Fermi and Astrophysics: Selected Papers with Commentaries and Translations

Edited by Remo Ruffini

with translations from Italian to English by

Emanuele Alesci, Donato Bini, Dino Boccaletti, Andrea Geralico,

Robert T. Jantzen, Simone Mercuri and Remo Ruffini

Preface

A meeting on Enrico Fermi and Astrophysics was held at the University of Rome “La Sapienza” and the ICRANet Center in Pescara in celebration of the hundredth anniversary of the birth of Enrico Fermi (1901–1954). During that anniversary year many events were organized covering the activities of Fermi in particle physics, nuclear physics, statistical mechanics and quantum statistics. Besides these fundamental fields of physics, amply documented in the existing literature, I thought Fermi had also played a key role by pioneering ideas which, directly or indirectly, became crucial for the understanding of some basic conceptual aspects of astrophysics and general relativity. This was the main focus of our meeting in Pescara, where a series of talks was presented dealing mainly with astrophysics, at the end of which I delivered a concluding lecture in the Aula Magna at the University of Rome “La Sapienza”: “Fermi, General Relativity, Astrophysics and Beyond.” The proceedings of that meeting were published as a special combined issue of *Il Nuovo Cimento B* [1].

I pointed out the paradoxical situation regarding a collaborative work by Fermi and Anthony L. Turkevich at the intersection of general relativity, cosmology and astrophysics: an article summarizing their findings was not published under the authors’ own names at the time but only later extensively quoted in a 1950 review written by others together with a declaration of its authenticity by the original authors. This unpublished Fermi-Turkevich article was not included in the collected papers of Fermi published in the West [2].

It has for the most part been ignored in the current scientific literature and in textbooks on cosmology and astrophysics. To the best of my knowledge, it has only been mentioned by Frank Wilczek in the opening talk at the Chicago celebration of Fermi’s 100th birthday. In the Russian edition of Fermi’s collected papers this article was included, thanks to the forceful request by Bruno Pontecorvo, as Bruno recalled to me many years later. Nevertheless this Fermi and Turkevich paper has indirectly greatly influenced developments in cosmology. It was well known to a small number of scientists and was certainly well known to Bob Dicke at Princeton, as I found out in 1968. Due to the beauty of its scientific approach, the numerical techniques adopted, and the importance of the results obtained, it has to be considered one of the fundamental contributions to relativistic cosmology, and since that time I have made a special effort to publicize it and assign it as mandatory reading for all my university students.

I then realized that a number of other articles by Fermi were equally insufficiently well known: a possible reason being that they had not yet been translated from Italian into English, especially those by the young Fermi when he was a student at the Scuola Normale Superiore in Pisa dealing mainly with electrodynamics and the special and general theories of relativity. This led to a lengthy process in which with the help of Emanuele Alesci, Donato Bini, Dino Boccaletti, Andrea Geralico, Robert Jantzen, and Simone Mercuri, we translated from Italian to English a selection of Fermi's papers, including the ones of the Pisa period. In the course of our work we also became aware of scientific results published in a series of six papers written by Fermi in 1922–1923 during his Pisa period while still a student and later in a temporary position at the University of Florence, results which he presented in Göttingen in 1924. This work, which has been overlooked in nearly all textbooks, is his solution of the infamous so called "4/3 problem" that plagued the classical theory of the electron introduced by Abraham and Lorentz during the first years of the life of special relativity and which was wrongly interpreted by Poincaré as due to unidentified internal stresses holding the electron together. I discussed this topic with Donato Bini, Andrea Geralico and Robert Jantzen over the period of a few years, resulting in our commentary article Appendix (A.1) and a shortened version (A.2) for the journal *General Relativity and Gravitation*.

While examining Fermi's early papers, we came across two important papers which we also translated. The first is a 1930 lecture delivered in Trento in which he clearly motivated his distrust toward approaching the internal constitutions of stars, an attitude which had negative consequences for the Italian development of astrophysics. The second was greatly rewarding: a crucial lecture that Fermi later delivered in Italian at the University of Rome in October 1949, "Theories on the origins of the elements," recorded by Ettore Pancini, which we have also translated into English. Through this I finally became aware of Fermi's deep knowledge of cosmology and derivation of the key equations, which allowed him to perform the computation in his work with Turkevich. There were also some other later papers related to astrophysics which, although they had been published in English, for a variety of reasons, had not yet reached the attention they deserved from the scientific community at large. We started assembling all of this material. Of course many books and even movies already exist which review the glorious achievements of the Fermi group in Rome on neutron physics, nuclear physics and statistical mechanics, but none of these overlap with our specific interest in the matter of general relativity and astrophysics. I first noticed with curiosity Fermi's apparent lack of interest in general relativity and also in astrophysics during the entire Rome period of this life. This was particularly surprising since many fundamental results were obtained in those years in England and in the United States which had great significance for astrophysics in the following decades. Many of the results were indeed obtained using Fermi's conceptual discoveries.

It became natural to ask why Fermi, one of the first scientists to reach a deep understanding of Einstein's theory of general relativity and to give profound contributions to that theory, already as a student in Pisa, never addressed any issue related to general relativity after transferring to Florence in 1924 and in 1926 to Rome. What could have happened during this Florence transition which inhibited his desire to pursue general relativity further?

While I was mulling over all these issues in the intervening years, I continued my work in relativistic astrophysics and was witnessing on a daily basis the tremendous relevance to the field of relativistic astrophysics of the classic work of three giants: Fermi, Einstein and Heisenberg. The greatest and most fundamental new results have come from the utilization of their ideas not in the isolation that they had created between themselves while alive but in a profound new interaction unhampered by their personal prejudices. From this thinking came the decision to contextualize this material with a companion book [3] dedicated to Einstein, Fermi, Heisenberg and the birth of relativistic astrophysics which took place due to both theoretical and observational advances that came one after the other in the 1960s, seen from my personal perspective as one of the participants in this story from its beginnings to the present time. I purposely avoided there entering into matters already extensively treated elsewhere, including in my own books, and have focused on a historical perspective regarding some particular events in the development of relativistic astrophysics which I have witnessed directly or have reconstructed in Rome, Princeton, Cambridge, Moscow and in locations where relativistic astrophysics after its inception flourished in the following years. I have privileged the indications on some current research which I consider particularly promising.

In the present volume the introductory Chapter 1 summarizes the contents of the remaining two chapters and appendices. We have reproduced and where necessary translated the fundamental contributions Fermi made which are relevant to astrophysics, starting from his early student days in Pisa (see Fig. 1) and continuing throughout his life. Chapter 2 contains those relevant papers from his Italian period before moving to America, followed by Chapter 3 which includes papers from his American period, including his paper on theories of element formation in the early universe from his 1949 Rome lecture as recorded by E. Pancini, and the Fermi-Turkevich work reproduced by Alpher and Herman. These are discussed in detail in the companion book. Appendix A contains some commentary articles regarding Fermi's early work in Italy, while Appendix B reproduces a selection of papers from the 2001 Meeting on Fermi and Astrophysics published in Nuovo Cimento in 2002.

In addition to remembering in this volume Fermi's contributions to fundamental physics starting from his student days in Pisa, continuing throughout his life, before closing, I recall here the influence Fermi had on science in China. This was commemorated in a special ceremony held in Beijing during the Fourth Galileo-Xu

Guangqi Meeting (GX4) in May 2015 (see Fig. 2) just preceding the Fourteenth Marcel Grossmann Meeting in Rome in July 2015. On that occasion both Fermi's former students C.N. Yang and T.D. Lee received Marcel Grossmann Awards (see Figs. 3, 4). Yang (see Fig. 5) then delivered a talk of his personal recollections of Fermi, including an exchange with Eugene Wigner, indicated by "W", as well as their final meeting in the hospital (accompanied oby Murray Gell'Mann) in the last minutes of Fermi's life. As the most unique Fermi reminiscence I have ever read and possibly the most touching words expressed by one human being for another, we reproduce them below.

— *Remo Ruffini*

Fig. 1 The young Enrico Fermi.

Fig. 2 The group photo for the GX4 Meeting (C.N. Yang and his wife at center of first row).

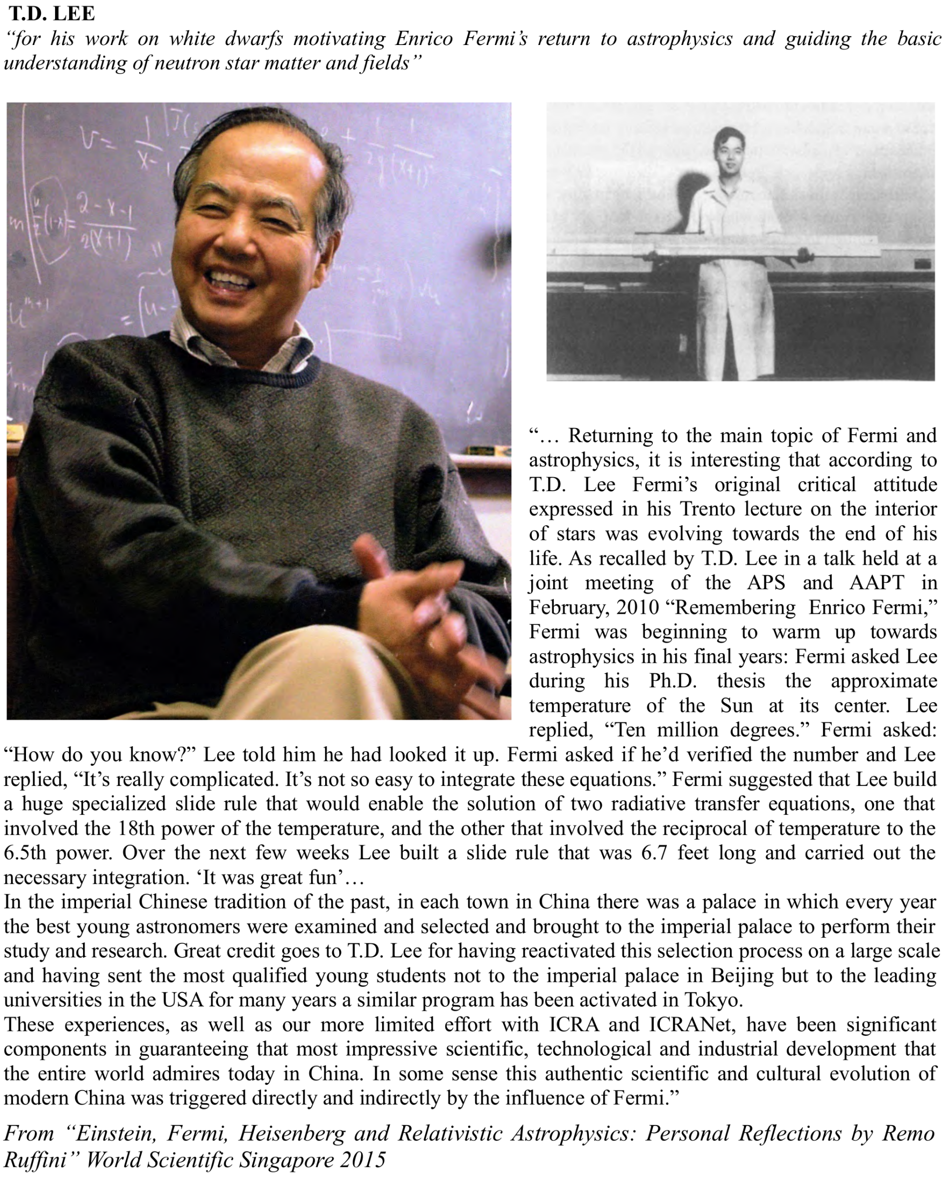
FRANK C.N. YANG

"for deepening Einstein's geometrical approach to physics in the best tradition of Paul Dirac and Hermann Weyl"

"… I would like to discuss some influence Fermi had in China: this is the case in which two of Fermi's Chinese students and collaborators had an unprecedented impact on science at the international level and triggered the scientific development of the largest nation in the world: China. During my second visit to China in 1979 I went to Kun Ming: it was quite an experience to see this beautiful location on the border of a lake so vividly described by Marco Polo. There was a train line constructed by the French reaching this town from Hanoi. There was also a beautiful university where two young students studied physics during World War II, there the professors from the Bei DA and Qing Hua university of Beijing and their families having escaped from the east of China ahead of the Japanese invasion. Their names were Chen Ning Yang and Tsung Dao Lee. At the end of the war they transferred to the USA: Frank C.N. Yang became Fermi's assistant and T.D. Lee was followed in his Ph.D. thesis by Fermi. The remarkable scientific career of these two young Chinese scientists is well recorded in the history of science. After Nixon's visit to China in 1972, Yang and Lee frequently went back to China to deliver lectures based on the Fermi tradition and today they are spending the greater part of their time in China organizing scientific centers and activities. In 1979 Yang gave a lecture at the second MG meeting in Trieste (see figure on the right: C.N. Yang speaking with a thoughtful Pam Dirac listening). During the Third Galileo-Xu Guangqi Meeting in 2011 I had another pleasant meeting with C.N. Yang. This also gave me the opportunity to see Beijing University again, having originally seen it in 1978 after the cultural revolution with all its libraries burned, now renewed and reaching a new splendor. Next to the Zhou Pei-Yuan Institute are the offices of the C.N. Yang Center. We talked about our common friend Isidor Rabi and his role in collaborating with Eisenhower as President of Columbia University prior to the latter's election as President of the USA. We also talked about Fermi's role in formulating his theory of beta decay, of the adventures of the A-bomb and H-bomb projects and many other topics. This also gave me the chance to introduce him to our ongoing projects with ICRANet in Brazil."

From "Einstein, Fermi, Heisenberg and Relativistic Astrophysics: Personal Reflections by Remo Ruffini" World Scientific Singapore 2015

Fig. 3 MG Awards booklet page [5] for C.N. Yang.



T.D. LEE

"for his work on white dwarfs motivating Enrico Fermi's return to astrophysics and guiding the basic understanding of neutron star matter and fields"

"… Returning to the main topic of Fermi and astrophysics, it is interesting that according to T.D. Lee Fermi's original critical attitude expressed in his Trento lecture on the interior of stars was evolving towards the end of his life. As recalled by T.D. Lee in a talk held at a joint meeting of the APS and AAPT in February, 2010 "Remembering Enrico Fermi," Fermi was beginning to warm up towards astrophysics in his final years: Fermi asked Lee during his Ph.D. thesis the approximate temperature of the Sun at its center. Lee replied, "Ten million degrees." Fermi asked: "How do you know?" Lee told him he had looked it up. Fermi asked if he'd verified the number and Lee replied, "It's really complicated. It's not so easy to integrate these equations." Fermi suggested that Lee build a huge specialized slide rule that would enable the solution of two radiative transfer equations, one that involved the 18th power of the temperature, and the other that involved the reciprocal of temperature to the 6.5th power. Over the next few weeks Lee built a slide rule that was 6.7 feet long and carried out the necessary integration. 'It was great fun'…

In the imperial Chinese tradition of the past, in each town in China there was a palace in which every year the best young astronomers were examined and selected and brought to the imperial palace to perform their study and research. Great credit goes to T.D. Lee for having reactivated this selection process on a large scale and having sent the most qualified young students not to the imperial palace in Beijing but to the leading universities in the USA for many years a similar program has been activated in Tokyo.

These experiences, as well as our more limited effort with ICRA and ICRANet, have been significant components in guaranteeing that most impressive scientific, technological and industrial development that the entire world admires today in China. In some sense this authentic scientific and cultural evolution of modern China was triggered directly and indirectly by the influence of Fermi."

From "Einstein, Fermi, Heisenberg and Relativistic Astrophysics: Personal Reflections by Remo Ruffini" World Scientific Singapore 2015

Fig. 4 MG Awards booklet page [5] for T.D. Lee.

Fig. 5 C.N. Yang receiving the Marcel Grossmann Award in Beijing at the GX4 in 2015.

Yang on Fermi

I remember that it was at the Second Marcel Grossman Meeting in Trieste in 1979, that I formulated the phrase "symmetry dictates interactions", which describes the principle that governs the structure of interactions. I am happy to receive this award from an organization based in Italy, the country I feel closest to, after China and the USA. Enrico Fermi was one of the great sons of Italy in her long history. Prometheus in Greek mythology, Suiren in Chinese mythology, taught mankind how to use chemical energy. Enrico Fermi in reality, taught mankind how to use nuclear energy.

Enrico Fermi was, of all the great physicists of the 20th century, among the most respected and admired. He was respected and admired because of his contributions to both theoretical and experimental physics, because of his leadership in discovering for mankind a powerful new source of energy, and above all, because of his personal character. He was always reliable and trustworthy. He had both of his feet on the ground all the time. He had great strength, but never threw his weight around. He did not play to the gallery. He did not practise one-up-manship. He exemplified, I always believe, the perfect Confucian gentleman.

Fermi from 1950 to 1951 was a member of the General Advisory Committee (GAC) of the Atomic Energy Committee (AEC) chaired by Oppenheimer. He then resigned with a quote:
"You know, I don't always trust my opinions about these political matters".

> Shakespeare's Sonnets No. 94
> *They that have power to hurt and will do none,*
> *That do not do the thing they most do show,*
> *Who, moving others, are themselves as stone,*
> *Unmoved, cold, and to temptation slow;*
> *They rightly do inherit heaven's graces,*
> *And husband nature's riches from expense;*
> *They are the lords and owners of their faces,*
> *Others but stewards of their excellence.*

In my years in Chicago, Fermi was personally very kind to me. I remember in June 1948, I had problems with the US Immigration Office. Fermi and Professor Allison , the Director of Chicago's Institute, went with me to the Immigration Office in Chicago. The Head of the office was so overwhelmed by the presence of Fermi that all my immigration problems were resolved immediately.

Fermi made many first rate contributions to physics. His contemporaries, including himself, considered his beta decay theory the most important. To bring out the great impact that paper had on physicists in the early 1930s, allow to me to tell you a story.

Y: What do you think was Fermi's most important contribution to theoretical physics?

W: Beta decay theory.
Y: How could that be? It is being replaced by more fundamental ideas. Of course it was a very important contribution which had sustained the whole field for some forty years: Fermi had characteristically swept what was unknowable at that time under the rug, and focused on what can be calculated. It was beautiful and agreed with experiment. But it was not permanent. In contrast the Fermi distribution is permanent.
W: No, no, you do not understand the impact it produced at the time. Von Neumann and I had been thinking about beta decay for a long time, as did everybody else. We simply did not know how to create an electron in a nucleus.
Y: Fermi knew how to do that by using a second quantized ψ?
W: Yes.
Y: But it was you and Jordan who had first invented the second quantized ψ?
W: Yes, yes. But we never dreamed that it could be used in real physics.

In the fall of 1954 Fermi was critically ill. Murray Gell-Mann and I went to the Billwigs Hospital to see him for a last time. He was thin, but not sad. He was reading a book full of stories about men who had succeeded, through shear will power, to overcome fantastic obstacles and misfortunes. As we bade goodbye and walked towards the door of his room, he said:
"Now I have to leave physics to your generation."
— Chen-Ning Franklin Yang

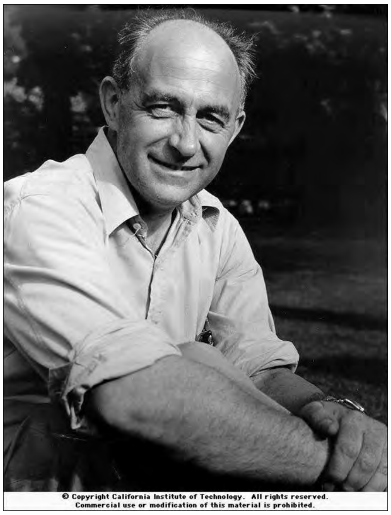


Fig. 6 Enrico Fermi (1901–1954).

Bibliography

[1] *Il Nuovo Cimento B* Vol. 117, Nos. 9–11, 2002; available online at:
`http://en.sif.it/journals/ncb/econtents/2002/117/09-11`,
Proceedings of the Ninth ICRA Network Workshop on Fermi and Astrophysics.

[2] *Enrico Fermi: Note e Memorie (Collected Papers)*, Accademia Nazionale dei Lincei and The University of Chicago Press, Vol. 1, 1961, Vol. 2, 1965; Volume 1 is available on-line at:
`http://www.archive.org/details/collectedpapersn007155mbp`.
Both volumes are available at the Accademia dei Lincei website:
`http://www.lincei.it/modules.php?name=Content&pa=showpage&pid=125`.

[3] *Einstein, Fermi, Heisenberg and the Birth of Relativistic Astrophysics*, Remo Ruffini, World Scientific, 2017.

[4] Rong-Gen Cai, Remo Ruffini, and Yue-Liang Wu, *Int. J. Mod. Phys. A* **30**, 1502005 (2015),
`http://www.worldscientific.com/toc/ijmpa/30/28n29`,
preface to the Proceedings of the 2015 Fourth Galileo-Xu Guangqi Meeting.

[5] The Fourth Marcel Grossmann Meeting Awards booklet
`http://www.icra.it/mg/mg14/mg14_awards.pdf`.

Contents

Chapter 1

Introduction

The present volume contains two chapters including translations and reproductions of key papers by Fermi relevant to astrophysics, together with three appendices of some historically relevant papers by other other authors and commentary on some of his articles.

1.1 Fermi's Italian Period

Chapter 2 contains the English translation of the papers originally published in Italian during Fermi's Pisa and Rome periods. The most famous of these introducing Fermi coordinates and Fermi transport (implicitly defining what later became known as Fermi-Walker transport, see Appendix B.2) was indeed a detour from Fermi's initial investigation of electromagnetic mass in special and general relativity that seems to have been largely ignored over the past ninety years. Credit for translation of Fermi's articles from Italian to English goes to: Emanuele Alesci for papers 4c) and 5), Donato Bini and Andrea Geralico for papers (1), (2), and (3), Dino Boccaletti for papers (7), (10), (12), (13), (30), (38) and (80a), and Simone Mercuri for paper (43), using the article labeling system from the two volume set of Fermi's collected works noted in the preface. Robert Jantzen edited these translations for English expression.

This section contains the English translations of a selection of papers from those Fermi published in Italian in the first part of his scientific career. The seminal papers selected are all related to relativity, astronomy and their applications. For a better account of the circumstances under which the papers were written, we also add excerpts of the presentations due to friends and collaborators of Fermi and published in Volume 1 of Fermi's *Note e Memorie*, 1961.

In paper FI 1 *On the Dynamics of a Rigid System of Electric Charges in Translational Motion* (1), Fermi calculates the inertial mass of a spherical distribution of charge with a constant acceleration by considering the reaction of the charge to its own average field. This leads to the formula $mc^2 = (4/3)U$ relating the inertial mass m to the classical electromagnetic energy U of the distribution. This value, in agreement with a calculation of the electromagnetic mass of a spherical homogeneous shell performed by Lorentz, contradicts the formula $mc^2 = U$ that one would expect from the principle of equivalence of mass and energy. Fermi considers the charge distribution at rest in a homogeneous gravitational field equal to the sign-reversed acceleration which appears to be in agreement with the relativistic formula. This topic is further examined in the subsequent article.

In paper FI 2 *On the Electrostatics of a Homogeneous Gravitational Field and on the Weight of Electromagnetic Masses* (2), Fermi reconsiders the calculation of the inertial mass of a spherical distribution of charge using for the first time general relativity, employing a Levi-Civita metric to describe a homogeneous gravitational field in the linear approximation. This approach has been expanded to what we now call today the Rindler metric.[1] His final result leads to the desired relation $mc^2 = U$. Another result derived in this paper is the value of the polarization of an infinitesimal conducting sphere at rest in a static gravitational field. An article by R. Ruffini[2] (see Appendix A.5) discusses some general relativistic developments that have taken place in the intervening years for describing electric charges in strong gravitational fields.

Paper FI 3 *On phenomena occurring close to a world Line* (3) is a classic result obtained by Fermi within the framework of general relativity expressing a system of space-time coordinates particularly suited to follow the behavior in time of phenomena happening in a small spatial region around the world line of a particle. Fermi explores the definition of the related coordinate transport which underlies it, later known as "Fermi transport," expressing the metric in the linear approximation for a general space-time. He also expresses Maxwell's equations in these coordinates, supporting the conclusions reached in the previous article.

The contribution by D. Bini and R. Jantzen (B.2) in Appendix B of this volume gives a summary of what we now call Fermi coordinates and Fermi transport with a historical update including Walker's contribution which led to the terminology of "Fermi-Walker transport." This article also discusses the geometry of the various relativistic contributions to the Fermi-Walker transport of vectors around circular orbits in black hole spacetimes and in their Minkowski limit.

In paper FI 4 *Correction of a Contradiction between Electrodynamic and Relativistic Electromagnetic Mass Theories* (4c), Fermi reconsiders the problem of the electromagnetic contribution to the mass of an elementary particle already discussed in the previous three articles. The discrepancy between the value $(4/3)(U/c^2)$, obtained by Lorentz for the inertial mass of a rigid, spherically symmetrical system of electric charges, and the value U/c^2 predicted by relativity was well known to Fermi from the previous articles. Such a discrepancy had been interpreted by Poincaré as due to the part of the stress-energy tensor contributed by internal non-

[1] See W. Rindler: *Essential relativity; special, general, and cosmological*, Van Nostrand Reinhold Co., 1969.

[2] *R. Ruffini: Charges in gravitational field: From Fermi, via Hanni-Ruffini-Wheeler, to the "electric Meissner effect", Nuovo Cimento* **119B**, 785–807, 2004.

electromagnetic stresses, whose existence was assumed to assure the equilibrium of the charged particles. A vast scientific literature of followers of this Poincaré's conjecture exists. Fermi shows that by assuming the accelerated charge distribution to be spherically symmetric in its rest frame instead of the laboratory frame, he obtains the correct inertial mass expected from the equivalence principle. This essentially reintroduces the crucial lapse factor between coordinate and physical components of the electric field which is responsible for the correction, to first order in the acceleration, of the approximation made in all of his "Fermi coordinate" system calculations. The results obtained by Fermi in this paper went unnoticed and for the most part remain that way today. Some of the crucial Fermi results in this paper and the historical developments of this most unique accident in physics are discussed in Appendix A by D. Bini, A. Geralico, R.T. Jantzen and R. Ruffini (see A.1) and by R.T. Jantzen and R. Ruffini in a brief summary of the key mathematics and their consequences (see A.2), as well as in a historical review by D. Boccaletti (see A.3).[3] Interestingly enough, related considerations were also put forward years later by B. Kwal without mentioning Fermi's work. Appendix A.4 reproduces this 1949 paper.

The paper FI 5 *Masses in the theory of relativity* (5)—a short contribution to a collective volume on the foundations of Einstein's theory of relativity—is evidence of the high reputation enjoyed by the young Fermi (age 22) in the physicists' community. Remarkable appears to be the prophetic premonition of things to come. A very favorable attitude toward Einstein theory by the young Fermi is clear at a time in which the older generation of Italian physicists was skeptical and hostile to relativity as recalled by Emilio Segré in Vol. 1, p. 33 of *Note e Memorie*.[4]

In paper FI 6 *On the mass of radiation in an empty space* (10), written in collaboration with Aldo Pontremoli, Fermi successfully applied the method used in FI 4 (4c) to the calculation of the mass of the radiation contained in a cavity with reflecting walls, for which the standard textbooks had an expression containing the same factor 4/3.

The papers FI 7 *The principle of adiabatics and the systems which do not admit angle coordinates* (12) and FI 8 *Some theorems of analytical mechanics of great importance for quantum theory* (13) are dedicated to the theory of adiabatic invariants. The interest of Fermi in the theory of adiabatic invariants, if we make reference to the published papers, goes from 1923 throughout 1926. As the other theoretical physicists in that period, he was convinced of the fundamental importance of the theory of adiabatic invariants for a rigorous formulation of quantum mechanics. On

[3] Boccaletti's review was written before the publication of the paper "The mass of the particles" by A. Bettini (*Rivista del Nuovo Cimento*, **32**, No. 7, 2009, pp. 295–337) where Fermi's priority in first resolving this problem is again noted and continuing ignorance of his result by many outstanding authors is recalled as well (see pp. 302–303).

[4] On this topic see also, e.g., Roberto Maiocchi: *Einstein in Italia—La scienza e la filosofia italiana fra le due guerre—Le Lettere*, Firenze, 1985.

the other hand, Max Born also shared the same opinion [5] .

Fermi also devoted a lecture in his university course on theoretical physics [6] to the theory of adiabatic invariants and he gave an elementary exposition of it in his book *Introduzione alla Fisica atomica.*[7] His interest was also awakened in conferences and seminars delivered at the University of Rome and in communications at the *Accademia Nazionale dei Lincei.* It was in those occasions that he sparked the interest of an outstanding listener: Tullio Levi-Civita.[8] The involvement of Levi-Civita was such that he, besides to giving a rigorous mathematical formulation of the subject,[9] also promoted astronomical applications of the theory. Those due to his collaborator Giulio Krall turned out to be particularly interesting. We must add that in those years James Jeans was also concerned with astronomical applications of the theory of adiabatic invariants.[10]

The paper FI 9 *A theorem of calculation of probability and some of its applications* (38b) is the second part of a Fermi's habilitation thesis to the "Scuola Normale Superiore" of Pisa (1922). It concerns the application of a theorem of calculation of probability to the dynamics of comets. The significance and the potentialities of this paper are well elucidated in the paper of C. Sigismondi and F. Maiolino (B.8) in Appendix B.

The paper FI 10 *Formation of images with Röntgen rays* (7) derives from a part of the degree thesis of Fermi at the University of Pisa. The thesis of Fermi was the most complete survey of X-rays physics in his time. He can also be considered a forerunner of techniques which are standard today. As Sigismondi and Mastroianni say in their article (B.9), although Fermi's seminal ideas are not among the sources investigated by Riccardo Giacconi and Bruno Rossi (1960) when they proposed a telescope using X-rays, Fermi's thesis was the most complete study of X-ray physics at his time. Fermi used the technique of 'mandrels' to form optical surfaces. He anticipated the technique used for the mirrors of Exosat, Beppo-SAX, Jet-X and XMM-Newton telescopes, which is now a mainstay of optical manufacturing. The paper by Sigismondi and Mastroianni discusses this noteworthy connection. It is appropriate here also to recall the comments of Franco Rasetti in the introduction of this article in Volume 1 of Fermi's Note e Memorie. Since at that time *"he had already published or at least completed several important theoretical papers, it may be asked why he did not present a theoretical thesis. It must be explained that at*

[5] See Max Born: *Vorlesungen über Atommechanik*, Berlin, 1925, pp. 58–67, 109-=114. English translation: *The mechanics of the atom*, London, 1927, pp. 52–59, 95–99.

[6] See A. De Gregorio, S. Esposito: Teaching theoretical Physics: The cases of Enrico Fermi and Ettore Majorana, Am. J. Phys. 75 (9), 781–790 (2007).

[7] Enrico Fermi: Introduzione alla Fisica Atomica, Zanichelli, Bologna, 1928, pp. 155–160.

[8] See P. Nastasi, R. Tazzioli: Tullio Levi-Civita, in Lettera Matematica pristem n. 57-58, Springer 2006

[9] We restrict ourselves to quote the last paper on the subject: T. Levi-Civita, A general survey of the theory of adiabatic invariants, *Journal of Math. and Physics*, Vol. 13, pp. 18–40 (1934).

[10] J.H. Jeans: Cosmogonic problems associated with a secular decrease of mass, MNRAS 85, 1, 2 (1924).
J.H. Jeans: The effect of varying mass on a binary system, MNRAS 85, 9, 912 (1925).

the time in Italy theoretical physics was not recognized as a discipline to be taught in universities, and a dissertation in that field would have been shocking at least to the older members of the faculty. Physicists were essentially experimentalists, and only an experimental dissertation would have passed as physics. The nearest subject to theoretical physics, mechanics, was taught by mathematicians as a field of applied mathematics, with complete disregard for its physical implications. These circumstances explain why such topics as the quantum theory had gained no foothold in Italy: they represented a "no man's land" between physics and mathematics. Fermi was the first in the country to fill the gap." (F. Rasetti, Vol. 1, pp. 55–56).

The paper FI 11 *On the quantization of an ideal monoatomic gas* (30) is the communication (to the Accademia Nazionale dei Lincei) in which Fermi expounds for the first time the statistical theory which will be named after him (together with P.A.M. Dirac). The enormous importance of the Fermi-Dirac statistics in astrophysics is recalled in Section 1.1 of Chapter 1.

In the following we give an excerpt from the presentation of Franco Rasetti *"... the present paper, probably his most famous theoretical contribution, where he formulated the theory of an ideal gas of particles obeying the Pauli exclusion principle, now designated in his honor as "fermion."*

There is conclusive evidence to show that Fermi had been concerned with the problem of the absolute entropy constant at least since January 1924, when he wrote a paper (Fermi 20) on the quantization of systems containing identical particles. He had also been discussing these problems with Rasetti several times in the following year. He told much later to Segré that the division of phase space into finite cells had occupied him very much and that had not Pauli discovered the exclusion principle he might have arrived at it a round-about way from the entropy constant (cfr. No. 20).

As soon as he read Pauli's article on the exclusion principle, he realized that he now possessed all the elements for a theory of the ideal gas which would satisfy the Nerst principle at absolute zero, give the correct Sackur-Tetrode formula for the absolute entropy in the limit of low density and high temperature, and be free of the various arbitrary assumptions that had been necessary to introduce in statistical mechanics in order to derive a correct entropy value. He does not seem to have been greatly influenced by Einstein's theory based on Bose's treatment of the black-body radiation as a photon gas, although he points out the analogy between the two forms of statistics. Apparently it took Fermi but a short time to develop the theory in the detailed and definitive form in which it was published in the German version." (F. Rasetti, Vol. 1, p. 178).

The paper FI 12 *A statistical method for the determination of some properties of the atom* (43), here translated, is the first of the papers Fermi devoted to the theory of what is today called the Thomas-Fermi atom. Fermi was unaware of the results previously reached by Thomas and his work went on independently for two years. Of great importance are the applications of the Thomas-Fermi model in astrophysics. He was, for example, quite familiar with the applications of his statistics (with the required relativistic modifications) to the theory of the structure of white dwarf stars: indeed, T.D. Lee, as a graduate student of Fermi, wrote his Ph.D. thesis on the *Hydrogen Content and Energy-Productive Mechanism of White Dwarf Stars* (*Ap. J.* **111**, 625, 1950). As we showed the general relativistic generalization of the Thomas-Fermi atom has recently led to a new theoretical framework to study both white dwarfs and neutron stars.

The paper FI 13 *An Attempt at a Theory of β Rays* (80a), translated here, can be described as the birth certificate of the theory of β-decay and weak interactions. Its importance is hardly questionable today. At that time (1933) things were not so easy (see Segré's report below).

"Fermi gave the first account of this theory to several of his Roman friends while we were spending the Christmas vacation of 1933 in the Alps. It was in the evening after a full day of skiing; we were all sitting on one bed in a hotel room, and I could hardly keep still in that position, bruised as I was after several falls on icy snow. Fermi was fully aware of the importance of his accomplishment and said that he thought he would be remembered for this paper, his best so far. He sent a letter to Nature advancing his theory but the editor refused it because he thought it contained speculations that were too remote from physical reality; and instead the paper ("tentative theory of beta rays") was published in Italian and in the Zeitschrift für Physik. Fermi never published anything else on this subject, although in 1950 he calculated matrix elements for beta decay as an application of the nuclear shell model." (Emilio Segré: *Enrico Fermi Physicists*, The University Chicago Press, 1970, p. 72).

In this paper there is the first mention to the possible existence of a massive neutrino.

1.2 Fermi's American Period

Chapter 3 reproduces some of Fermi's classic papers from his American period regarding the origin of cosmic rays and the mechanism of their acceleration, the interstellar magnetic field and its importance in astrophysics (in this field, Fermi was a pioneer), and the famous Fermi-Pasta-Ulam paper on nonlinear problems. Paper (240.3) is an article "The origin of the elements" of Fermi in Italian from Fermi's American period recorded by E. Pancini and translated by Dino Boccaletti, the third of nine lectures delivered in an Italian physics conference held in Rome and Milan in 1949, in response to Gamov's attempt to calculate the relative abundances of elements created in the early hot expanding universe. We also include the Fermi-Turkevich article which follows this same argument. A detailed discussion of the story behind these two papers and their relevance for relativistic cosmology can be found in the companion book *Einstein, Fermi, Heisenberg and the Birth of Relativistic Astrophysics* by Remo Ruffini.

We have selected seven of Fermi's papers from his American period to reproduce here, six of which are relevant to astrophysics. We have also added the famous paper "Study of non-linear problems." All of these have been quoted and commented on numerous times but we think that in order to have a clearer idea of their ideas, it is better to go back to the original sources. As in the preceding chapter, we also include some excerpts of commentary on those papers from Volume 2 of Fermi's *Note e Memorie.*

The first three papers, FA 1 *E. Fermi: On the Origin of the Cosmic Radiation* (237), FA 2 *E. Fermi: An Hypothesis on the Origin of the Cosmic Radiation* (238), FA 3 *E. Fermi: Galactic Magnetic Fields and the Origin of Cosmic Radiation* (265), tackle the problem of the origin of the cosmic rays formulating the hypothesis of a galactic origin and considering the role of the magnetic field. Comments on these papers can also be found in the Ames paper (B.1) in Appendix B.

As recalled by Anderson, *"Paper No. 237 was a direct outcome of heated disputes with Edward Teller on the origin of the cosmic rays. It was written to counter the view that cosmic rays were principally of solar origin and that they could not extend through all galactic space because of the very large amount of energy which would then be required. Taking up the study of the intergalactic magnetic fields, Fermi was able to find not only a way to account for the presence of the cosmic rays, but also a mechanism for accelerating them to the very high energies observed. He presented these same views on the origin of cosmic rays, though less extensively, in a talk at the Como International Congress on the Physics of Cosmic Rays (paper No. 238)."* (H.L. Anderson, Vol. 2, p. 655)

As Chandrasekhar recalls, *"In the fall of 1948, Edward Teller was advancing the view that cosmic rays are of solar origin. Fermi was want to say—half-jokingly—that this inspired him to take an opposing view and advocate a galactic origin of the cosmic rays."* (S. Chandrasekhar, Vol. 2, p. 924)

It is therefore appropriate to recall here Teller's point of view: *"Fermi mentioned to me his interest in the origin of cosmic rays as early as 1946. Several years before that time he mentioned the subject in some lectures in Chicago. He had the suspicion that magnetic fields could accelerate the cosmic particles. In 1948 Alfvén visited Chicago. He had been interested in electromagnetic phenomena on the cosmic scale for quite some time. At that time I was playing with the idea that cosmic rays might be accelerated in the neighborhood of the sun. I had discussed this question with Alfvén, and he visited us in Chicago in order to carry forward the discussion. During this visit Fermi learned from Alfvén about the probable existence of greatly extended magnetic fields in our galactic system. Since this field would necessarily be dragged along by the moving and ionized interstellar material, Fermi realized that here was an excellent way to obtain the acceleration mechanism for which he was looking. As a result he outlined a method of accelerating cosmic ray particles which serves today as a basis for most discussions on the subject. In his papers published in 1949 (No. 237 and 238) he explained most of the observed properties of cosmic rays with one important exception: it follows from his originally proposed mechanism that heavier nuclei will not attain as high velocities as protons do. This is in contradiction with experimental evidence. Fermi returned to this problem in his paper Galactic Magnetic Fields and the Origin of Cosmic Radiation (No. 264). Some details concerning the origin of cosmic rays have not been settled conclusively by Fermi's papers. Another competing theory has been proposed by Stirling Colgate and Montgomery Johnson according to which cosmic rays are produced by shock mechanism in exploding supernovae. The actual origin of cosmic rays continues to remain in doubt."* (E. Teller, Vol. 2, p. 655)

As Anderson recalls *"Fermi's interest in astrophysics was welcomed by the astrophysicists. They asked him to give the Sixth Henry Norris Russell Lecture of the American Astronomical Society. Fermi was quite pleased by this show of regard outside his own field and took the occasion to re-examine his earlier ideas about the origin of the cosmic rays in view of later developments in the knowledge of the strength and behavior of the magnetic fields."* (H.L. Anderson, Vol. 2, p. 970). (See also the introduction to paper No. 237.)

The paper FA 4 *E. Fermi: High energy nuclear events* (241) was published in the issue of the *Progress of Theoretical Physics* dedicated to the 15th anniversary of the Yukawa theory and considers a statistical description for pion production. As mentioned by Anderson in the comments to this paper in the collected work of Fermi,[11] the methods developed by Fermi were relatively simple, and moreover were deliberately simplified and therefore, were rather useful for experimentalists at that initial phase of high energy physics. Since pions are also bosons, at high energies when their rest mass can be neglected, the concept of temperature can be introduced and the energy density will be given by Stefan's law. Obtaining the temperature from the total energy within a given volume, the number densities of the produced

[11] *Fermi: Note e Memorie (Collected Papers)*, Vol. 2, 1965, p. 789.

pions and nucleons can then be estimated. The role played by thermalization in this paper has inspired us, even though the mechanism is different, namely astrophysical applications in the study of the spectra of gamma ray bursts (GRBs).

It is appropriate to recall here the comment of Isador Rabi in reaction to this paper as told by Anderson: *"Rabi's comment after hearing Fermi present this paper at an American Physical Society meeting in Chicago is worth recording here. 'If Fermi is right in saying that he can calculate what will happen at very high energies by purely statistical methods, then we will have nothing new to learn in this field.' Rabi should have had nothing to fear. Fermi's theory was greatly oversimplified as he intended it to be, and while it did not give very well the detailed results which were later found, it did serve as a standard against which one could make a first comparison of the experimental results of multiple production to reveal when something non-statistical was going on. In the later literature this made it appear that this theory was always wrong; a point that Fermi didn't enjoy at all. He had always stressed the purpose and limitations of his calculations and referred ironically to his own authority and to those who took his results beyond what he intended them to be."* (H.L. Anderson, Vol. 2, p. 789)

Fermi's theoretical papers rarely had co-authors. Among his few co-authors was Chandrasekhar, on two papers on magnetohydrodynamics, FA 5 *S. Chandrasekhar, E. Fermi: Magnetic Fields in Spiral Arms* (261) - FA 6 *S. Chandrasekhar, E. Fermi: Problems of Gravitational Stability in the Presence of a Magnetic Field* (262). Chandrasekhar's recollections on their joint work with remarkable details on Fermi's style of work are published in Volume 2 of Note e Memorie.[12] We give below some excerpts from them. D. Boccaletti comments on the two papers in an article (A.3) of Appendix A.

On paper 262) Chandrasekhar recalls: *"As I have already stated, Fermi and I discussed astrophysical problems regularly during 1952–53. The paper Problems of Gravitational Stability in the Presence of a Magnetic Field (No. 262) was an outcome of these discussions. Referring to this largely mathematical paper, several persons have remarked that it is "out of character" with Fermi. For this reason I may state that the problems which are considered in this paper were largely at Fermi's suggestion. The generalization of the virial theorem; the existence of an upper limit to the magnetic energy of a configuration in equilibrium under its own gravitation; the distortion of the spherical shape of a body in gravitational equilibrium by internal magnetic fields; the stabilization of the spiral arms of a galaxy by axial magnetic fields; all these were Fermi's ideas, novel at the time. But they had to be proved; for, as Fermi said: "It is so very easy to make mistakes in magneto-hydrodynamics that one should not believe in a result obtained after a long and complicated mathematical derivation if one cannot understand its physical origin; in the same way, one cannot also believe in a long and complicated piece of physical reasoning if one cannot demonstrate it mathematically." If only this dictum were followed by all!"* (S.

[12] *Fermi: Note e Memorie (Collected Papers)*, Vol. 2, 1965, p. 923–927.

Chandrasekhar, Vol. 2 p. 925)

And again Chandrasekhar: *"Fermi's interest in hydromagnetic turbulence led him to inquire into the physics of ordinary hydrodynamic turbulence. Confessing ignorance of this subject, Fermi asked me (early in 1950) to come to his office and tell him about the ideas of Kolmogoroff and Heisenberg which were then very much in the vogue. However, when I went to tell him, I found that it was not necessary for me to say beyond a few words: such as isotropy, the cascade of energy from large to small eddies etc. With only such words as clues, Fermi promptly went to the blackboard ("to see if I understand these words") and proceeded to derive the Kolmogoroff spectrum for isotropic turbulence (in the inertial range) and the basis of Heisenberg's elementary theory. Fermi's manner of arguing is worth recording for its transparent simplicity.*

Divide the scale of logk (where k denotes the wave number) into equal divisions, say $(\ldots, n, n+1, \ldots)$*. In a stationary state the rate of flow of energy across "n" must be equal to the rate of flow across "n + 1." Therefore:*

$$E_{n\,,n+1} = \rho \frac{v_n}{k_n}(v_n k_n)^2 - \rho \frac{v_{n+1}}{k_{n+1}}(v_{n+1}k_{n+1})^2\,, \tag{1}$$

if one remembers that the characteristic time associated with "eddies" with wave numbers in the interval $(n, n+1)$ *is* $(v_{n+1}k_{n+1})^{-1}$*. From this relation it follows that:*

$$v_n = Constant \times k_n^{-1/3}\,, \tag{2}$$

and this is equivalent to Kolmogoroff's law. For decaying turbulence, equation (1) should be replaced by:

$$\frac{d}{dt}(\rho v_n)^2 = E_{n\,,n+1} \tag{3}$$

and this equation expresses the content of Heisenberg's theory." (Chandrasekhar, Vol. 2, pp. 925–926)

The paper FA 7 *E. Fermi, J. Pasta, S. Ulam: Studies of Non-linear Problems* (266) (always quoted as F.P.U.) is outstanding for several reasons: (a) It represents the first computer study of a non-linear system; (b) the results contradicted the belief held since Poincaré, that any perturbed Hamiltonian system has to be chaotic. Fermi had considered it 'a little discovery' (as quoted by Ulam), thus immediately evaluating its extraordinary importance; (c) it was one of Fermi's last works, completed after his death in 1954; (d) remained unpublished for a decade; (e) coincides in time with Kolmogorov's theorem (1954), though FPU and Kolmogorov-Arnold-Moser (KAM) theory were linked to each other only in 1966; (f) inspired the discovery of solitons and numerous other studies; (g) its results are not fully understood till now and the FPU model continues its inspiring mission today, after half a century. In his recollections Ulam refers to Fermi's opinion on the importance of the "understanding of non-linear systems" for the future fundamental theories, and the "potentialities of the electronic computing machines" and even mentions

Fermi's learning of the actual coding (programming) during one summer. The FPU paper and its influence on various areas of astrophysics and stochastic dynamics are discussed in Appendix B (see the papers by A. Carati et al. (B.4), S. Ruffo (B.7) and G.M. Zaslavsky (B.11)). Here is the presentation written by S. Ulam.

"After the war, during one of his frequent summer visits to Los Alamos, Fermi became interested in the development and potentialities of the electronic computing machines. He held many discussions with me on the kind of future problems which could be studied through the use of such machines. We decided to try a selection of problems for heuristic work where in absence of closed analytic solutions experimental work on a computing machine would perhaps contribute to the understanding of properties of solutions. This could be particularly fruitful for problems involving the asymptotic-long time or "in the large" behavior of non-linear physical systems. In addition, such experiments on computing machines would have at least the virtue of having the postulates clearly stated. This is not always the case in an actual physical object or model where all the assumptions are not perhaps explicitly recognized.

Fermi expressed often a belief that future fundamental theories in physics may involve non-linear operators and equations, and that it would be useful to attempt practice in the mathematics needed for the understanding of non-linear systems. The plan was then to start with the possibly simplest such physical model and to study the results of the calculation of its long-time behavior. Then one would gradually increase the generality and the complexity of the problem calculated on the machine. The Los Alamos report LA–1940 (paper No. 266) presents the results of the very first such attempt. We had planned the work in the summer of 1952 and performed the calculations the following summer. In the discussions preceding the setting up and running of the problem on the machine we had envisaged as the next problem a two-dimensional version of the first one. Then perhaps problems of pure kinematics, e.g., the motion of a chain of points subject only to constraints but no external forces, moving on a smooth plane convoluting and knotting itself indefinitely. These were to be studied preliminary to setting up ultimate models for motions of system where "mixing" and "turbulence" would be observed. The motivation then was to observe the rates of mixing and "thermalization" with the hope that the calculational results would provide hints for a future theory. One could venture a guess that one motive in the selection of problems could be traced to Fermi's early interest in the ergodic theory. In fact, his early paper (No. 11a) presents an important contribution to this theory.

It should be stated here that during one summer Fermi learned very rapidly how to program problems for the electronic computers and he not only could plan the general outline and construct the so-called flow diagram but would work out himself the actual coding of the whole problem in detail. The results of the calculations (performed on the old MANIAC machine) were interesting and quite surprising to Fermi. He expressed to me the opinion that they really constituted a little discovery in providing intimations that the prevalent beliefs in the universality of "mixing and *thermalization" in non-linear systems may not be always justified.*

A few words about the subsequent history of this non-linear problem. A number of other examples of such physical systems were examined by calculations on the electronic computing machines in 1956 and 1957. I presented the results of the original paper on several occasions at scientific meetings; they seemed to have aroused considerable interest among mathematicians and physicists and there is by now a small literature dealing with this problem. The most recent results are due to N.J. Zabusky. (i) His analytical work shows, by the way, a good agreement of the numerical computations with the continuous solution up to a point where a discontinuity developed in the derivatives and the analytical work had to be modified. One obtains from it another indication that the phenomenon discovered is not due to numerical accidents of the algorithm of the computing machine, but seems to constitute a real property of the dynamical system.

In 1961, on more modern and faster machines, the original problem was considered for still longer periods of time. It was found by J. Tuck and M. Menzel that after one continues the calculations from the first "return" of the system to its original condition the return is not complete. The total energy is concentrated again essentially in the first Fourier mode, but the remaining one or two percent of the total energy is in higher modes. If one continues the calculation, at the end of the next great cycle the error (deviation from the original initial condition) is greater and amounts to perhaps three percent.[13] *Continuing again one finds the deviation increasing—after eight great cycles the deviation amounts to some eight percent; but from that time on an opposite development takes place! After eight more, i.e., sixteen great cycles altogether, the system gets very close better than within one percent to the original state! This supercycle constitutes another surprising property of our non-linear system."* (S.M. Ulam, Vol. 2, pp. 977–978)

Paper FA 8 *E. Fermi: Theories on the origin of the elements (240.3)* was a rough calculation of Fermi on the formation of the elements in the early hot big bang universe in response to Gamov's earlier attempt at solving this problem. It is followed by the later publication of the more detailed Fermi-Turkevich work on this problem, namely paper FA 9 *Fermi-Turkevich: An excerpt from "Theory of the origin and relative abundance distribution of the elements," by Ralph A. Alpher and Robert C. Herman.* These are discussed in detail in the companion book *Einstein, Fermi, Heisenberg and the Birth of Relativistic Astrophysics.*

[13] (i) Exact Solutions for the Vibrations of a non-linear continuous string. A. E. C. Research and Development Report, MATT-102, Plasma Physics Laboratory, Princeton University, October 1961.

