern Physics **(JMP)** is an international journal dedicated to the latest advancement of modern physics. The goal of this journal is to provide a platform for scientists and academicians all over the world to promote, share, and discuss various new issues and developments in different areas of modern physics.

Editor-in-Chief

Executive Editor-in-Chief

Subject Coverage

Journal of Modern Physics publishes original papers including but not limited to the following fields:

Biophysics and Medical Physics
Complex Systems Physics
Computational Physics
Condensed Matter Physics
Cosmology and Early Universe
Earth and Planetary Sciences
General Relativity
High Energy Astrophysics
High Energy/Accelerator Physics
Instrumentation and Measurement
Interdisciplinary Physics
Materials Sciences and Technology
Mathematical Physics
Mechanical Response of Solids and Structures
New Materials: Micro and Nano-Mechanics and Homogeneization
Non-Equilibrium Thermodynamics and Statistical Mechanics
Nuclear Science and Engineering
Optics
Physics of Nanostructures
Plasma Physics
Quantum Mechanical Developments
Quantum Theory
Relativistic Astrophysics
String Theory
Superconducting Physics
Theoretical High Energy Physics
Thermology

We are also interested in: 1) Short Reports—2-5 page papers where an author can either present an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data; 2) Book Reviews—Comments and critiques.

Notes for Intending Authors

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Paper submission will be handled electronically through the website. All papers are refereed through a peer review process. For more details about the submissions, please access the website.

Website and E-Mail

http://www.scirp.org/journal/jmp **E-mail: jmp@scirp.org**

What is SCIRP?

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

What is Open Access?

All original research papers published by SCIRP are made freely and permanently accessible online immediately upon publication. To be able to provide open access journals, SCIRP defrays operation costs from authors and subscription charges only for its printed version. Open access publishing allows an immediate, worldwide, barrier-free, open access to the full text of research papers, which is in the best interests of the scientific community.

- High visibility for maximum global exposure with open access publishing model
- Rigorous peer review of research papers
- Prompt faster publication with less cost
- Guaranteed targeted, multidisciplinary audience

Scientific Research Publishing

Website: http://www.scirp.org

Subscription: sub@scirp.org

Advertisement: service@scirp.org