1.3 Appendices

Appendix A includes some commentary articles on Fermi's resolution of this "4/3 problem" in the ratio between inertial mass and energy for the classical electron Coulomb field and a shorter journal article summarizing the natural completion of Fermi's original ideas about electromagnetic mass (see A.1–3), followed by a historical context commentary paper. We also reproduce the related article from 1949 by B. Kwal (see A.4) which seems to be the only one to touch upon this topic until the independent work of Rohrlich in 1960, after which Fermi's original contribution was rediscovered.

Appendix B contains a selection of the articles from the proceedings the meeting "Fermi and Astrophysics" organized at the University of Rome "La Sapienza" and at the ICRANet Center in Pescara October 3–6, 2001 and published in *Il Nuovo Cimento B* **117**, Nos. 9–11 (2002). The meeting was focused on the influence of Fermi on astrophysics and general relativity: his activities related to these topics were clustered at the beginning and end of his scientific career. These articles, selected because of their direct commentary on articles by Fermi or related applications of his ideas expressed in those articles, are presented in alphabetical order of their first authors.

Susan Ames discusses the historical background of Fermi's work on cosmic rays, along with current problems and further prospects for the physics of cosmic rays. In particular she points out how the frequently discussed ultra-high cosmic rays cannot be accelerated by the Fermi mechanism. Equipartition between the energy of matter and that of cosmic rays was among the initial points made by Fermi, and in that context Ames mentions also the role of the cosmic microwave background radiation.

Donato Bini and Robert Jantzen give a summary of Fermi's discussion of what we now call Fermi coordinates and Fermi transport with a historical update including Walker's contribution which led to the terminology of "Fermi-Walker transport." This article explicitly estimates the various relativistic contributions to the Fermi-Walker transport for vectors around circular orbits in black hole spacetimes and in their Minkowski limit.

Dino Boccaletti comments on the two papers which resulted from the collaboration of Fermi with Chandrasehkar (see papers 261, 262 of Chapter 4). The first paper is devoted to the study of light dispersion in the polarization plane and using the effect to derive the galactic magnetic field. The second paper contains the generalization of the virial theorem in the presence of a magnetic field. The commentary notes that Fermi was the first scientist to draw attention to the possible existence of a galactic magnetic field.

The review of Andrea Carati, Luigi Galgani, Antonio Ponno and Antonio Giorgilli is devoted to the equipartition problem in the Fermi-Pasta-Ulam paradox both in classical and quantum mechanics. Equipartition is discussed starting from Planck's work and Poincaré's theorem. Numerical results on the dependence of the existence of equipartition and the corresponding time scales on a certain critical energy are mentioned.

Piero Cipriani reviews the work of Fermi in the field of classical analytical mechanics. After a short historical introduction, he emphasizes some aspects of geometrical methods of the description of dynamics and the theory of stochastic differential equations. Interesting recollections on Fermi are quoted.

John G. Kirk reviews the Fermi acceleration mechanism in the context of galactic nuclei and gamma ray bursts, i.e., in processes involving relativistic motion. Diffusive and non-diffusive versions of Fermi's stochastic acceleration are considered, including those predicting a softer spectrum of accelerated particles. The appearance of anisotropy in the accelerated particles with increasing gamma factor is discussed for various astrophysical situations.

Stefano Ruffo reviews evidence for long relaxation time scales in Hamiltonian systems, and shows how complex and diverse is the dynamics of long-range systems. The 'quasi-states' of Fermi-Pasta-Ulam are discussed particularly in the context of two theoretical approaches developed by the author and collaborators, one based on the Vlasov-Poisson equation, and the other based on the averaging of fast oscillations.

Costantino Sigismondi and Francesca Maiolino review an early work by Fermi completed June 20, 1922, the year of his habilitation thesis on statistics at the Scuola Normale Superiore of Pisa, with an application to the case of comets. Fermi studied this case with a coplanar orbit to the one of Jupiter, neglecting the influence of other planets. The probability of ejection of the comet from the solar system (a parabolic or hyperbolic orbit) after interaction with Jupiter is calculated, as well as the probability of an impact with Jupiter. They apply Fermi's results to the case of the Earth in order to recover the time rate of collision of comets with our planet, which reliably produced the extinction of the dinosaurs. In this context the properties of the Oort cloud are discussed as well.

Costantino Sigismondi and Angelo Mastroianni recall that approximately in the same period Fermi studied the formation of X-ray images and presented his first experimental work as a dissertation at the University of Pisa in the spring of 1922. The need for Fermi to make an experimental essay was made mandatory since at that time theoretical physics was not yet considered sufficient to have independent validity. Although his seminal ideas are not among the bibliographical sources investigated by Riccardo Giacconi and Bruno Rossi (1960) when they proposed a telescope using X-rays, Fermi's thesis was the most complete study of X-ray physics in his time. Fermi used the technique of 'mandrels' to form optical surfaces. He anticipated the technique used for the mirrors of the Exosat, Beppo-SAX, Jet-X and XMM-Newton telescopes, a technique which is now a mainstay of optical manufacturing.

Alexei Yu. Smirnov reviews the neutrino flavor transformations in matter, as one of the authors of the original theoretical predictions and related observable effects. In particular, the Sudbury Neutrino Observatory results provide strong evidence of the neutrino flavor conversion. Neutrino conversion is discussed also in the context of supernova neutrinos and the corresponding predictions for the fluxes and energies at the Earth, including the role of the Earth matter effect. The author shows that the data of SN1987 can also be explained by the neutrino oscillations in the matter of Earth as conversions of muon and tau antineutrinos.

George M. Zaslavsky reviews the Fermi-Pasta Ulam problem with an attempt to find the transition from regular to chaotic dynamics. The Fermi acceleration mechanism is considered as a precursor of the Fermi-Pasta-Ulam problem. The Kepler map introduced by Roald Sagdeev and George Zaslavsky and several other problems are considered, demonstrating the role of the Fermi-Pasta-Ulam work in the discretization methods of differential equations and in the study of chaotic systems when the Lyapunov exponent method is not efficient.

Chapter 2

From Fermi's papers of the Italian period

1) On the dynamics of a rigid system of of electric charges on translational motion

"Sulla dinamica di un sistema rigido di cariche
elettriche in moto traslatorio,"
Nuovo Cimento ***22****, 199–207 (1921).*

§ 1.– When a rigid system of electric charges moves arbitrarily, the electric field it generates is different from that which Coulomb's law would predict. Now, the electric field produced by the entire system exerts some forces on each element of charge of the system. The resultant of these forces, namely the resultant of the internal electric forces, would of course be identically zero if Coulomb's law were valid, but it no longer is, however, at least in general, when the system moves, since in such a case that law is no longer valid.

This resultant gives the electromagnetic inertial reaction, and the aim of the present work is precisely its evaluation in the case of an arbitrary system in translational motion. In the case in which the system is a spherical distribution of surface electricity, as it is assumed in most electronic models, it is known that one finds [1] that such a resultant, at least in the first approximation, is given by

$$-\frac{2e^2}{3\mathrm{R}c^3}\,\mathbf{\Gamma}+\frac{2e^2}{3c^2}\,\dot{\mathbf{\Gamma}}\ , \tag{1}$$

where e, R denote the total charge and the radius of the system, c is the speed of light, $\mathbf{\Gamma}$ and $\dot{\mathbf{\Gamma}}$ are the acceleration and its derivative with respect to time. For quasi-stationary motions the second term of (1) becomes negligible, so that (1) reduces to

$$-m\mathbf{\Gamma}\ , \tag{2}$$

where m is the elecromagnetic mass.

In § 2 one finds the generalization of (1) to the case of any system, referring for example to molecular models, always assuming that the velocity is negligible with respect to the speed of light. If F_i $(i = 1, 2, 3)$ are the components of the resultant in question, one finds

$$\mathrm{F}_i=-\sum_k m_{ik}\Gamma_k+\sum_k \sigma_{ik}\dot{\Gamma}_k\ , \tag{3}$$

where m_{ik}, σ_{ik} are some quantities depending on the properties of the system. Therefore one can no longer refer to a scalar electromagnetic mass, but instead in its place one introduces the tensor m_{ik}.

§ 3 is devoted to the dynamical study of the law for quasi-stationary motions:

$$\mathrm{K}_i=\sum_k m_{ik}\Gamma_k\ , \tag{4}$$

[1] Richardson, *Electron Theory of Matter*, Chapter XIII. The difference between my formulas and those of Richardson is due to the fact that he adopts Heaviside units.

where K_i are the components of the external force. One shows that with such a law the fundamental kinetic energy theorem and Hamilton's principle continue to hold.

Finally in § 4 the law (4) for quasi-stationary motions, which holds only for small velocities, is generalized to the case of arbitrary velocity using special relativity.

With this the study of electromagnetic masses as inertial masses will be complete. In a forthcoming paper I will consider electromagnetic masses as masses endowed with weight from the point of view of the general theory of relativity.

§ 2.– It is known [2] that the electric force due to a point charge 1 in motion is the sum of two forces, which by assuming the ratio between the velocity v of the particle and the speed c of light to be negligible, are: the first one, $\mathbf{E}_1$, the force given by Coulomb's law; the second one $\mathbf{E}_2$ has the form

$$\mathbf{E}_2 = \frac{\mathbf{\Gamma}^* \cdot \mathbf{a}}{c^2 r}\, a - \frac{1}{c^2 r}\,\mathbf{\Gamma}^* \ . \tag{5}$$

In this formula r represents the distance between the particle M and the point P at which the force is calculated and $\mathbf{a}$ is a vector of magnitude 1 and orientation MP. Finally $\mathbf{\Gamma}^*$ is the acceleration of the particle at the time $t-(r/c)$. If instead of the charge 1 at M there is the charge $\rho\, d\tau$ (ρ is the electric density, $d\tau$ the volume element), the force at P will be $\rho\, d\tau(\mathbf{E}_1+\mathbf{E}_2)$, so that the force exerted at P by all charges will be $\int_\tau \rho(\mathbf{E}_1+\mathbf{E}_2)\, d\tau$, where the integration must be extended over the whole space τ occupied by charges. Now if at the point P there is the charge $\rho'\, d\tau'$, the force $\rho'\, d\tau' \int_\tau \rho(\mathbf{E}_1+\mathbf{E}_2)\, d\tau$ is acting on it.

The force acting on the entire system is therefore

$$\mathbf{F} = \iint \rho\rho'(\mathbf{E}_1+\mathbf{E}_2)\, d\tau\, d\tau' \ ,$$

where the two integrations must be extended over the same domain. On the other hand, one clearly has

$$\iint \rho\rho'\mathbf{E}_1\, d\tau\, d\tau' = 0 \ ,$$

so that

$$\mathbf{F} = \iint \rho\rho'\mathbf{E}_2\, d\tau\, d\tau' \ .$$

If we now denote by $\mathbf{\Gamma}$ and $\dot{\mathbf{\Gamma}}$ the acceleration and its derivative with respect to time, at the time t, if r is small enough, we can set

$$\mathbf{\Gamma}^* = \mathbf{\Gamma} - \frac{r}{c}\,\dot{\mathbf{\Gamma}} \ ,$$

obtaining finally

$$\mathbf{F} = \iint \left(\frac{\mathbf{\Gamma}\cdot\mathbf{a}}{c^2 r}\, a - \frac{\mathbf{\Gamma}}{c^2 r}\right)\rho\rho'\, d\tau\, d\tau' + \iint \left(\frac{\dot{\mathbf{\Gamma}}\cdot\mathbf{a}}{c^3}\, a - \frac{\dot{\mathbf{\Gamma}}}{c^3}\right)\rho\rho'\, d\tau\, d\tau' \ . \tag{6}$$

[2] See, e.g., RICHARDSON, op. cit.

We denote orthogonal Cartesian coordinates by x_1, x_2, x_3, and let (x_i) be the coordinates of M, (x'_i) those of P. The components of $\mathbf{a}$ are $a_i = \dfrac{x'_i - x_i}{r}$. Writing (6) in scalar form, one thus obtains

$$\mathrm{F}_i = -\sum_k m_{ik}\Gamma_k + \sum_k \sigma_{ik}\dot{\Gamma}_k \ , \tag{7}$$

noting that, under the assumption of translational motion, Γ_i and $\dot{\Gamma}_i$ are constant when the integration is performed.

Here one has set:

$$\begin{cases} m_{ii} = \dfrac{2\mathrm{U}}{c^2} - \displaystyle\iint \dfrac{\rho\rho'(x'_i - x_i)^2}{c^2 r^3}\, d\tau\, d\tau' \ , \\ m_{ik} = m_{ki} = -\displaystyle\iint \dfrac{\rho\rho'(x'_i - x_i)(x'_k - x_k)}{c^2 r^3}\, d\tau\, d\tau' \ , \qquad i \neq k \ , \end{cases} \tag{8}$$

$$\begin{cases} \sigma_{ii} = \dfrac{e^2}{c^3} - \displaystyle\iint \dfrac{\rho\rho'(x'_i - x_i)^2}{c^3 r^2}\, d\tau\, d\tau' \ , \\ \sigma_{ik} = \sigma_{ki} = -\displaystyle\iint \dfrac{\rho\rho'(x'_i - x_i)(x'_k - x_k)}{c^3 r^2}\, d\tau\, d\tau' \ , \qquad i \neq k \ . \end{cases} \tag{9}$$

In these formulae U represents the electrostatic energy of the system $= \dfrac{1}{2}\displaystyle\iint \dfrac{\rho\rho'}{r}\, d\tau\, d\tau'$, and e the total electric charge $= \displaystyle\int \rho\, d\tau = \int \rho'\, d\tau'$.

From the expressions (8), (9) it immediately follows that if the axes (x_i) are substituted by others (y_i) using the orthogonal substitution

$$y_i = \sum_k \alpha_{ik} x_k \ ,$$

the m_{ik} and σ_{ik} corresponding to the new axes are given by:

$$m'_{ik} = \sum_{rs} \alpha_{ir}\alpha_{ks} m_{ik} \ ,$$

$$\sigma'_{ik} = \sum_{rs} \alpha_{ir}\alpha_{ks} \sigma_{ik} \ .$$

Hence both m_{ik} and σ_{ik} are symmetric covariant tensors. Each of them will have three orthogonal principal directions such that, taking the axes to be parallel to them, one has either $m_{ik} = 0$ or $\sigma_{ik} = 0$ when $i \neq k$.

The principal axes of tensors m, σ, however, will be different in general. If the case that the system has spherical symmetry one can do the integrations (8) and (9), since instead of $\dfrac{(x'_i - x_i)(x'_k - x_k)}{r^2}$ one can put the mean value of this expression over all possible directions MP, since in this case to the two points MP correspond an infinite number of pairs which differ only by orientation. Now, this mean value if $i = k$ is given by $\dfrac{2\pi}{4\pi}\displaystyle\int_0^\pi \cos^2\theta \sin\theta\, d\theta$; if instead $i \neq k$, it is zero.

So one then has

$$m_{11} = m_{22} = m_{33} = \frac{4\mathrm{U}}{3c^2} \ ; \qquad m_{23} = m_{31} = m_{12} = 0 \ ;$$

$$\sigma_{11} = \sigma_{22} = \sigma_{33} = \frac{2}{3}\frac{e^2}{c^3} \ ; \qquad \sigma_{23} = \sigma_{31} = \sigma_{12} = 0 \ .$$

By substituting these values into (7), one obtains well known formulas if the system consists of a homogeneous spherical layer.

§ 3.– Returning to the general case, we note that for quasi-stationary motions (5) can be replaced by:

$$\mathrm{F}_i = -\sum_k m_{ik}\Gamma_k \ .$$

If one thinks of an external force (X_i) acting on the system, the total force will be ($\mathrm{X}_i + \mathrm{F}_i$). If one now supposes that the system has no material mass one must have $\mathrm{X}_i + \mathrm{F}_i = 0$, and so

$$\mathrm{X}_i = \sum_k m_{ik}\Gamma_k \ . \tag{10}$$

It is easy to show how with the law (10) the principle of the kinetic energy theorem and of Hamilton's principle are preserved. In fact, denoting the velocity by $\mathrm{V} \equiv (\mathrm{V}_1, \mathrm{V}_2, \mathrm{V}_3)$ and multiplying (10) by V_i, then summing with respect to i one obtains

$$\sum_i \mathrm{X}_i\mathrm{V}_i = \sum_{ik} m_{ik}\mathrm{V}_k\frac{d\mathrm{V}_i}{dt} \ .$$

Interchanging i and k in the second sum, and noting that $m_{ik} = m_{ki}$

$$\sum_i \mathrm{X}_i\mathrm{V}_i = \sum_{ik} m_{ik}\mathrm{V}_i\frac{d\mathrm{V}_k}{dt} \ ,$$

and summing

$$2\sum_i \mathrm{X}_i\mathrm{V}_i = \sum_{ik} m_{ik}\left(\mathrm{V}_i\frac{d\mathrm{V}_k}{dt} + \mathrm{V}_k\frac{d\mathrm{V}_i}{dt}\right) = \frac{d}{dt}\sum_{ik} m_{ik}\mathrm{V}_i\mathrm{V}_k \ .$$

The first left hand side is twice the potential P of the external forces. Thus one has

$$P = \frac{dT}{dt} \ , \qquad \text{where} \qquad T = \frac{1}{2}\sum_{ik} m_{ik}\mathrm{V}_i\mathrm{V}_k \ . \tag{11}$$

Multiplying, instead, the two sides of (10) by δx, and then summing, one similarly gets

$$\sum_i \mathrm{X}_i\delta x_i = \frac{1}{2}\sum_{ik} m_{ik}\left(\frac{d^2x_k}{dt^2}\,\delta x_i + \frac{d^2x_i}{dt^2}\,\delta x_k\right)$$

$$= \frac{d}{dt}\left\{\frac{1}{2}\sum_{ik} m_{ik}(\dot{x}_k\delta x_i + \dot{x}_i\delta x_k)\right\} - \frac{1}{2}\sum_{ik} m_{ik}(\dot{x}_k\delta\dot{x}_i + \dot{x}_i\delta\dot{x}_k)$$

$$= \frac{d}{dt}\left\{\frac{1}{2}\sum_{ik} m_{ik}(\dot{x}_k\delta x_i + \dot{x}_i\delta x_k)\right\} - \delta T \ .$$

Multiplying by dt and integrating between two limits t', t'' at which the variations δx_i are assumed to be zero, one obtains

$$\int_{t'}^{t''} \left(\delta T + \sum_i \mathrm{X}_i \delta x_i \right) = 0 \ , \tag{12}$$

expressing Hamilton's principle.

If one refers to the principal axes of the tensor m_{ik} instead of arbitrary ones, (10) takes the simple form:

$$\mathrm{X}_i = m_{ii}\Gamma_i \ . \tag{13}$$

§ 4. – This formula holds only if V/c is negligible. To generalize it to an arbitrary velocity let us denote by $\mathrm{S} \equiv (x_1, x_2, x_3, t)$ the indicated reference frame and by $\mathrm{S}^* \equiv (x, y, z, t)$ a frame fixed with respect to S with the $x-$axis orientated along the velocity of the system at a certain fixed but generic time $\bar{t}$, and finally let $\mathrm{S}' \equiv (x', y', z', t')$ be a system with spatial axes parallel to xyz which moves uniformly with respect to S^* with velocity equal to that of the moving one at time $\bar{t}$, whose magnitude is v. One will have

$$t' = \beta \left(t - \frac{v}{c^2}\, x \right) \ ; \quad x' = \beta\, (x - vt) \ ; \quad y' = y \ ; \quad z' = z \ ; \quad \beta = \frac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}} \ , \tag{14}$$

where, once $\bar{t}$ is fixed, v and hence β are constant.

Let us asumme that the forces acting on our system are due to an external electromagnetic field ($\mathbf{E}$, $\mathbf{H}$); since at the instant t the system has velocity zero with respect to S', (10) will hold for it, and so one will therefore have, with an obvious meaning for the symbols:

$$\begin{aligned} e\,\mathrm{E}'_x &= m_{xx}\Gamma'_x + m_{xy}\Gamma'_y + m_{xz}\Gamma'_z \\ e\,\mathrm{E}'_y &= m_{yx}\Gamma'_x + m_{yy}\Gamma'_y + m_{yz}\Gamma'_z \\ e\,\mathrm{E}'_z &= m_{zx}\Gamma'_x + m_{zy}\Gamma'_y + m_{zz}\Gamma'_z \ . \end{aligned}$$

But one has

$$e\,\mathrm{E}'_x = e\,\mathrm{E}_x \ , \qquad e\,\mathrm{E}'_y = e\beta \left(\mathrm{E}_y - \frac{v}{c}\,\mathrm{H}_z \right) \ , \qquad e\,\mathrm{E}'_z = e\beta \left(\mathrm{E}_z + \frac{v}{c}\,\mathrm{H}_y \right) \ .$$

So therefore setting

$$\mathbf{k} = e \left(\mathrm{E} + \frac{1}{c}\,\mathrm{V} \times \mathrm{H} \right) \ , \tag{15}$$

one finds

$$e\,\mathrm{E}'_x = e\,\mathrm{E}_x \ , \qquad e\,\mathrm{E}'_y = e\beta k_y \ , \qquad e\,\mathrm{E}'_z = e\beta k_z \ .$$

On the other hand:

$$\Gamma'_x = \frac{d^2x'\,dt' - d^2t'\,dx'}{dt'^3} \ ,$$

but at time $\bar{t}$, $\dfrac{dx'}{dt'} = 0$, hence $\Gamma'_x = \dfrac{d^2x'}{dt'^2}$. Taking t as the independent variable, and noting that $\dfrac{dx}{dt} = v$, then $\Gamma'_x = \beta^3\Gamma_x$. Analogously, $\Gamma'_y = \beta^2\Gamma_y$ and $\Gamma'_z = \beta^2\Gamma_z$. Substituting

$$\begin{cases} k_x = m_{xx}\beta^3\ddot{x} + m_{xy}\beta^2\ddot{y} + m_{xz}\beta^2\ddot{z} \\ k_y = m_{yx}\beta^3\ddot{x} + m_{yy}\beta\ddot{y} + m_{yz}\beta\ddot{z} \\ k_z = m_{zx}\beta^3\ddot{x} + m_{zy}\beta\ddot{y} + m_{zz}\beta\ddot{z}\ . \end{cases} \tag{16}$$

Denoting by α_{xi} the cosine of the angle between the $x-$axis and the x_i-axis, one has

$$k_i = \alpha_{xi}k_x + \alpha_{yi}k_y + \alpha_{zi}k_z\ .$$

On the other hand, being m_{i0} covariant, one has for instance

$$m_{xy} = \sum_r m_{rr}\alpha_{xr}\alpha_{yr}\ .$$

Analogously

$$\ddot{x} = \sum_j \ddot{x}_j\alpha_{xj}\ .$$

Multiplying then (16) by $\alpha_{xi}, \alpha_{yi}, \alpha_{zi}$ and summing, one finds

$$k_i = \sum_{rj} m_{rr}\ddot{x}_j \begin{bmatrix} \beta^3\alpha^2_{xr}\alpha_{xj}\alpha_{xi} + \beta^2\alpha_{xr}\alpha_{yr}\alpha_{yj}\alpha_{xi} + \beta^2\alpha_{xr}\alpha_{zr}\alpha_{zj}\alpha_{xi} \\ +\beta^2\alpha_{yr}\alpha_{xr}\alpha_{xj}\alpha_{yi} + \beta\alpha^2_{yr}\alpha_{yj}\alpha_{yi} + \beta\alpha_{yr}\alpha_{zr}\alpha_{zj}\alpha_{yi} \\ +\beta^2\alpha_{zr}\alpha_{xr}\alpha_{xj}\alpha_{zi} + \beta\alpha_{zr}\alpha_{yr}\alpha_{yj}\alpha_{zi} + \beta\alpha^2_{zr}\alpha_{zj}\alpha_{zi} \end{bmatrix}\ .$$

But one has $\alpha_{xi} = \dfrac{\dot{x}_i}{v}$. Taking into account the relations between the α's, one finally finds the sought after generalization of (13)

$$\begin{aligned} k_i = \beta\sum_{rj}\ddot{x}_j m_{rr}\Bigg\{&(\beta-1)^2\,\frac{\dot{x}_i\dot{x}_j\dot{x}_r^2}{v^4} \\ &+(\beta-1)\left[(jr)\,\frac{\dot{x}_i\dot{x}_r}{v^2} + (ir)\,\frac{\dot{x}_j\dot{x}_r}{v^2}\right] + (ir)\,(jr)\Bigg\}\ , \end{aligned} \tag{17}$$

where

$$(jr) = 1\ ,\quad \text{if}\quad j = r\ ;\qquad (jr) = 0\ ,\quad \text{if}\quad j \neq r\ .$$

In the case of spherical symmetry, setting $m_{11} = m_{22} = m_{33} = m$, one can evaluate the sum in (17), finding:

$$k_i = \beta m\ddot{x}_i + m\beta(\beta^2-1)\,\frac{\dot{x}_i}{v^2}\sum_j \dot{x}_j\ddot{x}_j\ ,$$

from which, recalling that

$$\beta = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \ ,$$

one recovers the well known formula of electronic dynamics

$$k_i = \frac{d}{dt} \frac{m\dot{x}_i}{\sqrt{1 - \frac{v^2}{c^2}}} \ .$$

Pisa, January 1921.

2) On the electrostatics of a homogeneous gravitational field and on the weight of electromagnetic masses

"Sull'elettrostatica di un campo gravitazionale uniforme
e sul peso delle masse elettromagnetiche,"
Nuovo Cimento ***22****, 176–188 (1921).*

INTRODUCTION

The aim of the present paper is to investigate in the framework of general relativity how a homogeneous gravitational field modifies the electrostatic phenomena occurring in it. Once the differential equation relating the electrostatic potential to the charge density, which corresponds to the Poisson equation in classical electrostatics, is established, one is able to integrate it at least when the gravitational field is weak enough (and certainly the gravitational field of the Earth amply satisfies this condition), obtaining in this way the corrections to Coulomb's law due to the presence of the gravitational field.

In a first application the distribution of the electric charges on a conducting sphere is studied, showing that the sphere polarizes by means of the gravitational field.

The second application is devoted to studying the weight of an electromagnetic mass, that is the force exerted on a fixed system of electric charges (e.g., sustained by a rigid dielectric), as a consequence of the presence of the gravitational field.

One finds that such a weight is given by the acceleration of gravity times u/c^2, where u denotes the electrostatic energy of the charges of the system, and c is the velocity of light. So the gravitational mass, namely the ratio between the weight and the acceleration of gravity, does not coincide in general with the inertial mass for the system under consideration, since the latter is given by $(4/3)u/c^2$ (with the same notation) if the system is endowed with spherical symmetry for example.

Besides it is known how special relativity leads us to take $\Delta u/c^2$ as the increase of the *inertial* mass of a system getting an energy Δu, and this fact can be easily related to the aforementioned result.

Finally, it is shown how to find a point having the same properties, with respect to the weight of the considered system of charges, as the center of gravity with respect to the weight of an ordinary system of material masses.

PART 1

ELECTROSTATICS IN A GRAVITATIONAL FIELD

§ 1.– Let us consider a portion of the spacetime where a homogeneous gravitational field is present, and assume the electrostatic phenomena that we think are taking place in it to be weak enough to neglect the effect they produce on the metric describing the region under consideration. Under this assumption, the line element of the spacetime manifold can be written as [1]

$$ds^2 = a\,dt^2 - dx^2 - dy^2 - dz^2 \ , \tag{1}$$

where a is a function only of z.

The variables t, x, y, z will also be denoted by x_0, x_1, x_2, x_3, and the coefficients of the quadratic form (1) by g_{ik}. Let φ_i be the vector potential, and F_{ik} the electromagnetic field. Then we have

$$\mathrm{F}_{ik} = \varphi_{i,k} - \varphi_{k,i} \ , \tag{2}$$

referring ourselves to the fundamental form (1).

By limiting our considerations to electrostatic fields, we can set $\varphi_1 = \varphi_2 = \varphi_3 = 0$, and, for the sake of brevity, $\varphi_0 = \varphi$. Thus one has:

$$\mathrm{F}_{ik} = \varphi_{i,k} - \varphi_{k,i} = \frac{\partial \varphi_i}{\partial x_k} - \frac{\partial \varphi_k}{\partial x_i} \ ,$$

that is

$$\begin{cases} \mathrm{F}_{01} = \dfrac{\partial \varphi}{\partial x} \ , \qquad \mathrm{F}_{02} = \dfrac{\partial \varphi}{\partial y} \ , \qquad \mathrm{F}_{03} = \dfrac{\partial \varphi}{\partial z} \ , \\ \mathrm{F}_{23} = \mathrm{F}_{31} = \mathrm{F}_{12} = 0 \ , \qquad \mathrm{F}_{ik} = -\mathrm{F}_{ki} \ , \qquad \mathrm{F}_{ij} = 0 \ . \end{cases} \tag{3}$$

In addition one has:

$$\mathrm{F}^{(ik)} = \sum_{hk} g^{(ih)} g^{(jk)} \mathrm{F}_{(hk)} = g^{(ii)} g^{(jj)} \mathrm{F}_{(ij)} \ ,$$

from which by noting that:

$$g^{(00)} = \frac{1}{a} \ , \qquad g^{(11)} = g^{(22)} = g^{(33)} = -1 \ ,$$

one obtains

$$\begin{cases} \mathrm{F}^{(01)} = -\dfrac{1}{a}\dfrac{\partial \varphi}{\partial x} \ , \qquad \mathrm{F}^{(02)} = -\dfrac{1}{a}\dfrac{\partial \varphi}{\partial y} \ , \qquad \mathrm{F}^{(03)} = -\dfrac{1}{a}\dfrac{\partial \varphi}{\partial z} \ , \\ \mathrm{F}^{(23)} = \mathrm{F}^{(31)} = \mathrm{F}^{(12)} = 0 \ , \qquad \mathrm{F}^{(ik)} = -\mathrm{F}^{(ki)} \ , \qquad \mathrm{F}^{(ii)} = 0 \ . \end{cases} \tag{4}$$

In the case under consideration here, the action can be written in the form

$$W = \int_\omega \sum_{ik} \mathrm{F}_{ik} \mathrm{F}^{(ik)}\, d\omega + \int de \int \varphi\, dx_0 \ , \tag{5}$$

where

$$d\omega = \sqrt{-||g_{ik}||}\, dx_0\, dx_1\, dx_2\, dx_3 = \sqrt{a}\, dx\, dy\, dz\, dt$$

[1] T. Levi-Civita, Note II. "Sui ds^2 einsteiniani". *Rend. Acc. Lincei*, **27**, 1° sem. N° 7.

is the hypervolume element of the manifold, and the integration corresponding to $d\omega$ has to be performed over a specific region of the manifold, while the integrations corresponding to de, dx_0 have to be extended to all the elements of electric charge whose world lines cross the region under consideration and to the portions of those world lines lying in it, respectively.

§ 2.– In the variation of W, φ can be arbitrarily varied, under the single condition that $\delta\varphi = 0$ on the boundary of the integration domain.

The variations δx, δy, δz instead, in addition to the condition $\delta x = \delta y = \delta z = 0$ on the boundary, could also be subjected to further conditions to be determined in each particular case. For example, inside a conducting body they will be quite arbitrary, while in a rigid dielectric they will have to represent the components of a rigid virtual displacement, and so on.

By putting the quantities (3), (4) into (5), one obtains:

$$W = -\frac{1}{2}\iiiint \frac{1}{\sqrt{a}}\left\{\left(\frac{\partial\varphi}{\partial x}\right)^2 + \left(\frac{\partial\varphi}{\partial y}\right)^2 + \left(\frac{\partial\varphi}{\partial z}\right)^2\right\} dx\,dy\,dz\,dt + \int de \int \varphi\,dt \ , \tag{6}$$

from which

$$\begin{aligned}\delta W = &\iiiint \delta\varphi\left[\frac{1}{\sqrt{a}}\Delta_2\,\varphi + \frac{\partial\varphi}{\partial z}\frac{d(1/\sqrt{a})}{dz} + \rho\right] dx\,dy\,dz\,dt \\ &+ \iiiint \rho\left(\frac{\partial\varphi}{\partial x}\,\delta x + \frac{\partial\varphi}{\partial y}\,\delta y + \frac{\partial\varphi}{\partial z}\,\delta z\right) dx\,dy\,dz\,dt\end{aligned} \tag{7}$$

as immediately follows by noting that $dx = dy = dz = 0$ along a given world line, as a consequence of our assumptions, and $\rho\,dx\,dy\,dz = de$, since ρ is the electric density.

Then in order for δW to vanish identically, since $\delta\varphi$ is arbitrary inside the integration domain, one finds that

$$\Delta_2\,\varphi - \frac{d\log\sqrt{a}}{dz}\frac{\partial\varphi}{\partial z} = -\rho\sqrt{a} \ . \tag{8}$$

Moreover, one must also have

$$\iiiint \rho\left(\frac{\partial\varphi}{\partial x}\,\delta x + \frac{\partial\varphi}{\partial y}\,\delta y + \frac{\partial\varphi}{\partial z}\,\delta z\right) dx\,dy\,dz\,dt = 0 \ , \tag{9}$$

for every system of values for δx, δy, δz satisfying the assumed constraints.

In equation (8) is contained the generalization of the Poisson's law, to which the (8) reduces if a is constant, that is if the gravitational field is absent.

§ 3.– If we indicate by G the acceleration of gravity of the field under consideration, namely the acceleration with which a free material point begins to move, one has:

$$\mathrm{G} = -\frac{1}{2}\frac{da}{dz} \ . \tag{10}$$

WIth this (8) becomes:

$$\Delta_2\,\varphi + \frac{\mathrm{G}}{a}\frac{\partial\varphi}{\partial z} = -\rho\sqrt{a}\ . \tag{11}$$

In order to find the solution of (11), given ρ at each point, we imagine the electric charges to be contained in a small region around the origin of the coordinates. Moreover, we will set $a = c^2$ at the origin (with c the velocity of light near the origin), and we will assume gravity to be so weak that those terms which contain the square of the ratio $l\,\mathrm{G}/c^2$ can be neglected, where l represents the maximum length entering into the problem under consideration. Under these assumptions, we can set:

$$\sqrt{a} = c + \frac{1}{2c}\frac{da}{dz}z = c\left(1 - \frac{\mathrm{G}}{c^2}z\right)\ .$$

Therefore (11) can be written as:

$$\Delta_2\,\varphi + \frac{\mathrm{G}}{c^2}\frac{\partial\varphi}{\partial z} = -c\left(1 - \frac{\mathrm{G}}{c^2}z\right)\rho\ . \tag{12}$$

The integral of that equation in this approximation, as can be directly verified, is given by:

$$\begin{aligned}\varphi_{\mathrm{P}} &= \frac{c}{4\pi}\int\left(1 - \frac{\mathrm{G}}{c^2}\right) z_{\mathrm{M}}\,d\tau_{\mathrm{M}}\left(\frac{1}{r} - \frac{\mathrm{G}}{2c^2}\frac{z_{\mathrm{P}} - z_{\mathrm{M}}}{r}\right)\\ &= \frac{c}{4\pi}\int \rho_{\mathrm{M}}\,d\tau_{\mathrm{M}}\left(\frac{1}{r} - \frac{\mathrm{G}}{2c^2}\frac{z_{\mathrm{P}} + z_{\mathrm{M}}}{r}\right)\ ,\end{aligned} \tag{13}$$

where M is the generic point of the region τ_{M} containing the electric charges, P is the point at which the potential φ is evaluated, and r is the distance MP.

Given the linearity of equation (12), any integral of the equation:

$$\Delta_2\,\varphi + \frac{\mathrm{G}}{c^2}\frac{\partial\varphi}{\partial z} = 0\ , \tag*{$(12)^*$}$$

obtained by setting $\rho = 0$ in (12), can be added to (13). This integral will represent the field due to causes external to ρ_{M}. For the application we have in mind it is convenient to consider a particular solution to $(12)^*$ given by

$$\varphi = -c\mathrm{E}^*_x x - c\mathrm{E}^*_y y + \frac{c^2}{\mathrm{G}}\mathrm{E}^*_z e^{-\frac{\mathrm{G}}{c^2}z}\ , \tag{14}$$

with $\mathrm{E}^*_x, \mathrm{E}^*_y, \mathrm{E}^*_z$ constants.

At the origin one has

$$\mathrm{E}_x = -\frac{1}{c}\,\mathrm{F}_{01}\ , \qquad \mathrm{E}_y = -\frac{1}{c}\,\mathrm{F}_{02}\ , \qquad \mathrm{E}_z = -\frac{1}{c}\,\mathrm{F}_{03}\ ,$$

since E is the electric force.

From this it follows that the electric force of the external field (14) has components

$$\mathrm{E}^*_x\ , \quad \mathrm{E}^*_y\ , \quad \mathrm{E}^*_z\ .$$

§ 4.– Let us now calculate the electric field due to a charge e concentrated at the origin of the coordinates. From (13) one has:

$$\varphi = \frac{ce}{4\pi}\left(\frac{1}{r} - \frac{\mathrm{G}}{2c^2}\frac{z}{r}\right) \ , \tag{15}$$

and this formula gives the generalization of Coulomb's law, as immediately follows by setting $\mathrm{G} = 0$. Recalling (3) one gets:

$$\begin{cases} \mathrm{F}_{01} = \dfrac{ce}{4\pi}\left(\dfrac{x}{r^3} - \dfrac{\mathrm{G}}{2c^2}\dfrac{zx}{r^3}\right) \ , \\[2ex] \mathrm{F}_{02} = \dfrac{ce}{4\pi}\left(\dfrac{y}{r^3} - \dfrac{\mathrm{G}}{2c^2}\dfrac{zy}{r^3}\right) \ , \\[2ex] \mathrm{F}_{03} = \dfrac{ce}{4\pi}\left(\dfrac{z}{r^3} - \dfrac{\mathrm{G}}{2c^2}\dfrac{z^2}{r^3} + \dfrac{\mathrm{G}}{2c^2}\dfrac{1}{r}\right) \ . \end{cases} \tag{16}$$

We can summarize all three of the preceding formulas in a single vector formula. In fact by indicating by F_0 the vector with components $\mathrm{F}_{01}, \mathrm{F}_{02}, \mathrm{F}_{03}$, with $\vec{a}$ a vector of magnitude 1 and orientation MP, and finally with $\vec{G}$ a vector of magnitude G and orientation z, (16) can be written as:

$$\mathrm{F}_0 = \frac{ce}{4\pi}\left\{\frac{\vec{a}}{r^2} + \frac{\vec{\mathrm{G}} \times \vec{a}}{2c^2 r}\,\vec{a} - \frac{1}{2c^2 r}\,\vec{\mathrm{G}}\right\} \ . \tag{17}$$

It is interesting to compare this formula with the one which gives the electric force exerted by an electric charge e which in the absence of gravitational attraction has acceleration $\vec{\Gamma}$, quasi-stationary motion and velocity negligible with respect to the speed of light. Such a force is expressed by

$$\mathrm{E} = \left\{\frac{\vec{a}}{r^2} + \frac{\vec{\Gamma} \times \vec{a}}{c^2 r}\,\vec{a} - \frac{1}{2c^2 r}\,\vec{\Gamma}\right\} \ , \tag{18}$$

with the same notation.

From here one sees that, by setting

$$\vec{\Gamma} = -\frac{\vec{\mathrm{G}}}{2} \tag{19}$$

in (18), one obtains

$$\mathrm{F}_0 = c\mathrm{E} \ .$$

This result can be put into words as follows, noting that $c\mathrm{E}$ is the electric part of the electromagnetic field generated by the charge in accelerated motion:

The electric part $(\mathrm{F}_{01}, \mathrm{F}_{02}, \mathrm{F}_{03})$ of the electromagnetic field (F_{ik}) generated by an electric charge at rest in a homogeneous field of strength G is equal to the electric part of the electromagnetic field which the same charge would produce in the absence of gravitational field if it moved under the conditions indicated above with acceleration G/2 in the direction opposite to the gravitational field.

§ 5.– Now, let us study how the distribution of the electricity over a conductor is modified by the gravitational field. To this end, let us note that since δx, δy, δz are arbitrary inside the conductor, from (9) it follows that $\varphi = constant$ inside, and so $\rho = 0$ by (8). Thus the electricity is completely at the surface. Then let us assume that our conductor is a sphere with center O at the origin of the coordinates and of radius R.

Let us try to satisfy the condition $\varphi = constant$ in the interior by assuming the following expression for the surface electric density at a generic point M of the surface:

$$\frac{e}{4\pi R^2} + \frac{e}{r}\, a \cos\theta \ , \tag{20}$$

where θ represents the angle spanned by the radius vector OM from the $z-$axis, and a is a constant to be determined, which we assume to be of the order of magnitude of G/c^2. From (13), the potential at a point P inside will be given by:

$$\varphi_\mathrm{P} = \frac{c}{4\pi} \int_\sigma \left(\frac{e}{4\pi r^2} + \frac{e}{r}\, a \cos\theta \right) \left(\frac{1}{r} - \frac{\mathrm{G}}{2c^2} \frac{z_\mathrm{P} + z_\mathrm{M}}{r} \right) d\sigma \ ,$$

where the integration must be extended over the whole surface σ of the sphere. By neglecting terms of order greater than G/c^2 one obtains:

$$\begin{aligned} \varphi_\mathrm{P} = {} & \frac{ce}{16\pi^2 r^2} \int \frac{d\sigma}{r} + \frac{cea}{4\pi r} \int \frac{\cos\theta d\sigma}{r} \\ & - \frac{ce\mathrm{G} z_\mathrm{P}}{32\pi^2 \mathrm{R}^2 c^2} \int \frac{d\sigma}{r} - \frac{ce\mathrm{G}}{22\pi^2 R^2 c^2} \int \frac{z_\mathrm{M} d\sigma}{r} \ . \end{aligned} \tag{21}$$

However, since P is inside, one has:

$$\int \frac{d\sigma}{r} = 4\,\pi\,\mathrm{S} \ , \qquad \int \frac{\cos\theta}{r}\, d\sigma = \frac{4}{3}\,\pi\, z_\mathrm{P} \ , \qquad \int \frac{z_\mathrm{M}}{r}\, d\sigma = \frac{4}{3}\,\pi\,\mathrm{R}\, z_\mathrm{P} \ .$$

Thus one finds:

$$\varphi_\mathrm{P} = \frac{ce}{4\pi\mathrm{R}} + \frac{c}{3} \left(\frac{e}{\mathrm{R}}\, a - \frac{e}{2\pi\mathrm{R}c^3} \right) z_\mathrm{P} \ . \tag{22}$$

So if we require φ_P to be constant, we will have to set

$$a = \frac{1}{2\pi} \frac{\mathrm{G}}{c^2} \ .$$

By substituting this value into (20), one finds the following expression for the surface density:

$$\frac{e}{4\pi\mathrm{R}^2} + \left(1 + \frac{2\mathrm{G}}{c^2}\, \mathrm{R} \cos\theta \right) \ . \tag{23}$$

Therefore, the fact of being in a gravitational field produces a polarization of the sphere with moment

$$\frac{2}{3} \frac{\mathrm{G}}{c^2}\, e\, \mathrm{R}^2 \ .$$

PART 2

WEIGHT OF ELECTROMAGNETIC MASSES

§ 6.– Suppose we have a system of charges held by a rigid support in such a way that the δx, δy, δz of § 2 have to be given the form corresponding to the components of a rigid displacement. Leaving the rotational displacements till later, let us consider now the translational ones, that is say assume that δx, δy, δz are arbitrary functions of time, but do not depend on x, y, z.

Then we will try to satisfy (9) by thinking of the potential φ_{P} at a generic point P as the sum of the potential given by (13) and one of the form (14). We will denote these two terms by $\varphi_{\mathrm{P}}{}'$ and $\varphi_{\mathrm{P}}{}''$, and suppose that the ratio between the derivatives of $\varphi_{\mathrm{P}}{}'$ and $\varphi_{\mathrm{P}}{}''$ with respect to any direction whatsoever is of order $l\,\mathrm{G}/c^2$, having decided to neglect the quadratic terms. Hence (9) can be written:

$$\int dt \Big\{ \int_{\tau_{\mathrm{P}}} \delta x \left(\frac{\partial \varphi'}{\partial x} + \frac{\partial \varphi''}{\partial x} \right) \rho_{\mathrm{P}}\, d\tau_{\mathrm{P}} + \delta y \left(\frac{\partial \varphi'}{\partial y} + \frac{\partial \varphi''}{\partial y} \right) \rho_{\mathrm{P}}\, d\tau_{\mathrm{P}} \\ + \delta z \left(\frac{\partial \varphi'}{\partial z} + \frac{\partial \varphi''}{\partial z} \right) \rho_{\mathrm{P}}\, d\tau_{\mathrm{P}} \Big\} = 0 \ .$$

Given that δx, δy, δz are arbitrary functions of time, independent of each other, this equation gives rise to the three equivalent equations:

$$\int_{\tau_{\mathrm{P}}} \left(\frac{\partial \varphi'}{\partial x} + \frac{\partial \varphi''}{\partial x} \right) \rho_{\mathrm{P}}\, d\tau_{\mathrm{P}} = \int_{\tau_{\mathrm{P}}} \left(\frac{\partial \varphi'}{\partial y} + \frac{\partial \varphi''}{\partial y} \right) \rho_{\mathrm{P}}\, d\tau_{\mathrm{P}} \\ = \int_{\tau_{\mathrm{P}}} \left(\frac{\partial \varphi'}{\partial z} + \frac{\partial \varphi''}{\partial z} \right) \rho_{\mathrm{P}}\, d\tau_{\mathrm{P}} = 0 \ . \qquad (24)$$

Now from the expression (13) for $\varphi_{\mathrm{P}}{}'$, by noting that

$$\frac{\partial r}{\partial x_{\mathrm{P}}} = \frac{x_{\mathrm{P}} - x_r}{r} \ ,$$

one immediately obtains:

$$\int_{\tau_{\mathrm{P}}} \frac{\partial \varphi'}{\partial x} \rho_{\mathrm{P}}\, d\tau_{\mathrm{P}} = -\frac{c}{4\pi} \int_{\tau_{\mathrm{P}}}\int_{\tau_{\mathrm{M}}} \rho_{\mathrm{P}} \rho_{\mathrm{M}}\, d\tau_{\mathrm{P}} d\tau_{\mathrm{M}} \left\{ \frac{x_{\mathrm{P}} - x_{\mathrm{M}}}{r^3} - \frac{\mathrm{G}}{2c^2} \frac{(x_{\mathrm{P}} - x_{\mathrm{M}})(z_{\mathrm{P}} + z_{\mathrm{M}})}{r^3} \right\} \ ,$$

where both integrals have to be performed over the region occupied by the charges. By interchanging P and M in the right hand side, which changes nothing, one obtains:

$$\int_{\tau_{\mathrm{P}}} \frac{\partial \varphi'}{\partial x} \rho_{\mathrm{P}}\, d\tau_{\mathrm{P}} = -\frac{c}{4\pi} \int_{\tau_{\mathrm{M}}}\int_{\tau_{\mathrm{P}}} \rho_{\mathrm{M}} \rho_{\mathrm{P}}\, d\tau_{\mathrm{M}} d\tau_{\mathrm{P}} \left\{ \frac{x_{\mathrm{M}} - x_{\mathrm{P}}}{r^3} - \frac{\mathrm{G}}{2c^2} \frac{(x_{\mathrm{M}} - x_{\mathrm{P}})(z_{\mathrm{M}} + z_{\mathrm{P}})}{r^3} \right\} \ ,$$

from which, by taking half the sum:

$$\int_{\tau_{\mathrm{P}}} \frac{\partial \varphi'}{\partial x} \rho_{\mathrm{P}}\, d\tau_{\mathrm{P}} = 0 \ . \qquad (25)$$

In an completely analogous way:

$$\int_{\tau_{\mathrm{P}}} \frac{\partial \varphi'}{\partial y} \rho_{\mathrm{P}} \, d\tau_{\mathrm{P}} = 0 \ . \tag{26}$$

On the other hand, similarly:

$$\begin{aligned}\int_{\tau_{\mathrm{P}}} \frac{\partial \varphi'}{\partial z} \rho_{\mathrm{P}} \, d\tau_{\mathrm{P}} = -\frac{c}{4\pi} \int_{\tau_{\mathrm{P}}}\!\int_{\tau_{\mathrm{M}}} \rho_{\mathrm{P}} \rho_{\mathrm{M}} \, d\tau_{\mathrm{P}} d\tau_{\mathrm{M}} \biggl\{ \frac{z_{\mathrm{P}} - z_{\mathrm{M}}}{r^3} - \\ - \frac{\mathrm{G}}{2c^2} \frac{(z_{\mathrm{P}} - z_{\mathrm{M}})(z_{\mathrm{P}} + z_{\mathrm{M}})}{r^3} + \frac{\mathrm{G}}{2c^2} \frac{1}{r} \biggr\} \ ,\end{aligned}$$

interchanging M and P:

$$\begin{aligned}\int_{\tau_{\mathrm{P}}} \frac{\partial \varphi'}{\partial z} \rho_{\mathrm{P}} \, d\tau_{\mathrm{P}} = -\frac{c}{4\pi} \int_{\tau_{\mathrm{M}}}\!\int_{\tau_{\mathrm{P}}} \rho_{\mathrm{M}} \rho_{\mathrm{P}} \, d\tau_{\mathrm{M}} d\tau_{\mathrm{P}} \biggl\{ \frac{z_{\mathrm{M}} - z_{\mathrm{P}}}{r^3} \\ - \frac{\mathrm{G}}{2c^2} \frac{(z_{\mathrm{M}} - z_{\mathrm{P}})(z_{\mathrm{M}} + z_{\mathrm{P}})}{r^3} + \frac{\mathrm{G}}{2c^2} \frac{1}{r} \biggr\} \ ,\end{aligned}$$

and by taking half the sum:

$$\int_{\tau_{\mathrm{P}}} \frac{\partial \varphi'}{\partial z} \rho_{\mathrm{P}} \, d\tau_{\mathrm{P}} = -\frac{c}{4\pi} \int_{\tau_{\mathrm{P}}}\!\int_{\tau_{\mathrm{M}}} \frac{\rho_{\mathrm{P}} \rho_{\mathrm{M}}}{r} \, d\tau_{\mathrm{P}} d\tau_{\mathrm{M}} = -\mathrm{G} \, \frac{\mathrm{U}}{c^2} \, e \ , \tag{27}$$

denoting by U the electrostatic energy of the system (apart from the gravitational correction terms). As a consequence of the assumptions made about the derivatives of φ_{P}'', we can certainly write, with our approximation:

$$\begin{cases} \displaystyle\int_{\tau} \frac{\partial \varphi''}{\partial x} \rho \, d\tau = -c \, \mathrm{E}_x^* \, e \ , \\ \displaystyle\int_{\tau} \frac{\partial \varphi''}{\partial y} \rho \, d\tau = -c \, \mathrm{E}_y^* \, e \ , \\ \displaystyle\int_{\tau} \frac{\partial \varphi''}{\partial z} \rho \, d\tau = -c \, \mathrm{E}_z^* \, e \ , \end{cases}$$

where $e = \displaystyle\int_{\tau} \rho \, d\tau$ indicates the total charge of the system. By substituting the expression just obtained into (24) one finds:

$$\begin{cases} e \, \mathrm{E}_x^* = 0 \ , \\ e \, \mathrm{E}_y^* = 0 \ , \\ e \, \mathrm{E}_z^* = -\mathrm{G} \, \dfrac{\mathrm{U}}{c^2} \ . \end{cases}$$

Our result is contained in these formulas. In fact, they tell us that in order to maintain our system in equilibrium an external field (E^*) is required exerting on the system a force given (in the first approximation) by $e \, \mathrm{E}^*$, which must be understood to balance the weight of the system, which is therefore given by $-e \, \mathrm{E}^*$, and so has components

$$0, \, 0, \, \mathrm{G} \, \frac{\mathrm{U}}{c^2} \ . \tag{28}$$

With this we conclude that *the weight of an electromagnetic mass always has the vertical direction and magnitude equal to the weight of a material mass u/c^2.*

§ 7.– In the preceding section we have taken δx, δy, δz to be the components of a translational displacement. If instead one takes the components of a virtual rotational displacement with the axis passing through the origin of the coordinates, namely setting

$$\delta x = qz - ry \ ; \qquad \delta y = rx - pz \ ; \qquad \delta z = py - qx \ , \tag{29}$$

the integral (9), apart from the contribution due to the external field φ'', becomes:

$$\int d\tau \Big\{ p \int_\tau \rho \left(y\frac{\partial\varphi}{\partial z} - z\frac{\partial\varphi}{\partial y}\right) d\tau + q \int_\tau \rho \left(z\frac{\partial\varphi}{\partial x} - x\frac{\partial\varphi}{\partial z}\right) d\tau \\ + r \int_\tau \rho \left(x\frac{\partial\varphi}{\partial y} - y\frac{\partial\varphi}{\partial x}\right) d\tau \Big\} \ . \tag{30}$$

The integrals between curly brackets are easily evaluated using (13) through methods similar to that used in the previous section. They have the values:

$$-\frac{\mathrm{G}}{8\pi c}\iint \frac{y_\mathrm{P}}{r}\rho_\mathrm{P}\rho_\mathrm{M}\, d\tau_\mathrm{P} d\tau_\mathrm{M} \ ; \quad +\frac{\mathrm{G}}{8\pi c}\iint \frac{x_\mathrm{P}}{r}\rho_\mathrm{P}\rho_\mathrm{M}\, d\tau_\mathrm{P} d\tau_\mathrm{M} \ ; \quad 0 \ . \tag{31}$$

By taking as the origin the point O′ defined by the point O and the vector

$$\mathrm{O}' \ - \ \mathrm{O} = \frac{1}{2\mathrm{U}}\iint \frac{\mathrm{P} \ - \ \mathrm{O}}{r}\rho_\mathrm{P}\rho_\mathrm{M}\, d\tau_\mathrm{P} d\tau_\mathrm{M} \ ,$$

one sees immediately that the three integrals vanish for *any* orientation of the system about O′. As a consequence, with respect to the new origin the integral (9) is identically zero, namely the moment of the weight with respect to O′ is zero for any orientation of the system; thus O′ enjoys the properties of the center of gravity.

Pisa, March 1921.

3) On Phenomena Occurring Close to a World Line

"Sopra i fenomeni che avvengono in vicinanza di una linea oraria,"
Rend. Lincei, ***31*** *(I), 21–23, 51–52, 101–103 (1922).* ([1])

Note I.

§ 1.– In order to study phenomena which occur close to a world line, i.e., in nonrelativistic language, in a region of space in the spacetime manifold, even varying in time, but always very small compared with the divergences from Euclidean space, it would be convenient to find a particular frame such that close to the line being studied, the spacetime ds^2 will assume a simple form. In order to find such a frame, we must begin with some geometrical considerations.

Let there be given a line L in a Riemannian manifold V$_n$ or in a manifold metrically connected in the sense of Weyl.[2] Let us associate at every point P of L a direction y perpendicular to L, with the rule that the direction $y+dy$, corresponding to the point P+dP, will be derived from that y associated to P in the following way: let η be the direction tangent to L at P; let y and η be parallel transported[3] from P to P+dP and let $y + \delta y$ and $\eta + \delta\eta$ be the directions obtained in this way, which because of the fundamental properties of parallel transport will be still orthogonal. If L is not geodesic $\eta+\delta\eta$ will not coincide with the direction $\eta+d\eta$ of the tangent to L at P+dP, and these two directions at P+dP will define a 2-dimensional subspace. Let us consider at P+dP the element of the S_{n-2} perpendicular to this subspace and let us rigidly rotate around this S$_{n-2}$ all the surrounding particle space until $\eta + \delta\eta$ is superposed on $\eta + d\eta$. Then $y + \delta y$ will be mapped to a position which we will consider to be the direction $y + dy$ relative to the point P+dP. It is clear that, arbitrarily fixing the direction y at a point of L, an integration process will allow it to be obtained at any point of L.

Let us now look for the analytic expressions which translate the indicated operations to a Riemannian manifold, which coincide with those valid for a Weyl metric manifold as long as the "Eichung" is choosen such that the measure of a segment, which moves rigidly around L, will be constant. Let

$$ds^2 = \sum_{ik} g_{ik} dx^i dx^k \qquad (i, k = 1, 2, \dots n) \tag{1}$$

and let y_i, $y^{(i)}$; η_i, $\eta^{(i)} = dx_i/ds$ be the co- and contravariant systems of the directions y, η. We will then have

$$\frac{\delta\eta^{(i)}}{ds} = -\sum_{hl} \begin{Bmatrix} h\,l \\ i \end{Bmatrix} \eta^{(h)} \frac{dx_l}{ds} = -\sum_{hl} \begin{Bmatrix} h\,l \\ i \end{Bmatrix} \frac{dx_h}{ds}\frac{dx_l}{ds} \ ,$$

[1] Presented by the Correspondent G. Armellini during the session of January 22, 1922.
[2] Weyl, *Space, Time, Matter*, p. 109. Berlin, Springer, 1921.
[3] T. Levi Civita, *Rend. Circ. Palermo*, Vol. XLII, p. 173 (1917).

and moreover $\frac{d\eta^i}{ds} = \frac{d}{ds}\frac{dx_i}{ds} = \frac{d^2x_i}{ds^2}$. Therefore one finds

$$\frac{\delta\eta^{(i)} - d\eta^{(i)}}{ds} = -\left(\frac{d^2x_i}{ds^2} + \sum_{hl} \begin{Bmatrix} h\,l \\ i \end{Bmatrix} \frac{dx_h}{ds}\frac{dx_l}{ds}\right) = -C^i \ .$$

The C^i are the contravariant components of the vector $\mathbf{C}$, the geodetic curvature, namely of a vector having the same orientation as the geodesic principal normal of L and a magnitude equal to its geodesic curvature.

On the other hand one has

$$\frac{\delta y^{(i)}}{ds} = -\sum_{hk} \begin{Bmatrix} h\,k \\ i \end{Bmatrix} y^{(k)} \frac{dx_k}{ds} \ . \tag{2}$$

Now since y is orthogonal to L, the displacement with which from $y + \delta y$ one gets $y + dy$ will be parallel to the tangent to L and will have magnitude equal to the projection onto the same y of $\delta\eta - d\eta$; that is to say, since y has magnitude 1, equal to the scalar product of $\delta\eta - d\eta$ and y, namely

$$\sum_i (\delta\eta_i - d\eta_i) y^{(i)} = -ds \sum_i C_i y^{(i)} \ .$$

Its contravariant components will be obtained therefore by multiplying its magnitude by the contravariant coordinates of the tangent to L, that is dx_i/ds. These are, in the final analysis, $-dx_i \sum_r C_r y^{(r)}$. From (2) it follows immediately that

$$\frac{dy^{(i)}}{ds} = -\sum_{hk} \begin{Bmatrix} h\,k \\ i \end{Bmatrix} y^{(k)} \frac{dx_k}{ds} - \frac{dx_i}{ds} \sum_h C_h y^h \ . \tag{3}$$

Eq. (3), written for $i = 1, 2, \ldots n$ gives a system of n first order differential equations for the n unknowns $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$, which are therefore determined once the initial data are assigned. It would also be easy to formally verify from (3) that, if the initial values of the $y^{(i)}$ satisfy the condition of perpendicularity to L, such a condition will remain satisfied all along the line.

§ 2.– Let us now assign at a point P_0 of L n mutually orthogonal directions $y_1, y_2, \ldots, y_n$ chosen at will, with the condition that y_n be tangent to L. The directions $y_1, y_2, \ldots, y_{n-1}$ will be perpendicular to L, and we can transport them along L by using the law given in the preceding section, which clearly from its definition preserves their orthogonality. We are then in a position to associate with every point of L n mutually orthogonal directions, the last one of which is tangent to L. Let us now consider our V_n embedded in a Euclidean $\mathrm{S_N}$ with a suitable number of dimensions. We can take as coordinates of a point of V_n the orthogonal Cartesian coordinates of its projection onto the $\mathrm{S_N}$ tangent to V_n at a generic point P of L, having P as the origin and the directions $y_1, y_2, \ldots, y_n$ relative to the point P as directions. In terms of these coordinates, the line element of V_n at P can be written in the form $ds_{\mathrm{P}}^2 = dy_1^2 + dy_2^2 + \cdots + dy_n^2$; in addition, they are geodesics at P, as one can immediately see. In other words, for the coordinates y it is possible in a neighbourhood of P to set $g_{ii} = 1$, $g_{ik} = 0$ $(i \neq k)$, up to infinitesimals of order greater than the first. Obviously we shall have such a reference frame at every point of L. Let us consider now a point Q_0 of V_n with coordinates $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{n-1}, 0$ in the reference frame corresponding to the point P_0 on L. For any other point P of L we can so determine a point Q having in the frame corresponding to P the same coordinates as Q_0 has in the frame corresponding to P_0. The point Q will therefore trace out a line parallel to L. Now we want to find the relation between ds_{Q} and ds_{P}, assuming that the point Q is infinitely close to P. In order to do so, we note that the displacement transporting Q to Q + dQ is composed of the displacements denoted in § 1 by δ and $d - \delta$, and that the first one gives $\delta s_{\mathrm{Q}} = ds_{\mathrm{P}}$ up to infinitesimals of greater order since it is a parallel displacement; the second one is a rotation, which gives $(d-\delta)s_{\mathrm{Q}} = ds_{\mathrm{P}}\,\mathbf{C} \cdot (\mathrm{Q}-\mathrm{P})$, as is seen from § 1, denoting by $\cdot$ the symbol of the scalar product, and with Q − P the vector with origin at P and endpoint at Q. Moreover, both ds_{Q} and $(d-\delta)s_{\mathrm{Q}}$ have the direction of the tangent to L. Hence, one has $ds_{\mathrm{Q}} = \delta s_{\mathrm{Q}} + (d-\delta)s_{\mathrm{Q}}$; namely

$$ds_{\mathrm{Q}} = ds_{\mathrm{P}}[1 + \mathbf{C} \cdot (\mathrm{Q}-\mathrm{P})] \ . \tag{4}$$

The trajectories of the points Q form a $(n-1)^{ple}$ infinity of lines, so at least with proper limitations through each point M of V_n will pass one of these lines; in this way, we can characterize M through the coordinates of the point Q, $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{n-1}$ corresponding to the line passing through M, and the arclength s_{P} of the line L marked off from an arbitrarily chosen origin to that point P corresponding to the Q one coinciding with M.

If M is infinitely close to L, ds_{Q} will be perpendicular to the hypersurface $s_{\mathrm{P}} =$ constant. Thus one will have

$$ds^2_{\mathrm{M}} = ds^2_{\mathrm{Q}} + d\bar{\mathrm{y}}^2_1 + d\bar{\mathrm{y}}^2_2 + \cdots + d\bar{\mathrm{y}}^2_{n-1} \ ;$$

and taking into account (4),

$$ds^2_{\mathrm{M}} = [1 + \mathbf{C} \cdot (\mathrm{M}-\mathrm{P})]^2 ds^2_{\mathrm{Q}} + d\bar{\mathrm{y}}^2_1 + d\bar{\mathrm{y}}^2_2 + \cdots + d\bar{\mathrm{y}}^2_{n-1} \ . \tag{5}$$

As a result, in the neighborhood of L we have found a very simple expression for ds^2.

Note II.

§ 3.− Before passing to the physical application of the results obtained above, we still want to make some geometrical observations. First of all, it is clear that the previous considerations, and so also the formula (5) representing their conclusion, which for any manifold whatsoever are only valid close to L, are instead completely rigorous for Euclidean spaces. So let us associate to the line L of V_n a line L^* in a Euclidean space S_n, in which we indicate the orthogonal cartesian coordinates by

x_i^*. If we indicate with asterisks the symbols referring to the line L*, we can write for S$_n$ the formula analogous to (5):

$$ds^2_{\mathrm{M}^*} = [1 + \mathbf{C}^* \cdot (\mathrm{M}^* - \mathrm{P}^*)]^2\, ds^2_{\mathrm{P}^*} + d\bar{\mathrm{y}}_1^{*2} + d\bar{\mathrm{y}}_2^{*2} + \cdots + d\bar{\mathrm{y}}_{n-1}^{*\;2} \; ; \qquad (5^*)$$

as in (5) $\mathbf{C}$ is a function of s_P, so in (5*) $\mathbf{C}^*$ is a function of s_{P^*}.

Let $\mathrm{K}^{(1)}$, $\mathrm{K}^{(2)}$, $\cdots$, $\mathrm{K}^{(n-1)}$ be the contravariant components of $\mathbf{C}$ with respect to $\bar{\mathrm{y}}_1$, $\bar{\mathrm{y}}_2$, $\cdots$, $\bar{\mathrm{y}}_{n-1}$, and $\mathrm{K}^{(1)\,*}$, $\mathrm{K}^{(2)\,*}$, $\cdots$, $\mathrm{K}^{(n-1)\,*}$ those of $\mathbf{C}^*$ with respect to the $\bar{\mathrm{y}}^*$. Let us try to determine L* in such a way that the functions $\mathrm{K}^{(r)\,*}(s_{\mathrm{P}^*})$ become equal to the $\mathrm{K}^{(r)}(s_\mathrm{P})$. In order so, we shall begin by imposing that $s_\mathrm{P} = s_{\mathrm{P}^*}$, i.e., by establishing between the points of L and L* a one-to-one correspondence preserving the arclength. We then note that $\mathrm{K}^{(r)\,*}$ is the projection of $\mathbf{C}^*$ on the r^{th} direction y^*. Namely, one has

$$\mathrm{K}^{(r)\,*} = \sum_{i=1}^{i=n} y^*_{i|r}\, \frac{d^2 x_i^*}{ds^2_\mathrm{P}} \qquad (r = 1, 2, \cdots, n-1). \qquad (6)$$

The $\mathrm{K}^{(r)}$ are then known functions of s_P. The condition $\mathrm{K}^{(r)} = \mathrm{K}^{(r\,*)}$ thus leads to the $(n-1)$ equations

$$\mathrm{K}^{(r)}(s_\mathrm{P}) = \sum_{i=1}^{i=n} y^*_{i|r}\, \frac{d^2 x_i^*}{ds^2_\mathrm{P}} \qquad (r = 1, 2, \cdots, n-1)\; . \qquad (7)$$

On the other hand, (3) once written for the S$_n$, gives us another $n(n-1)$ equations. If we add to these equations the following one

$$ds^2_\mathrm{P} = dx_1^{*2} + dx_2^{*2} + \cdots + dx_n^{*\;2}\; , \qquad (8)$$

we obtain a system of $n - 1 + n(n-1) + 1 = n^2$ equations for the n^2 unknowns x_i^*, $y^*_{i|r}$, which can be used to express them in terms of s_P. In this way we can determine the parametric equations $x_i^* = x_i^*(s_\mathrm{P})$ for L*. With that the formula (5*) becomes identical to (5), that is we have represented by applicability the neighborhood of the line L* onto that of L. In addition, since L* is in a Euclidean space, we can say that we have unfolded the neighborhood of L in a Euclidean space, i.e., we have found coordinates which are simultaneously geodesic at each point of L.

Note III.

§ 4.– In order to show the application to the theory of relativity of the results obtained above, we shall assume that V$_n$ is the V$_4$ spacetime and that L is a world line in whose neighborhood we want to study the phenomena. Setting $ds^2_\mathrm{M} = ds^2$ in (5) for the sake of brevity, one finds in this case:

$$ds^2 = [1 + \mathbf{C} \cdot (\mathrm{M} - \mathrm{P})]^2\, ds^2_\mathrm{P} + d\bar{\mathrm{y}}_1^2 + d\bar{\mathrm{y}}_2^2 + d\bar{\mathrm{y}}_3^2\; .$$

To avoid the appearance of imaginary terms and to restore the homogeneity, it is convenient to make the following change of variables:

$$s_\mathrm{P} = vt \quad ; \quad \bar{\mathrm{y}}_1 = ix \quad ; \quad \bar{\mathrm{y}}_2 = iy \quad ; \quad \bar{\mathrm{y}}_3 = iz\; ,$$

where v is a constant with dimensions of a velocity, so that t has the dimensions of time. Thus one obtains

$$ds^2 = a\,dt^2 - dx^2 - dy^2 - dz^2 \ , \tag{9}$$

where

$$a = v^2[1 + \mathbf{C}\cdot(\mathrm{M}-\mathrm{P})]^2 \ . \tag{10}$$

Hereafter, we refer to the space x, y, z using the ordinary symbols of vector calculus. And it is just in this sense that the scalar product which enters in (10) can be understood, provided that $\mathbf{C}$ is considered as the vector whose components are the covariant components of the geodesic curvature of the world line $x = y = z = 0$, and $\mathrm{M}-\mathrm{P}$ is the vector with components x, y, z. We will call x, y, z spatial coordinates, and t time. Sometimes for uniformity we will write x_0, x_1, x_2, x_3 in place of t, x, y, z, and we will also denote the coefficients of the quadratic form (9) by g_{ik}.

§ 5.– Let[4] F_{ik} be the electromagnetic field and $(\varphi_0, \varphi_1, \varphi_2, \varphi_3)$ the first rank tensor "potential" of F_{ik}, such that $\mathrm{F}_{ik} = \varphi_{i,k} - \varphi_{k,i}$. We set $\varphi_0 = \varphi$ and call $\mathbf{u}$ the vector with components $\varphi_1, \varphi_2, \varphi_3$. First of all, we have:

$$\left.\begin{array}{l}\mathrm{F}_{01}\\ \mathrm{F}_{02}\\ \mathrm{F}_{03}\end{array}\right\} = \mathrm{grad}\,\varphi - \frac{\partial\mathbf{u}}{\partial t} \quad , \quad \left.\begin{array}{l}\mathrm{F}_{23}\\ \mathrm{F}_{31}\\ \mathrm{F}_{12}\end{array}\right\} = -\mathrm{curl}\,\mathbf{u} \ , \ \mathrm{F}_{ii} = 0 \ , \ \mathrm{F}_{ik} = -\mathrm{F}_{ki} \ ;$$

analogously

$$\left.\begin{array}{l}\mathrm{F}^{(01)}\\ \mathrm{F}^{(02)}\\ \mathrm{F}^{(03)}\end{array}\right\} = \frac{1}{a}\left(-\mathrm{grad}\,\varphi + \frac{\partial\mathbf{u}}{\partial t}\right) \ , \ \left.\begin{array}{l}\mathrm{F}^{(23)}\\ \mathrm{F}^{(31)}\\ \mathrm{F}^{(12)}\end{array}\right\} = -\mathrm{curl}\,\mathbf{u} \ , \ \mathrm{F}^{(ii)} = 0 \ , \ \mathrm{F}^{(ik)} = -\mathrm{F}_{(ki)} \ ,$$

so that

$$\frac{1}{4}\sum_{ik}\mathrm{F}_{ik}\mathrm{F}^{(ik)} = \frac{1}{2}\left\{\mathrm{curl}^2\,\mathbf{u} - \frac{1}{a}\left(-\mathrm{grad}\,\varphi + \frac{\partial\mathbf{u}}{\partial t}\right)^2\right\} \ .$$

Let $d\omega$ be the hypervolume element of V_4. We will have

$$d\omega = \sqrt{-||g_{ik}||}\,dx_0\,dx_1\,dx_2\,dx_3 = \sqrt{a}\,dt\,d\tau \ ,$$

where $d\tau = dx\,dy\,dz$ is the volume element of the space.

One also has:

$$\sum\varphi_i dx_i = \varphi dx + \mathbf{u}\cdot d\mathrm{M} \ , \qquad d\mathrm{M} = (dx, dy, dz) \ .$$

[4] See Weyl, op. cit., pp. 186 and 208 for the notation and the Hamiltonian derivation of the laws of physics.

Apart from the action of the metric field, whose variation is zero since we consider it as given *a priori* by (9), the action will assume the following form:

$$W = \frac{1}{4}\int_\omega \sum_{ik} \mathrm{F}_{ik}\mathrm{F}^{(ik)}\, d\omega + \int_e de \int \varphi_i\, dx_i + \int_m dm \int ds \ ,$$

$$\begin{pmatrix} de & = \text{element of electric charge} \\ dm & = \text{element of mass} \end{pmatrix} .$$

By introducing the indicated notation, one finds

$$\begin{aligned} W = \frac{1}{2}\iint &\left\{\mathrm{curl}^2\,\mathbf{u} - \frac{1}{a}\left(-\mathrm{grad}\,\varphi + \frac{\partial \mathbf{u}}{\partial t}\right)^2\right\}\sqrt{a}\,dt\,d\tau \\ &+ \iint (\varphi + \mathbf{u}\cdot\mathbf{V}_\mathrm{L})\rho\,d\tau\,dt + \iint \sqrt{a - {\mathbf{V}_\mathrm{M}}^2}k\,d\tau\,dt \ , \end{aligned} \tag{11}$$

where ρ, k are respectively the density of electricity and of matter, so that $de = \rho\,d\tau$, $dm = k\,d\tau$, $\mathbf{V}_\mathrm{L}$ is the velocity of the electric charges, $\mathbf{V}_\mathrm{M}$ that of the masses.

The integrals on the right hand side can be extended to an arbitrary region τ between any two times t_1, t_2. Then one has the constraint that on the boundary of the region τ, and for the two times t_1, t_2, all variations are zero.

Apart from these conditions, the variations of φ and of $\mathbf{u}$ are completely arbitrary. Further conditions can be imposed on the variations of x, y, z thought of as coordinates of an element of charge or mass, expressing the constraints of the specific problem under consideration. Then writing that dW vanishes for any variation $\delta\varphi$ of φ whatsoever, one finds

$$0 = -\iint \left(\mathrm{grad}\,\varphi - \frac{\partial \mathbf{u}}{\partial t}\right)\cdot\delta\,\mathrm{grad}\,\varphi\frac{d\tau\,dt}{\sqrt{a}} + \iint \delta\varphi\rho\,dt\,d\tau \ .$$

Transforming the first integral by a suitable application of Gauss's theorem, and taking into account that $\delta\varphi$ vanishes at the boundary, we find

$$0 = \iint \delta\varphi\left\{\rho + \mathrm{div}\left[\frac{1}{\sqrt{a}}\left(\mathrm{grad}\,\varphi - \frac{\partial \mathbf{u}}{\partial t}\right)\right]\right\}dt\,d\tau \ ,$$

and since $\delta\varphi$ is arbitrary, we obtain the equation

$$\rho + \mathrm{div}\left[\frac{1}{\sqrt{a}}\left(\mathrm{grad}\,\varphi - \frac{\partial \mathbf{u}}{\partial t}\right)\right] = 0 \ . \tag{12}$$

Analogously, taking the variation of $\mathbf{u}$, one finds

$$\rho\mathbf{V}_\mathrm{L} + \mathrm{curl}(\sqrt{a}\,\mathrm{curl}\,\mathbf{u}) - \frac{\partial}{\partial t}\left[\frac{1}{\sqrt{a}}\left(\mathrm{grad}\,\varphi - \frac{\partial \mathbf{u}}{\partial t}\right)\right] = 0 \ . \tag{13}$$

These last two equations allow us to determine the electromagnetic field, once the charges and their motion are given.

Another set of equations can be obained by varying the trajectories of the charges and masses in W. Let $\delta\mathrm{P}_\mathrm{M}$ be the variation of the trajectory of the masses, $\delta\mathrm{P}_\mathrm{L}$ that of the charges. Moreover, since $\mathbf{u}$ is a vector function of the position and $\mathbf{V}$ a vector, let us denote by $(\partial\mathbf{u}/\partial\mathrm{P})(\mathbf{V})$ the vector with components $\dfrac{\partial u_x}{\partial x}V_x+\dfrac{\partial u_x}{\partial y}V_y+\dfrac{\partial u_x}{\partial z}V_z$, and so on. Now, writing that the variation of W is zero, one finds through the usual methods:

$$\iint\left(\delta\mathrm{P}_\mathrm{M}\cdot\operatorname{grad}\varphi-\delta\mathrm{P}_\mathrm{L}\left(\frac{\partial\mathbf{u}}{\partial t}+\frac{\partial\mathbf{u}}{\partial\mathrm{P}}(\mathrm{V}_\mathrm{L})\right)+\mathrm{V}_\mathrm{L}\cdot\frac{\partial\mathbf{u}}{\partial\mathrm{P}}(\delta\mathrm{P}_\mathrm{L})\right)\rho\,dt\,d\tau$$
$$+\iint\delta\mathrm{P}_\mathrm{M}\cdot\left\{\frac{dt}{ds}\,\frac{\operatorname{grad}a}{2}+\frac{d}{dt}\left(\frac{dt}{ds}\mathbf{V}_\mathrm{M}\right)\right\}k\,dt\,d\tau=0\ . \tag{14}$$

If the δP's at a given time do not depend on their values at other times, the coefficient of dt in (14) must be zero. So one finds:

$$\int\left\{\delta\mathrm{P}_\mathrm{M}\cdot\operatorname{grad}\varphi-\delta\mathrm{P}_\mathrm{L}\left[\frac{\partial\mathbf{u}}{\partial t}+\frac{\partial\mathbf{u}}{\partial\mathrm{P}}(\mathrm{V}_\mathrm{L})\right]+\mathrm{V}_\mathrm{L}\cdot\frac{\partial\mathbf{u}}{\partial\mathrm{P}}(\delta\mathrm{P}_\mathrm{L})\right\}\rho\,d\tau$$
$$+\iint\delta\mathrm{P}_\mathrm{M}\cdot\left\{\frac{1}{2}\frac{dt}{ds}\operatorname{grad}a+\frac{d}{dt}\left(\frac{dt}{ds}\mathbf{V}_\mathrm{M}\right)\right\}k\,d\tau\ , \tag{15}$$

which has to be satisfied for all systems of δP satisfying the constraints.

Pisa, March 1921.

4c) Correction of a contradiction between the electrodynamic theory and the relativistic theory of electromagnetic masses ([1])

"Correzione di una contraddizione tra la teoria elettrodinamica
e quella relativistica delle masse elettromagnetiche,"
Nuovo Cimento ***25****, 159–170 (1923)*

§ 1. – The theory of electromagnetic masses was studied for the first time by M. Abraham[2] before the discovery of the theory of relativity. Abraham therefore, as was natural, considered in his calculations the mass of a rigid system of charges in the sense of classical mechanics, and he found that, with the hypothesis that such a system had spherical symmetry, its mass varied with the speed and is precisely equal to[3] $\frac{4}{3}\frac{u}{c^2}$ (where u is the electrostatic energy of the system and c is the speed of light) for zero or very small speeds, but for speeds v comparable to c correction terms of order of magnitude v^2/c^2 appear which are a bit complicated. Even before the theory of relativity, FitzGerald introduced the hypothesis that solid bodies underwent a contraction in the direction of motion in the ratio

$$\sqrt{1-\frac{v^2}{c^2}}\ :1$$

and Lorentz redid Abraham's theory of electromagnetic masses, considering instead of rigid systems of electric charges in the sense of classical mechanics, systems that underwent this contraction. The result was that the rest mass, i.e., the limit of the mass for vanishing speed, was still $\frac{4}{3}\frac{u}{c^2}$, but the correction terms depending on v^2/c^2 changed. The experiences of Kaufmann, Bucherer and others with the mass of the β particles of radioactive bodies, and with high speed cathodic particles, decided in favor of the Lorentz theory, known as the contractile electron, against Abraham's theory of the rigid electron. This fact at the beginning was interpreted as a proof of the exclusively electromagnetic nature of the mass of electrons, because it was thought that otherwise their mass should be constant. Afterwards the discovery of the theory of relativity led to the consequence that all masses, electromagnetic or not, must vary with the speed like the mass of Lorentz's contractile electron; in this way the previous experiences left undecided the electromagnetic nature or not of the electron mass, being only a confirmation of the theory of relativity. On the other hand the special relativity theory first, and after the general theory, led to attribute a mass u/c^2 to a system with energy u and in this way arose a serious discrepancy between the Lorentz electrodynamic theory, which gives to a spherical

[1] On the same argument see my notes in *Rend. Acc. Lincei*, (5), 31, pp. 84, 306 (1922).

[2] Abraham, Theory of Electricity; Richardson, Electron Theory of Matter, Chapter XI; Lorentz, The Theory of Electrons, p. 37

[3] The electromagnetic mass of an homogeneous spherical shell of charge e, and radius r is $\frac{2}{3}\frac{e^2}{rc^2}$; but if we observe that the electrostatic energy is $u=\frac{1}{2}\frac{e^2}{r}$, we find the mass $\frac{4}{3}\frac{u}{c^2}$.

distribution of electricity the rest mass $\frac{4}{3}\frac{u}{c^2}$, and special relativity which attributes to this distribution the mass u/c^2. That difference[4] is particularly serious given the great importance of the notion of the electromagnetic mass as a foundation for the electronic theory of matter.

This discrepancy showed up dramatically in two recent articles[5] in one of which, using the ordinary electrodynamic theory I considered the electromagnetic masses of a system with arbitrary symmetry, finding that in general they are represented by tensors instead of scalars, that reduce to $\frac{4}{3}\frac{u}{c^2}$ in the spherical symmetry case; in the other one instead, starting from general relativity, I considered the weight of the same systems which was in every case equal to $\frac{u}{c^2}\,g$, where g is the acceleration of gravity.

In the present work we will demonstrate precisely: that the difference between the two values of the mass obtained in the two ways originates in a concept of a rigid body in contradiction with the principle of relativity, which is applied in the electromagnetic theory (as well as in the contractile electron) and leads to the mass $\frac{4}{3}\frac{u}{c^2}$, while a better justified notion of rigid body conforming to the theory of relativity leads to the value u/c^2.

We note that the relativistic dynamics of the electron was done by M. Born[6] who starting from a point of view not essentially different from the usual one naturally found the rest mass $\frac{4}{3}\frac{u}{c^2}$.

Our considerations will be based on Hamilton's principle as the most suitable one to study a problem subject to very complicated constraints; in fact our system of electric charges must satisfy a constraint of a nature that is different from those considered in ordinary mechanics, since it has to exhibit, depending on its speed, the Lorentz contraction, as a consequence of the principle of relativity. To avoid misunderstandings, we note that while Lorentz contraction is of order v^2/c^2, its influence on the electromagnetic mass is on the principal terms of this one, i.e., on the rest mass and therefore has a rather bigger importance, being appreciable for very small speeds as well.

§ 2. – So we consider a system of electric charges, sustained by a rigid dielectric that, under the action of an electromagnetic field generated partly from the system itself and partly from external sources, moves with a translation motion describing a world tube in the space-time.[7]

[4] The experiences of Kaufmann and others cannot be useful to understand which of the two results is right, because these allow only the measurement of the speed dependent correction terms which are the same in both theories, while the difference is between the rest masses.

[5] E. Fermi, *N. Cim.*, VI, 22, pp. 176, 192 (1921).

[6] Max Born, *Ann. d. Phys.*, 30, p. 1 (1909)

[7] In the following we consider a Euclidean space-time, because we suppose that the considered electromagnetic fields are small enough to not modify the metric structure.

Let's make precise what we mean by rigid translational motion. To do this we consider a Lorentz inertial frame and we suppose that in this frame at a certain time a point of the system of electric charges has zero speed; we will say that the motion is translational if in the same frame at the same time all the other points of the system have zero speed. This is equivalent to saying that the world lines of our system points are trajectories orthogonal to a family of linear spaces; in fact in a Lorentz-Einstein frame where the space is one of the spaces of the family and the time axis is perpendicular to it, the entire system is at rest at time zero, because the space cuts orthogonally all the world lines of all the points of the system. Using this definition of translational motion, which is essentially the one adopted by M. Born, the rigidity of the system is expressed by the fact that its shape in these spaces perpendicular to the tube remains invariable, i.e., all the sections of the tube are equal to each other.

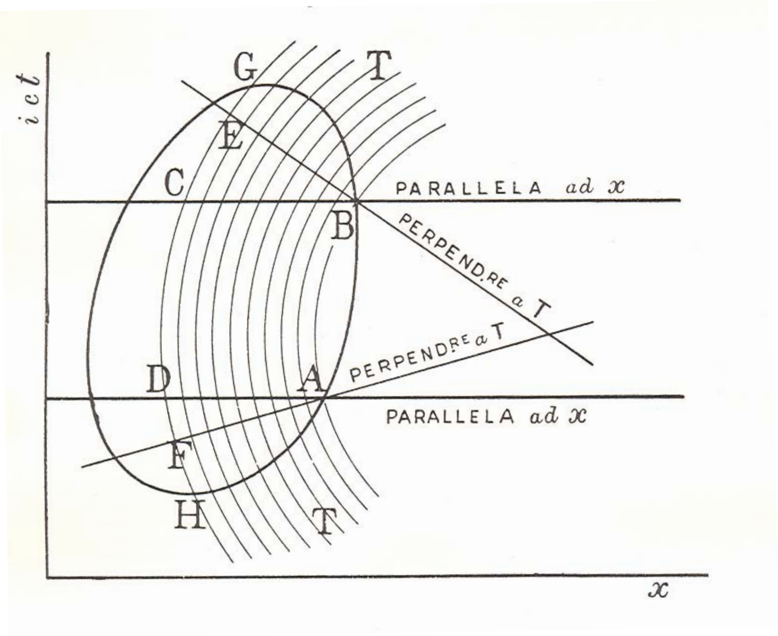


Fig. 1 Translator note: "parallel to x" and "perpendicular to T".

To be able to apply Hamilton's principle to our case there needs to be a variation of the movement of our system consistent with the constraints of the problem, i.e., with the rigidity, correctly interpreted. Now we will show that the value $\frac{4}{3}\frac{u}{c^2}$ or $\frac{u}{c^2}$ is obtained for the electromagnetic mass, if we use either one variation or another of the two that we are going to illustrate and that we distinguish from each other with the letters A and B. The variation A, however, as will immediately be clear, is to be discarded because it is in contradiction with the principle of relativity. Let T be the time tube described by the system. In the figure the space (x, y, z) is represented by only one dimension along the x axis, and the time t is substituted by ict to have a definite metric.

Variation A: one considers as a variation that satisfies the rigidity constraint an infinitesimal displacement, rigid in the ordinary kinematic sense, parallel to the space (x, y, z), of each section of the tube parallel to the same space. In the figure we will obtain such a variation by shifting each section t=const of the tube parallel to the x axis by an arbitrary infinitesimal segment. If we restrict ourselves to consider translational displacement, we will therefore have δx, δy, δz as arbitrary functions only of time, and $\delta t = 0$.

Variation B: one considers as a variation that satisfies the rigidity constraint an infinitesimal displacement perpendicular to the tube of each section normal to the same tube, rigid in the ordinary kinematic sense. In the figure we will obtain this variation by shifting each normal section of the tube parallel to itself by an arbitrary segment.

Of two such variations *A is in obvious contradiction with the principle of relativity* and must be discarded because, not even being Lorentz invariant, it depends on the particular frame (t, x, y, z) we have chosen and can't be the expression of any physical notion, like rigidity. The variation B instead, besides satisfying Lorentz invariance, since it only consists of elements of the tube T completely independent of the position of the frame axes, is the only one presents itself naturally, like that based on a rigid virtual displacement in the frame where at the instant considered the system of charges has zero speed. Now it would be wrong to think that the difference between the consequences of the two methods of variation A and B is significant only for high speeds, i.e., when the tube T has a big slope with respect to the time axis. Instead the calculations we are going to develop will demonstrate immediately that the difference is felt already at zero speed and that precisely A gives $\frac{4}{3}\frac{u}{c^2}$ as the electromagnetic mass the while B gives instead u/c^2.

§ 3.– We indicate the coordinates of time and space by (t, x, y, z) or (x_0, x_1, x_2, x_3) as convenient and let ϕ_i be the four-potential and

$$F_{ik} = \frac{\partial \phi_i}{\partial x_k} - \frac{\partial \phi_k}{\partial x_i}$$

the electromagnetic field, and **E** and **H** the electric and magnetic forces that derive from it.

Hamilton's principle that summarizes the laws of Maxwell Lorentz and those of mechanics says that:[8] the total action, i.e., the sum of the actions of the electromagnetic field and of the material and electric masses, has zero variation under the effect of an arbitrary variation of the ϕ_i and of the coordinates of the points of the electric charge world lines that respect the constraints and are zero on the

[8] Weyl, *Space, Time, Matter*, pp. 194–196; Berlin, Springer (1921).

boundary of the integration region. In our case there aren't material masses, and the only variable elements are the coordinates of the points on the world lines of the charges; therefore it is enough to consider only the action of the electric charges, i.e.:

$$W = \sum_i \int de \int \phi_i \, dx_i$$

where *de* is the generic element of electric charge and the second integral is calculated on the timeline arc described by *de* that is contained in the four-dimensional region G of integration. For each system of variations δx_i satisfying the constraints and that *vanishes on the boundary of G*, one must have $\delta W = 0$, i.e.:

$$\sum_{ik} \int \int de \; F_{ik} \delta x_i dx_k = 0 \; , \tag{1}$$

Now we must examine separately the results obtained substituting δx_i by the values given by the system of variations A or B.

§ 4.– *Consequences of the system of variations* A. — In this case the region of integration reduces to ABCD. The regions BCG, ADH give no contribution, because in them all the δx_i are zero since they have to vanish on the boundary of G, and therefore along the curves BG, AH and must be constants for t =const, i.e., on the straight lines parallel to the x axis. If we label the times of A and B by t_1 and t_2, the equation (1) can be written, since $\delta t = 0$ and δx, δy, δz are functions of the time only:

$$\sum_{ik} \int_{t_1}^{t_2} dt \; \delta x_i \int de \; F_{ik} \frac{dx_k}{dt} \qquad (i = 1, 2, 3) \qquad (k = 0, 1, 2, 3) \; .$$

Since the δx_i are arbitrary functions of t, we obtain the three equations

$$\int de \sum_k F_{ik} \frac{dx_k}{dt} = 0$$

i.e.,

$$\int de[E_x + \frac{dy}{dt} H_z - \frac{dz}{dt} H_y] = 0 \qquad \text{and the analogous two.}$$

If at the chosen instant the system has zero speed in the frame (t, x, y, z) the three equations can be summarized by a single vector equation:

$$\int \mathbf{E} \, de = 0 \; . \tag{2}$$

We could have obtained this equation without calculations if, as is usually done in the ordinary treatment and as M. Born essentially does in the cited work, we had set to zero from the beginning the total force acting on the system. We wanted deduce it using Hamilton's principle to show the fault of its origin, since it follows from the system of variations A that it is in contradiction with the relativity principle.

From (2) follows immediately the value $\frac{4}{3}\frac{u}{c^2}$ for the electromagnetic mass. Suppose in fact that $\mathbf{E}$ is the sum of a part $\mathbf{E}^{(i)}$ due to the system itself, plus a uniform field $\mathbf{E}^{(e)}$ due to external sources. (2) gives:

$$\int \mathbf{E}^{(i)}\, de + \int \mathbf{E}^{(e)}\, de = 0 \ .$$

Now $\int de = e =$ charge; and then $\mathbf{E}^{(e)} \int de = \mathbf{F} =$ external force. In the spherical symmetry case, both direct calculation, and the well known considerations of the electromagnetic moment[9] show that:

$$\int \mathbf{E}^{(i)}\, de = -\frac{4}{3}\frac{u}{c^2}\mathbf{\Gamma} \ ,$$

where $\mathbf{\Gamma}$ is the acceleration.

The previous equation then becomes:

$$\mathbf{F} = \frac{4}{3}\frac{u}{c^2}\mathbf{\Gamma}$$

that compared to the fundamental law of point dynamics, $\mathbf{F} = m\mathbf{\Gamma}$, gives:

$$m = \frac{4}{3}\frac{u}{c^2} \ .$$

§ 5.– *Consequences of the system of variations* B. — In this case the same considerations of the previous section demonstrate that the region of integration reduces to ABEF, i.e., to the region bounded by two normal sections of the tube T. By the use of infinite normal sections Decomposing it using an infinite number of normal sections into layers of infinitesimal thickness, and in order to calculate the contribution of one of these to the integral (1) we refer to its rest frame, by considering the space (x,y,z) parallel to the layer. For this $\delta t = 0$ will hold, while δx, δy, δz will be arbitrary constants. Moreover $dx = dy = dz = 0$, because the speed of all the points is zero, $dt =$ height of the layer, that will vary for each point, because the layer has for its faces two normal sections which in general are not parallel. If O is a generic point but fixed in the layer, for example the origin of coordinates, in which dt has the value dt_0, and $\mathbf{K}$ is the vector with the orientation of the principal normal to the timeline passing for O and size equal to its curvature, we have manifestly, since dt is the thickness at the generic point P of the layer:

$$dt = dt_0[1 - \mathbf{K} \cdot (P - O)] \ .$$

Since the speed is zero we have

$$\mathbf{K} = -\mathbf{\Gamma}/c^2 \ ,$$

and therefore:

$$dt = dt_0 \left(1 + \frac{\mathbf{\Gamma} \cdot (P - O)}{c^2}\right) \ .$$

[9] RICHARDSON loc. cit.

Substituting these values we find that the contribution of our layer to the integral (1) is:

$$-dt_0\Big\{\delta x\int\Big(1+\frac{\mathbf{\Gamma}\cdot(P-O)}{c^2}\Big)E_x de+\delta y\int\Big(1+\frac{\mathbf{\Gamma}\cdot(P-O)}{c^2}\Big)E_y de+ \\ +\delta z\int\Big(1+\frac{\mathbf{\Gamma}\cdot(P-O)}{c^2}\Big)E_z de\Big\}\ .$$

This expression must vanish for all the values of δx, δy, δz and we obtain from it three equations that can be summarized in the single vector equation:

$$\int\Big(1+\frac{\mathbf{\Gamma}\cdot(P-O)}{c^2}\Big)\mathbf{E}\,de=0 \tag{3}$$

A correct application of Hamilton's principle has then brought us to (3) instead of (2). Now it's easy to examine the consequences. Setting

$$\mathbf{E}=\mathbf{E}^{(i)}+\mathbf{E}^{(e)}$$

we find

$$\int\mathbf{E}^{(i)}de+\int\mathbf{E}^{(i)}\,\frac{\mathbf{\Gamma}\cdot(P-O)}{c^2}de+e\mathbf{E}^{(e)}+\mathbf{E}^{(e)}\int\frac{\mathbf{\Gamma}\cdot(P-O)}{c^2}de=0\ .$$

In the spherical symmetry case we have as before

$$\int\mathbf{E}^{(i)}\,de=-\frac{4}{3}\frac{u}{c^2}\mathbf{\Gamma}\ ;$$

substituting in the previous equation we find that $\mathbf{E}^{(e)}$ is compared only with the terms that contain $\mathbf{\Gamma}$. If we neglect the $\mathbf{\Gamma}^2$ terms[10], we can neglect the last integral, and we obtain:

$$-\frac{4}{3}\frac{u}{c^2}\mathbf{\Gamma}+\int\mathbf{E}^{(i)}\,\frac{\mathbf{\Gamma}\cdot(P-O)}{c^2}de+\mathbf{F}=0\ . \tag{4}$$

To calculate the integral which appears in (4) we observe that $\mathbf{E}^{(i)}$ is the sum of the Coulomb force

$$=\int\frac{P-P'}{r^3}de'$$

(P' is the point of charge de' and $r=\overline{PP'}$), and of a term containing $\mathbf{\Gamma}$ that can be neglected because it would give a contribution containing Γ^2. Our integral then becomes:

$$\int\int\frac{P-P'}{r^3}\,\frac{\mathbf{\Gamma}\cdot(P-O)}{c^2}de\,de'\ ;$$

or exchanging P with P', which doesn't change matters, and taking the half sum of the two values obtained in this way:

$$\frac{1}{2}\int\int\frac{P-P'}{cr^3}\,[\mathbf{\Gamma}\cdot(P-P')]de\,de'\ .$$

[10] More precisely the number compared to which the quadratic terms are negligible is $\Gamma\ell/c^2$, where ℓ is the largest length which appears in the problem. It is clear that such an approximation is more than justified in common situations.

We observe that, in our approximation $\mathbf{\Gamma}$ is constant for all the points and then can be taken out of the integrals. Therefore the x component of the previous integral is:

$$\frac{1}{2c^2}\Big\{\mathbf{\Gamma}_x\int\int\frac{(x-x')^2}{r^3}de\,de' + \mathbf{\Gamma}_y\int\int\frac{(y-y')(x-x')}{r^3}de\,de' + \\ + \mathbf{\Gamma}_z\int\int\frac{(z-z')(x-x')}{r^3}de\,de'\Big\}\ .$$

Now, since the system has spherical symmetry, to each segment PP' corresponds an infinite number of other segments differing only in orientation. In the three integrals we can therefore substitute

$$(x-x')^2, (x-x')(y-y'), (x-x')(z-z')$$

by their average values for all the possible orientations of PP', which are; $\frac{1}{3}r^2$, 0, 0.

With that the x component becomes:

$$\frac{\Gamma_x}{3c^2}\frac{1}{2}\int\int\frac{de\,de'}{r}$$

We now observe that the expression

$$\frac{1}{2}\int\int\frac{dede'}{r}$$

is the electrostatic energy u; going back to vector notation we find for the integral appearing in equation (4) the expression: $\dfrac{u}{3c^2}\mathbf{\Gamma}$. (4) becomes in this way:

$$\frac{u}{c^2}\mathbf{\Gamma} = \mathbf{F} \tag{5}$$

that says the electromagnetic mass is u/c^2.

5) Masses in the theory of relativity

"Le masses nella teoria della relatività,"
from A. Kopff, I fondamenti della relatività Einsteiniana,
Eds. R. Conti and T. Bembo, Hoepli, Milano, 1923, pp. 342–344

The grandiose conceptual importance of the theory of relativity as a contribution to a deeper understanding of the relationships between space and time and the often lively and passionate discussions to which it has as a consequence also given given rise outside of the scientific environment, have perhaps diverted attention away from another of its results that, even though less sensational and let's say, even less paradoxical, nevertheless has consequences for physics no less worthy of note, and whose interest is realistically destined to grow in the near term development of science.

The result to which we refer is the discovery of the relationship that ties the mass of a body to its energy. The mass of a body, says the theory of relativity, is equal to its total energy divided for the square of the speed of light. A superficial examination already shows us how, at least for the physics that is observed in the laboratories, the importance of this relationship between mass and energy is such that it considerably overshadows that of the other consequences, quantitatively much lighter, but to which the mind gets used to with more effort. This merits an example: a body one meter long that moves with the respectable enough speed of 30 km per minute (equal more or less to the speed of the earth through space) would always appear to be one meter long to an observer carried along by its motion, while to a fixed observer it would appear to be one meter long less five millionths of a millimeter; as one sees the result, however strange and paradoxical it can seem, is nevertheless very small, and it is hard to believe that the two observers would start quarreling over so little. The relationship between mass and energy brings us instead to enormous figures. For example if one succeeded in releasing the energy contained in a gram of matter, one would obtain an energy greater than that developed over three years of nonstop work by a motor of a thousand horse power (useless to comment!). One might say with reason that it doesn't appear possible, at least in the near future, to find a way to liberate these incredible quantities of energy, something that moreover one would hope not to be able to do, since the explosion of such an incredible quantity of energy would have as its first result reducing to pieces the physicist who had the misfortune to find a way to produce it.

But even if such a complete explosion of matter doesn't appear possible for now, there are already in progress during the past few years some experiments directed towards transforming the chemical elements into each other. Such a transformation, which happens naturally in radioactive bodies, has been recently done artificially by Rutherford who, bombarding some atoms with some α particles (corpuscles launched with huge speed by radioactive substances), has succeeded in obtaining their decomposition. Now to these transformations of the elements into each other are associated energy exchanges that the relationship between mass and energy allows us to study in a very clear way. To illustrate this it is worth another numerical example. We have reason to think that the nucleus of an atom of helium is composed of four nuclei of the hydrogen atom. Now the atomic weight of helium is 4.002 while that of hydrogen is 1.0077. The difference between four times the mass of hydrogen and the mass of the helium is therefore due to the energy of the bonds that unite the four nuclei of hydrogen to form the nucleus of helium. This difference is 0.029 corresponding, according to the relativistic relationship among mass and energy, to an energy of around six billion calories per gram-atom of helium. These figures show that the energy of the nuclear bonds is some million times greater than those of the most energetic chemical bonds and explains to us how against the problem of transformation of matter, the dream of alchemists, for so many centuries the efforts of the best minds have been useless, and how only now, using the most energetic means to our disposition, one has succeeded in obtaining this transformation; moreover in such a small quantity as to illude the most delicate analyses.

These brief indications are enough to show how the theory of relativity, besides giving us a clear interpretation of the relationships between space and time, will be, perhaps in the near future, destined to be the keystone for the resolution of the problem of the structure of matter, the last and more difficult problem of physics.

10) On the mass of radiation in an empty space

"Sulla massa della radiazione in uno spazio vuoto,"
with A. Pontremoli,
Rend. Lincei 32(1), 162–164 (1923)

Recently, one of us[1] had been able to demonstrate, by introducing a more correct concept of rigidity, that the standard electrodynamics allows us to reach a determination of the electron rest mass not different from that coming from the theory of relativity which, as is known, simply amounts to dividing the energy of the system by the squared speed of light. We have observed that a similar difference, between the value determined following from standard electrodynamics and the one given by the theory of relativity, occurs in the calculation of the mass of the radiation in an empty space.[2] We intend to demonstrate that this discrepancy can be removed by analogous arguments. The procedure followed until now for determining by electrodynamics the mass of the radiation in a cavity consisted first of all in evaluating the electromagnetic momentum $\boldsymbol{G}_0$ for slow and quasi-stationary motions, which, neglecting terms in v^2/c^2, results to be given by[3]

$$\boldsymbol{G}_0 = \frac{4}{3}\frac{W_0}{c^2}\boldsymbol{v}$$

where W_0 is the energy of the radiation for the cavity at rest, $\boldsymbol{v}$ is the actual velocity of the cavity, and c is the speed of light. From this, one deduced that the inertial reaction is given by

$$-\frac{d\boldsymbol{G}_0}{dt} = -\frac{4}{3}\frac{W_0}{c^2}\;\Gamma$$

where Γ is the acceleration; whence an apparent mass of the radiation equals $\frac{4}{3}\frac{W_0}{c^2}$, while, according to the theory of relativity, it should be simply $\frac{W_0}{c^2}$. In this procedure it is implicitly contained the assertion that the external force F is equal to the time derivative of the electromagnetic momentum, i.e., to the resultant of the electromagnetic forces $d\varphi$ acting on every single part of the system; in this way, one then puts:

$$F = \int d\varphi. \tag{1}$$

But this is not correct, because, if one considers the notion of rigidity discussed by one of us in the quoted paper, the external force is given instead by

$$F = \int d\varphi\left[1 + \frac{\Gamma(P-O)}{c^2}\right], \tag{2}$$

[1] E. Fermi, these "Rendiconti", Vol. XXXI, pp. 184 and 306 (1922), "Physikalische Zeit.", Vol. XXIII (1922), p. 340.

[2] F. Hasenöhrl, "Ann. der Physik", Vol. XV, p. 344 (1904) and Vol. XVI, p. 589 (1905); K. von Mosengeil, "Ann. der Physik", Vol XXII, p. 867 (1927); M. Planck, "Berlin. Sitzber.", p. 542 (1907); M. Abraham, Theorie der Elektrizität, Vol. II, p. 341 (1920).

[3] M. Abraham, loc. cit. p. 345.

$(P-O)$ being the vector from the point P, where the force $d\varphi$ is applied, to a fixed point O, which we can take as the center of coordinates, internal to the system. Now, $d\varphi$ is the resultant of force $d\varphi_1$, exerted by the radiation pressure which would exist if the cavity were at rest, and a force $d\varphi_2$, caused by the perturbations of this pressure due to the motion of the cavity. By applying (1), since evidently $\int d\varphi_1 = 0$, because $d\varphi_1$ is the force exerted by a homogeneous pressure on a closed surface, one finds that the external force is

$$F = \int d\varphi_2. \tag{3}$$

This force is exactly the one calculated as the inertial reaction by the quoted authors, whence

$$\int d\varphi_2 = -\frac{4}{3}\frac{W_o}{c^2}\ \Gamma \tag{4}$$

On the contrary, by applying (2), still taking into account that $\int d\varphi_1 = 0$, one finds

$$F = \int (d\varphi_1 + d\varphi_2)\left[1 + \frac{\Gamma(P-O)}{c^2}\right] = \int d\varphi_1 \frac{\Gamma(P-O)}{c^2} + \int d\varphi_2 + \int d\varphi_2 \frac{\Gamma(P-O)}{c^2}.$$

Neglecting terms in Γ^2 and observing that $d\varphi_2$ is proportional to Γ, one can simply put

$$F = \int d\varphi_1 \frac{\Gamma(P-O)}{c^2} + \int d\varphi_2. \tag{5}$$

In this case the difference between (3) and (5) is not *a priori* negligible, although it contains c^2 at the denominator, since $d\varphi_1/d\varphi_2$ can become considerably large, being the ratio between a force and its perturbation.[4] In fact $d\varphi_2 = p\boldsymbol{n}d\sigma$, where p is the radiation pressure which, as is known, equals $\dfrac{1}{3}\dfrac{W_o}{V}$, being V the volume of the cavity, and n a unit vector with the direction of the external normal to element $d\sigma$ of the surface of the cavity with coordinates *(x, y, z).* The x component of the first integral of ((5) is then

$$\left[\int d\varphi_1 \frac{\Gamma(P-O)}{c^2}\right]_x = \frac{W_o}{3c^2V}\int (\Gamma_x dx + \Gamma_y dy + \Gamma_z dz)\cos\widehat{nx}\, d\sigma =$$
$$= \frac{W_o}{3c^2V}\left(\Gamma_x \int dx\, \cos\widehat{nx}\, d\sigma + \Gamma_y \int dy\, \cos\widehat{nx}\, d\sigma + \Gamma_z \int dz\, \cos\widehat{nx}\, d\sigma\right);$$

but an immediate application of Gauss's theorem shows that

$$\int dx\, \cos\widehat{nx}\, d\sigma = V, \qquad \int dy\, \cos\widehat{nx}\, d\sigma = \int dz\, \cos\widehat{nx}\, d\sigma = 0.$$

[4] In the case of electromagnetic masses one has $d\varphi$ equal to the resultant of the Coulomb forces (which are the predominant part) and the forces due to the acceleration. For the former, evidently in this case the relation $\int d\varphi_1 = 0$ also holds; therefore these forces make their presence felt only if we apply (5) instead of (3).

Therefore our component is $(W_o\Gamma_x)/3c^2$ and

$$\int d\varphi_1 \frac{\Gamma(P-O)}{c^2} = \frac{W_o\Gamma_x}{3c^2}$$

Considering this relation and (4), it is easy to see that the ratio between the integrals of the right hand side of (5) is $-1/4$ and thus effectively not negligible. By substituting these values into (5), one finds

$$F = -\frac{W_o}{c^2}\ \Gamma$$

from which the requested rest mass results to be equal to W_o/c^2, in accordance with the principle of relativity.

12) The principle of adiabatics and the systems which do not admit angle coordinates

"Il principio delle adiabatiche ed i sistemi che non ammettono coordinate angolari,"
*Nuovo Cimento **25**, 171-175, (1923)*

§ 1. - The importance of the Ehrenfest's principle of adiabatics for the determination of the selection rules for the stationary orbits of a system, in the Bohr theory, is well-known [1]. This principle, as we know, can be enunciated as follows: Let us assume that, in a mechanical system, the forces or the constraints are continuously modified with time but very slowly in comparison with the periods of the system, or, accordig to the Ehrenfest's expression, adiabatically; the principle of adiabatics states that, if the system initially is in a quantum preferred orbit, it will still be there at the end of the transformation.

Let us consider, for instance, a pendulum and imagine to shorten its string at a very low rate in comparison with the period of the pendulum itself. Frequency ν of the pendulum will then grow slowly, but it is easy to realize that energy u also will grow and just so that the ratio u/ν mantains constant. In this way, if this ratio was initially an integer multiple of Planck constant h, it will ever remain the same and then the state of the system will remain quantum preferred during the whole transformation. For further examples we refer to the Ehrenfest's memoir.

The formal basis for the principle of the adiabatics is provided by Burger's theorem [2]. Let us consider a system that in certain general coordinates $q_1, q_2, ..., q_f$ allows the separation of variables [3]. Then put

$$I_K = \oint p_K dq_K \qquad\qquad (K = 1, 2, ..., f) \qquad (1)$$

being p_K the canonically conjugate momentum to q_K and the integral extended, according to the rules of quantum theory, to a complete oscillation of coordinate q_K; in this way the conditions in order that the considered orbit of the system be quantum preferred are:

$$I_1 = n_1 h \ ; \ I_2 = n_2 h \ ; ....; \ I_f = n_f h \qquad (2)$$

being $n_1, n_2, ..., n_f$ integers. Let us suppose, now, to modify adiabatically our system, but in a way it allows the separation of the variables at any instant. Burger's theorem states that in this case integrals $I_1, I_2, ..., I_f$ do not change during the

[1] Ehrenfest, Ann. d. Phys., 51, 327 (1916).

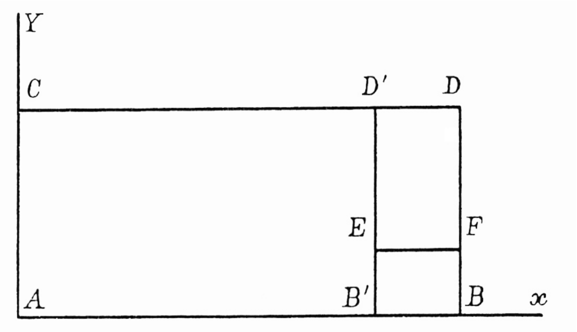
[2] Burgers, Versl. Akad. van. Wetensch. - Amsterdam 1916, 1917; Ann. d. Phys. 52, 195 (1917).

[3] For the validity of Burger's conclusions it is sufficient, more generally, that the system admits angle coordinates, i.e. it is possible to introduce in place of q_K, p_K new variables w_K, jk such that the q_K's, expressed by means of the (w_K, j_K) are periodical with period 1 in variables w_K, and the energy, in the new coordinates, results a function of the j's only. Then, because of the Hamilton equations, the j's result to be constant an the w's linear functions of the time; the q's as functions of the time can be expanded in Fourier series with f indexes.

transformation, i.e. that they are adiabatic invariants. Therefore, if conditions (2) are satisfied at the onset of the transformation, they will be also satisfied at the end; then the principle of the adiabatics is satisfied.

In this Note I intend to show by means of a simple example that if a system adiabatically transforms into another system and the initial and final states both admit the separation of variables, but the intermediate states do not, the I_K are no more adiabatic invariants. In this case the principle of adiabatics loses its basis.

§ 2. - Let us consider a mass point, moving on a plane inside a rectangle; we shall assume that no force acts on the point while it is inside the rectangle, but it bounces off the walls when it hits them. Consider sides AB and AC of the rectangle as coordinate axes x, y. Now, it is evident that our system admits the separation of variables in these coordinates. Calling a, b the lengths of sides AB, AC, coordinate x infact oscillates between values 0, a; coordinate y between values $0, b$.



Moreover, if at a certain instant the components of the velocity are u, v, at an instant whatever they will be $\pm u$, $\pm v$, where one must choose sign + or - according to, whether the relative coordinate is increasing or decreasing at the considered instant. The conjugate momenta to x and y will be $\pm mu$, $\pm mv$, being m the mass of the point; then one will have

$$I_x = \oint (\pm mu)\,dx = \int_0^a mu\,dx + \int_a^0 (-mu)\,dx = 2\,mua \tag{3}$$

and analogously

$$I_y = 2\,mvb \tag{3'}$$

Now we want to study how I_x and I_y change if we transform our system adiabatically. We just intend to transform rectangle ABCD into the other AB'CD'; we remark that such a transformation can be carried out in three ways:

(1) one parallelly shifts the segment BD until to arrive at B'D';
(2) one parallelly shifts the segment BB' until to arrive at DD', so that at an intermediate instant, the mass point can move inside concave polygon AB'EFDC;

(3) one deforms anyway the broken line B'BDD until to bring it to coincide with segment B'D'.

Keeping out the last case, really somewhat complicated, from our considerations, we shall limit ourselves to discuss the former two. As to the first one, we remark that in this case at any instant the point can always move inside a rectangle, therefore also in the intermediate instants it is always possible to have the separation of variables; according to Burger's theorem, in this case we must expect that I_x and I_y remain invariant. This is obviously evident for I_x, since neither b, nor v change during the transformation and then, due to (3'), nor I_y. As to I_x, instead, a decreases during the transformation, being reduced from a = AB to a' = AB'; but in the same time u increases in consequence of the bounces on the moving wall and an immediate consideration shows that things go just so that product au, and then also I_x, remains constant [4], obviously on condition that the transformation is realized slowly enough. If we pass on to consider case (2), it is easy to realize that now things are different. As to I_x, in fact one immediately sees that the x component of the velocity remains unchanged (except for the sign), since it could change its absolute value only hitting a moving wall parallelly to x axis, but the only moving wall, EF, moves parallelly to y; instead a decreases from AB to AB'. In all, therefore I_x reduces in the ratio a'/a and then does not remain constant. Likewise also I_y does not remain constant; in fact b remains unchanged whereas v increases due to the collisions on the moving wall EF. An immediate evaluation shows that v, and then also I_y, increases in the ratio a/a'. From the above considerations we can conclude that integrals I_K are adiabatic invariants only if in the intermediate states the system always admits the separation of variables or at least, according to Burger's theorems, always admits a system of angular coordinates. On the contrary, at least in general, this is not true if the system does not always own a multiperiodic motion. On the other hand, this fact is easyly understandable also from the point of view of quantum theory. In fact one knows, following Bohr, that a well defined quantization is possible only if the motion of the system is multiperiodic. Then one can realize that, if in the intermediate states the system cannot be quantized rigorously, this inexactitude transmits to the final state.

Göttingen, February 1923.

[4] In fact the number of knocks on the moving wall BD in time interval dt is obviously $\frac{u}{2a}dt$; on the other hand, if V is the velocity of wall BD, the velocity of the point will experience an increase of 2V at every knock; then the increase of u in time dt will be:

$$du = 2V\frac{u}{2a}dt = \frac{u}{a}V dt = -\frac{u}{a}da$$

since, obviously, $-da = V dt$. By integrating the preceding equation, we find exactly $ua = const.$, as said above.

13) Some theorems of analytical mechanics of great importance for quantum theory

"Alcuni teoremi di Meccanica Analitica importanti per la Teoria dei Quanti," Nuovo Cimento ***25****, 271–285, (1923)*

§ 1. - Ehrenfest's principle of adiabatics[1], as is known, states that, if a mechanical system is in a quantum orbit and its mechanism, forces or constraints, is changed in an infinitely slow way, the system remains in a quantum selected orbit during the whole transformation. In order that this principle have a definite sense, it is obviously necessary that the final orbit of the system only depends on the final mechanism and not on the one or another sequence of intermediate mechanisms followed during the transformation. Burgers[2] has shown that this is really the case, at least for that kind of systems which up to now has only been considered in quantum theory, i.e. for systems which, or admit a complete separation of variables, or at least can be represented by means of angular coordinates[3]. In this case, their motion can always be considered as resulting from periodic motions, generally having as many periods as many the degrees of freedom are or, in case of degeneracy, with a lower number. But, just at this moment, the study of the simplest atomic structures having been accomplished[4], some problems which do not admit angular coordinates continually occur, first of all the three-body problem which occurs in the study of hydrogen molecule. As is known, all the efforts made up to now to reduce the study of these systems to that of systems with angular coordinates were in vain. Then it is to be desired to investigate whether and how far is it possible to attempt an extension of the principle of adiabatics to the general systems, hoping that it can give some information which can help in the search for rules suitable to determine the preferred orbits of these more general systems.

§ 2. - First of all we shall have better to fix a classification of the systems to be studied. Therefore we turn to the usual representation of the state of the system by means of a point of a 2f-dimensional space Γ, which has $q_1, q_2, ...q_f$ as the general coordinates of the system and $p_1, p_2, ...p_f$ as their conjugate momenta. We have, through each point of this space, a trajectory which corresponds to the motion of the system having its initial position and velocity determined by the point itself. We shall assume the forces and the constraints of the system being time-independent and the forces deriving from a potential so that an integral of the energy conservation does exists. We call E hypersurfaces the ipersurfaces energy=constant; through each point of Γ, one of the E's is passing on which (as provided by the energy integral)

[1] P. Ehrenfest. Ann. d. Phys. 51, p. 327; 1916.

[2] Burgers. Versl. Akad. van Wetensch. Amsterdam, 25 November 1916. - Ann. d. Phys. 52, p. 195; 1917. - Phil. Mag. 33, p. 514; 1917.

[3] See for instance Sommerfeld. "Atombau und Spektrallinen, III ed. Zusatz 7.

[4] They are the hydrogen atom and its various perturbations (Zeeman effect, Stark effect, and Feinstruktur) and the ion of the hydrogen molecule H_2^+, when nucleus rotations are not present.

the trajectory through the point is located. The so called quasi-ergodic[5] mechanical systems enjoy the property that the trajectory generally passes infinitely close to every point of E, so to densely fill a 2f-1 dimensional manifold. However, it may be that our system, besides the energy integral, admits some other uniform integral independent of time. In this case the manifold filled by the trajectory will obviously have a lower number of dimensions. Then let us assume that our system have on the whole m uniform first integrals independent of time,

$$\Phi_1\,(p,q) = c_1; \Phi_2 = c_2; ....; \Phi_m = c_m$$

being c_i arbitrary constants. We shall have, through each point of Γ, a 2f-m dimensional manifold G, intersection of the m hypersurfaces $\Phi_i = c_i$; and the trajectory passing through that point will be wholly contained in G. In general it will not be possible to find, within G, a submanifold which contains the whole trajectory; on the contrary, on the analogy of quasi-ergodic systems, we shall assume for our systems that in general the whole G be densely filled by the trajectory, i.e. that the trajectory passes infinitely close to all the points of G. In this way, the trajectory will come out characterized, at least in its statistical elements, by the only knowledge of the values $\Phi_1, \Phi_2, ..., \Phi_m$ corresponding to it. Therefore we call these values *characteristics of the trajectory.* Then a quasi-ergodic system has only one characteristic, its energy. A system with its energy independent of time, which admits the separation of variables, has in general as many characteristics as degrees of freedom, corresponding to the f a constants of the Jacobi's complete integral; a higher number can only occur in case of degeneracy, i.e. when linear relations with integer coefficients between the fundamental frequencies exist. Let us consider, for instance, the motion of a point in a plane acted on by a force proportional to the distance from two orthogonal straight lines. If the two attraction coefficients are not commensurate, the point describes an open Lissajous' curve in the plane. And in the four-dimensional space Γ the representative point densely fills a two-dimensional surface G. Therefore the system has two characteristics; for them we can take the energies of the projections of the motion on the two orthogonal straight lines. If instead the attraction coefficients are commensurate, the Lissajous' curve degenerates in a closed curve and G becomes one-dimensional; this corresponds to three characteristics.

§ 3. - Now we shall assume to be able to change arbitrarily the forces, or the constraints of the system, i.e. what on the whole, with a happy naming due to P. Hertz [6], we shall call the *mechanism* of the system. If we change the mechanism in an infinitely slow way, we have what is said an adiabatic transformation; and, in § 5, we shall easily find a system of differential equations which shows how the characteristics of the system change when the guiding parameter of the mechanism

[5] The author recently demonstrated that the ordinary mechanical systems are, in general, quasi-ergodic, so that this is the most common case.

[6] P. Herz. Ann. d. Phys. 33, pp. 225, 537; 1910. Weber, Gans. Repertorium der Physik I, 2; 1916. We refer to these works for any explanations regarding the statistical part of the text.

μ, changes adiabatically. But, as we have already mentioned, one can speak of application of the Ehrenfest's principle to a definite system only if the values that its characteristics take at the end of an adiabatic transformation only depend on the final mechanism and not on the intermediate workings crossed during the transformation. To study this question, we shall assume afterwards that the mechanism, rather than depending on only one parameter, depend on two parameters λ and μ. The dependence of the characteristics on λ and μ, instead of being on a system of ordinary differential equations, will then be obviously expressed by a system of equations of total differentials; then the conditions for having the final values of the characteristics not depending on the path followed during the transformation in the λ, μ plane coincide with the integrability conditions for this system. We shall demonstrate that these conditions, for the quasi-ergodic system, are really satisfied. Instead, for the systems having more than one characteristic, in general they are not satisfied although important classes of exceptions exist.

§ 4. - Before passing to study the adiabatic transformations it is convenient to consider some formulae which are useful for calculating the probability that, at any instant, the representative point is in G. Then, for uniforming notations, differently from above we call $x_1, x_2, ..., x_{2f}$, the coordinates of Γ. Our problem can now be formulated in this way: calculate the probability that, at a certain instant, $x_1, x_2, ..., x_{2f-m}$ have values between x_1 and $x_1 + dx_1$, x_2 and $x_2 + dx_2$,....,x_{2f-m} and $x_{2f-m} + dx_{2f-m}$, while the remaining m x's obviously take the values necessary to maintain the representative point in G. As we know, statistical mechanics, through the Liouville's theorem, states that the necessary condition for having a stationary distribution of the points in the Γ space is that their density in Γ should have a constant value on any G. A volume element of Γ can be written $dx_1, dx_2, ..., dx_{2f}$, but also, taking as new variables $x_1, x_2, ..., x_{2f-m}, \Phi_1, \Phi_2, ..., \Phi_m$ as $\frac{1}{D} dx_1, dx_2, ..., dx_{2f-m}, d\Phi_1, d\Phi_2, ..., d\Phi_m$, where D is the functional determinant $\frac{\partial(\Phi_1, ..., \Phi_m)}{\partial(x_{2f-m+1}, ..., x_{2f})}$. And, since during the motion $d\Phi_1, d\Phi_2, ..., d\Phi_m$ obviously remain constant, the aforesaid volume element comes out to be proportional to $\frac{1}{D} dx_1, ..., dx_{2f-m}$. Therefore also the wanted probability is proportional to this expression; and since the total probability is obviously $= 1$, we finally find that the wanted probability is given by

$$\frac{\frac{d\sigma}{D}}{\int \frac{d\sigma}{D}} \tag{1}$$

where for short we put $d\sigma = dx_1, dx_2, ..., dx_{2f-m}$ and the integral is extended to all values of $x_1, x_2, ..., x_{2f-m}$, corresponding to points of G. Before leaving this subject, we also want to deduce a formula that will be useful in the case of quasi-ergodic systems. In this case G is a hypersurface, and we assume for the sake of simplicity it should be closed, and such to be intersected in only one point by the radii vectors coming out from a pole within it. This because a more general approach, even though it is not essentially different, would cause rather complicated calculations. We refer the space Γ to polar coordinates, by characterizing each point by means of its radius vector and the intersection of this one with the unit hypersphere having the pole as centre. We call H the only characteristic, i.e. the energy. In accordance with what said above, the probability that at a certain instant the representative point lie within an element of solid angle $d\omega$ is proportional to the hypervolume comprised between the two hypersurfaces $H\left(x_1, ..., x_{2f}\right) = H$, and $H\left(x_1, ..., x_{2f}\right) = H + dH$, and the solid angle $d\omega$. This volume, except for the constant factor dH, is evidently $\frac{r^{2f-1}d\omega}{H_r}$, where $H_r = \frac{\partial H}{\partial r}$. Since the total probability must be =1 , we find that the wanted probability is given by

$$\frac{r^{2f-1}\frac{d\omega}{H_r}}{\int r^{2f-1}\frac{d\omega}{H_r}} \tag{2}$$

where the integral is extended to the whole unit sphere.

§ 5. - In this section we assume the mechanism of our system as a function of a parameter μ and we aim to study how the characteristics change when this parameter changes adiabatically. Since the mechanism depends on the parameter μ, in general also the characteristics $\Phi_1, \Phi_2, ..., \Phi_m$ will depend on μ, besides the p's and q's. Then, if at a certain instant the parameter μ changes of $d\mu$, characteristic Φ_i will correspondingly undergo the change $\frac{\partial \Phi_i}{\partial \mu}\partial\mu$. Since we are in presence of an adiabatic change, to have the effective change of Φ_i, we must consider the average of this expression which, according to the results of the previous section, will be

$$d\mu\frac{\int \frac{\partial \Phi_i}{\partial \mu}\frac{d\sigma}{D}}{\int \frac{d\sigma}{D}} \tag{3}$$

which results only to be a function of μ and $\Phi_1, ..., \Phi_m$. The dependence of the characteristics on μ in an adiabatic transformation will then be expressed by the system of ordinary differential equations:

$$\frac{d\Phi_1}{d\mu} = \frac{\int \frac{\partial \Phi_1}{\partial \mu}\frac{d\sigma}{D}}{\int \frac{d\sigma}{D}}; \frac{d\Phi_2}{d\mu} = \frac{\int \frac{\partial \Phi_2}{\partial \mu}\frac{d\sigma}{D}}{\int \frac{d\sigma}{D}}; ....; \frac{d\Phi_m}{d\mu} = \frac{\int \frac{\partial \Phi_m}{\partial \mu}\frac{d\sigma}{D}}{\int \frac{d\sigma}{D}} \tag{4}$$

If we know the values of the Φ's, for instance for $\mu = 0$, the integration of this system gives us their values for any μ . In the particular case of the quasi-ergodic systems, system (4) reduces to the only equation:

$$\frac{dH}{d\mu} = \frac{\int \frac{H_\mu}{H_r} r^{2f-1} d\omega}{\int \frac{d\omega}{H_r} r^{2f-1}} \tag{5}$$

where $H_\mu = \frac{\partial H}{\partial \mu}$.

§ 6. - Now we want to study in which cases the final values of the characteristics are independent of the way followed in passing adiabatically from the initial mechanism to the final one. Therefore we shall represent the mechanism of the system as a function of two parameters, λ and μ . If one alters adiabatically these two parameters, of $d\lambda$ and $d\mu$ respectively, the same conclusion of the preceding section shows that the corresponding change of the characteristic is:

$$d\Phi_i = \frac{\int \frac{\partial \Phi_i}{\partial \lambda}\frac{d\sigma}{D}}{\int \frac{d\sigma}{D}} d\lambda + \frac{\int \frac{\partial \Phi_i}{\partial \mu}\frac{d\sigma}{D}}{\int \frac{d\sigma}{D}} d\mu \qquad (i = 1, 2, ..., m) \tag{6}$$

The coefficients of $d\lambda$ and $d\mu$ are evidently functions of only λ and $\Phi_1, ..., \Phi_m$, then m equations (6) represent a system of equations of total differentials; if it will result unlimitedly integrable, the final values of Φ's will be effectively independent of the way followed during the transformation, or else it will not be so. We want to demonstrate that, in the case of quasi-ergodic systems, the condition of unlimited integrability is satisfied. In fact, for these systems, system (6) reduces to only an equation of total differentials analogous to (5)

$$dH = Ld\lambda + Md\mu \tag{7}$$

where

$$L = \frac{\int \frac{r^{2f-1}H_\lambda d\omega}{H_r}}{\int \frac{r^{2f-1}d\omega}{H_r}}; \quad M = \frac{\int \frac{r^{2f-1}H_\mu d\omega}{H_r}}{\int \frac{r^{2f-1}d\omega}{H_r}} \tag{8}$$

and then L and M represent two functions of λ, μ and H. As we know, for obtaining the unlimited integrability of (7), it is necessary and sufficient that the total derivatives of L with respect to μ and of M with respect to λ be equal. Therefore it must be

$$\frac{\partial L}{\partial \mu} + M\frac{\partial L}{\partial H} = \frac{\partial M}{\partial \lambda} + L\frac{\partial M}{\partial H}. \tag{9}$$

To demonstrate that this equality is really satisfied, let us begin to calculate its first term. Therefore, let us imagine to give independent variations δH and $\delta\mu$ to H and μ, leaving λ unchanged; then we will have

$$\delta L = \frac{\partial L}{\partial H}\delta H + \frac{\partial L}{\delta \mu}\delta\mu. \tag{10}$$

On the other hand, from the first of (8), we remark that:

$$\delta L = \frac{1}{\left(\int \frac{r^{2f-1}d\omega}{H_r}\right)^2}\left\{ \left(\int \frac{r^{2f-1}d\omega}{H_r}\right)\delta\int \frac{r^{2f-1}H_\lambda d\omega}{H_r} - \right.$$
$$\left. - \left(\int \frac{r^{2f-1}H_\lambda d\omega}{H_r}\right)\delta\int \frac{r^{2f-1}d\omega}{H_r^2}\right\}. \tag{11}$$

In the calculation of the two variations of the integrals within the curly brackets, we can of course interchange symbols δ and $\int$, as the limits of the integral do not change since it is extended to the whole unit hypersphere. Then we have:

$$\delta\int \frac{r^{2f-1}d\omega}{H_r} = (2f-1)\int \frac{r^{2f-2}\delta r d\omega}{H_r} - \int \frac{r^{2f-1}\delta H_r d\omega}{H_r^2}. \tag{12}$$

On the other hand, from the invariance on the unit sphere, one has:

$$\delta H = H_r \delta r + H_\mu \delta\mu$$

wherefrom

$$\delta r = \frac{\delta H}{H_r} - \frac{H_\mu}{H_r}\delta\mu$$

and also

$$\delta H_r = H_{rr}\delta r + H_{r\mu}\delta\mu = \frac{H_{rr}}{H_r}\delta H + \left(H_{r\mu} - \frac{H_{rr}H_\mu}{H_r}\right)\delta\mu.$$

By substituting in (12) these expressions of δr, δH_r, one finds:

$$\delta\int\frac{r^{2f-1}d\omega}{H_r} = \delta H\biggl\{(2f-1)\int\frac{r^{2f-2}d\omega}{H_r^2} - \int\frac{r^{2f-1}H_{rr}d\omega}{H_r^3}\biggr\} - \\ -\delta\mu\biggl\{(2f-1)\int\frac{r^{2f-2}H_\mu d\omega}{H_r^2} + \int\frac{r^{2f-1}d\omega}{H_r^2}\left(H_{\mu r} - \frac{H_\mu H_{rr}}{H_r}\right)\biggr\}.$$

In a similar way one finds:

$$\delta\int\frac{r^{2f-1}H_\lambda d\omega}{H_r} = \delta H\biggl\{(2f-1)\int\frac{r^{2f-2}H_\lambda d\omega}{H_r^2} + \int\frac{r^{2f-1}H_{\lambda r}d\omega}{H_r^2} - \\ - \int\frac{r^{2f-1}H_\lambda H_{rr}}{H_r^3}d\omega\biggr\} + \delta\mu\biggl\{-(2f-1)\int\frac{r^{2f-2}H_\lambda H_\mu d\omega}{H_r^2} + \\ + \int\frac{r^{2f-1}d\omega}{H_r}\left(H_{\lambda r} - \frac{H_{\lambda r}H_\mu}{H_r}\right) - \int\frac{r^{2f-1}H_\lambda d\omega}{H_r^2}\left(H_{\mu r} - \frac{H_\mu H_{rr}}{H_r}\right)\biggr\}.$$

By substituting in (11) these two last expressions, and comparing with (10), one finally finds:

$$\frac{\partial L}{\partial H} = \frac{1}{\left(\int\frac{r^{2f-1}d\omega}{H_r}\right)^2}\biggl[\left(\int\frac{r^{2f-1}d\omega}{H_r}\right)\biggl\{(2f-1)\int\frac{r^{2f-2}H_\lambda d\omega}{H_r^2} + \\ + \int\frac{r^{2f-1}H_{\lambda r}d\omega}{H_r^2} - \int\frac{r^{2f-1}H_\lambda H_{rr}}{H_r^3}d\omega\biggr\} - \\ - \left(\int\frac{r^{2f-1}H_\lambda d\omega}{H_r}\right)\biggl\{(2f-1)\int\frac{r^{2f-2}d\omega}{H_r^2} - \int\frac{r^{2f-1}H_{rr}d\omega}{H_r^3}\biggr\}\biggr].$$

$$\frac{\partial L}{\partial \mu} = \frac{1}{\left(\int\frac{r^{2f-1}d\omega}{H_r}\right)^2}\biggl[\left(\int\frac{r^{2f-1}d\omega}{H_r}\right)\biggl\{-(2f-1)\int\frac{r^{2f-2}H_\lambda H_\mu d\omega}{H_r^2} + \\ + \int\frac{r^{2f-1}d\omega}{H_r}\left(H_{\lambda\mu} - \frac{H_{\lambda r}H_\mu}{H_r}\right) - \int\frac{r^{2f-1}H_\lambda d\omega}{H_r^2}\left(H_{\mu r} - \frac{H_\mu H_{rr}}{H_r}\right)\biggr\} - \\ - \int\frac{r^{2f-1}H_\lambda d\omega}{H_r}\biggl\{(2f-1)\int\frac{r^{2f-2}H_\mu d\omega}{H_r^2} + \int\frac{r^{2f-1}d\omega}{H_r^2}\left(H_{\mu r} - \frac{H_\mu H_{rr}}{H_r}\right)\biggr\}\biggr].$$

These two last equations, together with the second of (8), give us all the elements necessary to calculate the first term of (9). Once it has been calculated, it is immediate to recognize from its explicit expression that λ and μ appears symmetrically; then ((9) is verified.

Therefore we can conclude that, for the quasi-ergodic systems, the value assumed by the energy at the end of an adiabatic transformation does not depend at all on the intermediate mechanisms of the transformation.

§ 7. - Come back now to be interested in the systems with more than one characteristic. In order that, also for these systems, the final characteristics were independent of the intermediate mechanisms of the transformation, the conditions of unlimited integrability of system (6) should be satisfied. But, if through a calculation, obviously more complicated than that performed in the preceding section but not essentially different from it, we effectively build up these conditions, we find that in general they are not satisfied. Rather than to report here this lengthy calculation, we prefer to show the argument through an example of a system with two characteristics. The example we choose is very similar to another one I have recently used in a note on the principle of adiabatics. From an origin O, we draw in a plane two orthogonal axes x, y. Then we take in the first quadrant two points P, Q and draw the perpendiculars from them to axes (PA, PB, QC, QD). We shall assume that P be internal to the rectangle OCQD. Now let us suppose that inside concave polygon APBDQCA a mass point is moving not acted on by forces and elastically bouncing off the walls of the polygon. Absolute values u, v of the components of the velocity of the point on axes x, y keep evidently constant during the motion, therefore the system has two characteristics. Let us suppose then to keep point Q (of coordinates a, b) fixed and to move point P (of coordinates λ, μ). In this way we shall have accomplished a mechanical system with two characteristics u, v and depending on two parameters λ, μ. By easy arguments, analogous to the ones carried out in the note quoted above, one finds that, changing adiabatically the position of point P, u and v change following the rule:

$$d\log u = \frac{2\mu\, d\lambda}{ab - \lambda\mu}; \quad d\log v = \frac{2\lambda\, d\mu}{ab - \lambda\mu}$$

obviously none of these two equations is unlimitedly integrable; therefore the values that u and v take at the end of a transformation also depend on the path followed by point P. Then, in general, it is not possible to apply Ehrenfest's principle to systems with more characteristics.

§ 8. - However, some important classes of exceptions to this rule exist. We aim to study them in this section. The first one, and also the most important, is that of the systems with angular coordinates. Of these systems, according to Burgers' theorems, we not only know that Ehrenfest's principle can be applied (in the sense that it leads in any case to definite final conditions) but also that for them the aforesaid principle results to be *verified, by experience* as a logical consequence of Sommerfeld's conditions which are supported by all the theory and the experience made on the hydrogen atom. Another remarkable class of exceptions to the conclusions of § 7 is the following: Let us assume that of the m characteristics of our system only one, the energy, depends explicitly on parameters λ, μ of the mechanism. I say that for these systems, at the end of every adiabatic transformation, the energy takes a value independent of the intermediate mechanisms, while the other characteristics even stay unchanged. The fact that all the characteristics, but the energy, stay unchanged comes out evident from the circumstance that, since they do not contain the parameters explicitly, stay unchanged in all the elementary processes of the transformation; the same conclusion can be drawn from system (6) since, if Φ_i is one of these characteristics, one has by hypothesis $\frac{\partial \Phi_i}{\partial \lambda} = \frac{\partial \Phi_i}{\partial \mu} = 0$. For demonstrating that the final value of the energy does not depend on the path followed during the transformation in the plane λ, μ, one could put forward a consideration analogous to that of § 6. But it is easier to remark that, on the basis of the hypothesis, by means of a canonical transformation independent of the parameters, one can try to transform the characteristics independent of the parameters into coordinates of Γ. After this, the consideration of § 6 can be repeated word for word and the constant characteristics simply stand for constant parameter. Systems of this kind occur very frequently in applications; for instance, of this kind are all the systems which have, as only uniform integrals besides the energy (and not dependent on the energy), some integral of the conservation of momentum, or angular momentum, since the latter are always independent of the parameters of the mechanism.

§ 9. - As regards a possible application of what said to the theory of quanta, we remark the following: On the basis of our conclusions, the possibility of an extension of Ehrenfest's principle is ruled out, save the mentioned exceptions. Instead, for quasi-ergodic systems, or the exceptions studied in § 8, such an application is not a priori ruled out, though obviously it is not possible now to foresee if the experience will confirm its results. All the same, one might try if, going on this way, some useful information on the rules for the determination of the quantum orbits of the systems without angular coordinates could be obtained. Of course, Ehrenfest's principle by itself, even if in case the experience should confirm it in this more general application, is not sufficient for the determination of such rules. It only allows us, when we know the selected orbits of a certain system, to deduce the orbits for all the systems which can be obtained from it by means of an adiabatic transformation. Therefore perhaps it might be useful, apart from the complexity of calculations, for finding the quantitative relations between the spark spectra, for instance of the alkaline metals, and the arc spectra of the noble gases. In fact, the systems which emit these spectra only differ in the charge of the nucleus and then can be easily transformed the one into the other.

Göttingen, April 1923.

38b) A theorem of calculus of probability and some applications

"Un teorema di calcolo delle probabilitá ed alcune sue applicazioni,"
Teacher's Diploma Thesis of the Scuola Normale di Pisa.
Presented on June 20, 1922.

§ 1. The theorem we want to deal with concerns the properties of the sums of many incoherent addenda having a known stastistical distribution. The fundamental theorem on these sums is due to Laplace[1]. We announce the theorem together with a short account of its demonstration from which we shall start for establishing a new theorem. Let n be a very great number and $y_1, y_2 \ldots y_n$ represent n unknowns, of which we know the statistical distribution; that is, we know that the probability that y_i has a value ranging between y_i and $y_i + dy_i$ is $\varphi_i(y_i)dy_i$, being φ_i a known function for which, obviously

$$\int_{-\infty}^{\infty} \varphi_i(y)dy = 1, \tag{1}$$

which means that y_i has certainly a value between $-\infty$ and $+\infty$. In addition we will assume that the statistical distribution of y_i is not affected by the values that the other y's can assume, that is, we assume the y_i's are completely incoherent among them. Then we take y_i having a vanishing average, that is:

$$\bar{y}_i = \int_{-\infty}^{\infty} y\varphi_i(y)dy = 0. \tag{2}$$

Finally the average of the squared y_i is put as

$$\bar{y}_i^2 = \int_{-\infty}^{\infty} y^2\varphi_i(y)dy = k_i^2 \tag{3}$$

and assume that, for any i, k_i^2 is negligible with respect to $\sum_1^n k_i^2$. Under these assumptions, the Laplace's theorem holds which says that: The probability that inequalities

$$x \leq \sum_1^n y_i \leq x + dx \tag{4}$$

hold at the same time is given by

$$F(x)dx = \frac{1}{\sqrt{2\pi \sum_1^n k_i^2}} e^{-\frac{x^2}{2\sum_1^n k_i^2}} dx. \tag{5}$$

To demonstrate it, we call r a number $\leq n$ and let $F(r, x)dx$ be the probability that inequalities

$$x \leq \sum_1^r y_i \leq x + dx \tag{6}$$

[1] Théorie analytique des probabilités, Oeuvres, VII, p. 309.

hold true. Now, if p is any value, let us look for the probability that inequalities

$$\sum_1^{r-1} y_i < p < \sum_1^{r} y_i \tag{7}$$

hold together, that is, that the addition of y_r to $\sum_1^{r-1} y_i$ does not exceed p. This probability is obviously given by

$$\int_0^\infty d\xi F(r-1, p-\xi) \int_\xi^\infty \varphi_r(y) dy.$$

Analogously, the probability that inequalities

$$\sum_1^{r-1} y_i > p > \sum_1^{r} y_i \tag{8}$$

hold together is

$$\int_0^\infty d\xi F(r-1, p+\xi) \int_\xi^\infty \varphi_r(y) dy.$$

The difference of these two probabilities is obviously given by the difference between the probability that $\sum_1^r y_i > p$ and the probability that $\sum_1^{r-1} y_i > p$, that is by

$$\int_p^\infty F(r,x)dx - \int_p^\infty F(r-1,x)dx \quad .$$

Then we have

$$\int_p^\infty F(r,x)dx - \int_p^\infty F(r-1,x)dx = \int_0^\infty d\xi F(r-1, p-\xi) \int_\xi^\infty \varphi_r(y) dy -$$

$$- \int_0^\infty d\xi F(r-1, p+\xi) \int_\xi^\infty \varphi_r(y) dy.$$

In the r.h.s. we can reverse the integrations by formulae

$$\int_0^\infty d\xi \int_\xi^\infty dy = \int_0^\infty dy \int_0^y d\xi \quad ; \quad \int_0^\infty d\xi \int_{-\infty}^{-\xi} dy = \int_{-\infty}^0 dy \int_0^{-y} d\xi$$

and it becomes, also changing in the second term ξ with $-\xi$

$$\int_{-\infty}^\infty \varphi_r(y) dy \int_0^y F(r-1, p-\xi) d\xi.$$

We put, as an approximation

$$F(r-1, p-\xi) = F(r-1,p) - \xi \frac{\partial F(r-1,p)}{\partial p}.$$

Thus the above expression becomes

$$F(r-1,p) \int_{-\infty}^\infty \varphi_r(y) dy \int_0^y d\xi - \frac{\partial F(r-1,p)}{\partial p} \int_{-\infty}^\infty \varphi_r(y) dy \int_0^y \xi d\xi =$$

$$= F(r-1,p)\int_{-\infty}^{\infty} y\varphi_r(y)dy - \frac{1}{2}\frac{\partial F(r-1,p)}{\partial p}\int_{-\infty}^{\infty} y^2\varphi_r(y)dy$$

i.e., remembering (2) and (3):

$$-\frac{k_r^2}{2}\frac{\partial F(r-1,p)}{\partial p}.$$

In this way we obtain equality

$$\int_p^{\infty} F(r,x)dx - \int_p^{\infty} F(r-1,x)dx = -\frac{k_r^2}{2}\frac{\partial F(r-1,p)}{\partial p}. \qquad (9)$$

Differentiating it with respect to p we obtain

$$-F(r,p) + F(r-1,p) = -\frac{k_r^2}{2}\frac{\partial^2 F(r-1,p)}{\partial p^2}. \qquad (10)$$

Let us change in it $r-1$ with r, p with x, and, in our approximation, put

$$F(r+1,x) - F(r,x) = \frac{\partial}{\partial r}F(r,x).$$

Then (10) gives, for $F(r,x)$, differential equation

$$\frac{\partial}{\partial r}F(r,x) = -\frac{k_{r+1}^2}{2}\frac{\partial^2}{\partial x}F(r,x). \qquad (11)$$

Changing r with the other variable

$$t = \int_0^{r+1} k_i^2 di \qquad (12)$$

(11) becomes

$$\frac{\partial F}{\partial t} = \frac{1}{2}\frac{\partial^2 F}{\partial x^2}\ . \qquad (13)$$

Then one has, obviously, the condition that, for any t

$$\int_{-\infty}^{\infty} F dx = 1 \qquad (14)$$

and that, for $t = 0$, F has a non vanishing value only when $|x|$ is infinitesimal. It is known that these conditions are more than sufficient to determine F. They are satisfied by putting

$$F = \frac{1}{\sqrt{2\pi t}}e^{-\frac{x^2}{2t}}.$$

By giving to t its value, which at our degree of approximation is $\sum_1^r k_i^2$, we find

$$F(r,x) = \frac{1}{\sqrt{2\pi\sum_1^r k_i^2}}e^{-\frac{x^2}{2\sum_1^r k_i^2}}. \qquad (15)$$

Then one obviously has $F(x) = F(n,x)$, and then

$$F(x) = \frac{1}{\sqrt{2\pi\sum_1^n k_i^2}}e^{-\frac{x^2}{2\sum_1^n k_i^2}} \qquad \text{q.e.d.}$$

§ 2. Let us mantain the notations and the assumptions made at the beginning of the previous section and in addition assume that all $\varphi_i(y)$ are equal (as a consequence we will cancel their index). Then let us indicate with a a positive value whatever. Thus we can state the following

Theorem 2.1. *The probability that at least one among the quantities*

$$y_1, y_1 + y_2, y_1 + y_2 + y_3, \ldots, \sum_1^n y_n$$

exceeds a is given by

$$\frac{2}{\sqrt{\pi}} \int_{\frac{a}{\sqrt{2nk^2}}}^{\infty} e^{-x^2} dx$$

provided that a is great enough with respect to k.

In particular, if n tends to infinity, such probability tends to 1, i.e. to certitude. To demonstrate it, let us indicate with $F(r,x)dx(x < a)$ the probability that the inequalities (6) are fulfilled and in addition all r quantities

$$y_1, y_1 + y_2, \ldots, \sum_1^r y_i \tag{16}$$

are lower than a. At the same time, the same arguments of the previous section show us that $F(r,x)$ still will satisfy the differential equation (11) which, in this case, can be written as

$$\frac{\partial F}{\partial r} = \frac{k^2}{2} \frac{\partial^2 F}{\partial x^2} \tag{17}$$

The boundary conditions will be changed instead. In fact, we observe that

$$\int_{-\infty}^{a} F(r,x)dx$$

gives the probability that none of quantities (16) exceeds a and then

$$-\int_{-\infty}^{a} F(r+1,x)dx + \int_{-\infty}^{a} F(r,x)dx$$

gives the proability that, because of the addition of y_{r+1}, $\sum_1^r y_i$ arrives at exceeding a. A calculation analogous to that performed in the previous section shows us that this probability is

$$\int_0^{\infty} F(r, a - \xi)d\xi \int_{\xi}^{\infty} \varphi(y)dy$$

i.e., at our degree of approximation, neglecting ξ with respect to a

$$F(r,a) \int_0^{\infty} d\xi \int_{\xi}^{\infty} \varphi(y)dy$$

that is, by reversing the quadratures

$$F(r,a)\int_0^\infty \varphi(y)dy\int_0^y d\xi = F(r,a)\int_0^\infty y\varphi(y)dy.$$

By putting now

$$h=\int_0^\infty y\varphi(y)dy \tag{18}$$

we find

$$\int_{-\infty}^{a}\{F(r+1,x)-F(r,x)\}\,dx=-hF(r,a).$$

But, at our usual degree of approximation, we can put

$$F(r+1,x)-F(r,x)=\frac{\partial F(r,x)}{\partial r}$$

and the previous equation becomes

$$\frac{\partial}{\partial r}\int_{-\infty}^{a}F(r,x)dx=-hF(r,a). \tag{19}$$

After all, our unknown function F must fulfill differential equation (17) in interval $-\infty, a$; fulfill equation (19) in extreme a; then it must vanish together with its derivatives in extreme $-\infty$ and, for $r=0$, have a non-vanishing value only for $|x|$ very small, but with the condition that the area comprised between it and x axis is $=1$. It is easy to prove that at least when h is positive, as in our case, these conditions are sufficient to determine F. Therefore, we observe that, by multiplying (17) by dx and integrating it between $-\infty$ and a, one finds

$$\frac{k^2}{2}\left(\frac{\partial F}{\partial x}\right)_a=\frac{\partial}{\partial r}\int_{-\infty}^{a}F(r,x)dx$$

as a consequence, (19) becomes

$$\frac{k^2}{2h}\left(\frac{\partial F(r,x)}{\partial x}\right)_a+F(r,a)=0. \tag{19}$$

Then, for our purpose, it is evidently sufficient to prove that, if a function $\Phi(r,x)$ is $=0$ for $r=0$ and fulfilles equations

$$\frac{\partial\Phi}{\partial r}=\frac{k^2}{2}\frac{\partial^2\Phi}{\partial x^2}\quad;\quad\frac{k^2}{2h}\left(\frac{\partial\Phi}{\partial x}\right)_{x=a}+\phi(r,a)=0 \tag{20}$$

and, for $x=-\infty$, it is always $=0$, it is certainly identically zero. In fact one has

$$\int_{-\infty}^{a}\left(\frac{\partial\Phi}{\partial x}\right)^2dx=\int_{-\infty}^{a}\frac{\partial}{\partial x}\left(\Phi\frac{\partial\Phi}{\partial x}\right)dx-\int_{-\infty}^{a}\Phi\frac{\partial^2\Phi}{\partial x^2}dx$$

that is, owing to (20)

$$\int_{-\infty}^{a}\left(\frac{\partial\Phi}{\partial x}\right)^2dx=\left(\Phi\frac{\partial\Phi}{\partial x}\right)_{-\infty}^{a}-\frac{2}{k^2}\int_{-\infty}^{a}\Phi\frac{\partial\Phi}{\partial r}dx=$$

$$= \Phi(r,a)\left(\frac{\partial\Phi}{\partial x}\right)_{x=a} - \frac{1}{k^2}\frac{\partial}{\partial r}\int_{-\infty}^{a}\Phi^2 dx = -\frac{2h}{k^2}\Phi^2(r,a) - \frac{1}{k^2}\frac{\partial}{\partial r}\int_{-\infty}^{a}\Phi^2 dx$$

i.e.

$$\int_{-\infty}^{a}\left(\frac{\partial\Phi}{\partial x}\right)^2 dx + -\frac{2h}{k^2}\Phi^2(r,a) + \frac{1}{k^2}\frac{\partial}{\partial r}\int_{-\infty}^{a}\Phi^2(r,x)dx = 0\ . \tag{21}$$

Let us now suppose that, for some value of r and x, Φ could be different from zero; then for some value $\overline{r}$ of r $\int_{-\infty}^{a}\Phi^2 dx$ would be certainly positive; in addition, since for $r = 0$ is $\phi = 0$, and then $\int_{-\infty}^{a}\Phi^2(0,x)dx = 0$, there will be certainly between zero and $\overline{r}$ some value of r for which $\frac{d}{dr}\int_{-\infty}^{a}\Phi^2(r,x)dx$ is positive. Now, the first two terms in (21) cannot be negative; the first one is, at least in some cases, positive and this is absurd. Then it will certainly be always $\phi(r,x) = 0$.
q.e.d.
Granted that, it will be enough for us to find a solution whatever fulfilling the imposed conditions for being sure it is the solution we were looking for. Let us try if our conditions can be satisfied by putting

$$F(r,x) = \frac{1}{k\sqrt{2\pi r}}e^{-\frac{x^2}{2rk^2}} - \frac{1}{k\sqrt{2\pi}}\int_0^r \frac{u(\rho)e^{-\frac{(a-x)^2}{2(r-\rho)k^2}}}{\sqrt{r-\rho}}d\rho \tag{22}$$

being $u(\rho)$ a function to be determined. With this position, differential equation (17) and the limit conditions for $x = -\infty$ and $r = 0$ are certainly satisfied. Then it remains to determine $u(\rho)$ so that (19) is satisfied too. Now, from (22) we have

$$F(r,a) = \frac{1}{k\sqrt{2\pi r}}e^{-\frac{a^2}{2rk^2}} - \frac{1}{k\sqrt{2\pi}}\int_0^r \frac{u(\rho)d\rho}{\sqrt{r-\rho}}$$

$$\int_{-\infty}^{a} F(r,x)dx = \frac{1}{k\sqrt{2\pi r}}\int_{-\infty}^{a} e^{-\frac{x^2}{2rk^2}}dx - \frac{1}{k\sqrt{2\pi}}\int_0^r \frac{u(\rho)d\rho}{\sqrt{r-\rho}}\int_{-\infty}^{a} e^{-\frac{(a-x)^2}{2(r-\rho)k^2}}dx =$$
$$= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\frac{a}{k\sqrt{2r}}} e^{-x^2}dx - \frac{1}{2}\int_0^r u(\rho)d\rho \tag{23}$$

and then

$$\frac{\partial}{\partial r}\int_{-\infty}^{a} F(r,x)dx = -\frac{ae^{-\frac{a^2}{2rk^2}}}{2k\sqrt{2\pi r^3}} - \frac{1}{2}u(r)$$

in this way (19) becomes

$$\frac{e^{-\frac{a^2}{2rk^2}}}{k\sqrt{2\pi r}}\left(h - \frac{a}{2r}\right) = \frac{h}{k\sqrt{2\pi}}\int_0^r \frac{u(\rho)d\rho}{\sqrt{r-\rho}} + \frac{u(r)}{2} \tag{24}$$

that is an integral equation of second kind for the unknown function $u(\rho)$. In spite of all our efforts, we have not suceeded to solve it exactly; we only have an approximate solution. We shall deal with this in a little while. We want to prove first, without approximations, that one has

$$\int_0^\infty u(r)dr = 1\ .$$

Therefore, let ϑ be an arbitrary positive quantity and let us multiplicate both sides of (24) by $\sqrt{\theta}e^{-\theta r}dr$ and integrate then from $r=0$ to $r=\infty$. One finds

$$\frac{\sqrt{\theta}h}{k\sqrt{2\pi}}\int_0^\infty \frac{e^{-\theta r-\frac{a^2}{2rk^2}}}{\sqrt{r}}dr - \frac{a\sqrt{\theta}}{2k\sqrt{2\pi}}\int_0^\infty \frac{e^{-\theta r-\frac{a^2}{2rk^2}}}{r^{3/2}}dr =$$

$$= \frac{h\sqrt{\theta}}{k\sqrt{2\pi}}\int_0^\infty e^{-\theta r}dr\int_0^r \frac{u(\rho)d\rho}{\sqrt{r-\rho}} + \frac{\sqrt{\theta}}{2}\int_0^\infty e^{-\theta r}u(r)dr =$$

$$= \frac{h\sqrt{\theta}}{k\sqrt{2\pi}}\int_0^\infty u(\rho)d\rho\int_\rho^\infty \frac{e^{-\theta r}dr}{\sqrt{r-\rho}} + \frac{\sqrt{\theta}}{2}\int_0^\infty e^{-\theta r}u(r)dr =$$

$$= \frac{h}{k\sqrt{2}}\int_0^\infty e^{-\theta\rho}u(\rho)d\rho + \frac{\sqrt{\theta}}{2}\int_0^\infty e^{-\theta r}u(r)dr\ .$$

In addition one has

$$\sqrt{\theta}\int_0^\infty \frac{e^{-\theta r-\frac{a^2}{2rk^2}}}{\sqrt{r}}dr = 2\int_0^\infty e^{-x^2-\frac{a^2\theta}{2k^2x^2}}dx = \sqrt{\pi}e^{-\frac{a\sqrt{2\theta}}{k}}\ .$$

Passing to the limit for $\theta = 0$ the above equation then becomes

$$\frac{h}{k\sqrt{2}} = \frac{h}{k\sqrt{2}}\int_0^\infty u(\rho)d\rho.$$

From which

$$\int_0^\infty u(\rho)d\rho = 1 \tag{25}$$

q.e.d.

At this point we can already get an interesting result. In fact, from (23) we have

$$\lim_{r=\infty}\int_{-\infty}^a F(r,x)dx = \lim_{r=\infty}\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\frac{a}{k\sqrt{2r}}} e^{-x^2}dx - \frac{1}{2}\int_0^\infty u(r)dr = 0\ . \tag{26}$$

If we remember the meaning of $F(r,x)$ this result can be read: The probability that at least one of values (16) exceeds a becomes certitude when r tends to infinity. We remark that this result holds true independently of the approximation we are going to make to solve (24). Let us pass now to the approximate solution of (24). For this we observe that, as one can immediately verify,

$$w(r) = \frac{ae^{-\frac{a^2}{2rk^2}}}{k\sqrt{2\pi r^3}} \tag{27}$$

is a solution of the integral equation of second kind

$$\frac{e^{-\frac{a^2}{2rk^2}}}{k\sqrt{2\pi r}}\left(h+\frac{a}{2r}\right) = \frac{h}{k\sqrt{2\pi}}\int_0^r \frac{w(\rho)d\rho}{\sqrt{r-\rho}} + \frac{1}{2}w(r) \tag{28}$$

which differs from (24) only in the sign inside the bracket of the left-hand side. Now, owing to the assumptions we have made, whenever r is great enough so that

$e^{-\frac{a^2}{2rk^2}}$ is not too small $a/2r$ is negligible with respect to h and then we shall be allowed to assume $w(r)$ as an approximate solution of (24), by putting

$$u(r) = \frac{ae^{-\frac{a^2}{2rk^2}}}{k\sqrt{2\pi r^3}} \tag{29}$$

It is easy to check that from (29) it is $\int_0^\infty u(r)dr = 1$.
Now, from (23), we get

$$\int_{-\infty}^{a} F(r,x)dx = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\frac{a}{k\sqrt{2r}}} e^{-x^2}dx - \frac{1}{2}\frac{ae^{-\frac{a^2}{2\rho k^2}}}{k\sqrt{2\pi\rho^3}}d\rho =$$

$$= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\frac{a}{k\sqrt{2r}}} e^{-x^2}dx - \frac{1}{\sqrt{\pi}}\int_{\frac{a}{k\sqrt{2r}}}^{\infty} e^{-x^2}dx = 1 - \frac{2}{\sqrt{\pi}}\int_{\frac{a}{k\sqrt{2r}}}^{\infty} e^{-x^2}dx \ .$$

And then

$$1 - \int_{-\infty}^{a} F(r,x)dx = \frac{2}{\sqrt{\pi}}\int_{\frac{a}{k\sqrt{2r}}}^{\infty} e^{-x^2}dx \ . \tag{30}$$

Remembering now the meaning of $F(r,x)$ one immediately realizes that

$$1 - \int_{-\infty}^{a} F(r,x)dx$$

represents the probability that at least one of expressions (16) is greater than a. Therefore (30) completely demonstrates the theorem we have enunciated.

§ 3. The theorem just proved is susceptible of an immediate application to a famous theorem of calculus of probability: Peter and Paul make a game of chance. In each game each one has probability 1/2 to win; the stake is always of k lire. Now Peter is infinitely rich, on the contrary Paul owns only a lire. If at a certain moment Peter is able to win all the substance of Paul, the latter is ruined and is obliged to stop the game. So we are in the case considered in the above theorem and we can conclude that, , after a sufficient number of games Peter will certainly ruin Paul; moreover, if a is much greater than k the probability that this fact happens in n games is

$$\frac{2}{\sqrt{\pi}}\int_{\frac{a}{k\sqrt{2n}}}^{\infty} e^{-x^2}dx$$

§ 4. We want now to apply the above theorem to an astronomic problem. Let us consider an elliptic comet which intersects Jupiter's orbit. The cometary orbit will be obviously perturbed by the action of Jupiter, and this particularly when Jupiter and the comet pass very close. Now it may happen that in these continuous transformations the comet's orbit ends by changing into a parabolic or hyperbolic orbit; then the comet will go away forever escaping from the attraction of Jupiter and the Sun. I want to study what is the probability that this happens in a certain time. As far as I know the theory of the influence of Jupiter on the cometary orbits has never been studied from this point of view; people only dealt with this matter[2] looking for an explanation of the capture of comets with parabolic orbits when passing by chance close to Jupiter. We will make the following simplifying assumptions, the same of the restricted 3-body problem: The comet has a negligible mass, so that it does not perturb nor Jupiter neither the Sun. The mass of Jupiter (m) is negligible with respect to the mass of the Sun (M). In this way we are allowed to assume the Sun as fixed and to consider the orbit of the comet being appreciably perturbed only when passing in the close neighbourhood of Jupiter. Jupiter's orbit is circular. Comet's orbit is coplanar with Jupiter's orbit. We call u the velocity of Jupiter and V the velocity of the comet when it crosses Jupiter's orbit with respect to a reference frame moving along this orbit with velocity u; we indicate with θ the angle between the direction of V and Jupiter's orbit. If v is the absolute velocity of the comet, when it is crossing Jupiter's orbit one will have

$$v^2 = u^2 + \mathrm{V}^2 + 2u\mathrm{V}\cos\theta \tag{31}$$

Let us suppose that once, while the comet is crossing Jupiter's orbit, it passes very close this planet. Then it will be affected by a strong perturbation. Let b be the smallest distance between the two bodies if they were not attracted to one another. According to our assumptions, in order that the perturbation is considerable b must be very small if compared with the curvature radii of the two unperturbed orbits so that, during this "collision", the comet will appreciably describe a keplerian hyperbolic orbit during its motion around Jupiter.

§ 5. Thus, let us consider this relative motion, referring to polar coordinates (r, φ) having Jupiter as a pole and the polar axis parallel to the direction of the incoming comet. Since the motion is a Kepler motion, we have

$$\frac{1}{r} = \mathrm{A} - \mathrm{B}\cos(\varphi - \varphi_0) \tag{32}$$

being A, B, φ_0, constant. Moreover, for $\varphi = 0$, r must be infinite, that is

$$\mathrm{A} - \mathrm{B}\cos\varphi_0 = 0\ . \tag{33}$$

then it must be

$$b = \lim_{r=\infty} r\sin\varphi = \lim_{\varphi=0} \frac{\sin\varphi}{\mathrm{A} - \mathrm{B}\cos(\varphi - \varphi_0)} = -\frac{1}{\mathrm{B}\sin\varphi_0} \tag{34}$$

The areas constant is then evidently Vb and owing to the well known formulae of the Kepler motion one has

$$\mathrm{A} = \frac{m}{\mathrm{V}^2 b^2} \tag{35}$$

[2] TISSERAND, ≪*Traité de mécanique céleste*≫, Tome IV, pp. 198-216; CALLANDREAU, ≪Ann. de l'observatoire≫ T. 22; A. NEWTON, ≪Mem. of the Nat. Acad. of Sci.≫, T. 6.

From (33) and (34) we can now obtain the other two constants. One finds exactly

$$\tan\varphi_0 = -\frac{\mathrm{V}^2 b}{m} \quad , \quad \mathrm{B} = \frac{1}{b}\sqrt{1+\frac{m^2}{b^2\mathrm{V}^4}} \tag{36}$$

Now, let ψ be the angle between the direction of the comet when approaching and its direction when going away. Obviously one will have:

$$\psi = 2\varphi_0 - \pi$$

and then

$$\tan\frac{\psi}{2} = -\cot\varphi_0 = \frac{m}{\mathrm{V}^2 b} \tag{37}$$

We can conclude that the perturbation consists in keeping V unchanged and in altering θ of the angle ψ given by (37). Now it is convenient to calculate the averages of the squares of ψ. Therefore we observe that one has:

$$\psi = 2\arctan\frac{m}{\mathrm{V}^2 b}$$

and then

$$\int_{-\infty}^{\infty}\psi^2 db = 4\int_{-\infty}^{\infty}\left(\arctan\frac{m}{\mathrm{V}^2 b}\right)^2 db =$$

$$= \frac{4m}{\mathrm{V}^2}\int_{-\infty}^{\infty}\left(\arctan\frac{1}{x}\right)^2 dx = \frac{8m}{\mathrm{V}^2}\int_{0}^{\infty}\left(\arctan\frac{1}{x}\right)^2 dx$$

by putting

$$h = \int_{0}^{\infty}\left(\arctan\frac{1}{x}\right)^2 dx \approx 2.5$$

then one has

$$\int_{-\infty}^{\infty}\psi^2 db = \frac{8mh}{\mathrm{V}^2} \tag{38}$$

Now, b being very small, the probability that its value is comprised between b and $b + db$ is obviously

$$\frac{db}{2\pi\mathrm{R}\sin\theta}$$

R being the radius of Jupiter's orbit. The average of the squares of ψ therefore is

$$\bar{\psi^2} = \int_{-\infty}^{\infty}\psi^2\frac{db}{2\pi\mathrm{R}\sin\theta} = \frac{4mh}{\pi\mathrm{R}\ \mathrm{V}^2\sin\theta} \tag{39}$$

§ 6. In its motion around the Sun the energy constant of our comet is given by

$$\frac{v^2}{2} - \frac{\mathrm{M}}{\mathrm{R}} = \mathrm{W}\ .$$

As it is well known, a Kepler orbit is elliptic, parabolic or hyperbolic according as the energy constant is negative, null or positive; now, remembering (31) we find for our comet:

$$\mathrm{W} = \frac{1}{2}\left(u^2 + V^2 + 2uV\cos\theta - 2\frac{\mathrm{M}}{\mathrm{R}}\right)$$

but since for Jupiter we have the relation:

$$\frac{u^2}{\mathrm{R}} = \frac{\mathrm{M}}{\mathrm{R}^2}$$

we can write

$$2\mathrm{W} = \mathrm{V}^2 + 2u\mathrm{V}\cos\theta - \frac{\mathrm{M}}{\mathrm{R}}\ .$$

Since in the subsequent perturbations V is not changed and only θ changes, in order that the comet can become hyperbolic it is necessary that W, negative at present, can become positive in correspondence to suitable values of θ. Then it must be

$$\mathrm{V}^2 + 2u\mathrm{V} > \frac{\mathrm{M}}{\mathrm{R}}$$

but we remark that

$$u = \sqrt{\frac{\mathrm{M}}{\mathrm{R}}}$$

therefore the above inequality can be written:

$$\left(\mathrm{V} + \sqrt{\frac{\mathrm{M}}{\mathrm{R}}}\right)^2 > \frac{2\mathrm{M}}{\mathrm{R}}$$

from which [*] and reduces at the end to

$$\mathrm{V} > \left(\sqrt{2} - 1\right)\sqrt{\frac{\mathrm{M}}{\mathrm{R}}} = \left(\sqrt{2} - 1\right)u\ . \tag{40}$$

Then we will assume this inequality as certainly fulfilled. Moreover, for some values of θ, W must certainly be negative, otherwise the cometary orbit could not be elliptic; so it will be:

$$\mathrm{V}^2 + 2u\mathrm{V} < \frac{\mathrm{M}}{\mathrm{R}}$$

From which as above

$$\mathrm{V} > \left(\sqrt{2} + 1\right)\sqrt{\frac{\mathrm{M}}{\mathrm{R}}} = \left(\sqrt{2} + 1\right)u\ . \tag{41}$$

Therefore let us assume that V fulfil (40) and (41) and indicate with θ_0 that particular value of θ for which the comet's orbit is hyperbolic, i.e. one has $\mathrm{W} = 0$, that is

$$\mathrm{V}^2 + 2u\mathrm{V}\cos\theta_0 = \frac{\mathrm{M}}{\mathrm{R}}$$

[*] Editor's Note: At this point, in the Fermi's manuscript there is a blank line which, obviously, would have contained the expansion of the square of the last formula.

and then

$$\cos\theta_0 = \frac{\frac{\mathrm{M}}{\mathrm{R}} - \mathrm{V}^2}{2u\mathrm{V}} = \frac{u^2 - \mathrm{V}^2}{2u\mathrm{V}} \ . \tag{42}$$

When θ is greater than θ_0, one has W ¡ 0 and then the comet describes an elliptic orbit; on the contrary, when θ is less then θ_0 the orbit is hyperbolic.

Now we will suppose that initially the orbit is elliptic and very stretched, so that θ is very close to θ_0, and precisely slightly greater. We call θ^* this initial value. Whenever the comet goes beyond Jupiter's orbit θ is changed of an amount ψ; the average of the squares of ψ depends indeed on θ, as (39) shows, but since we have supposed that θ remains always very close to θ_0 we can put

$$\bar{\psi}^2 = \frac{4mh}{\pi\mathrm{R}\ \mathrm{V}^2 \sin\theta_0} \tag{43}$$

if after a certain time θ became ¡ θ_0 the comet should become hyperbolic and should go away forever. Therefore we are in condition of being able to apply the theorem of §2. Then we must put $a = \theta^* - \theta_0$; $k^2 = \frac{4mh}{\pi\mathrm{R}\ \mathrm{V}^2 \sin\theta_0}$. And the theorem we proved says us that: The probability that the comet will be changed in hyperbolic after having crossed n times Jupiter's orbit is:

$$\frac{2}{\sqrt{\pi}} \int_{\frac{\theta^* - \theta_0}{\sqrt{\frac{8mhn}{\pi\mathrm{R}\ \mathrm{V}^2 \sin\theta_0}}}}^{\infty} e^{-x^2}\, dx \tag{44}$$

and then tends to 1 when n tends to infinity. In the strict sense one could object that the above calculations would fail if the value of V were such that, when the orbit is parabolic, the comet took the same time as Jupiter to go from A to B, being A the point where the comet enters Jupiter's orbit, and B the point where it goes out. In Figure 1.3, S is the Sun, AJB Jupiter's orbit, AKB the orbit of the comet. But it is easy to realize that this case certainly cannot happen if the comet describes its trajectory with direct motion. In fact, if v is the absolute velocity in A of the comet in its parabolic orbit, one has

$$v^2 = u^2 + \mathrm{V}^2 + 2u\mathrm{V}\cos\theta_0$$

and then from (42)

$$v^2 = 2u^2$$

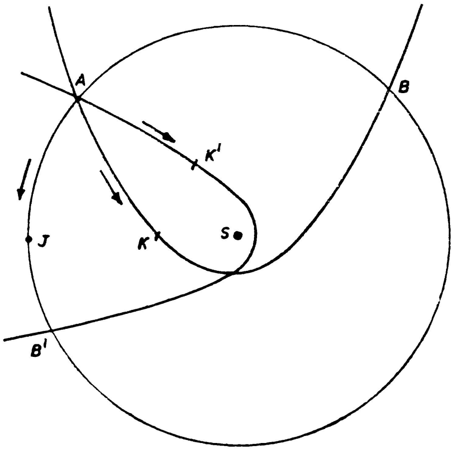
that is:

$$v > u \ . \tag{45}$$

Now, the velocity of the comet is not constant, but in whole tract AKB it is always greater than in the extremes A and B, thus inequality (45) holds true with all the more reason in whole tract AKB. On the other hand, if the motion is direct one has that arc AKB is shorter than arc AJB, and since it is covered with even higher velocity it is certain that the comet will arrive at B before Jupiter. If on the contrary the motion of the comet were retrograde, and it described for instance the orbit AK'B' in the sense indicated by the arrow one would have

$$\text{arc AK'B'} > \text{arc AJB'}$$

and then, though (45) still holds, it is evident that for a particular value of the parameter of the cometary orbit it can happen that the two heavenly bodies take the same time to go from A to B'; of course this can only happen for a particular value of V.



Now if this happened it could occur that the comet, elliptic at first, crossed Jupiter when passing through A and got changed in a parabolic one; but in this case it would meet Jupiter again when passing through B and could in case have a new perturbation which would change it in an elliptic comet again. For this reason we consider this particular value of V ruled out from our calculations.

§ 7. At last we want to consider the possibility that before being changed in hyperbolic the comet can crash into Jupiter and then be destroyed. What is the probability of this event? For this let us look first for the probability that the comet, crossing once Jupiter's orbit, collides with the planet. If we indicate with ρ the sum of the radii of Jupiter and the comet, to have the collision it is necessary that the perihelian distance of Jupiter from the comet, as calculated though the formulae of the Kepler motion is smaller than ρ. Call δ this perihelian distance; from the formulae of §5 it results

$$\frac{1}{\delta} = \mathrm{A} + \mathrm{B}$$

and then from (35) and (36)

$$\frac{1}{\delta} = \frac{m}{\mathrm{V}^2 b^2} + \frac{1}{b}\sqrt{1 + \frac{m}{\mathrm{V}^4 b^2}}$$

If we want the collision occurs it must be $\delta < \rho$ and then

$$\frac{m}{\mathrm{V}^2 b^2} + \frac{1}{b}\sqrt{1 + \frac{m}{\mathrm{V}^4 b^2}} > \frac{1}{\rho}$$

by multiplying this inequality by the quantity, certainly positive

$$\rho \left(\frac{1}{b}\sqrt{1 + \frac{m}{\mathrm{V}^4 b^2}} > \frac{1}{\rho} - \frac{m}{\mathrm{V}^2 b^2} \right)$$

we find

$$\frac{\rho}{b^2} \frac{1}{b}\sqrt{1 + \frac{m}{\mathrm{V}^4 b^2}} > \frac{1}{\rho} - \frac{m}{\mathrm{V}^2 b^2}$$

and summing the last two inequalities

$$\left(\frac{2m}{\mathrm{V}^2} + \rho \right) \frac{1}{b^2} > \frac{1}{\rho}$$

wherefrom finally

$$|b| < \sqrt{\rho^2 + \frac{2m\rho}{\mathrm{V}^2}} \tag{46}$$

We recall now that the probability that the value of b is comprised between b and $b + db$ is $\frac{db}{2\pi \mathrm{R} \sin\theta_0}$ and then probability p that the collision occurs in only one crossing of Jupiter's orbit is given by

$$p = \frac{1}{\pi \mathrm{R} \sin\theta_0} \sqrt{\rho^2 + \frac{2m\rho}{\mathrm{V}^2}} \tag{47}$$

We will assume p very small, and this obviously is equivalent to consider Jupiter's radius negligible if compared with the radius of its orbit. Let us now look for the probability that the collision occurs at the n-th time the comet crosses Jupiter's orbit. Therefore it is evidently necessary that the collision has not occurred before and the probability of this is obviously $(1 - p)^{n-1}$, that is in our approximation

$$e^{-pn} \ .$$

That the comet has not yet been changed in hyperbolic; and, having supposed p extremely small, remembering (44) and putting for the sake of brevity:

$$\frac{\theta^* - \theta_0}{\sqrt{\frac{8mh}{\pi \mathrm{R}\mathrm{V}^2 \sin\theta_0}}} = \mathrm{H}$$

we can hold that the probability of this event is given by

$$1 - \frac{2}{\sqrt{\pi}} \int_{\frac{\mathrm{H}}{\sqrt{n}}}^{\infty} e^{-x^2} dx = \frac{2}{\sqrt{\pi}} \int_0^{\frac{\mathrm{H}}{\sqrt{n}}} e^{-x^2} dx \ .$$

And finally that the collision really occurs, for which we have the probability p. After all the probability that the collision occurs the n-th time is

$$\frac{2e^{-pn} p}{\sqrt{\pi}} \int_0^{\frac{\mathrm{H}}{\sqrt{n}}} e^{-x^2} dx$$

and therefore the probability that the collision occurs a time whatsoever will be the sum of the above expression from $n = 1$ to $n = \infty$, or replacing the sum by an integral

$$\frac{2p}{\sqrt{\pi}} \int_0^\infty e^{-pn} dn \int_0^{\frac{\mathrm{H}}{\sqrt{n}}} e^{-x^2} dx \ .$$

In this expression it is convenient to reverse the integration by the formula

$$\int_0^\infty dn \int_0^{\frac{\mathrm{H}}{\sqrt{n}}} dx = \int_0^\infty dx \int_0^{\frac{\mathrm{H}}{x^2}} dn$$

and in this way one finds for the wanted probability the expression:

$$\frac{2p}{\sqrt{\pi}} \int_0^\infty e^{-x^2} dx \int_0^{\frac{\mathrm{H}}{x^2}} e^{-pn} dn = \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-x^2} \left(1 - e^{-\frac{p\mathrm{H}}{x^2}}\right) dx =$$

$$= 1 - \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-x^2 - \frac{p\mathrm{H}}{x^2}} dx = 1 - e^{-2\sqrt{p\mathrm{H}}} \ .$$

The probability that the collision never occurs is then:

$$e^{-2\sqrt{p\mathrm{H}}} \ .$$

7) Formation of images with Röntgen rays

"Formazione di immagini coi raggi Röntgen,"
Nuovo Cimento ***25****, 63-68 (1923)*

Röntgen rays do not undergo reflections nor refractions, at least in the usual sense of the word, since the reflection of diffraction occurs only under well definite incidence angles. As a consequence in the X-ray optics the problem of obtaining images cannot be solved, as in the ordinary optics, by means of spherical lenses or mirrors. Gouy* suggested theoretically a method for obtaining monochromatic images with X-rays, by means of a cylinder of mica. In a few words it is the following. Let us consider a circular cylinder of mica and suppose that in a point of its axis there is a source S of monochromatic Röntgen rays. They will be reflected on the mica in those points where Bragg's relation is fulfilled: these points obviously are on circular sections of the cylinder. And the rays reflected on one of these circles will gather in a point I on the axis, symmetric of S with respect to the plain of the reflecting circle, where one will have a real monochromatic image of S. If S were in the neighborhood of the axis, still an image of it will be formed in the neighborhood of the axis†. Suppose now to have, in the neighborhood of the axis, a planar figure from which points monochromatic X-rays come out, and to place a plate in the position where its image is formed. Let r be the mirror-object distance, R the radius of the cylinder of mica, θ the Bragg incidence angle, r' the image-mirror distance. If we project everything on a plane orthogonal to the axis of the cylinder of mica, the projections of r and r' will be $r \cos\theta$, $r' \cos\theta$; and then, according to the usual formulae of the spherical mirrors it will be

$$\frac{1}{r\cos\theta}+\frac{1}{r'\cos\theta}=\frac{2}{r} \quad ,$$

from which

$$r'=\frac{Rr}{2r\cos\theta-R} \quad .$$

The linear coefficient of enlargement of the segments orthogonal to r and the axis of the cylinder will be

$$\mu_1=\frac{r'}{r}=\frac{R}{2r\cos\theta-R} \tag{1}$$

If the object is close to the axis we have approximately $\mu_1 = 1$. To calculate the enlargement of the segments parallel to the plane of the axis and of r, let us call φ and φ' the angles that the lines orthogonal to the plane of the object and of the plate form with r and r' respectively. Then one immediately sees that the looked for enlargement is

$$\mu_2=\frac{\cos\varphi'}{\cos\varphi} \tag{2}$$

*C. R. GOUY, ≪C. R.≫, **161**, 176 (1915).

†Of course, provided that the cylinder is confined in a region small enough comprised between two generatrices.

Suppose now to photograph an aperture placed orthogonally to the plane of r and the axis by a flat plate of mica of length l. If h is the length of the aperture, the length of its image will be $2l + k$. If instead we bend the mica in order that the image is formed in the focus, the length will become h. The intensities of the two images will be obviously approximately in the inverse ratio of their lengths. Their ratio is then

$$\frac{2l+k}{h} \quad .$$

If, for instance, $h = 1$ *cm*, $l = 4$ *cm* the ratio is 9. Then the intensity is almost decupled. I shall now describe the way in which I have actually succeeded in obtaining these images. The source of the rays consisted in a tube of the shape and size approximately indicated in Figure 1. I created the vacuum by a rotational pump Cacciari, type Gaede. Cathode K was concave, with a radius of 6 or 7 *cm* when one wanted to concentrate the rays on the anticathode as much as possible; if instead one wanted the whole surface of the anticathode be hit by the rays, the cathode was made with a smaller radius. The anticathode was generally of iron and sometimes was cut almost orthogonally to the cathode rays, in order to do without the slit. Instead, in other experiments it was cut as the spout of a flute, in order to present a large surface to the detecting instruments.

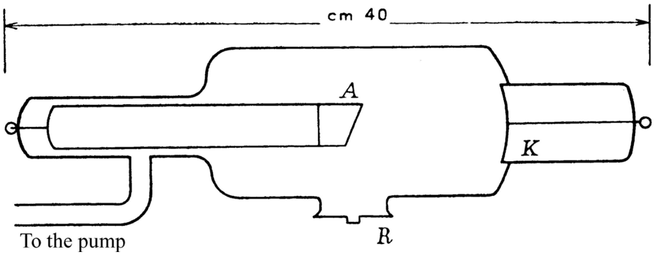


Fig. 1

Since the radiations typical of the iron are largely absorbed by the glass of the bulb, I thought it right to equip the tube with little window of aluminium R. During the work the tube was kept attached to the pump, so that after a short time, it assumed a running regular enough. The tube was driven by a big induction coil with a Wehnelt switch; in ordinary conditions the equivalent spark was 10 or 12 *cm* long. The tube was contained in a small wooden box sheathed by lead 6 *mm* thick on the side of the instruments and 3 *mm* thick on the other sides. To obtain fairly precise images it was necessary the reflecting plate of mica be regular as much as possible. Therefore, it was carefully chosen among many samples; nevertheless I have never succeeded in finding plates that, in reflecting the light, were more regular than an ordinary windowpane. This is the cause of the irregularities and smudges we can observe in the reported images. The mica was bended by binding it fast on a turned brass cylinder. Then a layer of sealing wax (little more than half a centimetre thick) was spread on the convex part. When the sealing wax had cooled one could remove the fastenings and detach the mirror from the cylinder. In this way I succeeded in obtaining cylindrical mirrors relatively precise given the limit imposed by the natural irregularity of the plates used. They had mostly dimensions of 4 × 6 *cm* but usually their aperture was reduced to 4 × 2 *cm* for making use of the less irregular parts, which were judged by trying the mirrors by the reflection of the ordinary light. The mirror was mounted on a graduate circle in order to be able to put it right. (The angle of which was turned for the study of the third order of the K_α of the iron was of $16°50'$). The detection of the rays was performed photographically. I carried out first a few experiments of orientation with planar crystals to verify the nature of the anticathode and the intensities of the reflections of the various orders. It resulted that the double K_α $K_{\alpha'}(\lambda = 1.932; 1.928)$, scarcely resolvable in the experimental conditions in which I was, the $K_\beta(\lambda = 1.748)$ were emitted. The K_γ was scarcely visible due to the low intensity. The most intense orders were the first and the third. I preferred to work in the third in order not to be obliged to use incidence angles too much close to 90°. Then I experienced the indicated method to obtain images first on the anticathode which was also working as an aperture. The distances anticathode crystal and crystal image varied from 18 to 22 *cm*. The exposure lasted about ten minutes.

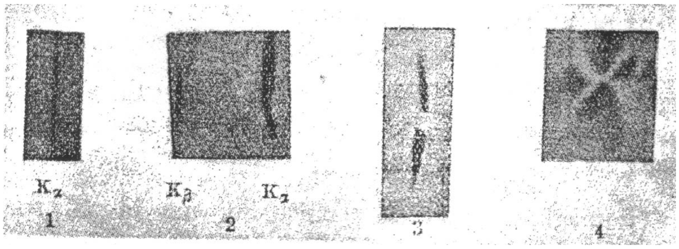


Fig. 2 1-4

I could immediately ascertain the very strong increase of intensity which can be obtained in this way. A rough idea of this is given by Figs. 2, 1, and 2, 2 which represent two photographs of the 3rd order of iron K_α obtained approximately in the same conditions of exposure and operation of the tube, the first one with flat mica and the second one with curved mica. The increase in intensity was indeed such that, particularly using mirrors of 6 *cm* of aperture, accustoming a few minutes the eyes to the darkness of the room, it was possible to see clearly the images on a screen of barium platinum cyanide. From Fig. 2, 2 it is clearly visible that the emission intensity of the central part of the anticathode, where the cathode rays were concentrated, is considerably greater than that of the side parts. It is possible to see this because the method of images allows to observe the slit " Lockyer's art", that is, to observe point by point what happens in the slit. To put this more in evidence I made the following experience: I placed before the window of aluminium a leaden thread of about 1 *mm* of diameter and shifted the photographic plate to carry it in the point where the image of the aluminium window was forming. Fig. 2, 3 gives the result of this experiment; in the figure the gap in the image produced by the leaden thread is clearly visible. Finally Fig. 2, 4 represents an attempt of obtaining an image of an object in two dimensions. The anticathode of iron was therefore cut as the spout of a flute and two cross shaped furrows were cut in it and inside them two copper wires were driven in order to form a sort of X. In Fig. 2, 4, one can see the image of this X, obviously together with several irregularities due to the irregularity of the reflector.

This work was carried out at the Institute of Physics of the University of Pisa in Winter 1922.

30) On the quantization of an ideal monatomic gas

"Sulla quantizzazione
del gas perfetto monoatomico,"
*Rend. Lincei **3**, 145–149 (1926).*

§ 1.– In classical thermodynamics one takes (referring to a single molecule) as specific heat at constant volume of an ideal monatomic gas $c = 3/2k$. However it is clear that, if one wants to admit the validity of the Nerst principle also in the case of an ideal gas, one must think that the above expression of c is only an approximation valid at high temperatures and that, as a matter of fact, c tends to zero for $T = 0$, so that one can extend up to the absolute zero the integral expressing the value of entropy without the indeterminacy of the constant. And for realizing how such a variation of c can occur, it is necessary to admit that the motions of an ideal gas must be quantized as well. Then one realizes that such a quantization, besides the energy content of the gas, will influence the equation of state as well, thus giving rise to the so called phenomena of degeneration of the ideal gas at low temperatures.

The purpose of this work is the exposition of a method for carrying out the quantization of an ideal gas which, in our opinion, is as much as possible independent of unjustified hypotheses on the statistical behaviour of the molecules of the gas.*

Recently various attempts have been made for arriving to establish an equation of state for the ideal gas.

The formulae given by the various authors differ from ours and from the classical equation of state only for very low temperatures and very high densities; unfortunately these are the same circumstances in which the deviations of the laws of the real gases from the ones of ideal gases are more important; and since, on conditions one can easily carry out experimentally, the deviations from the equation of state $pV = kT$ due to the degeneration of the gas, even if not negligible, are always considerably smaller than those due to the fact that the gas is real and not ideal, the former have been so far hidden by the latter. This does not exclude the possibility that, in a more or less near future, and with a more profound knowledge of the forces which act among the molecules of a real gas, one can pull the two deviations apart, thus arriving to choose experimentally among the different theories of the degeneration of the ideal gases.

§ 2.– For being able to carry out the quantization of the motions of the molecules of an ideal gas one must be in such a condition to be able to apply Sommerfield's rules to their motion; and this can be made in an infinite number of ways all of which, besides, lead to the same result. One can, for instance, suppose the gas contained in a parallelepiped vessel with elastic walls, quantizing the three fold periodic motion

*See for instance A. Einstein, *Sitzber. d. Pr. Akad. d. Wiss.* **22**, 261 (1924); **23**, 3, 18 (1925); M. Planck, *Sitzber. d. Pr. Akad. d. Wiss.* **23**, 49, (1925).

of the molecule bouncing off the six walls; or, more generally, one can subject the molecules to a system of forces such as their motion becomes periodic and then can be quantized. The hypothesis that the gas is ideal allow us in all these cases to neglect the forces acting among the molecules, so that the mechanical motion of each of them happens as if the other ones should not exist. Nevertheless one can recognize that the mere quantization, following Sommerfield's rules, of the motion of the molecules, considered mutually independent, is not sufficient for obtaining correct outcomes; since, even if in this case the specific heat tends to zero for $T = 0$, yet his value , besides on temperature and density, comes to depend on the total quantity of gas as well, and tends, at any temperature, to the limit $3/2k$ when, even if the density remains constant, the quantity of gas tends to infinite. Then it appears necessary to admit that some complement to Sommerfield's rules is needed, when calculating systems which, as ours, contain elements indistinguishable between them.[†]

To have an hint on how to formulate the most plausible hypothesis, it is worth to consider how things go in other systems which, as our ideal gas, contain indistinguishable elements; and precisely we want to examine the behaviour of the atoms heavier than hydrogen which all contain more than an electron. If we consider the deep parts of a heavy atom, we are in such conditions that the forces acting among the electrons are very small in comparison with the ones exerted by the nucleus. In these circumstances the mere application of the Sommerfield's rules would lead to expect that, in the normal state of the atom, a considerable number of electrons should lie in an orbit of total quantum number 1. As a matter of fact, instead one sees that the ring K is already saturated when contains two electrons, and likewise the ring becomes saturated when contains 8 electrons, etc. This fact has been interpreted by Stoner,[‡] and in an even still more precise way by Pauli,[§] as follows: let us characterize an electronic orbit possible in a complex atom by means of 4 quantum numbers; n, k, j, m, which have respectively the meanings of total quantum, azimuthal quantum, internal quantum and magnetic quantum. Given the inequalities to which these 4 numbers must satisfy, one finds that, for $n = 1$, only two triplets of values exist of k, j, m: for $n = 2$, there are 8, etc. To realize the above fact, therefore it is sufficient to admit that in the atom two electrons whose orbits are characterized by the same quantum numbers cannot exist; in other words one must admit that an electronic orbit is already "occupied" when contains only one electron.

§ 3. – We now intend to investigate if such hypothesis can give good outcomes in the problem of the quantization of the ideal gas as well: therefore we shall admit that in our gas almost a molecule whose motion is characterized by certain quantum numbers can exist, and we shall show that this hypothesis leads to a

[†] E. Fermi, *N. Cimento* **1**, 145 (1924).

[‡] E. C. Stoner, *Phil. Mag.* **48**, 719 (1924).

[§] W. Pauli, *Zs. f. Phys.* **31**, 765 (1925).

perfectly consequent theory of the quantization of the ideal gas, and in particular it gives reasons for the expected decrease of the specific heat at low temperatures, and leads to the exact value for the constant of entropy of the ideal gas.

Putting off the publication of the mathematical details of the present theory to a next occasion, in this Note we limit ourselves to expose the principles of the method we have followed and the results obtained.

First of all we must put our gas in such a condition that the motion of its molecules results to be quantizable. As we have seen, this can be made in an infinity of ways; but, since the result is independent of the particular way one adopts, we shall choose the most convenient for the calculation; and precisely we shall admit that our molecules are attracted by a fixed point O, with a force proportional to the distance r of the molecule from O; so that each molecule will be a spatial harmonic oscillator whose frequency we call ν. The orbit of the molecule will be characterized by three quantum numbers, s_1, s_2, s_3, which are linked to its energy through the relation

$$w = h\nu(s_1 + s_2 + s_3) = sh\nu \ . \tag{1}$$

Then the energy of a molecule can take all the values integer multiple of $h\nu$, and the value $sh\nu$ can be assumed $Q = \frac{1}{2}(s+1)(s+2)$ ways.

Therefore the zero energy can be realized in only one way, the energy $h\nu$ in 3 ways, the energy $2h\nu$ in 6 ways, etc. To realize the influence of our hypothesis, i.e. that to given quantum numbers can correspond only one molecule, let us consider the extreme case of N molecules to the absolute zero. At this temperature the gas must lie in the state of minimum energy. If we had no limitation to the number of molecules which can have a certain energy, all the molecules would lie in the state of zero energy, and all the three quantum numbers of each of them would be zero. On the contrary, as provided by our hypothesis, the existence of more than one molecule with all the three quantum numbers equal to zero is forbidden; therefore if $N = 1$, the only one molecule will occupy the place of zero energy; if instead $N = 4$, one of the molecules will occupy the place of zero energy, and the other three the place of energy $h\nu$; if $N = 10$, one of the molecules will occupy the place of zero energy, three of them the places of energy $h\nu$, and the remaining six the six places of energy $2h\nu$, etc. Now let us suppose to have to distribute the total energy $W = Eh\nu$ ($E =$ integer) among our molecules; and call $N_s \leq Q_s$ the numbers of molecules of energy $sh\nu$. We find easily that the most probable values of N_s are

$$N_s = \frac{\alpha Q_s}{e^{\beta s} + \alpha} \ , \tag{2}$$

where α and β are constants depending on W and N. To find the relation between these constants and the temperature, we observe that, as a consequence of the attraction toward O, the density of our gas will be a function of r, which must tend to zero for $r = 8$. Accordingly, for $r = 8$ the phenomena of degeneration must cease, and in particular the distribution of velocities, easily deducible from

(2), must change into Maxwell law. Thus one finds that it must be

$$\beta = \frac{h\nu}{kT} \, . \tag{3}$$

Now we are able to deduce from (2) the function $n(L)dL$, which represents, for a given value of r, the density of the molecules of energy between L and $L+dL$ (analogous to the Maxwell law), and from this we can deduce the mean kinetic energy $\bar{L}$ of the molecules at distance r, which is a function, besides of the temperature, of the density n as well. One finds precisely

$$\bar{L} = \frac{3h^2 n^{2/3}}{4\pi m} P\left(\frac{2\pi mkT}{h^2 n^{2/3}}\right) \, . \tag{4}$$

In (4) we have called $P(x)$ a function, of a bit complicated analytic definition, which for values of x either very large or very small, can be calculated through the asymptotic formulae

$$\begin{aligned} P(x) &= x\left(1 + \frac{1}{2^{5/2}x^{3/2}} + \dots\right) \; ; \\ P(x) &= \frac{1}{5}\left(\frac{9\pi}{2}\right)^{1/3}\left[1 + \frac{5}{9}\left(\frac{4\pi^4}{3}\right)^{1/3} x^2 + \dots\right] \, . \end{aligned} \tag{5}$$

To deduce from (4) the equation of state, we apply the virial relation. Then we find that the pressure is given by

$$p = \frac{2}{3} n\bar{L} = \frac{h^2 n^{5/3}}{2\pi m} P\left(\frac{2\pi mkT}{h^2 n^{2/3}}\right) \, . \tag{6}$$

At the limit for high temperatures, that is for small degeneration, the equation of state takes then the form

$$p = nkT\left[1 + \frac{1}{16}\frac{h^3 n}{(\pi mkT)^{3/2}} + \dots\right] \, . \tag{7}$$

Then the pressure results higher than the one coming from the classical equation of state. For an ideal gas having the atomic weight of the helium, at the temperature of absolute 5° and at pressure of 10 atmospheres, the difference would be of 15%. From (4) and (5) one can also deduce the expression of the specific heat for low temperatures. One finds

$$c_v = \left(\frac{16\pi^8}{9}\right)^{1/3} \frac{mk^2}{h^2 n^{2/3}} + \dots \tag{8}$$

Likewise we can find the absolute value of entropy. Carrying out the calculations, at high temperatures one finds

$$S = \int_0^T \frac{1}{T} d\bar{L} = n\left[\frac{5}{2}\log T - \log p + \log\frac{(2\pi m)^{3/2} k^{5/2} e^{5/2}}{h^3}\right] \, , \tag{9}$$

which coincides with the value of entropy given by Tetrode and Stern.

43) A statistical method for the determination of some properties of the atom (*)

*"Un metodo statistico per la determinazione di alcune proprietà dell'atomo," Rend. Lincei **6**, 602–607 (1927).*

The purpose of this work is to show some results about the distribution of electrons in a heavy atom which can be obtained dealing with these electrons, given their great number, using a statistical method; or in other words, considering them as a gas formed by electrons surrounding the nucleus.

Naturally this gas of electrons comes to find itself in a state of complete degeneracy, so much so that we cannot deal with it using classical statistics; on the contrary we must use the form of statistics proposed by the author ([†]) and based on the application of Pauli's exclusion principle to the theory of gas. This has the effect that the kinetic energy of the electrons, in the conditions in which they come to find themselves inside the atom, actually turns out to be bigger than it would have been according to the principle of equipartition of energy and practically independent of the temperature, at least as long as it does not go beyond certain limits.

In this Note we shall show first of all how the distribution of electrons around the nucleus can be calculated statistically; and based on this we shall then calculate the necessary energy to ionize completely the atom, that is to tear off all the electrons from it. The calculation of the distribution of electrons around the nucleus also allows the determination of the behavior of the potential at various distances from the nucleus and therefore to know the electric field in which the electrons of the atom come to find themselves. I hope to be able to show in a future work the application of this to the approximate calculation of the binding energies of single electrons and to some questions about the structure of the periodic system of elements.

To determine the distribution of electrons, we must first of all search for the relation between their density and the electric potential at every point. If V is the potential, the energy of an electron will be $-eV$ and therefore according to classical statistics, the density of electrons would have to then be proportional to $e^{eV/kT}$. But, according to the new statistics, the relation between density and temperature is the following one:

$$n = \frac{(2\pi mkT)^{3/2}}{h^3} F(\alpha e^{eV/kT}) \tag{1}$$

where α is constant for the whole gas; the function F in our case (complete degeneracy), has the asymptotic expression

$$F(A) = \frac{4}{3\sqrt{\pi}} (\log A)^{3/2}. \tag{2}$$

Then in our case we find

$$n = \frac{2^{7/2}\,\pi m^{3/2}\,e^{3/2}}{3h^3} v^{3/2} \tag{3}$$

* Presented in the session of December 4, 1927 by the Fellow O.M. Corbino.

[†] E. Fermi, *Zs. f. Phys.* **36**, 902 (1926).

where

$$v = V + \frac{kT}{e} \log \alpha \tag{4}$$

represents the potential, apart from an additional constant. Now we observe that since in our case we are dealing with a gas of electrons, we must take into account the fact (‡) that the statistical weight of the electron is 2 (corresponding to the two possibilities for the orientation of the spinning electron); and so for the density of electrons we must actually take a value equal to twice the value (3); namely we have:

$$n = \frac{2^{9/2}\,\pi m^{3/2}\,e^{3/2}}{3h^3} v^{3/2} \ . \tag{5}$$

If in our case classical statistics were valid, we would have the average kinetic energy of the electrons $= \frac{3}{2}kT$. On the contrary according to the new statistics it turns out to be

$$L = \frac{3}{2}kTG(\alpha e^{eV/kT})/F(\alpha e^{eV/kT})$$

where G represents a function that, in the case of complete degeneracy, takes the asymptotic expression

$$F(A) = \frac{8}{15\sqrt{\pi}}(\log A)^{5/2}.$$

Therefore we find for our case

$$L = \frac{3}{5}\,ev \ . \tag{6}$$

Now we observe that the electric density at a point is evidently given by $-ne$ so the potential v satisfies the equation

$$\Delta v = 4\,\pi\,ne = \frac{2^{13/2}\,\pi^2 m^{3/2}\,e^{5/2}}{3h^3} v^{3/2} \ . \tag{7}$$

Since in our case it will then evidently be only a function of the distance r from the nucleus; then (7) can be written

$$\frac{d^2v}{dr^2} + \frac{2}{r}\frac{dv}{dr} = \frac{2^{13/2}\,\pi^2 m^{3/2}\,e^{5/2}}{3h^3} v^{3/2} \ . \tag{8}$$

If we indicate by Z the atomic number of our atom we shall evidently have

$$\lim_{r=0} rv \ = \ Ze \tag{9}$$

$$\int n d\tau = 4\pi \int_0^\infty r^2\, ndr \ = \ Z \qquad (d\tau = \text{volume element}) \ .$$

[‡] W. Pauli, *Zs. f. Phys.* **41**, 81 (1927).

This last equation, taking into account (5) can be written:

$$\frac{2^{13/2}\pi^2 m^{3/2} e^{5/2}}{3h^3}\int_0^\infty v^{3/2} r^2 dr \;=\; Ze\ . \tag{10}$$

So the potential v will be obtained searching for a function which satisfies Eq. (8) with the two conditions (9) and (10).

To simplify the search for v we change the variables r, v into two others x, ψ proportional to them, setting

$$r = \mu x \qquad , \qquad v = \gamma\psi \tag{11}$$

where we have

$$\mu = \frac{3^{2/3}\, h^2}{2^{13/3}\,\pi^{4/3}\, me^2\, Z^{1/3}} \qquad , \qquad \gamma = \frac{2^{13/3}\,\pi^{4/3}\, mZ^{4/3}\, e^3}{3^{2/3}\, h^2}\ . \tag{12}$$

Equations (8), (9) and (10) thus become

$$\begin{cases} \psi'' + \frac{2}{x}\,\psi' = \psi^{3/2} \\ \lim\limits_{x=0} x\psi = 1 \\ \int\limits_0^\infty \psi^{3/2} x^2 dx = 1\ . \end{cases} \tag{13}$$

These equations simplify further by setting

$$\varphi = x\psi\ . \tag{14}$$

Indeed they become

$$\begin{cases} \varphi'' = \varphi^{3/2}/\sqrt{x} \\ \varphi(0) = 1 \\ \int\limits_0^\infty \varphi^{3/2}\sqrt{x} dx = 1\ . \end{cases} \tag{15}$$

It is easy to see that the last condition is certainly satisfied if φ goes to zero for $x = \infty$. So it remains only to search for a solution to the first of (15), with the conditions at its limits $\varphi(0) = 1$, $\varphi(\infty) = 0$.

Since I did not succeed in finding the general integral of the first of (15), I have solved it numerically. The graph in Figure 1 represents $\varphi(x)$; for x close to zero we have

$$\varphi(x) = 1 - 1.58\, x + \frac{4}{3}\, x^{3/2} + ... \tag{16}$$

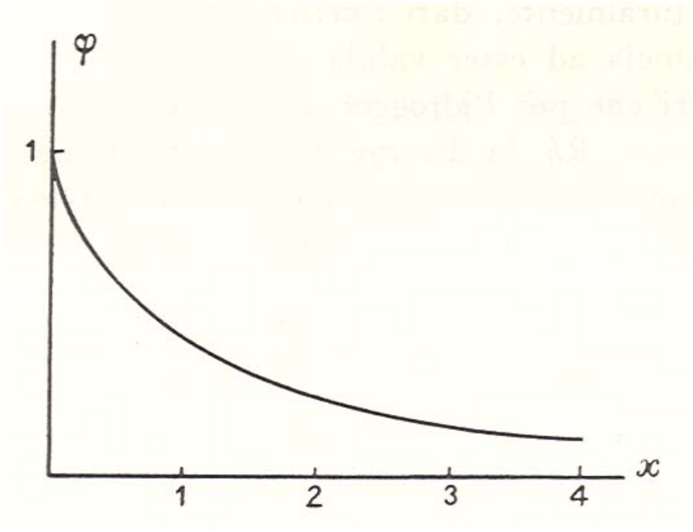


Fig. 1

Thus the problem of the determination of the electric potential of the atom at a fixed distance from the nucleus is solved. Its result is given by

$$v = \gamma \frac{\varphi(x)}{x} = \frac{\gamma\mu}{r}\varphi(x) = \frac{Ze}{r}\varphi\left(\frac{r}{\mu}\right) . \tag{17}$$

So we can therefore say that the potential at every point is equal to that produced by an effective charge

$$Ze\,\varphi\left(\frac{r}{\mu}\right) .$$

Now we move on to calculate the total energy of the atom; this should be calculated as the sum of the kinetic energy of all the electrons and the potential energy of the nucleus and electrons. However, it is easier taking into account the fact that in an atom the total energy is equal, except for the sign, to the kinetic energy (which anyway in our case can be verified with an easy calculation). Thus we have

$$W = -\int L\, n d\tau$$

and taking into account (5), (6), (11), (12), (14) we find

$$W = -\frac{3}{5}\int_0^\infty r^2\, nv\, dr = -\frac{2^{13/3}3^{1/3}\pi^{4/3}me^4Z^{7/3}}{5h^2}\int_0^\infty \frac{\varphi^{5/2}}{\sqrt{x}}\, dx .$$

The last integral can be evaluated taking into account that φ satisfies (15) and (16); one finds

$$\int_0^\infty \frac{\varphi^{5/2}}{\sqrt{x}}\, dx = -\frac{5}{7}\left(\frac{d\varphi}{dx}\right)_{x=0} = \frac{5}{7}\,1.58$$

and therefore we have

$$W = -1.58\,\frac{2^{13/3}3^{1/3}\pi^{4/3}me^4Z^{7/3}}{7h^2} = -1.58\,\frac{2^{1/3}3^{1/3}}{7\pi^{2/3}}Rh\,Z^{7/3}$$

that is

$$W = -1.54\,RhZ^{7/3} \tag{18}$$

where by R we indicate Rydberg's number, so that $-Rh$ is the energy of the fundamental state of hydrogen.

(18) gives us the necessary energy to tear off from an atom all its electrons. Naturally given the statistical criteria which it has been deduced from, it begins to be valid only for considerable values of Z; in fact we find that for hydrogen (18) gives $W = -1.54\,Rh$, while we actually have $W = -\,Rh$; the discrepancy is thus 54%. For helium the energy to produce complete ionization is obviously equal to the sum of the ionization energies of He and He^+; so we have

$$-W = (1.8 + 4)\,Rh = 5.8\,Rh$$

but from the theory we obtain $1.54 \cdot 2^{7/3} = 7.8\,Rh$; therefore the discrepancy in this case comes down to 35%. For the elements immediately following helium (Li, Be, B, C), nearly all of the atomic energy is due only to the two K electrons (for carbon about 86%) so the statistical method of course must still certainly give considerable discrepancies. For C in fact we still find a discrepancy close to 34%.

But we must expect that for elements of considerable atomic weight, the discrepancies between the statistical theory and empirical data are very much reduced; unfortunately the data is lacking for a precise comparison and we can base ourselves only on a rough valuation of the shield numbers for various orbits; such an evaluation, however, shows much better agreement.

80a) An attempt at a theory of β rays

"Tentativo di un a teoria dei raggi β,"
*Nuovo Cimento **11**, 1–19 (1934)*

ABSTRACT

A quantitative theory of the emission of β rays is proposed in which the existence of the "neutrino" is assumed and the emission of electrons and neutrinos in β decay is treated in a way similar to the one followed in the theory of radiation for describing the emission of a quantum of light from an excited atom. We deduce the formulas for the lifetime and for the shape of the continuous spectrum of β rays and compare them with experimental data.

The fundamental hypotheses of the theory

§ 1.– In the attempt to construct a theory of the nuclear electrons and the emission of β rays, one encounters, as is known, two principal difficulties. The first depends on the fact that the primary β rays are emitted from nuclei with a continuous velocity distribution. If we do not want to abandon the energy conservation principle, we are obliged to admit that a fraction of the energy which is released in the process of β decay escapes our present possibilities of observation. According to Pauli's proposal one can for instance assume the existence of a new particle, the so called "neutrino", having vanishing electric charge and mass on the order of magnitude of the electron mass or less. Thus we assume that in any β process are simultaneously emitted an electron, which is detected as a ray, and a neutrino which eludes the observation carrying a part of the energy away. In the present theory, we shall adopt the neutrino hypothesis.

A second difficulty for a theory of nuclear electrons depends on the fact that the present relativistic theories of the light particles (electrons or neutrinos) do not give a satisfactory explanation for the possibility that these particles are bound in orbits of nuclear size.

Consequently it seems more appropriate to agree with Heisenberg* and assume that all nuclei consist only of heavy particles, protons and neutrons. Then with the aim of understanding the possibility of emission of β rays, we will attempt to construct a theory of the emission of light particles from a nucleus in analogy with the theory of the emission of a quantum of light from an excited atom in the usual process of radiation. In the theory of radiation, the total number of the light quanta is not constant; the quanta are created when being emitted from an excited atom

*W. Heisenberg, *ZS. für Phys.* **77**, 1 (1932); E. Majorana, *ZS. für Phys.* **82**, 137 (1933).

and disappear when absorbed. In analogy with that we will try to establish the theory of β rays on these assumptions:

(a) The total number of electrons and neutrinos is not necessarily constant. Electrons (or neutrinos) can be created or destroyed. On the other hand this possibility has no analogy with the possibility of the creation or destruction of an electron-positron pair; in fact if we interpret a positron as a Dirac "hole", we can simply consider this latter process as a quantum jump of an electron from a state of negative energy to a state of positive energy, conserving the total number (infinitely large) of the electrons.

(b) The heavy particles, neutron and proton, can be considered, following Heisenberg, as two different internal states of the heavy particle. We shall formulate this fact by introducing an internal coordinate ρ of the heavy particle, which can assume only two values: $\rho = +\,1$, if the particle is a neutron; $\rho = -\,1$, if the particle is a proton.

(c) The Hamiltonian function of the overall system, consisting of heavy and light particles, must be chosen so that every transition from neutron to proton be accompanied by the creation of an electron and a neutrino; and the inverse process, transformation of a proton into a neutron, be accompanied by the disappearance of an electron and a neutrino. It must be remarked that in this way the conservation of the electric charge is assured.

The operators of the theory

§ 2. – A mathematical formalism which allows us to construct a theory in agreement with the three points of the preceding section can be easily constructed by using the method of Dirac-Jordan-Klein[†] called "the method of second quantization." Then we shall consider the probability amplitudes ψ and φ of the electrons and neutrinos in ordinary space, and their complex conjugates ψ^* and φ^* as operators; while for describing the heavy particles we shall use the usual representation in configuration space, in which obviously also ψ will be considered as a coordinate.

We introduce first two operators Q and Q^* which operate on the functions of the two-valued variable ρ as the linear substitutions

$$Q = \begin{vmatrix} 0 & 1 \\ 0 & 0 \end{vmatrix} \; ; \qquad Q^* = \begin{vmatrix} 0 & 0 \\ 1 & 0 \end{vmatrix} \; . \tag{1}$$

One immediately realizes that Q determines the transitions from proton to neutron, and Q^* the inverse transitions from neutron to proton.

[†] Cf. e.g. P. Jordan and O. Klein, *ZS. für Phys.* **45**, 751 (1927); W. Heisenberg, *Ann. d. Phys.* **10**, 888 (1931).

The meaning of the probability amplitudes ψ and φ interpreted as operators is, as we know, the following. Let

$$\psi_1\psi_2\ldots\psi_s\ldots$$

be a system of individual quantum states of the electrons. Then put

$$\psi=\sum_s\psi_s a_s\ ;\qquad \psi^*=\sum_s\psi_s^* a_s^*\ . \tag{2}$$

The amplitudes a_s and the conjugate complex quantities a_s^* are operators which act on the functions of the occupation numbers N_1, N_2,. . .,N_s,. . . of the individual quantum states. If the Pauli principle holds, each of the N_s can assume only one of the values 0, 1; and the operators a_s and a_s^* are defined as follows:

$$\begin{aligned} a_s\Psi\ \ &(N_1,N_2,\ldots,N_s,\ldots)\\ &=(-1)^{N_1+N_2+\ldots+N_{s-1}}\left(1-N_s\right)\Psi\left(N_1,N_2,\ldots,1-N_s,\ldots\right) \qquad (3)\\ a_s^*\Psi\ \ &(N_1,N_2,\ldots,N_s,\ldots)\\ &=(-1)^{N_1+N_2+\ldots+N_{s-1}}\left(1-N_s\right)\Psi\left(N_1,N_2,\ldots,N_s,\ldots\right)\ . \end{aligned}$$

The operator a_s^* determines the creation, while the operator a_s determines the disappearance of an electron in the quantum state s.

Corresponding to (2), for the neutrinos we shall set:

$$\varphi=\sum\varphi_\sigma b_\sigma\qquad ;\qquad \varphi^*=\sum\varphi_\sigma^* b_\sigma^*\ . \tag{4}$$

The conjugate complex operators b_σ and b_σ^* operate on the functions of the occupation numbers M_1,M_2,. . .,M_σ,. . . of the individual quantum states φ_1, φ_2,. . .,φ_σ,. . . of the neutrinos. If we assume that the Pauli principle also holds for these particles, the numbers M_σ can only assume the two values 0, 1; and one has

$$\begin{aligned} b_\sigma\ \ &\Phi\left(M_1,M_2,\ldots,M_\sigma,\ldots\right)\\ &=(-1)^{M_1+M_2+\ldots+M_{\sigma-1}}\left(1-M_\sigma\right)\Phi\left(M_1,M_2,\ldots,1-M_\sigma,\ldots\right) \qquad (5)\\ b_\sigma^*\ \ &\Phi\left(M_1,M_2,\ldots,M_\sigma,\ldots\right)\\ &=(-1)^{M_1+M_2+\ldots+M_{\sigma-1}}\left(1-M_\sigma\right)\Phi\left(M_1,M_2,\ldots,M_\sigma,\ldots\right)\ . \end{aligned}$$

The operators b_σ and b_σ^* determine the disappearance and the creation of a neutrino in the state σ, respectively.

The Hamiltonian function

§ 3.– The energy of the overall system constituted by the heavy and the light particles is the sum of the energy H_{hea} of the heavy particles + the energy H_{lig} of the light particles + the interaction energy $\mathcal{H}$ between the light and heavy particles.

Limiting ourselves for the sake of simplicity to consider only the heavy particle, we shall write the first term in the form

$$H_{\text{hea}}=\frac{1+\rho}{2}\mathcal{N}+\frac{1-\rho}{2}\mathcal{P} \tag{6}$$

in which $\mathcal{N}$ and $\mathcal{P}$ are the operators which represent the energy of the neutron and the proton. We notice in fact that, for $\rho = +\,1$ (neutron), (6) reduces to $\mathcal{N}$; while for $\rho = -\,1$ (proton) it reduces to $\mathcal{P}$.

To write the energy H_{lig} in the simplest way, we shall consider the quantum states ψ_s and φ_σ of the electrons and neutrinos to be stationary states. For the electrons we shall take the eigenfunctions in the Coulomb field of the nucleus (conveniently shielded in order to take into account the action of the atomic electrons); for the neutrinos we simply shall take De Broglie plane waves, since possible forces acting on neutrinos are certainly very weak. Let $H_1,H_2,\ldots,H_s,\ldots$ and $K_1,K_2,\ldots,K_\sigma,\ldots$ be the energies of the stationary states of the electrons and the neutrinos; then we shall have

$$H_{\text{lig}} = \sum_s H_s N_s + \sum_\sigma K_\sigma M_\sigma \ . \tag{7}$$

There still remains to write the interaction energy. It consists first of the Coulomb energy between proton and electrons; however, in the case of heavy nuclei the attraction exercised by only a proton has no importance[‡] and in any case does not contribute in any way to the process of β decay. In order not to uselessly complicate the problem, we shall neglect this term. We must instead add a term to the Hamiltonian such that it satisfies the condition c) of § 1.

A term which necessarily joins the transformation of a neutron into a proton with the creation of an electron and a neutrino has, according with the results of § 2, the form

$$Q^* a_s^* b_\sigma^* \tag{8}$$

while the conjugate complex operator

$$Q a_s b_\sigma \tag{8}$$

joins together the inverse processes (transformation of a proton into a neutron and disappearance of an electron and a neutrino).

An interaction term satisfying the condition c) will then have the following form

$$\mathcal{H} = Q \sum_{s\sigma} c_{s\sigma} a_s b_\sigma + Q^* \sum_{s\sigma} c_{s\sigma}^* a_s^* b_\sigma^* \ , \tag{9}$$

where $c_{s\sigma}$ and $c_{s\sigma}^*$ are quantities which may depend on the coordinates, the momenta, etc.. . . of the heavy particle.

A further determination of $\mathcal{H}$ must necessarily follow the principle of greatest simplicity; in any case the choices for $\mathcal{H}$ are restricted by the fact that $\mathcal{H}$ must be invariant with respect to a change of coordinates and moreover it must also satisfy momentum conservation.

If at first we neglect spin and relativistic effects, the simplest choice for (9) is the following

$$\mathcal{H} = g\,[Q\psi(x)\varphi(x) + Q^*\psi^*(x)\varphi^*(x)] \ , \tag{10}$$

[‡] The Coulomb attraction due to the many other protons must obviously be taken into account as a static field.

where g is a constant with dimensions L^5MT^{-2}; x represents the coordinates of the heavy particle; ψ, φ, ψ^*, φ^* are given by (2) and (4) and must be evaluated at the position x, y, z of the heavy particle.

Obviously (10) is not the only possible choice for $\mathcal{H}$; any scalar expression as

$$L(p)\psi(x)M(p)\varphi(x)N(p) + \text{compl. conj.}$$

where $L(p)$, $M(p)$, $N(p)$, represent convenient functions of the momentum of the heavy particle, would have been admissible. On the other hand, since until now the consequences of (10) have been in agreement with experience, there is no need to resort to more complicated expressions.

On the contrary, it is essential to generalize (10) in such a way to be able to treat relativistically at least the light particles. Of course, also in this generalization, it does not seem possible to eliminate all arbitrariness. However, the most natural solution of the problem appears to be the following: Relativistically we have, in place of ψ and φ, two sets $\psi_1\psi_2\psi_3\psi_4$ and $\varphi_1\varphi_2\varphi_3\varphi_4$ of four Dirac functions. Let us consider the 16 independent bilinear combinations of $\psi_1\psi_2\psi_3\psi_4$ and $\varphi_1\varphi_2\varphi_3\varphi_4$. When the frame of reference undergoes a Lorentz transformation, the 16 bilinear combinations undergo a linear substitution which gives a representation of the Lorenz group. In particular the four bilinear combinations

$$\begin{aligned} A_0 &= -\psi_1\varphi_2 + \psi_2\varphi_1 + \psi_3\varphi_4 - \psi_4\varphi_3 \\ A_1 &= \psi_1\varphi_3 - \psi_2\varphi_4 - \psi_3\varphi_1 + \psi_4\varphi_2 \\ A_2 &= i\psi_1\varphi_3 + i\psi_2\varphi_4 - i\psi_3\varphi_1 - i\psi_4\varphi_2 \\ A_3 &= -\psi_1\varphi_4 - \psi_2\varphi_3 + \psi_3\varphi_2 + \psi_4\varphi_1 \end{aligned} \tag{11}$$

transform like the components of a four-vector, that is like the components of the electromagnetic four-potential. Then it is natural to introduce in the Hamiltonian of the heavy particle the four quantities

$$g\,(QA_i + Q^*A_i^*)$$

in a situation corresponding to that of the components of the four-potential. Here we run into a problem depending on the fact that we do not know a relativistic wave equation for the heavy particles. However, in the case in which the velocity of the heavy particle is small compared to c, one can limit oneself to the term corresponding to eV (V the scalar potential) and write

$$\mathcal{H} = g\left[Q\left(-\psi_1\varphi_2 + \psi_2\varphi_1 + \psi_3\varphi_4 - \psi_4\varphi_3\right) + Q^*\left(\psi_1^*\varphi_2^* + \psi_2^*\varphi_1^* + \psi_3^*\varphi_4^* - \psi_4^*\varphi_3^*\right)\right]\ . \tag{12}$$

To this term one must add other ones of the order of magnitude v/c. At the moment, however, we shall neglect these terms, since the velocities of the neutrons and protons inside the nuclei are in general small compared to c (Cf. § 9).

In matrix language, (12) can be written

$$\mathcal{H} = g\left[Q\tilde{\psi}^*\delta\varphi + Q^*\tilde{\psi}\delta\varphi^*\right]\ , \tag{13}$$

where ψ and φ are meant as matrices with one column and the symbol $\sim$ transforms a matrix into its transposed conjugate; and moreover

$$\delta = \begin{vmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{vmatrix} . \tag{14}$$

With this notation, one finds by comparing (12) with (9)

$$c_{s\sigma} = g\tilde{\psi}_s^* \delta \varphi_\sigma \ ; \qquad c^*_{s\sigma} = g\tilde{\psi}_s \delta \varphi^*_\sigma \ , \tag{15}$$

where ψ and φ represent the four-component normalized eigenfunctions of the states s of the electron and σ of the neutrino, considered as functions of the position x, y, z occupied by the heavy particle.

The perturbation matrix

§ 4.– With the Hamiltonian we have established one can develop a theory of β decay in complete analogy with the theory of radiation. In that theory, as is known, the Hamiltonian consists of the sum: Energy of the atom + Energy of the radiation field + Interaction between atom and radiation; the latter term is considered as a perturbation of the other two. Analogously we shall take

$$H_{\text{hea}} + H_{\text{lig}} \tag{16}$$

as the unperturbed Hamiltonian. The perturbation is represented by the interaction term (13).

The quantum states of the unperturbed system can be enumerated in the following way:

$$(\rho, n, N_1, N_2 \ldots N_s \ldots M_1, M_2 \ldots M_\sigma \ldots) \ , \tag{17}$$

where the first number ρ takes one of the values ± 1 and indicates if the heavy particle is a neutron or a proton. The second number n indicates the quantum state of the neutron or the proton. For $\rho = +\ 1$ (neutron) let the corresponding eigenfunction be

$$u_n(x) \ , \tag{18}$$

while for $\rho = -\ 1$ (proton) let the eigenfunction be

$$v_n(x) \ . \tag{19}$$

The other numbers $N_1, N_2 \ldots N_s \ldots M_1, M_2 \ldots M_\sigma \ldots$ can only take the values 0, 1 and indicate what states of the electrons and neutrinos are occupied.

By an examination of the general form (9) of the perturbation energy, one immediately realizes that it has nonvanishing matrix elements only for transitions in which either the heavy particle passes from neutron to proton, while in the meantime one electron and one neutrino are created, or viceversa.

Through (1), (3), (5), (9), (18), (19) one easily finds that the corresponding matrix element is

$$\mathcal{H}\,{}^{1nN_1N_2\ldots0_sM_1M_2\ldots0_\sigma\ldots}_{-1mN_1N_2\ldots1_sM_1M_2\ldots1_\sigma\ldots} = \pm\int v_m^* c_{s\sigma}^* u_n d\tau \ , \tag{20}$$

where the integration must be extended over the entire configuration space of the heavy particle (with the exception of the coordinate ρ); the ± sign means more precisely

$$(-1)^{N_1+N_2+\ldots+N_{s-1}+M_1+M_2+\ldots M_{\sigma-1}}$$

and in any case does not enter into the calculations that will follow. To the inverse transition corresponds a matrix element which is the conjugate complex of (20).

Taking (15) into account, (20) becomes

$$\mathcal{H}\,{}^{1n0_s0_\sigma}_{-1m1_s1_\sigma} = \pm\int v_m^* u_n \tilde{\psi}_s \delta\varphi_\sigma^* d\tau \ , \tag{21}$$

where for the sake of brevity in the left hand side we have omitted writing all the indexes which do not change.

Theory of β decay

§ 5.– A β decay consists of a process in which a nuclear neutron transforms into a proton, while at the same time, in the way we have described, an electron, which is observed as a β particle, and a neutrino are emitted. To calculate the probability of this process, we shall assume that, at the time $t = 0$, a neutron is in a nuclear state of eigenfunction $u_n(x)$, and furthermore the electron state s and the neutrino state σ are free, that is $N_s = M_\sigma = 0$. Then for $t = 0$ we shall put the probability amplitude of the state $(1, n, 0_s, 0_\sigma)$ equal to 1, that is

$$a_{1,n,0_s,0_\sigma} = 1 \ , \tag{22}$$

whereas we shall put the probability amplitude of the state $(-1, m, 1_s, 1_\sigma)$, in which the neutron has been transformed into a proton with eigenfunction $v_m(x)$ emitting an electron and a neutrino in the states s and σ initially equal to zero.

By applying the usual formulas of perturbation theory, for a time short enough to still consider (22) approximately valid one finds

$$\dot{a}_{-1,m,1_s,1_\sigma} = -\frac{2\pi i}{h}\mathcal{H}\,{}^{1n0_s0_\sigma}_{-1m1_s1_\sigma}\, e^{\frac{2\pi i}{h}(-W+H_s+K_\sigma)t} \ , \tag{23}$$

where W stands for the difference in energy between the neutron state and the proton state.

By integrating (23) we obtain (since for $t = 0$, $a_{-1m1_s1_\sigma} = 0$)

$$a_{-1m1_s1_\sigma} = -\mathcal{H}^{\;\;1n0_s0_\sigma}_{-1m1_s1_\sigma} \frac{e^{\frac{2\pi i}{h}(-W+H_s+K_\sigma)t} - 1}{-W + H_s + K_\sigma} \; . \tag{24}$$

The probability of the transition we consider is then

$$|a_{-1m1_s1_\sigma}|^2 = 4 \left| \mathcal{H}^{\;\;1n0_s0_\sigma}_{-1m1_s1_\sigma} \right|^2 \frac{\sin^2 \frac{\pi t}{h} (-W + H_s + K_\sigma)}{(-W + H_s + K_\sigma)^2} \; . \tag{25}$$

To calculate the lifetime of the neutron state u_n it is necessary to sum (25) with respect to all unoccupied states of the electrons and neutrinos. A strong reduction of this sum can be obtained by observing that the De Broglie wave length for electrons or neutrinos having energies of some millions of volts is much larger than the nuclear sizes. Thus one can, as a first approximation, consider the eigenfunctions ψ_s and φ_σ to be constants inside the nucleus. Thus (21) becomes

$$\mathcal{H}^{\;\;1n0_s0_\sigma}_{-1m1_s1_\sigma} = \pm g \tilde{\psi}_s \delta \varphi^*_\sigma \int v^*_m u_n d\tau \; , \tag{26}$$

where here and below ψ_s and φ_σ are meant to be taken in the nucleus (Cf. § 8). From (26) we draw:

$$\left| \mathcal{H}^{\;\;1n0_s0_\sigma}_{-1m1_s1_\sigma} \right|^2 = g^2 \left| \int v^*_m u_n d\tau \right|^2 \tilde{\psi}_s \delta \varphi^*_\sigma \tilde{\varphi}^*_\sigma \tilde{\delta} \psi_\sigma \; . \tag{27}$$

States σ of the neutrino are characterized by their momentum p_σ and by the spin direction. If, for the convenience of normalization, we quantize inside a volume Ω, whose size later on will be made to tend to infinity, the normalized neutrino eigenfunctions are Dirac plane waves having density $1/\Omega$. Then simple algebraic considerations allow us to perform in (27) an average with respect to all the orientations of p_σ and of the spin. (And in this only the states of positive energy must be considered; the negative energy states must be eliminated through a device like the Dirac hole theory). One finds

$$\overline{\left| \mathcal{H}^{\;\;1n0_s0_\sigma}_{-1m1_s1_\sigma} \right|^2} = \frac{g^2}{4\Omega} \left| \int v^*_m u_n d\tau \right|^2 \left(\tilde{\psi}_s \psi_s - \frac{\mu c^2}{K_\sigma} \tilde{\psi}_s \beta \psi_s \right) \; , \tag{28}$$

where μ is the rest mass of the neutrino and β the Dirac matrix

$$\beta = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{vmatrix} \tag{29}$$

By observing that the number of positive energy neutrino states with momentum between p_σ and $p_\sigma + dp_\sigma$ is $8\pi\Omega p_\sigma^2 dp_\sigma / h^3$, that furthermore $\partial K_\sigma / \partial p_\sigma$ is the neutrino velocity for the state σ, and finally that (25) has a strong maximum for the value of p_σ for which there is no variation of the unperturbed energy, that is

$$-W + H_s + K_\sigma = 0 \; , \tag{30}$$

one can perform the sum of (25) with respect to σ in the usual way[§] and one finds

$$t\frac{8\pi^3 g^2}{h^4}\left|\int v_m^* u_n d\tau\right|^2 \frac{p_\sigma^2}{v_\sigma}\left(\tilde{\psi}_s\psi_s - \frac{\mu c^2}{K_\sigma}\tilde{\psi}_s\beta\psi_s\right) \ , \tag{31}$$

where p_σ is the value of the momentum of the neutrino for which (30) holds.

Determining elements of the transition probability

§ 6.– (31) expresses the probability that in a time t a β decay takes place in which the electron is emitted in the state s. As must be the case, this probability turns out to be proportional to the time (t has been considered small with respect to the lifetime); the coefficient of t gives the transition probability for the process we consider; it turns out to be

$$P_s = \frac{8\pi^3 g^2}{h^4}\left|\int v_m^* u_n d\tau\right|^2 \frac{p_\sigma^2}{v_\sigma}\left(\tilde{\psi}_s\psi_s - \frac{\mu c^2}{K_\sigma}\tilde{\psi}_s\beta\psi_s\right) \ . \tag{32}$$

Note that:

(a) For the free states of the neutrinos one always has $K_\sigma \geq \mu c^2$. Then it is necessary, in order that (30) can be satisfied, that

$$H_s \leq W - \mu c^2 \tag{33}$$

The upper limit of the β ray spectrum corresponds to the = sign.

(b) Secondly, since for the unoccupied electron state one has $H_s \geq mc^2$, we obtain, in order that the decay be possible, the following condition:

$$W \geq (m+\mu)c^2 \tag{34}$$

Then, in order that the β decay be possible, one must have a rather high occupied neutron state over a free proton state.

(c) According to (32), P_s depends on the eigenfunctions u_n and v_m of the heavy particle in the nucleus, through the matrix element

$$Q_{mn}^* = \int v_m^* u_n d\tau \tag{35}$$

This matrix element plays a role, in the in the theory of β rays, which is analogous to that of the matrix element of the electric moment in the theory of radiation. The matrix element (35) has normally the order of magnitude 1; nevertheless it often happens that, due to particular symmetries of the eigenfunctions u_n and v_m, Q_{mn}^* exactly vanishes. In that case we shall speak of "forbidden β transitions". On the other hand, one should not expect that the forbidden transitions are really impossible, since (32) is only an approximate formula. We shall come back to this matter in § 9.

[§] For a description of the methods used for performing such sums, cf. any expository article on the theory of radiation. For instance, E. Fermi *Rev. of Mod. Phys.* **4**, 87, (1932).

The mass of the neutrino

§ 7.– The transition probability (32) determines among other things the shape of the continuous spectrum of β rays. We will discuss here how the shape of this spectrum depends on the rest mass of the neutrino, in order to be able to determine this mass through a comparison with the experimental shape of the spectrum itself. The mass μ also enters into (32) through the factor p_σ^2/v_σ. The dependence of the shape of the curve of the energy distribution on μ is particularly pronounced in the proximity of the maximum energy E_0 of the β rays. It is easy to recognize that the distribution curve for energies E close to the maximum value E_0, behaves, apart from a factor independent of E, as

$$\frac{p_\sigma^2}{v_\sigma} = \frac{1}{c^3}\left(\mu c^2 + E_0 - E\right)\sqrt{\left(E_0 - E\right)^2 + 2\mu c^2\left(E_0 - E\right)}\ . \qquad (36)$$

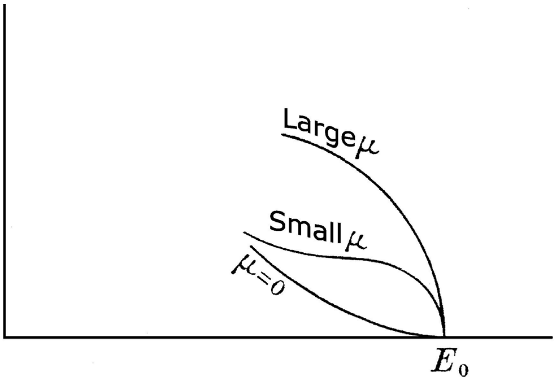


Fig. 1

In Figure 1 the end of the distribution curve is represented for $\mu = 0$, and for a small value and a large value of μ. The closest similarity of the theoretical curve to the experimental curves corresponds to $\mu = 0$. Thus we arrive at concluding that the mass of the neutrino is equal to zero or, in any case, much smaller than the mass of the electron[¶]. In the calculations below, for the sake of simplicity, we always set $\mu = 0$.

Then we have, also taking (32) into account

$$v_\sigma = c\ ; \qquad K_\sigma = cp_\sigma\ ; \qquad p_\sigma = \frac{K_\sigma}{c} = \frac{W - H_s}{c} \qquad (37)$$

and the inequalities (33) and (34) become

$$H_s \leq W\ ; \qquad W \geq mc^2\ . \qquad (38)$$

[¶] In a recent note F. Perrin, *C.R.*, **197**, 1625 (1933), by means of quantitative arguments arrives at a similar conclusion.

Finally the transition probability takes the form

$$P_s = \frac{8\pi^3 g^2}{c^3 h^4} \left| \int v_m^* u_n d\tau \right|^2 \tilde{\psi}_s \psi_s \left(W - H_s\right)^2 \ . \tag{39}$$

Lifetime and shape of the energy distribution curve for allowed transitions

§ 8.– From (39) one can derive a formula which expresses how many β transitions in which a β particle gets a momentum ranging from $mc\eta$ to $mc(\eta + d\eta)$ take place in unit time. For this it is necessary to calculate the sum of the values of $\tilde{\psi}_s \psi_s$ in the nucleus, extended to all the states (of the continuum) which belong to the indicated range of momentum. In this regard we point out that the relativistic eigenfunctions in the Coulomb field for the states with j=1/2 ($^2s_{1/2}$ e $^2p_{1/2}$) become infinite in the center. On the other hand the Coulomb law does not hold up to the center of the nucleus, but only up to a distance from it larger than R, where R is the nuclear radius. At this point, a tentative calculation shows that, if we make plausible assumptions on the behavior of the electric potential inside the nucleus, the value of $\tilde{\psi}_s \psi_s$ in the center of the nucleus turns out to be very close to the value which $\tilde{\psi}_s \psi_s$ should assume if the Coulomb law were valid at a distance R from the center. Applying the known formulas[‖] for the relativistic eigenfunctions of the continuum spectrum in a Coulomb field, after a rather long but easy calculation, one finds

$$\sum_{d\eta} \tilde{\psi}_s \psi_s = d\eta \cdot \frac{32\pi m^3 c^3}{h^3 \left[\Gamma\left(3+2S\right)\right]^2} \left(\frac{4\pi mcR}{h}\right)^{2S} \eta^{2+2S} e^{\pi\gamma \frac{\sqrt{1+\eta^2}}{\eta}} \times$$
$$\times \left| \Gamma\left(1 + S + i\gamma \frac{\sqrt{1+\eta^2}}{\eta}\right) \right|^2 \ , \tag{40}$$

where we have set

$$\gamma = Z/137 \ ; \qquad S = \sqrt{1-\gamma^2} - 1 \ . \tag{41}$$

The transition probability in an electric state in which the momentum has a value in the interval $mc\, d\eta$ then becomes (see (39))

$$P(\eta)d\eta = d\eta \cdot g^2 \frac{256\pi^4}{\left[\Gamma\left(3+2S\right)\right]^2} \frac{m^5 c^4}{h^7} \left(\frac{4\pi mcR}{h}\right)^{2S} \left| \int v_m^* u_n d\tau \right|^2 \eta^{2+2S} \times$$
$$\times e^{\pi\gamma \frac{\sqrt{1+\eta^2}}{\eta}} \left| \Gamma\left(1 + S + i\gamma \frac{\sqrt{1+\eta^2}}{\eta}\right) \right|^2 \left(\sqrt{1+\eta_0^2} - \sqrt{1+\eta^2}\right)^2 \ , \tag{42}$$

where η_0 is the maximum momentum of the emitted β rays, as measured in units of mc.

[‖] R.H. Hulme, *Proc. Roy. Soc.* **133**, 381 (1931).

For a numerical evaluation of (42) we refer to the particular value $\gamma = 0.6$, which corresponds to $Z = 82.2$ since the atomic numbers of the radioactive substances are not far from this value. For $\gamma = 0.6$, we have from (41) $S = -0.2$. Moreover one finds that, for $\eta < 10$ it is possible to set, with a sufficient approximation

$$\eta^{1.6} e^{0.6\pi \frac{\sqrt{1+\eta^2}}{\eta}} \left| \Gamma\left(0.8 + 0.6i\frac{\sqrt{1+\eta^2}}{\eta}\right)\right|^2 \cong 4.5\eta + 1.6\eta^2 \; . \tag{43}$$

With this, (42) becomes, setting $R = 9 \cdot 10^{-13}$ in it,

$$P(\eta)d\eta = 1.75 \cdot 10^{95} g^2 \left| \int v_m^* u_n d\tau \right|^2 \left(\eta + 0.355\eta^2\right) \left(\sqrt{1+\eta_0^2} - \sqrt{1+\eta^2}\right)^2 \; . \tag{44}$$

The inverse of the lifetime is obtained by integrating (44) from $\eta = 0$ to $\eta = \eta_0$; one finds

$$\frac{1}{\tau} = 1.75 \cdot 10^{95} g^2 \left| \int v_m^* u_n d\tau \right|^2 F(\eta_0) \; , \tag{45}$$

where we have set

$$\begin{aligned} F(\eta_0) &= \frac{2}{3}\sqrt{1+\eta_0^2} - \frac{2}{3} + \frac{\eta_0^4}{12} - \frac{\eta_0^2}{3} + \\ &+ 0.355\left[-\frac{\eta_0}{4} - \frac{\eta_0^3}{12} + \frac{\eta_0^5}{30} + \frac{\sqrt{1+\eta_0^2}}{4}\log\left(\eta_0 + \sqrt{1+\eta_0^2}\right)\right] \; . \end{aligned} \tag{46}$$

For small values of the argument, $F(\eta_0)$ behaves like $\eta_0^6/24$; for larger values of the argument, the values of F are gathered together in the following table.

Table 1

η_0	$F(\eta_0)$	η_0	$F(\eta_0)$	η_0	$F(\eta_0)$	η_0	$F(\eta_0)$
0	$\eta_0^6/24$	2	1.2	4	29	6	185
1	0.03	3	7.5	5	80	7	380

The forbidden transitions

§ 9.– Before moving on to a comparison of the theory with experience, we still want to illustrate some properties of the forbidden transitions.

As we have already said, a transition is forbidden when the corresponding matrix element (35) vanishes. If the representation of the nucleus by means of individual quantum states of the protons and neutrons turns out to be a good approximation, the matrix element Q_{mn}^* vanishes, due to symmetry, when

$$i = i' \tag{47}$$

does not hold, where i and i' are the angular momentum, in units $h/2\pi$, of the neutron state u_n and the proton state v_m, respectively. When the individual quantum states do not turn out to be a good approximation, to the selection rule (47) coresponds the other one

$$I = I' \ , \tag{48}$$

where I and I' represent the angular momentum of the nucleus before and after the β decay.

The selection rules (47) and (48) are much less rigorous than the selection rules of optics. It is possible to find exceptions to them, particularly with the two following processes:

(a) Formula (26) has been obtained by neglecting the variations of ψ_s and φ_s inside the region of the nucleus. If on the contrary these variations are taken into account, one has the possibility of obtaining β transitions even when Q^*_{mn} vanishes. It is easy to recognize that the intensity of these transitions has a ratio, as an order of magnitude, with the intensity of the allowed processes given by $(R/\lambda)^2$, where λ is the De Broglie wave length of the light particles. It must be noted that, if the electron and the neutrino have the same energy, when the former is near the nucleus it has a higher kinetic energy, due to the electrostatic attraction and so the most important effect comes from the variations of ψ_s. An evaluation of the order of magnitude of the intensity of these forbidden processes show that, at the same energy of the emitted electrons, they must have an intensity of one hundredth of the intensity of the normal processes. Besides the relatively small intensity, a characteristics of the forbidden transitions of this type can be found in the different shape of the curve of the energy distribution of β rays, which, for the forbidden transitions, must give a number of particles with small energy lower than in the normal case.

(b) A second possibility to have β transitions forbidden by the rule (48) depends on the fact, already pointed out at the end of § 3, that when the velocity of neutrons and protons is not negligible in comparison with the velocity of light we must add to the interaction term (12) other terms of order v/c. If e.g. one would assume a relativistic wave equation of the Dirac type also for the heavy particles, one could add to (12) terms like

$$gQ\,(\alpha_x A_1 + \alpha_y A_2 + \alpha_z A_3) \ + \ \text{complex conjugate} \ , \tag{49}$$

where $\alpha_x \alpha_y \alpha_z$ are the usual Dirac matrices for the heavy particle and $A_1 A_2 A_3$ the spatial components of the four vector defined by (12). A term of the type (49) allows also β transitions which do not satisfy the selection rule (48), and their intensity is, with respect to that of normal processes, of the order of magnitude $(v/c)^2$, that is about 1/100. Thus we find a second possibility for forbidden transitions nearly 100 times less intense than the normal ones.

Comparison with experience

§ 10.– (45) establishes a relation between the maximum momentum η_0 of the β rays emitted by a substance and its lifetime. In this relation, really, also an unknown element enters, the integral

$$\int v_m^* u_n d\tau \tag{50}$$

whose evaluation requires knowledge of the nuclear eigenfunctions u_n and v_m of the neutron and the proton. However, in the case of the allowed transitions, (50) is of the order of magnitude of unity. Then we expect the product

$$\tau F(\eta_0) \tag{51}$$

to have the same order of magnitude in all the allowed transitions. Instead, for the forbidden transitions, the lifetime will be, as an order of magnitude, one hundred times larger, and correspondingly also the product (51) will be larger. In the following table we collect the products $\tau F(\eta_0)$ for all the substances which disintegrate by emitting β rays and for which we have sufficiently exact data.

Table 2

Element	τ(hours)	η_0	$F(\eta_0)$	$\tau F(\eta_0)$
UX_2	0.026	5.4	115	3.0
RaB	0.64	2.04	1.34	0.9
ThB	15.3	1.37	0.176	2.7
ThC''	0.076	4.4	44	3.3
AcC''	0.115	3.6	17.6	2.0
RaC	0.47	7.07	398	190
RaE	173	3.23	10.5	1800
ThC	2.4	5.2	95	230
$MsTh_2$	8.8	6.13	73	640

In this table the two groups we have expected are certainly recognizable; moreover such a division of the elements which emit primary β rays into two groups had been already observed experimentally by Sargent.** The values of η_0 have been taken from the quoted paper of Sargent (for a comparison, note that: $\eta_0 = (H\rho)_{max/1700}$). Besides the data in this table, Sargent gives the data for three other elements, warning that they are not as reliable as the other ones. They are UX_1 for which $\tau = 830$; $\eta_0 = 0.76$; $F(\eta_0) = 0.0065$; $\tau F(\eta_0) = 5.4$; then this element appears to be attributable to the first group. For AcB one has: $\tau = 0.87$; $\eta_0 = 1.24$; $F(\eta_0) = 0.102$; $\tau F(\eta_0) = 0.09$; then one finds a value of $\tau F(\eta_0)$ about ten

B.W. Sargent, *Proc. Roy. Soc.* **139, 659, (1933).

times smaller than those of the first group. Finally for RaD one has: $\tau = 320000$; $\eta_0 = 0.38$ (largely uncertain); $F(\eta_0) = 0.00011$; $\tau F(\eta_0) = 35$. Then this element can be put roughly half-way between the two groups. I have not succeeded in finding data for the other elements which emit primary β rays, that is Ms, Th_1, UY, Ac, AcC, UZ, RaC''.

On the whole one can conclude from this comparison between theory and experience that the agreement is certainly as good as one would have expected. The discrepancies observed for the elements with uncertain experimental data, RaD and AcB, can be explained well partly by the lack of precision of the measures, partly also by oscillations, quite plausible, in the value of the matrix element (50). Moreover one must notice that the fact that the majority of β decays are accompanied by emission of γ rays indicates that the larger part of the β processes can leave the proton in different excitation states and this gives a further mechanism which can determine oscillations in the value of $\tau F(\eta_0)$.

From the data of Table 2 one can infer an evaluation, even if rough, of constant g. If we admit, for instance, that when the matrix element (50) has the value 1, one has $\tau F(\eta_0) = 1\,hour = 3600\,s$; one finds from (45)

$$g = 4 \cdot 10^{-50} cm^3 \cdot erg$$

which gives nothing more than the order of magnitude.

Let us move on to discuss the shape of the curve of the velocity distribution of β rays. In the case of allowed processes, the distribution curve, as a function of η (that is, apart from a factor 1700, of $H\rho$) is represented in Fig. 2, for values of the maximum momentum η_0.

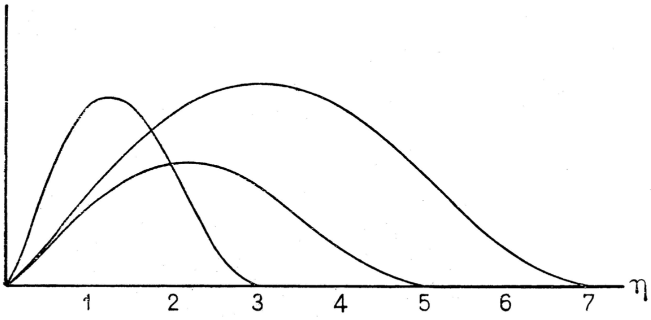


Fig. 2

The curves are satisfactorily similar to experimental ones collected by Sargent.[††] Only in the range of small energy Sargent's curves are a little lower than the theoretical ones, and this is more easily evident in the curves of Fig. 3 where the abscissas are the energies instead of the momenta. But we must remark that the part of

[††] B.W. Sargent, *Proc. Camb. Phil. Soc.* **28**, 538 (1932).

the curves of small energy is not perfectly known experimentally.[‡‡] Moreover, for the forbidden transitions, also theoretically, in the range of small energies the curve must be lower than the curves of the allowed transitions, represented in Figs. 2 and 3.

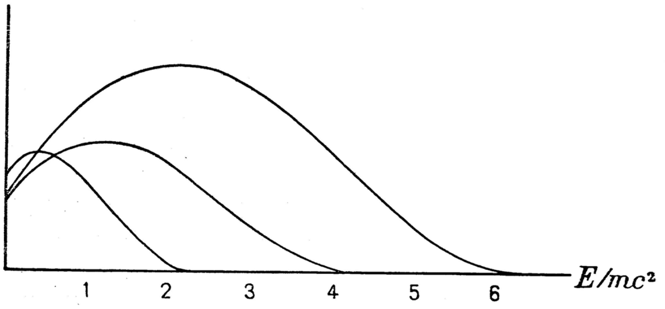


Fig. 3

Of this fact one must particularly take into account for the case of RaE, which is the best known from an experimental point of view. The emission of β rays from this element, as results from the abnormally large value of $\tau F(\eta_0)$ (Cf. Table 2), is certainly forbidden, or better it is possible that it is allowed only in the second approximation. I hope, in a future article, to be able to better specify the behavior of distribution curves for the forbidden transitions.

To summarize, it seems justified to assert that the theory in the form described here does agree with the experimental data, which in any case are not always sufficiently accurate. On the other hand, even if in a further comparison of the theory with experience, one should arrive at some discrepancy, it would be always be possible to modify the theory without changing its conceptual foundations in an essential way. It would be possible precisely to keep equation (9) but choose the $c_{s\sigma}$ in a different way. This will carry us, in particular, to a different form of the selection rule (48) and to a different form of the curve of the energy distribution.

Only a further development of the theory, as also an increase in the precision of the experimental data, will be able to indicate if such a change will be necessary.

[‡‡] Cf. e.g., Rutherford, Ellis and Chadwick, *Radiation from Radio-active Substances*, Cambridge, 1930. See, in particular p. 407.

FI 14 - E. Fermi: Artificial Radioactivity Produced by Neutron Bombardment (Nobel lecture: December 12, 1938)

ENRICO FERMI

Artifical radioactivity produced by neutron bombardment

Nobel Lecture, December 12, 1938

Although the problem of transmuting chemical elements into each other is much older than a satisfactory definition of the very concept of chemical element, it is well known that the first and most important step towards its solution was made only nineteen years ago by the late Lord Rutherford, who started the method of the nuclear bombardments. He showed on a few examples that, when the nucleus of a light element is struck by a fast α-particle, some disintegration process of the struck nucleus occurs, as a consequence of which the α-particle remains captured inside the nucleus and a different particle, in many cases a proton, is emitted in its place. What remains at the end of the process is a nucleus different from the original one; different in general both in electric charge and in atomic weight.

The nucleus that remains as disintegration product coincides sometimes with one of the stable nuclei, known from the isotopic analysis; very often, however, this is not the case. The product nucleus is then different from all "natural" nuclei; the reason being that the product nucleus is not stable. It disintegrates further, with a mean life characteristic of the nucleus, by emission of an electric charge (positive or negative), until it finally reaches a stable form. The emission of electrons that follows with a lag in time the first practically instantaneous disintegration, is the so-called artificial radioactivity, and was discovered by Joliot and Irene Curie at the end of the year 1933.

These authors obtained the first cases of artificial radioactivity by bombarding boron, magnesium, and aluminium with α-particles from a polonium source. They produced thus three radioactive isotopes of nitrogen, silicon and phosphorus, and succeeded also in separating chemically the activity from the bulk of the unmodified atoms of the bombarded substance.

The neutron bombardment

Immediately after these discoveries, it appeared that α-particles very likely did not represent the only type of bombarding projectiles for producing artificial radioactivity. I decided therefore to investigate from this point of view the effects of the bombardment with neutrons.

Compared with α-particles, the neutrons have the obvious drawback that the available neutron sources emit only a comparatively small number of neutrons. Indeed neutrons are emitted as products ofnuclear reactions, whose yield is only seldom larger than 10^{-4}. This drawback is, however, compensated by the fact that neutrons, having no electric charge, can reach the nuclei of all atoms, without having to overcome the potential barrier, due to the Coulomb field that surrounds the nucleus. Furthermore, since neutrons practically do not interact with electrons, their range is very long, and the probability of a nuclear collision is correspondingly larger than in the case of the α-particle or the proton bombardment. As a matter of fact, neutrons were already known to be an efficient agent for producing some nuclear disintegrations.

As source of neutrons in these researches I used a small glass bulb containing beryllium powder and radon. With amounts of radon up to 800 millicuries such a source emits about 2 x 10^{7} neutrons per second. This number is of course very small compared to the yield of neutrons that can be obtained from cyclotrons or from high-voltage tubes. The small dimensions, the perfect steadiness and the utmost simplicity are, however, sometimes very useful features of the radon + beryllium sources.

Nuclear reactions produced by neutrons

Since the first experiments, I could prove that the majority of the elements tested became active under the effect of the neutron bombardment. In some cases the decay of the activity with time corresponded to a single mean life; in others to the superposition of more than one exponential decay curve.

A systematic investigation of the behaviour of the elements throughout the Periodic Table was carried out by myself, with the help of several collaborators, namely Amaldi, d’Agostino, Pontecorvo, Rasetti, and Segré. In most cases we performed also a chemical analysis, in order to identify the chemical element that was the carrier of the activity. For short living substances, such an analysis must be performed very quickly, in a time of the order of one minute.

The results of this first survey of the radioactivities produced by neutrons can be summarized as follows: Out of 63 elements investigated, 37 showed an easily detectable activity; the percentage of the activatable elements did not show any marked dependence on the atomic weight of the element. Chemical analysis and other considerations, mainly based on the distribution of the isotopes, permitted further to identify the following three types of nuclear reactions giving rise to artificial radioactivity :

$$ {}^{M}_{Z}A + {}^{1}_{0}n = {}^{M-3}_{Z-2}A + {}^{4}_{2}He \qquad (1)$$

$$ {}^{M}_{Z}A + {}^{1}_{0}n = {}^{M}_{Z-1}A + {}^{1}_{1}H \qquad (2)$$

$$ {}^{M}_{Z}A + {}^{1}_{0}n = {}^{M+1}_{Z}A \qquad (3)$$

where ${}^{M}_{Z}A$ is the symbol for an element with atomic number Z and mass number M; n is the symbol of the neutron.

The reactions of the types (1) and (2) occur chiefly among the light elements, while those of the type (3) are found very often also for heavy elements. In many cases the three processes are found at the same time in a single element. For instance, neutron bombardment of aluminium that has a single isotope ^{27}Al, gives rise to three radioactive products: ^{24}Na, with a half-period of 15 hours by process (1); ^{27}Mg, with a period of IO minutes by process (2); and ^{28}A1 with a period of 2 to 3 minutes by process (3).

As mentioned before, the heavy elements usually react only according to process (3) and therefore, but for certain complications to be discussed later, and for the case in which the original element has more than one stable isotope, they give rise to an exponentially decaying activity. A very striking exception to this behaviour is found for the activities induced by neutrons in the naturally active elements thorium and uranium. For the investigation of these elements it is necessary to purify first the element as thoroughly as possible from the daughter substances that emit β-particles. When thus purified, both thorium and uranium emit spontaneously only α-particles, that can be immediately distinguished, by absorption, from the β-activity induced by the neutrons.

Both elements show a rather strong, induced activity when bombarded with neutrons; and in both cases the decay curve of the induced activity shows that several active bodies with different mean lives are produced. We attempted, since the spring of 1934, to isolate chemically the carriers of these activities, with the result that the carriers of some of the activities of uranium are neither isotopes of uranium itself, nor of the elements lighter than uranium down to the atomic number 86. We concluded that the carriers were one or more elements of atomic number larger than 92 ; we, in Rome, use to call the elements 93 and 94 Ausenium and Hesperium respectively. It is known that O. Hahn and L. Meitner have investigated very carefully and extensively the decay products of irradiated uranium, and were able to trace among them elements up to the atomic number 96.*

It should be noticed here, that besides processes (1), (2), and (3) for the production of artificial radioactivity with neutrons, neutrons of sufficiently high energy can react also as follows, as was first shown by Heyn: The primary neutron does not remain bound in the nucleus, but knocks off instead, one of the nuclear neutrons out of the nucleus; the result is a new nucleus, that is isotopic with the original one and has an atomic weight less by one unit. The final result is therefore identical with the products obtained by means of the nuclear photoeffect (Bothe), or by bombardment with fast deuterons. One of the most important results of the comparison of the active products obtained by these processes, is the proof, first given by Bothe, of the existence of isomeric nuclei, analogous to the isomers UX_2 and UZ, recognized long since by O. Hahn in his researches on the uranium family. The number of well-established cases of isomerism appears to increase rather rapidly, as investigation goes on, and represents an attractive field of research.

The slow neutrons

The intensity of the activation as a function of the distance from the neutron source shows in some cases anomalies apparently dependent on the objects that surround the source. A careful investigation of these effects led to the unexpected result that surrounding both source and body to be activated with masses of paraffin, increases in some cases the intensity of activation by a very large factor (up to 100). A similar effect is produced by water, and in general by substances containing a large concentration of hydrogen. Substances not containing hydrogen show sometimes similar features, though extremely less pronounced.

The interpretation of these results was the following. The neutron and the

* The discovery by Hahn and Strassmann of barium among the disintegration products of bombarded uranium, as a consequence of a process in which uranium splits into two approximately equal parts, makes it necessary to reexamine all the problems of the transuranic elements, as many of them might be found to be products of a splitting of uranium.

proton having approximately the same mass, any elastic impact of a fast neutron against a proton initially at rest, gives rise to a distribution of the available kinetic energy between neutron and proton; it can be shown that a neutron having an initial energy of 10^6 volts, after about 20 impacts against hydrogen atoms has its energy already reduced to a value close to that corresponding to thermal agitation. It follows that, when neutrons of high energy are shot by a source inside a large mass of paraffin or water, they very rapidly lose most of their energy and are transformed into "slow neutrons". Both theory and experiment show that certain types of neutron reactions, and especially those of type (3), occur with a much larger cross-section for slow neutrons than for fast neutrons, thus accounting for the larger intensities of activation observed when irradiation is performed inside a large mass of paraffin or water.

It should be remarked furthermore that the mean free path for the elastic collisions of neutrons against hydrogen atoms in paraffin, decreases rather pronouncedly with the energy. When therefore, after three or four impacts, the energy of the neutron is already considerably reduced, its probability of diffusing outside of the paraffin, before the process of slowing down is completed, becomes very small.

To the large cross-section for the capture of slow neutrons by several atoms, there must obviously correspond a very strong absorption of these atoms for the slow neutrons. We investigated systematically such absorptions, and found that the behaviour of different elements in this respect is widely different; the cross-section for the capture of slow neutrons varies, with no apparent regularity for different elements, from about 10^{-24} cm^2 or less, to about a thousand times as much. Before discussing this point, as well as the dependence of the capture cross-section on the energy of the neutrons we shall first consider how far down the energy of the primary neutrons can be reduced by the collisions against the protons.

The thermal neutrons

If the neutrons could go on indefinitely diffusing inside the paraffin, their energy would evidently reach finally a mean value equal to that of thermal agitation. It is possible, however, that, before the neutrons have reached this lowest limit of energy, either they escape by diffusion out of the paraffin, or are captured by some nucleus. If the neutron energy reaches the thermal value, one should expect the intensity of the activation by slow neutrons to depend upon the temperature of the paraffin.

Soon after the discovery of the slow neutrons, we attempted to find a temperature dependence of the activation, but, owing to insufficient accuracy, we did not succeed. That the activation intensities depend upon the temperature was proved some months later by Moon and Tillman in London; as they showed, there is a considerable increase in the activation of several detectors, when the paraffin, in which the neutrons are slowed down, is cooled from room temperature to liquid-air temperature. This experiment definitely proves that a considerable percentage of the neutrons actually reaches the energy of thermal agitation. Another consequence is that the diffusion process must go on inside the paraffin for a relatively long time.

In order to measure, directly at least, the order of magnitude of this time, an experiment was attempted by myself and my collaborators. The source of neutrons was fastened at the edge of a rotating wheel, and two identical detectors were placed on the same edge, at equal distances from the source, one in front and one behind with respect to the sense of rotation. The wheel was then spun at a very high speed inside a fissure in a large paraffin block. We found that, while, with the wheel at rest, the two detectors became equally active, when the wheel was in motion during the activation, the detector that was behind the source became considerably more active than the one in front. From a discussion of this experiment was deduced, that the neutrons remain inside the paraffin for a time of the order of 10^{-4} seconds.

Other mechanical experiments with different arrangements were performed in several laboratories. For instance Dunning, Fink, Mitchell, Pegram, and Segré: in New York, built a mechanical velocity selector, and proved by direct measurement, that a large amount of the neutrons diffusing outside of a block of paraffin, have actually a velocity corresponding to thermal agitation.

After their energy is reduced to a value corresponding to thermal agitation, the neutrons go on diffusing without further change of their average energy. The investigation of this diffusion process, by Amaldi and myself, showed that thermal neutrons in paraffin or water can diffuse for a number of paths of the order of 100 before being captured. Since, however, the mean free path of the thermal neutrons in paraffin is very short (about 0.3 cm) the total displacement of the thermal neutrons during this diffusion process is rather small (of the order of 2 or 3 cm). The diffusion ends when the thermal neutron is captured, generally by one of the protons, with production of a deuteron. The order of magnitude for this capture probability can be calculated, in good agreement with the experimental value, on the assumption that the transition from a free-neutron state to the state in which the neutron is bound in the deuteron is due to the magnetic dipole moments of the proton and the neutron. The binding energy set free in this process, is emitted in the form of γ-rays, as first observed by Lea.

All the processes of capture of slow neutrons by any nucleus are generally accompanied by the emission of γ-rays : Immediately after the capture of the neutron, the nucleus remains in a state of high excitation and emits one or more γ-quanta, before reaching the ground state. The γ-rays emitted by this process were investigated by Rasetti and by Fleischmann.

Absorption anomalies

A theoretical discussion of the probability of capture of a neutron by a nucleus, under the assumption that the energy of the neutron is small compared with the differences between neighbouring energy levels in the nucleus, leads to the result that the cross-section for the capture process should be inversely proportional to the velocity of the neutron. While this result is in qualitative agreement with the high efficiency of the slow-neutron bombardment observed experimentally, it fails on the other hand to account for several features of the absorption process, that we are now going to discuss.

If the capture probability of a neutron were inversely proportional to its velocity, one would expect two different elements to behave in exactly the same way as absorbers of the slow neutrons, provided the thicknesses of the two absorbers were conveniently chosen, so as to have equal absorption for neutrons of a given energy. That the absorption obeys instead more complicated laws, was soon observed by Moon and Tillman and other authors who showed that the absorption by a given element appears, as a rule, to be larger when the slow neutrons are detected by means of the activity induced in the same element. That the simple law of inverse proportionality does not hold, was also proved by a direct mechanical experiment by Dunning, Pegram, Rasetti, and others in New York.

In the winter of 1935-1936 a systematic investigation of these phenomena was carried out by Amaldi and myself The result was, that each absorber of the slow neutrons has one or more characteristic absorption bands, usually for energies below 100 volts. Besides this or these absorption bands, the absorption coefficient is always large also for neutrons of thermal energy. Some elements, especially cadmium, have their characteristic absorption band overlapping with the absorption in the thermal region. This element absorbs therefore very strongly the thermal neutrons, while it is almost transparent to neutrons of higher energies. A thin cadmium sheet is therefore used for filtering the thermal neutrons out of the complex radiation that comes out of a paraffin block containing a neutron source inside.

Bohr and Breit and Wigner proposed independently to explain the above anomalies, as due to resonance with a virtual energy level of the compound nucleus (i.e. the nucleus composed of the bombarded nucleus and the neutron). Bohr went much farther in giving also a qualitative explanation of the large probability for the existence of at least one such level, within an energy interval of the order of magnitide of 100 volts corresponding to the energy band of the slow neutrons. This band corresponds, however, to an excitation energy of the compound nucleus of many million volts, representing the binding energy of the neutron. Bohr could show that, since nuclei, and especially heavy nuclei, are systems with a very large number of degrees of freedom, the spacing between neighbouring energy levels decreases very rapidly with increasing excitation energy. An evaluation of this spacing shows that whereas for low excitation energies the spacing is of the order of magnitude of 10^5 volts, for high excitation energies, of the order of ten million volts, it is reduced (for elements of mean atomic weight) to less than one volt. It is therefore a very plausible assumption that one (or more) such level lies within the slow-neutron band, thus explaining the large frequency of the cases in which absorption anomalies are observed.

Before concluding this review of the work on artificial radioactivity produced by neutrons, I feel it as a duty to thank all those who have contributed to the success of these researches. I must thank in particular all my collaborators that have already been mentioned; the Istituto di Sanità Pubblica in Rome and especially Prof. G. C. Trabacchi, for the supply of all the many radon sources that have been used; the Consiglio Nazionale delle Richerche for several grants.

Chapter 3

From Fermi's papers of the American period

On the origin of cosmic radiation (237)

Phys. Rev. ***75****, 1169–1174 (1949).*

PHYSICAL REVIEW VOLUME 75, NUMBER 8 APRIL 15, 1949

On the Origin of the Cosmic Radiation

ENRICO FERMI
Institute for Nuclear Studies, University of Chicago, Chicago, Illinois

(Received January 3, 1949)

A theory of the origin of cosmic radiation is proposed according to which cosmic rays are originated and accelerated primarily in the interstellar space of the galaxy by collisions against moving magmetic fields. One of the features of the theory is that it yields naturally an inverse power law for the spectral distribution of the cosmic rays. The chief difficulty is that it fails to explain in a straightforward way the heavy nuclei observed in the primary radiation.

I. INTRODUCTION

In recent discussions on the origin of the cosmic radiation E. Teller[1] has advocated the view that cosmic rays are of solar origin and are kept relatively near the sun by the action of magnetic fields. These views are amplified by Alfvén, Richtmyer, and Teller.[2] The argument against the conventional view that cosmic radiation may extend at least to all the galactic space is the very large amount of energy that should be present in form of cosmic radiation if it were to extend to such a huge space. Indeed, if this were the case, the mechanism of acceleration of the cosmic radiation should be extremely efficient.

I propose in the present note to discuss a hypothesis on the origin of cosmic rays which attempts to meet in part this objection, and according to which cosmic rays originate and are accelerated primarily in the interstellar space, although they are assumed to be prevented by magnetic fields from leaving the boundaries of the galaxy. The main process of acceleration is due to the interaction of cosmic particles with wandering magnetic fields which, according to Alfvén, occupy the interstellar spaces.

Such fields have a remarkably great stability because of their large dimensions (of the order of magnitude of light years), and of the relatively high electrical conductivity of the interstellar space. Indeed, the conductivity is so high that one might describe the magnetic lines of force as attached to the matter and partaking in its streaming motions. On the other hand, the magnetic field itself reacts on the hydrodynamics[3] of the interstellar matter giving it properties which, according to Alfvén, can pictorially be described by saying that to each line of force one should attach a material density due to the mass of the matter to which the line of force is linked. Developing this point of view, Alfvén is able to calculate a simple formula for the velocity V of propagation of magneto-elastic waves:

$$V=H/(4\pi\rho)^{\frac{1}{2}}, \tag{1}$$

where H is the intensity of the magnetic field and ρ is the density of the interstellar matter.

One finds according to the present theory that a particle that is projected into the interstellar medium with energy above a certain injection threshold gains energy by collisions against the moving irregularities of the interstellar magnetic field. The rate of gain is very slow but appears capable of building up the energy to the maximum values observed. Indeed one finds quite naturally an inverse power law for the energy spectrum of the protons. The experimentally observed exponent of this law appears to be well within the range of the possibilities.

The present theory is incomplete because no satisfactory injection mechanism is proposed except for protons which apparently can be regenerated at least in part in the collision processes of the cosmic radiation itself with the diffuse interstellar matter. The most serious difficulty is in the injection process for the heavy nuclear component of the radiation. For these particles the injection energy is very high and the injection mechanism must be correspondingly efficient.

II. THE MOTIONS OF THE INTERSTELLAR MEDIUM

It is currently assumed that the interstellar space of the galaxy is occupied by matter at extremely low density, corresponding to about one atom of hydrogen per cc, or to a density of about 10^{-24} g/cc. The evidence indicates, however, that this matter is not uniformly spread, but that there are condensations where the density may be as much as ten or a hundred times as large and which extend to average dimensions of the order of 10 parsec. (1 parsec. $=3.1\times10^{18}$ cm $=3.3$ light years.) From the measurements of Adams[4] on the Doppler effect of the interstellar absorption lines one knows the radial velocity with respect to the sun of a sample of such clouds located at not too great distance from us. The root mean square of the radial velocity, corrected for the proper motion of the sun with respect to the neighboring stars, is about 15 km/sec. We may assume that the root-mean-square velocity

[1] Nuclear Physics Conference, Birmingham, 1948.

[2] Alfvén, Richtmyer, and Teller, Phys. Rev., to be published.

[3] H. Alfvén, Arkiv Mat. f. Astr., o. Fys. **29B**, 2 (1943).

[4] W. S. Adams, A.p.J. **97**, 105 (1943).

is obtained by multiplying this figure by the square root of 3, and is therefore about 26 km/sec. Such relatively dense clouds occupy approximately 5 percent of the interstellar space.[5]

Much less is known of the much more dilute matter between such clouds. For the sake of definiteness in what follows, the assumption will be made that this matter has a density of the order of 10^{-25}, or about 0.1 hydrogen atoms per cc. Even fairly extensive variations on this figure would not very drastically alter the qualitative conclusions. If the assumption is made that most of this material consists of hydrogen atoms, it is to be expected that most of the hydrogen will be ionized by the photoelectric effect of the stellar light. Indeed, one can estimate that some kind of dissociation equilibrium is established under average interstellar conditions, outside the relatively dense clouds, for which

$$n_+{}^2/n_0 \approx (T_1)^{\frac{1}{2}}, \tag{2}$$

where n_+ and n_0 are the concentrations of ions and neutral atoms per cc, and T_1 is the absolute kinetic temperature in degrees K. Putting in this formula $n_+ = 0.1$, one finds that the fraction of undissociated atoms is of the order of 1 percent, even assuming a rather low kinetic temperature of the order of 100°K.

It is reasonable to assume that this very low density medium will have considerable streaming motions, since it will be kept stirred by the moving heavier clouds passing through it. In what follows, a root-mean-square velocity of the order of 30 km/sec. will be assumed. According to Alfvén's picture, we must assume that the kinetic energy of these streams will be partially converted into magnetic energy, that indeed, the magnetic field will build up to such a strength that the velocity of propagation of the magneto-elastic waves becomes of the same order of magnitude as the velocity of the streaming motions. From (1) it follows then that the magnetic field in the dilute matter is of the order of magnitude of 5×10^{-6} gauss, while its intensity is probably greater in the heavier clouds. The lines of force of this field will form a very crooked pattern, since they will be dragged in all directions by the streaming motions of the matter to which they are attached. They will, on the other hand, tend to oppose motions where two portions of the interstellar matter try to flow into each other, because this would lead to a strengthening of the magnetic field and a considerable increase of magnetic energy. Indeed, this magnetic effect will have the result to minimize what otherwise would be extremely large friction losses which would damp the streaming motions and reduce them to disordered thermal motions in a relatively short time.

III. ACCELERATION OF THE COSMIC RAYS

We now consider a fast particle moving among such wandering magnetic fields. If the particle is a proton having a few Bev energy, it will spiral around the lines of force with a radius of the order of 10^{12} cm until it "collides" against an irregularity in the cosmic field and so is reflected, undergoing some kind of irregular motion. On a collision both a gain or a loss of energy may take place. Gain of energy, however, will be more probable than loss. This can be understood most easily by observing that ultimately statistical equilibrium should be established between the degrees of freedom of the wandering fields and the degrees of freedom of the particle. Equipartition evidently corresponds to an unbelievably high energy. The essential limitation, therefore, is not the ceiling of energy that can be attained, but rather the rate at which energy is acquired. A detailed discussion of this process of acceleration will be given in Section VI. An elementary estimate can be obtained by picturing the "collisions" of the particles against the magnetic irregularities as if they were collisions against reflecting obstacles of very large mass, moving with disordered velocities averaging to $V = 30$ km/sec. Assuming this picture, one finds easily that the average gain in energy per collision is given as order of magnitude by

$$\delta w = B^2 w, \tag{3}$$

where w represents the energy of the particle inclusive of rest energy, and $B = V/c \approx 10^{-4}$. This corresponds, therefore, for a proton to an average gain of 10 volts per collision in the non-relativistic region, and higher as the energy increases. It follows that except for losses the energy will increase by a factor e every 10^8 collisions. In particular, a particle starting with non-relativistic energy will attain, after N collisions, an energy

$$w = Mc^2 \exp(B^2 N). \tag{4}$$

Naturally, the energy can increase only if the losses are less than the gain in energy. An estimate to be given later (see Section VII) indicates that the ionization loss becomes smaller than the energy gain for protons having energy of about 200 Mev. For higher energy the ionization loss practically becomes negligible. We shall discuss later the injection mechanism.

IV. SPECTRUM OF THE COSMIC RADIATION

During the process of acceleration a proton may lose most of its energy by a nuclear collision. This process is observed as absorption of primary cosmic radiation in the high atmosphere and occurs with a mean free path of the order of magnitude of 70

[5] B. Stromgren, A.p.J. **108**, 242 (1948).

g/cm², corresponding to a cross section of about

$$\sigma_{\text{abs}} \approx 2.5\times 10^{-26}\ \text{cm}^2 \tag{5}$$

per nucleon.

In a collision of this type most of the kinetic energy of the colliding nucleons is probably converted into energy of a spray of several mesons.

It is reasonable to assume that the cosmic rays will occupy with approximately equal density all the interstellar space of the galaxy. They will be exposed, therefore, to the collisions with matter of an average density of 10^{-24}, leading to an absorption mean free path

$$\Lambda = 7\times 10^{25}\ \text{cm}. \tag{6}$$

A particle traveling with the velocity of light will traverse this distance in a time

$$T = \Lambda/c = 2\times 10^{15}\ \text{sec}. \tag{7}$$

or about 60 million years.

The cosmic-ray particles now present will therefore, in the average, have this age. Some of them will have accidentally escaped destruction and be considerably older. Indeed, the absorption process can be considered to proceed according to an exponential law. If we assume that original particles at all times have been supplied at the same rate, we expect the age distribution now to be

$$\exp(-t/T)dt/T. \tag{8}$$

During its age t, the particle has been gaining energy. If we call τ the time between scattering collisions, the energy acquired by a particle of age t will be

$$w(t) = Mc^2 \exp(B^2t/\tau). \tag{9}$$

Combining this relationship between age and energy with the probability distribution of age given previously, one finds the probability distribution of the energy. An elementary calculation shows that the probability for a particle to have energy between w and $w+dw$ is given by

$$\pi(w)dw = (\tau/B^2T)(Mc^2)^{\tau/B^2T}dw/w^{1+\tau/B^2T}. \tag{10}$$

It is gratifying to find that the theory leads naturally to the conclusion that the spectrum of the cosmic radiation obeys an inverse power law. By comparison of the exponent of this law with the one known from cosmic-ray observations, that is, about 2.9, one finds a relationship which permits one to determine the interval of time τ between collisions. Precisely, one finds: $2.9 = 1+\tau/B^2T$, from which follows

$$\tau = 1.9B^2T. \tag{11}$$

Using the previous values of B and T, one finds $\tau = 4\times 10^7 \approx 1.3$ years. Since the particles travel with approximately the velocity of light, this corresponds to a mean distance between collisions of the order of a light year, or about 10^{18} cm. Such a collision mean free path seems to be quite reasonable.

The theory explains quite naturally why no electrons are found in the primary cosmic radiation. This is due to the fact that at all energies the rate of loss of energy by an electron exceeds the gain. At low energies, up to about 300 Mev, the loss is mainly due to ionization. Above this energy radiative losses due to the acceleration of the electrons in the interstellar magnetic field play the dominant role. This last energy loss is instead quite negligible for protons. Also, the inverse Compton effect discussed by Feenberg and Primakoff[6] will contribute to eliminate high energy electrons.

V. THE INJECTION MECHANISM. DIFFICULTIES WITH THE INJECTION OF HEAVY NUCLEI

In order to complete the present theory, the injection mechanism should be discussed.

In order to keep the cosmic radiation at the present level it is necessary to inject a number of protons of at least 200 Mev, to compensate for those that are lost by the absorption process. According to recent evidence,[7] the primary cosmic radiation contains not only protons but also some relatively heavy nuclei. Their injection energy is much higher than that of protons, primarily on account of their large ionization loss. (See further Section VII.) Such high energy protons and heavier nuclei conceivably could be produced in the vicinity of some magnetically very active star.[8] To state this, however, merely means to shift the difficulty from the problem of accelerating the particles to that of injecting them unless a more precise estimate can be given for the efficiency of this or of some equivalent mechanism. With respect to the injection of heavy nuclei I do not know a plausible answer to this point.

For the production of protons, however, one might consider also a simple mechanism which, if the present theory is at all correct in its general features, should be responsible for at least a large fraction of the total number of protons injected. According to this mechanism the cosmic radiation regenerates itself as follows. When a fast cosmic-ray proton collides in the interstellar space against a proton nearly at rest, a good share of the energy will be lost in the form of a spray of mesons, and two nucleons will be left over with energy much less than that of the original cosmic ray. Estimates indicate that in some cases both particles may have an energy left over above the injection threshold of

[6] E. Feenberg and H. Primakoff, Phys. Rev. **73**, 449 (1948).

[7] Freier, Lofgren, Ney, and Oppenheimer, Phys. Rev. **74**, 1818 (1948); H. L. Bradt and B. Peters, Phys. Rev. **74**, 1828 (1948).

[8] See for example W. F. G. Swann, Phys. Rev. **43**, 217 (1933) and Horace W. Babcock, Phys. Rev. **74**, 489 (1948).

200 Mev, in some cases one and in some cases none. We can introduce a reproduction factor k, defined as the average number of new protons above the injection energy arising in a collision of an original cosmic-ray particle. As in a chain reaction, if k is greater than one the over-all number of cosmic rays will increase; if k is less than one it will decrease; if k is equal to one it will stay level.

Apparently the reproduction factor under interstellar conditions is rather close to one. This is perhaps not a chance, but may be due in part to the following self-stabilizing mechanism. The motions of the interstellar matter are not quite conservative, in spite of the reduced friction, caused by the magnetic fields. One should assume, therefore, that some source is present which steadily delivers kinetic energy into the streaming motions of the interstellar matter. Probably such a source of energy ultimately involves conversion of energy from the large supplies in the interior of the stars. The motions of the interstellar medium are in a dynamic equilibrium between the energy delivered by this source and the energy losses caused by friction and other causes. In this balance the amount of energy transferred by the interstellar medium to cosmic radiation is by no means irrelevant, since the total cosmic ray energy is comparable to the kinetic energy of the streaming, irregular motions of the galaxy. One should expect, therefore, that if the general level of the cosmic radiation should increase, the kinetic energy of the interstellar motion would decrease, and vice versa. The reproduction factor depends upon the density. As the density increases, the ionization losses will increase proportionally to it. This tends to increase the injection energy and consequently to decrease the reproduction factor. On the other hand, also, the rate of energy gained will change by an amount which is hard to define unambiguously. One might perhaps assume, however, that the velocity of the wandering magnetic fields increases with the $\frac{1}{3}$ power of the density, as would correspond to the virial theorem, and that the collision mean free path is inversely proportional to the $\frac{1}{3}$ power of the density, as one might get from geometrical similitude. One would find that the rate of energy increase is proportional only to the $\frac{2}{3}$ power of the density. The net effect is an increase of the injection energy and a decrease of the reproduction factor with increasing density. If the reproduction factor had been initially somewhat larger than one, the general level of the cosmic radiation would increase, draining energy out of the kinetic energy of the galaxy. This would determine a gravitational contraction which would increase the density and decrease k until the stable value of one is reached. The opposite would take place if k initially had been considerably less than one.

But even if this stabilizing mechanism is not adequate to keep the reproduction factor at the value one, and therefore an appreciable change in the general level of the cosmic radiation occurs over periods of hundreds of millions of years, the general conclusions reached in Section IV would not be qualitatively changed. Indeed, if k were somewhat different from one, the general level of the cosmic radiation would increase or decrease exponentially, depending on whether k is larger than or less than one. Consequently the number of cosmic particles injected according to the mechanism that has been discussed will not be constant in time but will vary exponentially. Combining this exponential variation with the exponential absorption (8), one still finds an exponential law for the age distribution of the cosmic particles at the present time, the only difference being that the period of this exponential will be changed by a small numerical factor.

The injection mechanism here proposed appears to be quite straightforward for protons, but utterly inadequate to explain the abundance of the heavy nuclei in the primary cosmic radiation. The injection energy of these particles is of several Bev, and it is difficult to imagine a secondary effect of the cosmic radiation on the diffuse interstellar matter which might produce this type of secondary with any appreciable probability. One might perhaps assume that the heavy particles originate at the fringes of the galaxy where the density is probably lower and the injection energy is therefore probably smaller. This, however, would require extreme conditions of density which are not easily justifiable. It seems more probable that heavy particles are injected by a totally different mechanism, perhaps as a consequence of the stellar magnetism.[8]

If such a mechanism exists one would naturally expect that it would inject protons together with heavier nuclei. The protons and perhaps to a somewhat lesser extent the α-particles would be further increased in numbers by the "chain reaction" which in this case should have $k<1$. Indeed their number would be equal to the number injected during the lifetime T increased by the factor $1/(1-k)$. Heavy particles instead would slowly gain or slowly loose energy according to whether their initial energy is above or below the injection threshold. They would, however, have a shorter lifetime than protons because of the presumably larger destruction cross section. Their number should be approximately equal to the number injected during their lifetime.

One should remark in this connection that a consequence of the present theory is that the energy spectrum of the heavy nuclei of the cosmic radiation should be quite different from the spectrum of the protons, since the absorption cross section for a heavy particle is presumably several times larger than that of a proton. One would expect, therefore, that the average age of a heavy particle is shorter than the age of a proton, which leads to an energy spectrum decreasing much more rapidly with energy for a heavy particle than it does for protons. An experimental check on this point should be possible.

VI. FURTHER DISCUSSION OF THE MAGNETIC ACCELERATION

In this section the process of acceleration of the cosmic-ray protons by collision against irregularities of the magnetic field will be discussed in somewhat more detail than has been done in Section III.

The path of a fast proton in an irregular magnetic field of the type that we have assumed will be represented very closely by a spiraling motion around a line of force. Since the radius of this spiral may be of the order of 10^{12} cm, and the irregularities in the field have dimensions of the order of 10^{18} cm, the cosmic ray will perform many turns on its spiraling path before encountering an appreciably different field intensity. One finds by an elementary discussion that as the particle approaches a region where the field intensity increases, the pitch of the spiral will decrease. One finds precisely that

$$\sin^2\vartheta/H \approx \text{constant}, \tag{12}$$

where ϑ is the angle between the direction of the line of force and the direction of the velocity of the particle, and H is the local field intensity. As the particle approaches a region where the field intensity is larger, one will expect, therefore, that the angle ϑ increases until $\sin\vartheta$ attains the maximum possible value of one. At this point the particle is reflected back along the same line of force and spirals backwards until the next region of high field intensity is encountered. This process will be called a "Type A" reflection. If the magnetic field were static, such a reflection would not produce any change in the kinetic energy of the particle. This is not so, however, if the magnetic field is slowly variable. It may happen that a region of high field intensity moves toward the cosmic-ray particle which collides against it. In this case, the particle will gain energy in the collision. Conversely, it may happen that the region of high field intensity moves away from the particle. Since the particle is much faster, it will overtake the irregularity of the field and be reflected backwards, in this case with loss of energy. The net result will be an average gain, primarily for the reason that head-on collisions are more frequent than overtaking collisions because the relative velocity is larger in the former case.

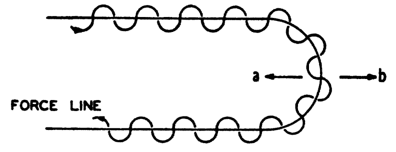


FIG. 1. Type B reflection of a cosmic-ray particle.

Somewhat similar processes take place when the cosmic-ray particle spirals around a curve of the line of force as outlined in Fig. 1 ("Type B" reflection). Here again, the energy of the particle would not change if the magnetic field were static. On the other hand, the lines of force partake of the streaming motions of the matter, and it may happen that the line of force at the bottom of the curve moves in the direction indicated by the arrow a, or that it moves in the direction indicated by the arrow b. In the former case there will be an energy gain (head-on collision) while in case b (overtaking collision) there will be an energy loss. Gain and loss, however, do not average out completely, because also in this case a head-on collision is slightly more probable than an overtaking collision due to the greater relative velocity.

The amount of energy gained or lost in a collision of the two types described can be estimated with a simple argument of special relativity, without any reference to the detailed mechanism of the collision. In the frame of reference in which the perturbation of the field against which the collision takes place is at rest, there is no change of energy of the particle. The change of energy in the rest frame of reference is obtained, therefore, by first transforming initial energy and momentum from the rest frame to the frame of the moving perturbation. In this frame an elastic collision takes place whereby the momentum changes direction and the energy remains unchanged. Transforming back to the frame of reference at rest, one obtains the final values of energy and momentum. This procedure applied to a head-on collision, gives the following result,

$$\frac{w'}{w} = \frac{1 + 2B\beta\cos\vartheta + B^2}{1 - B^2}, \tag{13}$$

where βc is the velocity of the particle, ϑ is the angle of inclination of the spiral, and Bc is the velocity of the perturbation. It is assumed that the collision is such as to produce a complete reversal of the spiraling direction by either of the two mechanisms outlined previously. For an overtaking collision, one finds a similar formula except that the sign of B must be changed. We now average the results of head-on and overtaking collisions, taking into account that the probabilities of these two types of events are proportional to the relative velocities and are given therefore by $(\beta\cos\vartheta + B/2\beta\cos\vartheta)$ for a head-on collision and $(\beta\cos\vartheta - B/2\beta\cos\vartheta)$ for an overtaking collision. The result for

TABLE I. Energy loss per g/cm² of material traversed.

Energy	Loss/g/cm²	Gain/g/cm²
10^7 ev	94×10^6 ev	7.8×10^6 ev
10^8	15×10^6	8.6×10^6
10^9	4.6×10^6	16.1×10^6
10^{10}	4.6×10^6	91×10^6

the average of $\ln(w'/w)$ up to terms of the order of B^2 is:

$$\langle \ln(w'/w)\rangle_{\mathrm{Av}} = 4B^2 - 2B^2\beta^2 \cos^2\vartheta \tag{14}$$

which confirms the order of magnitude for the average gain of energy adopted in Section III.

As a result of the extreme complication of the magnetic field and of its motion, it does not appear practical to attempt an estimate by more than the order of magnitude.

One might expect that after a relatively short time the angle ϑ will be reduced to a fairly low value so that Type A reflections will become infrequent. This is due to the fact that when ϑ is large, fairly large increases in energy and decreases of ϑ may occur, if the particle should be caught between two regions of high field moving against each other along a force line. One can prove that Type B reflections change gradually and rather slowly the average pitch of the spiral. It appears, therefore, that except for the beginning of the acceleration processes the Type A will not give as large a contribution as one otherwise might expect.

VII. ESTIMATE OF THE INJECTION ENERGY

Acceleration of a cosmic-ray particle will not be possible unless the energy gain is greater than the ionization loss. Since this last is very large for protons of low velocity, only protons above a certain energy threshold will be accelerated. In Section III a value of 200 Mev for this "injection energy" has been given; a justification for this assumed value will be given now. In estimating the injection energy we will assume that the particle, during its acceleration, finds itself both inside relatively dense clouds and in the more dilute material outside of the clouds for lengths of time proportional to the volumes of these two regions. The ionization loss will be due, therefore, to a material of an average density equal to the average density of the interstellar matter, which has been assumed to be 10^{-24} g/cm², consisting mostly of hydrogen. In Table I the energy loss per g/cm² of material traversed is given as a function of the energy of the proton. In the third column of the table the corresponding energy gain is given. It is seen that the loss exceeds the gain for particles of energy less than about 200 Mev, as has been stated.

A similar estimate yields for the acceleration of α-particles an injection energy of about 1 Bev, for the acceleration of oxygen nuclei the initial energy required is about 20 Bev, and for an iron nucleus it would amount to about 300 Bev. As already stated, it does not appear probable that the heavy nuclei found in the cosmic radiation are accelerated by the process here described, unless they should originate at some place in the galaxy where the interstellar material is extremely dilute.

I would like to acknowledge the help that I had from several discussions with E. Teller on the relative merits of the two opposing views that we are presenting. I learned many facts on cosmic magnetism from a discussion with H. Alfvén, on the occasion of his recent visit to Chicago. The views that he expressed then were quite material in influencing my own ideas on the subject.

An hypothesis on the origin of the cosmic radiation (238)

Nuovo Cimento, Suppl. ***22****, 317–323 (1949).*

N° 238.

For the introduction to this paper see paper N° 237.

238.

AN HYPOTHESIS ON THE ORIGIN OF THE COSMIC RADIATION

« Nuovo Cimento », Suppl. 317–323 (1949).

Cosmic rays are more and more being recognised as a phenomenon of cosmic importance. As an introduction I would like to give a few figures that stress this importance. We know the intensity of the cosmic radiation that comes from the outside into the atmosphere. The number of particles with an energy of the order of or greater than four billions of electron volt is about 0.1 particles per square centimetre per steradian per second. From this figure we can estimate the energy present per cm^3 in the form of cosmic rays of over 4 Gev. One finds $6 \cdot 10^{-13}$ erg/cm^3.

Very probably, particles of lower energy are also present and may be cut of by magnetic field action, perhaps by the magnetic field of the sun. By a rather uncertain estimate one may be led to increase the previous figure by a factor 3. The cosmic rays represent therefore an energy density of $22 \cdot 10^{-12}$ erg/cm^3. This energy should be compared with other astronomical or cosmic energies.

If one assumes that radiation with this average density occupies all the interstellar space of the galaxy, one obtains the result that the overall energy of the cosmic radiation is of the same order of magnitude as the kinetic energy of the disordered motions of the stars. The amount of energy is so large that one might legitimately doubt whether or not it is possible to find a mechanism capable of producing cosmic radiation in such a staggering amount. For this reason Teller has recently proposed a " Non-Cosmic Theory " of the cosmic radiation by assuming that the cosmic radiation instead of extending to the insterstellar space is confined to the immediate vicinity of the sun. This hypothesis was later developed by Teller, Richtmyer and Alfvén.

I will not discuss it now because I believe that Prof. Alfvén will do so next, and I will not even discuss the hypothesis considered in Bagge's report concerning a possible stellar origin of the cosmic radiation. I would like instead to discuss a different possibility according to which cosmic rays acquire most of their energy while travelling through space.

I want to assume: first, that the cosmic radiation is a galactic phenomenon, whereby I mean that the cosmic radiation fills with more or less uniform energy distribution all the space of our galaxy. This assumption requires a mechanism capable of holding the cosmic ray particles within the galaxy. It has been often assumed that this may be due to a galactic magnetic field with closed in lines of force. Before making further assumptions, I would like to investigate what one can deduce from this hypothesis.

Our galaxy comprises stars and matter. The diffuse matter has an average density of about 10^{-24} g/cm³, a figure easy to remember because it corresponds approximately to one hydrogen atom per cm³. A simple calculation shows that the probability of collision of a cosmic ray against a star is extremely small. However, the probability that the cosmic ray particle may have a nuclear collision is not at all negligible. Indeed, we can make a crude estimate of this probability as follows: We know directly from cosmic ray experiments that when cosmic ray particles enter from the outside into the earth's atmosphere they soon collide with air nuclei. The mean free path for this collision is of the order of magnitude of one hundred grams per cm². Since the density is 10^{-24}, the corresponding mean free path for a cosmic ray particle travelling through the galaxy will be about 10^{26} cm. Since the particle travels with almost the velocity of light, the time taken for traversing this distance will be $10^{26}/3 \cdot 10^{10} = 3 \cdot 10^{15}$ sec $= 10^{8}$ years. In the following calculations I have used slightly different figures yielding:

$$\text{(1)} \qquad \mathrm{T} = 7 \cdot 10^{7} \text{ years,}$$

for the average time that a cosmic ray particle travels before a nuclear collision happens that effectively destroys it.

This time is rather short compared to the age of the universe estimated to be two or three billion years. We are therefore led to the conclusion that only very few of the cosmic ray particles that we now observe can be as old as the galaxy. It seems necessary, therefore, to assume the existence of a mechanism that continuously produces new cosmic ray particles.

Without discussing yet what this mechanism may be, we want to introduce as a second assumption that the production is uniform in time. Since a particle has a mean life of 70 million years, its probability of survival after a time, *t*, will be

$$\text{(2)} \qquad \exp\left(\frac{-t}{\mathrm{T}}\right) \cdot$$

This expression gives the age distribution law of the particles that are now in existence.

One can now make two alternate assumptions: one is that the cosmic radiation particles are originally produced with a energy equal to or higher than their present energy. The other one is that the cosmic ray particles are originally produced at a relatively low energy and are gradually accelerated. In what follows we shall take this second point of view which has the advantage to require an injection mechanism less powerful than the one that would be required for the first assumption.

We assume, therefore, the existence of an accelerating process whereby the energy of a particle gradually increases as its age increases and is a function of the age. The dependence of the energy upon the age may then be determined from the knowledge of the energy distribution of the cosmic radiation. The experimentally known energy distribution of cosmic ray particles is rather complicated at low energies but takes the form of a simple power law for energies above a few Gev. We will assume this simplified law:

$$\text{I}\,(\text{E})\,d\text{E} = k\text{E}^{-2.9}\,d\text{E}. \tag{3}$$

the exponent 2.9 is chosen to fit the observations.

We have assumed that the energy of a particle is a function of its age, t

$$\text{E} = f\,(t) \tag{4}$$

From the knowledge of the age distribution (2), and the energy distribution (3), one can determine $f\,(t)$. Indeed, the number of particles with age between t and $t + dt$ is proportional to $\exp\left[\frac{-t}{\text{T}}\right] dt$, and the number of particles with energy between E and E + dE is proportional to

$$\frac{d\,\text{E}}{\text{E}^{2.9}} = \frac{df}{f^{2.9}}$$

we find, therefore,

$$\frac{df}{f^{2.9}} = a\exp\left[\frac{-t}{\text{T}}\right], \tag{5}$$

where a is a proportional constant. Integration yields

$$\frac{1}{1.9 f^{1.9}} = \frac{a}{\text{T}}\exp\left[\frac{-t}{\text{T}}\right], \tag{6}$$

where the integration constant has been set to equal zero because for large t, f becomes infinite. This equation can be rewritten in the form:

$$f\,(t) = \text{E}_0 \exp\left[\frac{t}{(1.9\,\text{T})}\right], \tag{7}$$

where E_0 is a new constant that represents the initial energy of the particle. From our assumptions follows a very specific law (7) whereby the energy of the cosmic ray particles must increase with time. According to (7), the energy must increase every year by a fraction of about 10^{-8} of its value, so for a proton with energy equal to its rest energy, the energy will increase at the rate of about only 10 ev per year and will increase correspondingly faster for protons of higher energy. It is clear, however, that in any case the rate of increase of the energy will be quite slow since it takes about 100 million years to double the initial value of the energy.

A very simple process that leads to the acceleration law (7), is due to the collision against large moving objects. Without specifying yet what particular objects will be considered as likely obstacles against which the collisions take place, we want to assume that a cosmic ray frequently collides against large moving obstacles. That the energy of the cosmic ray will on the average increase in such collisions is clear from the fact that ultimately statistical equilibrium would be established with equipartition of energy between the obstacles and the cosmic ray particles. This corresponds, of course, to an extremely high energy very many orders of magnitude beyond the maximum energy observed in cosmic rays. What limits the efficiency of this process in increasing the energy of the cosmic ray particles is, therefore, not the maximum energy attainable which is effectively infinite, but rather the rate at which the energy is transmitted.

Not all collisions will accelerate the particle. Actually, head-on collisions will produce an acceleration and over-taking collisions will produce a deceleration. On the average, there is acceleration primarily because head-on collisions are somewhat more probable than over-taking collisions since the relative velocity is larger in the former case. One can compute in an elementary way the order of magnitude of the average increase, δE, per collision of a particle of energy E (including rest energy) colliding against objects moving with velocity, V. The result is:

$$\delta E \sim \frac{EV^2}{c^2}\,. \tag{8}$$

If we assume that the collision cross-section is independent of the energy, the number of collisions per unit time will also be approximately independent of the energy since the velocity of the cosmic ray particles is practically constant and equal to c. It follows from (8) that the gain in energy per unit time is proportional to the energy. The energy therefore increases according to an exponential law.

We shall take V of the order of 30 km/sec. This gives (8) $\delta E \sim 10^{-8}$ E. That is again an average gain of energy per collision of one part in 10^8.

After N collisions the energy will be:

$$E = E_0 \exp\,[NB^2]\,, \tag{9}$$

where

$$B^2 = \frac{V^2}{c^2} = 10^{-8}\,.$$

In order to estimate the number of collisions, we introduce a scattering mean free path λ. The number of collisions after a time, t, since the initiation of the process will be $N = \frac{ct}{\lambda}$, since the particle travels practically at the velocity of light. Consequently we get:

$$E = E_0 \exp\left|\frac{ct\,B^2}{\lambda}\right|\,. \tag{10}$$

Comparing (10) with (7) one obtains

$$\lambda = 1.9\,B^2\,cT \sim 10^{18}\text{ cm} = 1 \text{ light year.} \tag{11}$$

We will now further specialize our assumptions and introduce the hypothesis that the collisions responsible for the increase in energy are against moving irregularities in a cosmic magnetic field.

The idea of the existence of such a moving magnetic field is due to Dr. Alfvén who has made a thorough magneto hydro-dynamic study of the influence that the extremely tenuous interstellar matter has on the propagation of a magnetic field that penetrates it.

Unfortunately, I do not have the time to explain in detail his very stimulating ideas on this subject. I shall only mention that due to the relatively high electrical conductivity of the interstellar medium, the lines of force are practically attached to the matter so that they are dragged by the turbulent motion of the interstellar matter. A cosmic ray particle will be deflected by the action of the magnetic field and will gain energy in the process as previously discussed. In order to obtain agreement with the experimentally observed spectrum, we must assume that the size of the minimum vortices which drag the lines of force is of the order of one light year, a value which does not appear implausible.

Nothing has been said so far of the injection mechanism of the particles of relatively low energy which will be further accelerated by the proposed method. In order that a particle so injected may eventually become an energetic cosmic ray, it is necessary that its initial energy be above a certain limit which will be called the injection threshold. Indeed, the accelerating mechanism will function only when the energy gained by the accelerating mechanism is greater that the energy lost by ionization. An estimate of the injection threshold yields for various particles the following values:

Protons	–	100 Mev
α Particles	–	1 Gev
Oxygen nuclei	–	1 Gev per nucleon
Iron nuclei	–	5 Gev per nucleon

It is seen that the injection threshold is quite large for heavy nuclei and this represents the most serious difficulty for the proposed theory. Even without special assumption as to the origin of the initial protons and α particles the injection of these light components of the cosmic radiation may be understood perhaps as due to the collisions of the cosmic radiation itself with the nuclei of the interstellar matter. From this point of view, cosmic radiation would be a self regenerating process.

No such simple explanation, however, seems adequate for the injection of heavier nuclei like, for instance, oxygen or iron, since no known mechanism could yield an iron nucleus with 5 Gev per nucleon without destroying the nucleus itself. One must assume, therefore, a special injection mechanism for these particles, perhaps like the one suggested recently by Spitzer [1].

In conclusion, the proposed theory seems to be quite adequate for understanding the main features of the proton components of the cosmic radiation and perhaps also of the α particle component. It does not seem adequate to understand the presence in the cosmic radiation of a significant fraction of heavier nuclei. If the general features of the present theory should prove correct, there should be an independent and very powerful injection mechanism, of the heavy component of the cosmic radiation.

BIBLIOGRAPHY.

1] LYMAN SPITZER, Jr., « Phys. Rev. », *76*, 583 (1949).

DISCUSSIONE E OSSERVAZIONI.

E. BAGGE, *Hamburg*:

Bei dem von Fermi diskutierten Mechanismus für die Beschleunigung und Aufsammlung der Höhenstrahlungsteilchen innerhalb der Milchstrasse scheint es wichtig zu sein, die Frage zu diskutieren, welcher Teil kosmischer Strahlung an der Oberfläche des galaktischen Systems in den Weltraum hinauswandert. Da das von Fermi postulierte galaktische Magnetfeld grosse räumliche Schwankungen besitzen muss, wenn der Beschleunigungsprozess überhaupt wirksam werden soll, wird es unter anderem auch nahezu feldfreie Bereiche geben und dies kann besonders an den Randbereichen der Milchstrasse das Entweichen der Höhenstrahlungsteilchen ermöglichen.

E. FERMI, *Chicago*:

In order to avoid a large loss of particles out of the boundaries of the galaxy it is sufficient to assume that the lines of force are closed or at least that very few of them escape to the outer space.

L. JANOSSY, *Dublin*:

The injection process of protons is helped by the circumstance that a nucleon colliding with a nucleus is likely to retain an appreciable fraction of its energy and thus remain very much above the injection energy. The question is raised whether heavy fragments arising out of nuclear collisions can serve to inject heavier nuclei.

E. FERMI, *Chicago*:

Naturally I hope that you are right. On the other hand it seems to me very difficult to understand the fact from the theoretical point of view. But, of course, if the fact should prove to be true it will have some theoretical explanation.

Galactic magnetic fields and the origin of cosmic radiation (265)

*ApJ **119**, 1–5 (1954).*

THE ASTROPHYSICAL JOURNAL

AN INTERNATIONAL REVIEW OF SPECTROSCOPY AND ASTRONOMICAL PHYSICS

VOLUME 119 JANUARY 1954 NUMBER 1

GALACTIC MAGNETIC FIELDS AND THE ORIGIN OF COSMIC RADIATION*

E. Fermi
Institute for Nuclear Studies, University of Chicago

Received September 11, 1953

I became interested in the possible existence of magnetic fields extending through the volume of the galaxy in connection with a discussion on the origin of the cosmic radiation a few years ago.[1] The hypothesis was discussed then that the acceleration of cosmic-ray particles to extremely high energies was due to their interaction with a galactic magnetic field that was postulated to pervade the galactic space. According to Alfvén's ideas on magnetohydrodynamics, this field would be strongly influenced by the turbulent motions of the diffuse matter within the galaxy. Indeed, the electric conductivity of this matter is so large that any lateral shift of the magnetic lines of force with respect to the matter is effectively prevented. A strong magnetic field quenches the transverse components of the displacement due to the turbulent motion. A weak field yields to the material motions, so that its lines of force are soon bent into a very crooked pattern.

The observation by Hiltner and Hall of an appreciable polarization of the light coming to us from distant stars has been interpreted[2] as due to the orientation of nonspherical dust grains by a magnetic field. If this general type of interpretation is correct, the polarization gives us some information on the strength and the direction of the magnetic field. Hiltner's measurements[3] indicate that in the vicinity of the earth the magnetic field is approximately parallel to the direction of the spiral arm. This fact suggests that we may perhaps think that the spiral arms are magnetic tubes of force. In the following discussion we will assume that this is the case.

The direction of polarization of the stellar light indicates further that in our vicinity the magnetic lines of force show irregular deviations from parallelism of the order of 10°. This fact excludes the hypothesis that the lines of force yield completely to the turbulent motions of interstellar matter, because then they would be rapidly bent into shapes much more irregular than those observed. One is rather led to the conclusion that the field is

* Sixth Henry Norris Russell Lecture of the American Astronomical Society delivered at Boulder, Colorado, on August 28, 1953.

[1] E. Fermi, *Phys. Rev.*, **75**, 1169, 1949; cited hereafter as "R."

[2] Davis and Greenstein, *Ap. J.*, **114**, 206, 1951; Spitzer and Tukey, *Ap. J.*, **114**, 187, 1951.

[3] *Ap. J.*, **114**, 241, 1951.

sufficiently strong to yield only a little to the transverse component of the turbulent motion. Indeed, as was pointed out by Davis, the small deviations from parallelism of the field enable one to estimate that the intensity of the magnetic field must be of the order of 10^{-5} gauss. Recently Chandrasekhar and I[4] have re-examined this problem, considering, in particular, the balance between magnetic and gravitational effects in the spiral arm. Our conclusion is that the field intensity is about 6×10^{-6} gauss. Owing to the turbulence, the lines of force are irregularly pushed sidewise until the magnetic stress increases to the point of forcing a reversal of the material motion and of pushing back the diffuse matter, impressing on it some kind of very irregular oscillatory motion. One expects, therefore, that the lines of force sway back and forth and also that the field intensity will fluctuate along the same line of force.

A cosmic-ray particle spiraling around these moving lines of force is gradually accelerated. The acceleration mechanism was discussed in R, although the shape of the lines of force assumed then was quite different from what we now believe it to be. A cosmic-ray proton of 10 bev energy is bent in a magnetic field of 6×10^{-6} gauss in a spiral having a radius of the order of one-third the radius of the earth's orbit. This is very small on the galactic scale. The motion of a proton with this energy and also of one with much greater energy is, therefore, properly described as a very small radius spiral around a line of force. Apart from the very rapid changes due to the spiraling, the general direction of motion may change for two reasons. One is that the line of force around which the particle spirals may be curved. In R a change of direction of this kind was called a "collision of type *b*." A second type of event, called a "collision of type *a*," takes place when the particle in its spiraling encounters a region of high field strength.

Let ϑ be the angle between the direction of the line of force and the direction of motion of the spiraling particle. The angle ϑ will be called the "angle of pitch." One can prove that, in a static magnetic field, the quantity

$$q = \frac{\sin^2 \vartheta}{H} \tag{1}$$

is approximately constant with time. For this reason, in a static field, the particle cannot enter a region where

$$H > \frac{1}{q}. \tag{2}$$

When the particle approaches such a region, its pitch decreases until $\vartheta = 90°$, at which moment the particle is reflected and spirals backward in the direction whence it came.

In a variable magnetic field both *a*- and *b*-type collisions may cause changes in energy. As a rule, the energy will increase or decrease according to whether the irregularity of the field that causes the collision moves toward the particle (head-on collision) or away from it (overtaking collision). It was shown in R that, on the average, the energy tends to increase primarily because the head-on collisions are more probable than the overtaking collisions. In the present discussion the same general acceleration mechanism will be assumed. The details, however, will be quite different from those previously assumed.

It was shown in R that through this mechanism the energy of the particle increases at a rate that, for extreme relativistic particles, is proportional to their energy. The energy E, therefore, increases exponentially with time:

$$E(t) = E_0 e^{t/A}. \tag{3}$$

According to this law, the oldest particles should have the highest energy.

The time A needed for an energy increase by a factor e was estimated in R to be about 100 million years. If this estimate was correct, A would be comparable to the time B for

[4] *Ap. J.*, **118**, 113, 1953.

nuclear collisions of the particle. Assuming that a nuclear collision effectively destroys the particle, the probability that a particle has the age t should be

$$e^{-t/B}\frac{dt}{B}. \tag{4}$$

Combining conditions (3) and (4), one readily finds that the probability that a particle observed now has energy E should be proportional to

$$\frac{dE}{E^n}, \tag{5}$$

with

$$n = 1 + \frac{A}{B}. \tag{6}$$

An exponent law like (5), with n of the order of 2 or 3, seems to fit fairly well the observed energy spectrum of the cosmic radiation.

Two main objections can be raised against the theory proposed in R. One is that, according to present evidence, the structure of the galactic magnetic field is much more regular than was assumed in R. We shall try to make plausible the conclusion that, in spite of this fact, the acceleration may still take place at an adequate rate. Indeed, as will be discussed below, it will be necessary to provide an acceleration process five to ten times more efficient than was previously supposed. The second difficulty arises from the fact that the protonic and the nuclear components of the cosmic radiation have very much the same energy spectrum. Heavy nuclei have a larger nuclear collision cross-section than protons. Their mean life B, therefore, should be shorter, and the exponent n given by equation (6) should be larger.

This difficulty would be removed if the process that eliminates the particles were equally effective against protons and against larger nuclei, because B would then be the same for both kinds of particle. For example, collisions against stars or planets do not differentiate between protons and nuclei. On the other hand, a simple estimate shows that the probability of collisions against such massive objects is quite negligible, even in a time equal to the age of the universe. Another means of removing cosmic-ray particles that is equally effective for protons and nuclei is diffusion outside the galaxy. Assume, for example, that the lines of force follow the spiral arms. The stretched-out length of the galactic spiral is about a million light-years, and the particles travel with a velocity very close to that of light. They could, therefore, escape in a time of the order of a million years. The escape time, of course, would be longer if the particle occasionally reversed its direction, owing, for example, to collisions of type a.

In a theory that yields essentially the same energy spectrum for cosmic-radiation protons and nuclei it will be necessary to assume that the escape time for diffusion outside the galaxy is appreciably shorter than the nuclear collision time. Then escape will be dominant with respect to nuclear collisions. We will assume that this is the case, and we will take, somewhat arbitrarily, in the numerical examples the escape time,

$$B = 10 \text{ million years}. \tag{7}$$

The excape time is then about ten times shorter than the nuclear collision time.

Assuming 2.5 as an average value of the exponent n in quantity (5), we have, from equation (6),

$$A = 1.5B = 15 \text{ million years}, \tag{8}$$

in which A is the time through which the energy of the particle increases, on the average, by the factor e. This time, 15 million years, is appreciably shorter than was estimated in R.

It therefore appears necessary to modify the acceleration mechanism of R in two ways. The mean free path must be much longer, in order to allow the escape of the particles from the galaxy in a relatively short time. And the process of acceleration must be much faster. At first sight, these two requirements seem to be contradictory, and perhaps they are. On the other hand, there is an acceleration mechanism that is potentially much more efficient than the others. This process was discussed in R and was then dismissed as of little importance for certain reasons to be mentioned shortly. I propose to criticize those reasons and to make a case in favor of this acceleration mechanism.

First of all, we observe that the knowledge that we now have of the general shape of the magnetic field makes the collisions of type *b* rather unimportant. We shall therefore concentrate our attention on type-*a* collisions. They take place, as will be remembered, when the particle encounters a region of large field strength, where condition (2) is fulfilled. A particle that finds itself between two such regions will be trapped on the stretch of line of force comprised between them. When this happens, the energy of the particle will change with time at a rate much faster than usual. It will decrease or increase according to whether the jaws of the trap move away from or toward each other.

Let H be the average value of the magnetic field and $H_{\max}$ the maximum field along the line of force that may be likely to cause a type-*a* reflection. If ϑ is the angle of pitch of the spiral where the field is average, reflections will occur only for particles having

$$\vartheta > \chi, \tag{9}$$

where

$$\sin \chi = \sqrt{\frac{H}{H_{\max}}}. \tag{10}$$

A simple calculation shows that when a particle with $\vartheta > \chi$ is caught in a trap, both its energy and its angle of pitch will change with time, but the product,

$$E \sin \vartheta, \tag{11}$$

remains approximately constant. The particle can escape the trap only when ϑ has decreased to the point that condition (9) is no longer fulfilled. In this process the energy must increase by a factor

$$\frac{\sin \vartheta}{\sin \chi}. \tag{12}$$

This process may lead to a sizable energy gain in a relatively short time. For example, if the jaws of the trap are 10 light-years apart and move toward each other at 10 km/sec, the time needed for a 10 per cent energy increase is only a few tens of thousands of years.

To be sure, the jaws will occasionally move away from each other, causing a loss instead of a gain of energy. But in this case ϑ will increase, making the particle more easily caught in similar traps. The process ends only when the energy has increased to the point that ϑ has become less than χ, because only then will the particle be capable of passing without reflection through occasional maxima of the field intensity that it may encounter along its path.

When this condition is reached, the process of acceleration becomes exceedingly slow. Indeed, if q in equation (1) were exactly a constant of motion, the particle would always keep on spiraling in the same direction and would eventually escape from the galaxy. It is for this reason that the process of acceleration in a trap was not considered of major importance in R. The process may become important only if there is machinery whereby the angle of pitch, after having been reduced to a small value in the process of acceleration in a trap, can be increased. If this is the case, the trap mechanism again becomes operative, and the energy may be increased by a further factor.

Now one finds that q is almost exactly a constant as long as the particle is not caught in a trap and there are no sharp variations in the magnetic field either in time or in space. When I first discussed the acceleration of cosmic rays by magnetic fields, I was not aware of the possibility of sharp discontinuities of the field and for this reason did not think that the traps could be the dominant factor. I propose now to show that discontinuities in the direction of the magnetic field should not be too exceptional in the galaxy.

Recently de Hoffmann and Teller[5] have discussed the features of magnetohydrodynamic shocks. They show, in particular, that at a shock front sudden variations in direction and intensity of the field are likely to occur. One is tempted to identify the boundaries of many clouds of the galactic diffuse matter with shock fronts. If this is correct, we have a source of magnetic discontinuities. Probably many of these discontinuities will be rather small. However, either their cumulative effect or the effect of some occasional major discontinuity will tend to convert the angle of pitch that a previous trap acceleration has reduced to a small value back to a statistical distribution corresponding to isotropy of direction. At this moment the particle is ready for a new trap acceleration.

Probably our knowledge of the galactic magnetic field is still inadequate for a realistic discussion of the process here proposed. I would like, nevertheless, to list here in a purely hypothetical way a set of parameters that may be compatible with our present knowledge. We assume that a particle for part of the time has $\vartheta < \chi$ and that it spirals in the same direction along the line of force without being caught in traps. Let λ be the average distance of travel while the particle is in this state, measured along the line of force. After the mean path λ, the particle will change ϑ back to a high value and for a period of time will be frequently caught in traps, until its energy increases by a factor f and ϑ decreases again to $\vartheta < \chi$. After this the process repeats itself. Let T be the average duration of one such cycle. The time A for acceleration by a factor e, then, is

$$A = \frac{T}{\ln f}. \tag{13}$$

The escape time B can be estimated as follows. Let L be the stretched-out length of the galaxy. The motion of the particle along L can be described as a random walk, with steps of duration T. The mean displacement in a step is given in first approximation by λ, because the particle, during the acceleration phase, changes its direction very frequently and does not move far. Estimating B with the diffusion theory, one finds

$$B = \frac{T\,(L/\lambda)^2}{\pi^2}. \tag{14}$$

A set of parameters that yields $A = 15$ million years and $B = 10$ million years is the following.

$$L = 1.3 \times 10^{24}\ \text{cm}, \qquad \lambda = 2 \times 10^{23}\ \text{cm};$$
$$T = 2.3 \times 10^{6}\ \text{years}, \qquad f = 1.17.$$

Naturally, these values are given here merely as an indication of possible orders of magnitude. Only a much more thorough discussion of the actual conditions in the galaxy may enable one to find reliable values of these quantities.

Two important questions should be discussed further. One is the injection mechanism that should feed into interstellar space an adequate number of particles of energy large enough for the present acceleration mechanism to take over. This problem was discussed in R, but no definite conclusion was reached there. The fact that the acceleration by the galactic magnetic field discussed here is appreciably faster than in R makes the requirements of the injection somewhat less stringent. Nevertheless, one still needs a very

[5] *Phys. Rev.*, **80**, 692, 1950.

powerful injection mechanism. Recent evidence that cosmic-ray-like particles are emitted by the sun indicates the stars, or perhaps stars of special types, as the most likely injectors.

A second question has to do with the energy balance of the turbulence of the interstellar gas. If it is true that the cosmic radiation leaks out of the galaxy in a time of the order of 10 million years, it is necessary that its energy be replenished a few hundred times during a time equal to the age of the universe. A simple estimate shows that the energy present in the galaxy in the form of cosmic rays is comparable to the kinetic energy due to the turbulence of the intergalactic gas. According to the present theory, the cosmic rays are accelerated at the expense of the turbulent energy. This last, therefore, must be continuously renewed by some very abundant source, perhaps like a small fraction of the radiation energy of the stars.

In conclusion, I should like to stress the fact that, regardless of the details of the acceleration mechanism, cosmic radiation and magnetic fields in the galaxy must be counted as very important factors in the equilibrium of interstellar gas.

High energy nuclear events (241)

Progr. Theor. Phys. ***5****, 570–583 (1950).*

Progress of Theoretical Physics, Vol. 5, No. 4, July~August, 1950

High Energy Nuclear Events

Enrico Fermi

Institute for Nuclear Studies
University of Chicago
Chicago, Illinois

(Received June 30, 1950)

Abstract

A statistical method for computing high energy collisions of protons with multiple production of particles is discussed. The method consists in assuming that as a result of fairly strong interactions between nucleons and mesons the probabilities of formation of the various possible numbers of particles are determined essentially by the statistical weights of the various possibilities.

1. Introduction.

The meson theory has been a dominant factor in the development of physics since it was announced fifteen years ago by Yukawa. One of its outstanding achievements has been the prediction that mesons should be produced in high energy nuclear collisions. At relatively low energies only one meson can be emitted. At higher energies multiple emission becomes possible.

In this paper an attempt will be made to develop a crude theoretical approach for calculating the outcome of nuclear collisions with very great energy. In particular, phenomena in which two colliding nucleons may give rise to several π-mesons, briefly called hereafter pions, and perhaps also to some anti-nucleons, will be discussed.

In treating this type of processes the conventional perturbation theory solution of the pioduction and destruction of pions breaks down entirely. Indeed, the large value of the interaction constant leads quite commonly to situations in which higher approximations yield larger results than do lower approximations. For this reason it is proposed to explore the possibilities of a method that makes use of this fact. The general idea is the following:

When two nucleons collide with very great energy in their center of mass system this energy will be suddenly released in a small volume surrounding the two nucleons. We may think pictorially of the event as of a collision in which the nucleons with their surrounding retinue of pions hit against each other so that all the portion of space occupied by the nucleons and by their surrounding pion field will be suddenly loaded with a very great amount of energy. Since the interactions of the pion field are strong we may expect that rapidly this energy will be distributed among the various degrees of freedom present in this volume according to statistical laws. One can then compute statistically the probability that in this tiny volume a certain number of pions will be created with a given energy distribution. It is then assumed that the concentration of energy will rapidly dissolve and that the particles into which the energy has been converted will fly out in all directions.

It is realized that this description of the phenomenon is probably as extreme, although in the opposite direction, as is the perturbation theory approach. On the other hand, it might be helpful to explore a theory that deviates from the unknown truth in the opposite direction from that of the conventional theory. It may then be possible to bracket the correct state of fact in between the two theories. One might also make a case that a theory of the kind here proposed may perhaps be a fairly good approximation to actual events at very high energy, since then the number of possible states of the given energy is large and the probability of establishing a state to its average statistical strength will be increased by the very many ways to arrive at the state in question.

The statement that we expect some sort of statistical equilibrium should be qualified as follows. First of all there are conservation laws of charge and of momentum that evidently must be fulfilled. One might expect further that only those states that are easily reachable from the initial state may actually attain statistical equilibrium. So, for example, radiative phenomena in which photons could be created will certainly not have time to develop. The only type of transitions that are believed to be fast enough are the transitions of the Yukawa theory. A succession of such transitions starting with two colliding nucleons may lead only to the formations of a number of charged or neutral pions and also presumably of nucleon-anti-nucleon pairs. The discussion shall be limited, therefore, to these particles only. Notice the additional conservation law for the difference of the numbers of the nucleons and the anti-nucleons.

The proposed theory has some resemblance to a point of view that has been adopted by Heisenberg[1] who describes a very high energy collision of two nucleons by assuming that the pion "fluid" surrounding the nucleons is set in some sort of turbulent motion by the impact energy. He uses qualitative ideas of turbulence in order to estimate the distribution of energy of this turbulent motion among eddies of different sizes. Turbulence represents the beginning of an approach to thermal equilibrium of a fluid. It describes the spreading of the energy of motion to the many states of larger and larger wave number. One might say, therefore, in a qualitative way that the present proposal consists in pushing the Heisenberg point of view to its extreme consequences of actually reaching statistical equilibrium.

The multiple meson production has also been investigated in an interesting paper by Lewis, Oppenheimer and Wouthuysen[2]. These authors stress the importance of the strong coupling expected in the pseudoscalar meson theory for the production of processes of high multiplicity.

In the theory here proposed there is only one adjustable parameter, the volume Ω, into which the energy of the two colliding nucleons is dumped. Since the pion field surrounding the nucleons extends to a distance of the order $\hbar/\mu c$ where μ is the pion mass, Ω is expected to have linear dimensions of this order of magnitude. As long as the Lorentz contraction is neglected one could take for example a sphere of radius $\hbar/\mu c$. However, when the two nucleons approach each other with very high energy in the center of gravity system, their surrounding pion clouds will be Lorentz contracted and the volume will be correspondingly reduced.

For this reason the volume Ω will be taken energy dependent according to the relationship:

$$\Omega=\Omega_0\frac{2Mc^2}{W}, \tag{1}$$

where Ω_0 is the volume without Lorentz contraction. W is the total energy of the two colliding nucleons in the center of gravity system and M is the nucleon mass. The factor $2Mc^2/W$ is the Lorentz contraction. The uncontracted volume Ω_0 may be taken as a sphere of radius R:

$$\Omega_0=4\pi R^3/3. \tag{2}$$

It is found in the applications that one seems to get an acceptable agreement with known facts by assuming:

$$R=\hbar/\mu c=1.4\times 10^{-13}\text{cm}. \tag{3}$$

This choice of the volume, although plausible as order of magnitude, is clearly arbitrary and could be changed in order to improve the agreement with experiment. One finds that an increase of Ω_0 would tend to favor processes in which a large number of particles is created.

According to this point of view the total collision cross-section of the two nucleons will be always of the order of magnitude of the geometrical cross-section of the pion cloud. In the numerical calculations actually, the total cross-section has been taken equal to area of a circle of radius R, namely,

$$\sigma_{tot}=\pi R^2. \tag{4}$$

Assuming (3) one finds $\sigma_{tot}=6\times 10^{-26}\text{cm}^2$. In order to compute the partial cross-section for a phenomenon in which for example three pions are produced in the collision, one will multiply the total cross-section (4) by the relative probability that three pions instead of any other possible number and kind of particles are produced.

The probability of transition into a state of a given type is proportional to the square of the corresponding effective matrix element and to the density of states per unit energy interval. Our assumption of a statistical equilibrium consists in postulating that the square of the effective matrix element is merely proportional to the probability that, for the state in question, all particles are contained at the same time inside Ω. For example in the case of a state that describes n completely independent particles with momenta $p_1, p_2, \cdots p_n$ this probability is $(\Omega/V)^n$ where V is the large normalization volume. The number of states per unit energy interval is

$$\left(\frac{V}{8\pi^3\hbar^3}\right)^n \frac{d}{dW}Q(W),$$

where $Q(W)$ is the volume of momentum space corresponding to the total energy W. The probability for the formation of the state in question is therefore assumed to be proportional to the product:

$$S(n)=\left(\frac{\Omega}{8\pi^3\hbar^3}\right)^n \frac{dQ(W)}{dW}. \tag{5}$$

There are some complications arising from the fact that the particles are not independent.

a) In the center of mass system the positions and momenta of only $n-1$ of the n particles are independent variables. For this reason the exponent of Ω will be $n-1$ instead of n. Also the momentum space $Q(W)$ will be $3(n-1)$-dimensional instead of $3n$-dimensional.

b) Some of the particles may be identical and this fact should be taken into account in computing $Q(W)$.

c) Some of the particles may carry a spin and one should then allow for the corresponding multiplicity of the states.

d) The conservation of angular momentum restricts the statistical equilibrium to states with angular momentum equal to that of the two colliding nucleons. In all cases considered $\lambda\!\!\!^{-}$ for the nucleons is smaller than the radius $\hbar/\mu c$ of the sphere of action. It is then meaningful to discuss separately collisions with various values of the impact parameter b (b is the distance of the two straight lines along which the nucleons move before the collision). In units of $\hbar$ the angular momentum is $l=b/\lambda\!\!\!^{-}$. The cross-section for collisions with impact parameter between b and $b+db$ is $2\pi b db=2\pi\lambda\!\!\!^{-2} l dl$. One should treat separately collisions with different values of the impact parameter and compute for each of them the probability of the various possible events. The cross-section for a special event is then obtained by adding the contribution of the various l-values.

It is found in most cases that the results so obtained differ only by small numerical factors from those obtained by neglecting the conservation of angular momentum.

This has been done as a rule in order to simplify the mathematics. The corrections arising from the conservation of angular momentum have been, however, indicated in typical cases.

2. Example. Pion Production in Low Energy Nucleon Collisions.

As a first example the production of pions in a collision of two nucleons with relative energy barely above the threshold needed for emission of a pion will be discussed. This example is chosen because it is the simplest possible. It is, however, a case in which the statistical approach may be misleading, since only few states of rather low energy are involved. We will first simplify this example by disregarding the spin of the nucleons as well as the possible existence of a spin of the pions and by disregarding also the various possible electric charges of the particles in question. In the center of gravity system we will have therefore two nucleons colliding against each other. $T/2$ is the kinetic energy of each of the two nucleons. A pion can be emitted when $T > \mu c^2$. We shall assume that this inequality is fulfilled; that the kinetic energy, however, exceeds the threshold by only a small amount, so that both the two nucleons and the pion that may be formed will have non-relativistic energies.

Conservation of energy in this case allows only two types of states; Type (a) in which the two nucleons are scattered elastically without formation of pions; and type (b) in which a pion is formed and three particles, two nucleons and a pion, emerge after the collision.

The statistical weight of the states of type (a) is obtained as follows: Since the momenta of the two nucleons are equal and opposite the momentum space will be three-dimensional. We can compute the statistical weight with (5). s will be taken$=1$ because the momentum of one particle determines that of the other. The reduced mass is $M/2$, the momentum $p=\sqrt{MT}$ and the phase space volume $Q(T)=4\pi p^3/3$. According to (5) the statistical weight of this state has therefore the familiar expression:

$$S_2=\frac{\Omega M^{3/2}}{4\pi^2\hbar^3}\sqrt{T}. \qquad (6)$$

S_2 should be compared with the statistical weight S_3 for case (b) in which three particles, two nucleons and one pion, emerge. Since the total momentum is zero, only the momenta of two of the three particles are independent and therefore in (5) $s=2$. The calculation of the momentum volume involves some slight complication on account of the conservation of momentum. Let p be the momentum of the pion and let the momenta of the two nucleons be $-\frac{1}{2}p\pm q$. The kinetic energy will then be:

$$T_1=\left(\frac{1}{2\mu}+\frac{1}{4M}\right)p^2+\frac{1}{M}q^2 \qquad (7)$$

where $T_1=T-\mu c^2$ is the kinetic energy left over after a pion has been formed. Formula (7) represents an ellipsoid in the six dimensional momentum space of the two vectors p and q. Its volume is:

$$Q_3=\frac{\pi^3}{3!}\left(\frac{4M^2\mu}{2M+\mu}\right)^{3/2}T_1^3. \tag{8}$$

The factor $\pi^3/3!$ is, for a six-dimensional sphere, tne analog of the factor $4\pi/3$ in the volume of an ordinary sphere. Substituting in (5) one finds:

$$S_3=\frac{\Omega^2}{16\pi^3\hbar^6}\left(\frac{M^2\mu}{2M+\mu}\right)^{3/2}T_1^2\approx\frac{\Omega^2 M^{3/2}\mu^{3/2}T_1^2}{32\sqrt{2}\,\pi^3\hbar^6}. \tag{9}$$

The last expression is simplified by assuming $\mu \ll M$. The probabilities of the two events (a) and (b) are proportional to S_2 and S_3. Since S_3 is very small, we may take the ratio S_3/S_2 to be the probability that the collision leads to pion formation. This is given by:

$$\frac{S_3}{S_2}=\frac{\Omega\mu^{3/2}}{8\sqrt{2}\,\pi\hbar^3}\frac{(T-\mu c^2)^2}{\sqrt{T}}\approx\frac{\Omega\mu(T-\mu c^2)^2}{8\sqrt{2}\,\pi\hbar^3 c}.$$

Since T is barely larger than the thresholo energy μc^2, this value has been substituted for T in the denominator. In this case also the Lorentz contraction of the two colliding nucleons is negligible and we can therefore substitute for Ω the value Ω_0 given by (2) and (3). One finds:

$$\frac{S_3}{S_2}=\frac{1}{6\sqrt{2}}\left(\frac{T}{\mu c^2}-1\right)^2. \tag{11}$$

The cross-section for pion formation is given by the product of the total cross-section (4) and the probability (11). For example, in a bombardment of nucleons at rest with 345 MeV nucleons, (the proton energy available at Berkeley) one finds that the energy available in the center of gravity system is $T=165$MeV. On the other hand $\mu c^2=140$MeV and the previous formula gives therefore $S_3/S_2=.0038$. This means that at this bombarding energy a pion will be formed in about 0.4 per cent of the nucleon collisions.

If one examines the process more in detail one will recognize that in a collision of two protons the probability of emission of a positive pion is twice (11) namely, .0076. Because if a positive pion is formed a proton and a neutron, instead of two protons, will also emerge. Their statistical weight is twice that for two protons because they are not identical particles. Similarly in the collision of a proton and a neutron the probability of emission of a positive pion is one-half of (10) namely, .0019. The probability of emission of a negative pion is the same.

For example, when a carbon target is bombarded by 345 MeV protons the probabilities that the collision takes place between the proton and another proton or a neutron are the same. Hence, the probability of emission of a positive pion will be .0076/2+.0019/2; that is, .0048; and the probability of emission of a negative pion will be .0019/2=.001. Since the nuclear cross-section of carbon is about 3×10^{-25}cm^2, one will obtain the expected values of the cross-sections for emission of a positive and a negative pion by multiplying the nuclear cross-section by the above probabilities. The results are $1.4\times10^{-27}\text{cm}^2$ for the positive and $3\times10^{-28}\text{cm}^2$ for the negative pions. Considering the extremely crude calculation these values are in surprisingly good agreement with the experimental results.

In the above discussion the conservation of angular momentum has been disregarded. When a pion is produced the kinetic energy of the three emerging particles is small and they will therefore escape in an s-state. Consequently, only the initial states of zero angular momentum can be contribute to this type of final state. Their maximum cross-section has the well known expression $\pi\lambda\!\!\!^{-2}$ which is appreciably smaller than (4). However, also the competition of elastic scattering versus pion production is less since only the scattering states of zero angular momentum will contribute.

By carrying out the calculation one finds that the two effects almost cancel each other and that the conservation of angular momentum changes the previous results for the cross-section for pion production by only a factor 2/3. As long as the conservation of angular momentum is neglected one expects the scattering of the two nucleons to be spherically symmetrical in the center of mass system. This is no longer the case when the angular momentum is conserved. One finds then that the elastic scattering cross-section per steradian in the center of mass system instead of being constant is approximately proportional to $1/\sin\theta$, where θ is the scattering angle in the same system.

3. Formulas for the Statistical Weights.

Some standard formulas expressing the statistical weights, S, for a number of simple cases will be collected here.

First, the case will be considered that after a collision n particles emerge with masses m_1, $m_2\cdots\cdots m_n$. Neglecting spin properties and assuming that the particles are statistically independent, and disregarding also the momentum conservation, one finds for S the following two formulas corresponding to the classical and to the extreme relativistic case:

$$S_n=\frac{(m_1m_2\cdots m_n)^{3/2}\Omega^n}{2^{3n/2}\pi^{3n/2}\hbar^{3n}}\frac{T^{3n/2-1}}{(3n/2-1)\,!}, \tag{12}$$

(classical case)

$$S_n=\frac{\Omega^n}{\pi^{2n}\hbar^{3n}c^{3n}}\frac{W^{3n-1}}{(3n-1)\,!}. \tag{13}$$

(extr. relativistic case)

In (12) T is the total classical kinetic energy of the n particles and in (13) W is the total energy including rest energy of the n particles. One can also compute a formula for the case that s of the particles, usually the nucleons, are classical and n, usually the pions, are extreme relativistic. Neglecting again spin statistics and momentum conservation and assuming further that all the classical particles have the nucleon mass, M, one finds:

$$S(s,n)=\frac{M^{3s/2}\Omega^{n+s}}{2^{3s/2}\pi^{2n+3s/2}\hbar^{3s+3n}c^{3n}}\frac{(W-sMc^2)^{3n+3s/2-1}}{(3n+3s/2-1)!}. \tag{14}$$

It is sometimes convenient to re-write (14) in the following form:

$$S(s,n)=\frac{M^{3s/2}\Omega^{s/2+1/3}}{2^{3s/2}\pi^{s/2+2/3}\hbar^{3s/2+1}c^{-3s/2+1}}\frac{\left(\frac{\Omega^{1/3}(W-sMc^2)}{\pi^{2/3}\hbar c}\right)^{3n+3s/2-1}}{(3n+3s/2-1)!}, \tag{15}$$

since it is thus easy to obtain an approximate expression for the sum of the statistical weights $S(s,n)$ over all values of n. The approximation applies to the cases when the average value of n is $\gg 1$. One finds then:

$$\sum_{n=0}^{\infty} S(s,n)\approx\frac{M^{3s/2}\Omega^{s/2+1/3}}{3\times 2^{3s/2}\pi^{s/2+2/3}\hbar^{3s/2+1}c^{-3s/2+1}}\exp\left(\frac{\Omega^{1/3}(W-sMc^2)}{\pi^{2/3}\hbar c}\right). \tag{16}$$

The numerical values of (15) and (16) adopting for Ω (1), (2) and (3) are:

$$Mc^2S(s,n)=\frac{6.31}{w^{1/3}}\left(\frac{98.8}{w}\right)^{s/2}\frac{\left(6.31\frac{w-s}{w^{1/3}}\right)^{3n+3s/2-1}}{(3n+3s/2-1)!} \tag{17}$$

and:

$$Mc^2\sum_{n=0}^{\infty} S(s,n)\approx\frac{2.10}{w^{1/3}}\left(\frac{98.8}{w}\right)^{s/2}\exp\left(6.31\frac{w-s}{w^{1/3}}\right) \tag{18}$$

where one has put

$$w=W/Mc^2 \tag{19}$$

and it has further been assumed $\mu/M=.15$.

In the previous formulas the momentum conservation has been disregarded. The formulas, however, can be generalized without difficulty so as to introduce at least approximately the requirement that the total momentum be zero. The approximation consists in assuming that the mass of the pion is very small compared to the nucleon mass. The momentum of the nucleons will then be much greater than the momentum of the pions since the kinetic energy is approximately equipartitioned among the various particles. Therefore, one can approximately apply the condition that the sum of the momenta vanishes to the nucleons only. One recognizes then that formula (15) must be changed as follows. a) Instead of s one will write $s-1$ at all places except in the term $W-sMc^2$ since we have now $s-1$ independent momenta of the heavy particles. b) The factor $M^{3s/2}$ must be changed because instead of the mass one should substitute an expression which is the analog of a reduced mass. It is found that the factor in question must be substituted by $M^{3(s-1)/2}/s^{3/2}$. When the conservation of momentum is approximately taken into account formulas (17) and (18) should then be changed as follows:

$$Mc^2S(s, n) = \frac{6.31}{s^{3/2}w^{1/3}}\left(\frac{98.8}{w}\right)^{(s-1)/2}\frac{(6.31(w-s)/w^{1/3})^{3n+3s/2-1}}{(3n+3s/2-1)\,!}, \tag{20}$$

$$Mc^2\sum_{n=0}^{\infty} S(s, n) \approx \frac{2.10}{s^{3/2}w^{1/3}}\left(\frac{98.8}{w}\right)^{(s-1)/2}\exp(6.31(w-s)/w^{1/3}). \tag{21}$$

In all the preceding formulas the particles have been assumed to be statistically independent. As long as few nucleons and pions are involved, the error is not great. Larger errors are expected for high multiplicity. The formulas, however, become quite involved since there are at least three kinds of pions and four kinds of nucleons and anti-nucleons. No attempt has been made to introduce these complications for phenomena of relatively low energy. They have been calculated as if there were only one type of pions, one of nucleons and one of anti-nucleons statistically independent. This procedure is certainly inadequate and will give a too high multiplicity at high energy. For phenomena of extremely high energy it becomes simple to introduce the statistical correlations by substituting the statistical by a thermo-dynamical model. This case will be treated in Section 6.

In the previous expressions also the conservation of angular momentum has been neglected. The error introduced with this omission will be discussed in Section 6, where an appropriate correction factor for it will be given.

4. Transition from Single to Multiple Production of Pions.

In Section 2 the emission of a single low energy pion has been discussed. Collisions of higher energy in which besides the two original nucleons also several pions may be produced will be considered now. A rough indication of the features of this process may be obtained by computing the relative probabilities for the emission of 0, 1, 2,······n··· pions with (20). In that formula one will put $s=2$. Statistical correlations and conservation of angular momentum will be negelcted. Omitting a common factor, the probabilities of the various values of n are proportional to:

$$\left\{\frac{251}{w}(w-2)^3\right\}^n\Big/\left(\frac{3}{2}\times\frac{5}{2}\times\cdots\times\frac{6n+1}{2}\right). \tag{22}$$

Table I gives the probabilities of pion production of different multiplicities calculated according to this formula. The first column of the table gives the energy, w, of the two nucleons in the center of gravity system in units of Mc^2. The second column gives the energy, w', of the primary particle in the laboratory frame of reference. The next eight columns are labeled by the number, n, of pions produced and give the probabilities of various events in per cents. The last column gives the average number of pions produced.

TABLE I

w	w'	$n=0$	1	2	3	4	5	6	7	$\bar{n}$
2.5	2.1	49	47	4						.6
3	3.5	9	59	30	2					1.2
3.5	5.1	2	31	46	18	3				1.9
4	7.0		13	40	33	11	2			2.5
5	11.5		2	15	34	31	14	3	1	3.5

Notice that already for a bombarding energy of about 1 BeV corresponding to the first line of the table the probability of elastic collision of the two nucleons is 50%. This probability decreases rapidly and drops below one per cent for bombarding energies of about 5 BeV. As the bombarding energy increases the probability of multiple phenomena increases as indicated in the table. The most probable value of n according to (22) should be given approximately by $2.1(w-2)/w^{1/3}$.

It will be seen in Section 6 that at high energy very appreciable errors are are introduced by neglecting the angular momentum conservation and the statistical correlations. Table I gives only a qualitative indication of the transition from elastic scattering to single and then multiple pion production. The quantitative features of the multiple production, however, should be more reliably represented by formula (32).

5. Production of Anti-Nucleons.

When the two colliding pions have a total energy $>4Mc^2$ in the center of gravity system, competition with processes in which a nucleon-anti-nucleon pair is formed becomes possible. When the energy is barely above the $4Mc^2$-threshold, no pions can be formed accompanying the pair. As the energy increases, however, the pair will be as a rule accompanied by a number of pions. For moderate energy $w<10$ one will use formula (20). Substituting in it $s=4$ we obtain the statistical weight for nucleon pair formation associated with the emission of n pions. Substituting $s=2$ we obtain an expression proportional to the probability that no pair is formed and the two original nucleons plus n pions emerge.

Omitting the common factor Mc^2 one obtains from (20) :

$$S(4,\,n)=\frac{775}{w^{11/6}}\,\frac{(6.31\,(w-4)/w^{1/3})^{3n+7/2}}{(3n+7/2)\,!}. \tag{23}$$

In normalizing these probabilities to total probabilitiy$=1$, one can make use of the fact that the probability of pair formation in the range of energies here discussed is always less than one per cent. One can therefore disregard the pair formation in the normalization factor which reduces to $\sum\limits_n S(2,n)$. In calculating this sum one can use (21). One obtains in the end the following expression for the probability of pair formation accompanied by n pions :

$$P(4,\,n)=\frac{105}{w}\left(6.31\,\frac{w-4}{w^{1/3}}\right)^{3n+7/2}\frac{\exp(-6.31(w-2)/w^{1/3})}{(3n+7/2)\,!}.$$

TABLE II

w	w'	$n=0$	$n=1$	$n=2$	$n=3$	$n=4$	Total
$4+\varepsilon$	7.0	$1000\,\varepsilon^{7/2}$					$.1\varepsilon^{7/2}$
4.5	9.1	14	.6				15×10^{-4}
5	11.5	27	8	.7			36×10^{-4}
5.5	14.3	21	21	5	.5		47×10^{-4}
6	16.9	12	25	14	3	.3	54×10^{-4}

Columns 3 to 7 inclusive $\times\ 10^{-4}$

Table II is calculated with this formula. Again the first and second columns represent in units of Mc^2 the total energy in the center of gravity system and the total energy of the bombarding particle in the laboratory system. The next five columns give the probabilities of pair formation accompanied by n pions. These probabilities have been multiplied by a factor 10^4. The eighth column is the total probability of pair formation.

Again, in computing this table the statistical correlations mentioned in Section 3 and the conservation of angular momentum have been disregarded. For this reason the data of the table are merely indicative of the results that would be given by a more correct computation.

At the highest energy here considered the probability of anti-nucleon formation is 0.005. Since in a collision of this energy probably two or three pions are formed in the average one concludes that at these energies the ratio of anti-nucleons to pions formed is about 0.002. Therefore, anti-nucleons will be hard to find even in fairly high energy collisions.

6. Collisions of Extremely High Energy.

In discussing the collision of two nucleons with extremely high energy one can simplify the calculations by assuming that all the various particles produced are extreme relativistic and that thermo-dynamics may be applied instead of a detailed statistical computation of the probabilities of the various events.

In this discussion the conservation of angular momentum will first be neglected. Its effect will be given at the end of this section.

The extremely high energy density that is suddenly formed in the volume Ω will give rise to multiple production of pions. and of pairs of nucleons and anti-nucleons. Since both kinds of particles are extreme relativistic, the energy density will be proportional to the fourth power of the temperature, T, as in Stefan's law.

The pions, like the photons, obey the Bose-Einstein statistics. Since we further assume that the temperature is so high that the rest mass is negligible, their energy momentum relationship will be the same as for the photons. Consequently the Stefan's law for the pions will be quite similar to the ordinary Stefan's law of the black body radiation. The difference is only in a statistical weight factor. For the photons the statistical weight is the factor, 2, because of the two polarization directions. If we assume that the pions have spin zero and differ only by their charge $\pm e$ or 0, their statistical weight will be 3. Consequently, the energy density of the pions will be obtained by multiplying the energy density of the ordinary Stefan's law by the factor 3/2. This energy density is therefore:

$$\frac{3\times 6.494(kT)^4}{2\pi^2\hbar^3c^3}. \tag{25}$$

The numerical factor $6.494=\pi^4/15$ is six times the sum of the inverse fourth powers of the integral numbers.

The contribution of the nucleons and anti-nucleons to the energy density is given by a similar formula. The differences are that the statistical weight of the nucleons is eight since we have four different types of nucleons and anti-nucleons and for each, two spin orientations. A further difference is due to the fact that these particles obey the Pauli principle. In the extreme relativistic case their energy density is:

$$\frac{4\times 5.682(kT)^4}{\pi^2\hbar^3c^3}. \tag{26}$$

Here the numerical factor 5.682 is $6\sum_1^\infty(-1)^{n+1}/n^4$.
The temperature is obtained by equating the total energy to the product of the volume Ω times the sum of the two energy densities (25) and (26). Making use of (1) one obtains the temperature from the following equation:

$$(kT)^4=.152\frac{\hbar^3c^3W^2}{Mc^2\Omega_0}. \tag{27}$$

In order to compute the number of pions, nucleons and anti-nucleons produced we need formulas for the density of the various particles. These are computed according to standard procedures of statistical mechanics. In the extreme relativistic case the density of the particles turns out to be proportional to the third power of the temperature. The total densities of the pions and of the nucleons are given by the following two expressions:

$$n_\pi=.367\frac{(kT)^3}{\hbar^3c^3}; \qquad n_N=.855\frac{(kT)^3}{\hbar^3c^3}. \tag{28}$$

The total numbers of pions and nucleons are obtained by multiplying the expressions (28) by the volume Ω and by substituting in them the temperature calculated from (27). The result must be finally corrected in order to take into account the conservation of angular momentum. Only the result of this correction will be given. It is found that conservation of angular momentum has the effect of reducing the numbers of pions and nucleons by a factor that has been calculated numerically to be about .51. The conservation of angular momentum has the further effect that the angular distribution of particles produced is no longer isotropical but tends to favor somewhat, particles moving parallel to the original direction of the two colliding nucleons. Introducing these corrections one finds that the number of pions is:

$$\text{No. of pions}=.091\left(\frac{\Omega_0 MW^2}{c\hbar^3}\right)^{1/4}=.54\sqrt{W/Mc^2} \tag{29}$$

and the number of nucleons plus anti-nucleons is:

$$\text{No. of nucleons and anti-nucleons}=.21\left(\frac{\Omega_0 MW^2}{c\hbar^3}\right)^{1/4}=1.3\sqrt{WMc^2}\,. \tag{30}$$

From this follows that the number of charged particles that emerge out of an extremely high energy collision is given by:

$$1.2(W'/Mc^2)^{1/4}. \tag{30a}$$

(IV'=energy in the laboratory system)

In these tormulas Ω_0 has been substituted by its value, (2), (3).

These formulas apply only to extreme high energies. Substituting the value, (2), (3), for Ω_0 one finds from (27) that the relationship between temperature and energy can be written in the form:

$$kT/Mc^2=.105\sqrt{W/Mc^2}\,. \tag{31}$$

Relativistic conditions for the nucleons will be achieved therefore only when $W>100Mc^2$. This corresponds in the laboratory system to an energy of the bombarding particles of more than 5×10^{12} eV. At somewhat lower energies the number of anti-nucleon pairs formed will decrease very rapidly, especially since an energy $2Mc^2$ is needed in order to form a pair. In this energy range the formation of paris is probably better represented by the computation of Section 5.

A comparison of (29) and (30) indicates that in such collisions of extremely high energy the number of nucleons and anti-nucleons produced exceeds that of the pions. Naturally, the anti-protons which are the particles in which we are most interested from the experimental point of view are only one-fourth of the particles (30). Therefore, a somewhat larger number of pions than of anti-protons is formed, even at these high energies. The reason why so many nucleons of all kinds are formed compared to the pions is their statistical weight (8 for the nucleons, 3 for the pions).

In an intermediate energy range where the multiple production of pions is the relevant phenomenon one can still apply the thermo-dynamic method restricting, however, the thermo-dynamic equilibrium to the pion gas only, and assuming that the activation energy of the pairs is too high for producing a sizeable numbet of these particles at the given temperature. The energy density in this case will be given by (25). The numerical coefficient in formula (27) will be reduced for this reason from .152 to .046. Also in the same tormula one will substitute W by $W-2Mc^2$ since the energy of the two nucleons does not contribute to the energy of the pion gas. Introducing also the factor .51 for the conseivation of angular momentum one finds that the number of pions in this approximation is given by:

$$\text{No. of pions} = .323\frac{M^{1/4}R^{3/4}(W-2Mc^2)^{3/2}}{\hbar^{3/4}c^{1/4}W} = 1.34\frac{(w-2)^{3/2}}{w}, \tag{32}$$

where $w=W/Mc^2$.

In the intermediate energy range of bombarding particles from 10 to 100 BeV this formula probably gives a better estimate of the multiplicities than do the computations of Table I. In particular it would appear that especially the multiplicities given in the last two lines of Table I are too large. According to (32) one would expect for these two energies multiplicities of about 2 instead of the considerably higher values given in Table I. The difference is due to two effects which have been disregarded in computing Table I; namely, the statistical correlation between various types of pions and the angular momentum conservation. Both factors are approximately taken into account in formula (32).

Since no observation of multiple production of an isolated nucleon is available at present, the comparison of these findings with experimental results is only tentative. The present theory seems to give rather low multiplicities except at extremely high energies of the order of 10^{12} to 10^{13} eV. As more experimental results become available it may be possible to improve the agreement of the theory with experiment by changing the choice (3) of R. If experimentally the multiplicities should turn out to be larger than according to the theory, one would increase R or decrease it in the opposite case.

In the present theory we have considered only one type of mesons, the pions. If mesons of larger mass strongly bound to nucleons should exist, as seems to be indicated by the recent experiments of Anderson[3], these particles also could reach statistical equilibrium. Since their rest energy is large, however, they would compete unfavorably with the production of pions except at very high energies. One would expect therefore in most collisions that the number of pions produced should be appreciably larger than that of the heavier mesons.

References

1) W. Heisenberg, Nature, **164** (1949), 65, ZS. f. Phys., **126** (1949), 569.

2) H. W. Lewis, J. R. Oppenheimer and S. A. Wouthuysen, Phys. Rev. **73** (1948), 127.

3) A. J. Seriff, R. B. Leighton, C. Hsiao, E. W. Cowan and C. D. Anderson, Phys. Rev. **78** (1950), 290.

Magnetic fields in spiral arms (261)

*ApJ **118**, 113–115 (1953).*

MAGNETIC FIELDS IN SPIRAL ARMS

S. Chandrasekhar and E. Fermi
University of Chicago

Received March 23, 1953

ABSTRACT

In this paper two independent methods are described for estimating the magnetic field in the spiral arm in which we are located. The first method is based on an interpretation of the dispersion (of the order of 10°) in the observed planes of polarization of the light of the distant stars; it leads to an estimate of $H = 7.2 \times 10^{-6}$ gauss. The second method is based on the requirement of equilibrium of the spiral arm with respect to lateral expansion and contraction: it leads to an estimate of $H = 6 \times 10^{-6}$ gauss.

The hypothesis of the existence of a magnetic field in galactic space[1] has received some confirmation by Hiltner's[2] observation of the polarization of the light of the distant stars. It seems plausible that this polarization is due to a magnetic orientation of the interstellar dust particles;[3] for such an orientation would lead to different amounts of absorption of light polarized parallel and perpendicular to the magnetic field and, therefore, to a polarization of the light reaching us. On this interpretation of the interstellar polarization we should expect to observe no polarization in the general direction of the magnetic lines of force and a maximum polarization in a direction normal to the lines of force. And if we interpret from this point of view the maps[4] of the polarization effect as a function of the direction of observation, it appears that the direction of the galactic magnetic field is roughly parallel to the direction of the spiral arm in which we are located. In this paper we shall discuss some further consequences of this interpretation of interstellar polarization, in an attempt to arrive at an estimate of the strength of the interstellar magnetic field.

As we observe distant stars in a direction approximately perpendicular to the spiral arm, it appears that the direction of polarization is only approximately parallel to the arm. There are indeed quite appreciable and apparently irregular fluctuations in the direction of polarization of the distant stars.[4] This would indicate that the magnetic lines of force are not strictly straight and that they may be better described as "wavy" lines. The mean angular deviation of the plane of polarization from the direction of the spiral arm appears to be about $\alpha = 0.2$ radians.[4] There must clearly be a relation between this angle, α, and the strength of the magnetic field, H. For, if the magnetic field were sufficiently strong, the lines of force would be quite straight and α would be very small; on the other hand, if the magnetic field were sufficiently weak, the lines of force would be dragged around in various directions by the turbulent motions of the gas masses in the spiral arm and α would be large. To obtain the general relation between α and H, we proceed as follows:

The velocity of the transverse magneto-hydrodynamic wave is given by

$$V = \frac{H}{\sqrt{(4\pi\rho)}}, \tag{1}$$

[1] E. Fermi, *Phys. Rev.*, **75**, 1169, 1949.

[2] W. A. Hiltner, *Ap. J.*, **109**, 471, 1949.

[3] Of the two theories which have been proposed (L. Spitzer and J. W. Tukey, *Ap. J.*, **114**, 187, 1951, and L. Davis and J. L. Greenstein, *Ap. J.*, **114**, 206, 1951), that by Davis and Greenstein appears to be in better accord with the facts.

[4] W. A. Hiltner, *Ap. J.*, **114**, 241, 1951.

where ρ is the density of the diffused matter. In computing the velocity, V, we should not include in ρ the average density due to the stars, since the stars may be presumed to move across the lines of force without appreciable interaction with them, whereas the diffused matter in the form of both gas and dust has a sufficiently high electrical conductivity to be effectively attached to the magnetic lines of force in such a way that only longitudinal relative displacements are possible.

According to equation (1), the transverse oscillations of a particular line of force can be described by an equation of the form

$$y = a \cos k\,(x - Vt), \tag{2}$$

where x is a longitudinal co-ordinate and y represents the lateral displacement. We take the derivatives of y with respect to x and t and obtain

$$y' = -\,ak \sin k\,(x - Vt)$$

and

$$\dot{y} = -\,akV \sin k\,(x - Vt). \tag{3}$$

From these equations it follows that

$$V^2\overline{y'^2} = \overline{\dot{y}^2}\,. \tag{4}$$

The lateral velocity of the lines of force must be equal to the lateral velocity of the turbulent gas. If v denotes the root-mean-square velocity of the turbulent motion, we should have

$$\overline{\dot{y}^2} = \tfrac{1}{3}\,v^2\,. \tag{5}$$

The factor $\frac{1}{3}$ arises from the fact that only one component of the velocity is effective in shifting the lines of force in the y-direction. The quantity y', on the other hand, represents the deviation of the line of force from a straight line projected on the plane of view. Hence,

$$\overline{y'^2} = a^2\,. \tag{6}$$

Now, combining equations (1), (4), (5), and (6), we obtain

$$H = (\tfrac{4}{3}\pi\rho)^{1/2}\,\frac{v}{a}\,. \tag{7}$$

In equation (7) we shall substitute the following numerical values, which appear to describe approximately the conditions prevailing in the spiral arm in which we are located:[5]

$$\rho = 2 \times 10^{-24}\ \text{gm/cm}^3,\ v = 5 \times 10^5\ \text{cm/sec, and } a = 0.2\ \text{radians}. \tag{8}$$

With these values equation (7) gives

$$H = 7.2 \times 10^{-6}\ \text{gauss}. \tag{9}$$

An alternative procedure for estimating the intensity of the magnetic field is based on the requirement of equilibrium of the spiral arm with respect to lateral expansion and contraction. As an order of magnitude, we may expect to obtain the condition for this equilibrium by equating the gravitational pressure in the spiral arm to the sum of the material pressure and the pressure due to the magnetic field. In computing the gravita-

[5] For ρ, the estimate of J. H. Oort (cf. *Ap. J.*, **116**, 233, 1952) from observations of the 21-cm line is used; while the value of v adopted is that of A. Blaauw, *B.A.N.*, **11**, 405, 1952.

tional pressure, we should allow for the gravitational force due to all the mass present, i.e., of the stars as well as of the diffused matter. We are interested, however, in computing the gravitational pressure exerted on the diffused matter only. Assuming for simplicity that the spiral arm is a cylinder of radius R with uniform density, one finds for the gravitational pressure:

$$p_{\text{grav}} = \pi G \rho \rho_t R^2 \,, \tag{10}$$

where G denotes the constant of gravitation, ρ is the density of the diffused matter only, and ρ_t is the total mean density, including the contribution of the stars. The kinetic pressure of the turbulent gas is given by

$$p_{\text{kin}} = \tfrac{1}{3}\rho v^2 \tag{11}$$

while the magnetic pressure is given by

$$p_{\text{mag}} = \frac{H^2}{8\pi} \,. \tag{12}$$

And for the equilibrium we must have

$$p_{\text{grav}} = p_{\text{kin}} + p_{\text{mag}} \,. \tag{13}$$

In computing p_{grav} we shall assume a radius of the spiral arm of 250 parsecs or $R = 7.7 \times 10^{20}$ cm. As before, we shall take $\rho = 2 \times 10^{-24}$ gm/cm^3; and for ρ_t we shall assume[6] 6×10^{-24} gm/cm^3. For these values of R, ρ, and ρ_t equation (9) gives $p_{\text{grav}} = 1.5 \times 10^{-12}$ dynes, while p_{kin} computed with the values already given is 0.2×10^{-12} dynes. We attribute the difference to the magnetic pressure. Hence

$$\frac{H^2}{8\pi} = 1.3 \times 10^{-12} \,, \tag{14}$$

or

$$H = 6 \times 10^{-6} \text{ gauss.} \tag{15}$$

The two independent methods of estimating H therefore agree in giving essentially the same value for the field strength. A field of about 7×10^{-6} gauss indicated by these estimates is ten times smaller than that which Davis and Greenstein[3] have estimated as necessary for producing an adequate orientation of the dust particles to account for the interstellar polarization. If the present estimate of 7×10^{-6} gauss is correct, one should conclude that the mechanism of orientation is somewhat more effective than has been assumed by Davis and Greenstein.

Since this paper was written, our attention has been drawn to the fact that the idea underlying the first of the two methods by which we estimate the magnetic field in the spiral arm is contained in an earlier paper by Leverett Davis, Jr. (*Phys. Rev.*, **81,** 890, 1951). We are sorry that we were not aware of this paper when we wrote ours. However, since with the better estimates of the astronomical parameters now available the value of H derived is a great deal different from Davis' value and since further the value we have derived is in accord with our second independent estimate, we have allowed the paper to stand in its original form.

[6] Cf. J. H. Oort, *Ap. J.*, **116,** 233, 1952.

Problems of gravitational stability in the presence of a magnetic field (262)

ApJ ***118****, 116–141 (1953).*

PROBLEMS OF GRAVITATIONAL STABILITY IN THE PRESENCE OF A MAGNETIC FIELD

S. Chandrasekhar and E. Fermi
University of Chicago

Received March 23, 1953

ABSTRACT

In this paper a number of problems are considered which are related to the gravitational stability of cosmical masses of infinite electrical conductivity in which there is a prevalent magnetic field. In Section I the virial theorem is extended to include the magnetic terms in the equations of motion, and it is shown that when the magnetic energy exceeds the numerical value of the gravitational potential energy, the configuration becomes dynamically unstable. It is suggested that the relatively long periods of the magnetic variables may be due to the magnetic energy of these stars approaching the limit set by the virial theorem. In Section II the adiabatic radial pulsations of an infinite cylinder along the axis of which a magnetic field is acting is considered. An explicit expression for the period is obtained. Section III is devoted to an investigation of the stability for transverse oscillations of an infinite cylinder of incompressible fluid when there is a uniform magnetic field acting in the direction of the axis. It is shown that the cylinder is unstable for all periodic deformations of the boundary with wave lengths exceeding a certain critical value, depending on the strength of the field. The wave length of maximum instability is also determined. It is found that the magnetic field has a stabilizing effect both in increasing the wave length of maximum instability and in prolonging the time needed for the instability to manifest itself. For a cylinder of radius $R = 250$ parsecs and $\rho = 2 \times 10^{-24}$ gm/cm^3 a magnetic field in excess of 7×10^{-6} gauss effectively removes the instability. In Section IV it is shown that a fluid sphere with a uniform magnetic field inside and a dipole field outside is not a configuration of equilibrium and that it will tend to become oblate by contracting in the direction of the field. Finally, in Section V the gravitational instability of an infinite homogeneous medium in the presence of a magnetic field is considered, and it is shown that Jeans's condition is unaffected by the presence of the field.

1. *Introduction.*—In this paper we shall consider a number of problems relating to the dynamical and gravitational stability of cosmical masses in which there is a prevalent magnetic field. In the discussion of these problems, the assumption will be made that the medium is effectively of infinite electrical conductivity. This latter assumption implies only that the conductivity is large enough for the magnetic lines of force to be considered as practically attached to the matter during the length of time under consideration; it has been known for some time that this is the case in most astronomical connections.[1]

The abstract gives an adequate summary of the paper.

I. THE VIRIAL THEOREM AND THE CONDITION FOR DYNAMICAL STABILITY

2. *The virial theorem.*—In a subject such as this it is perhaps best that we start by establishing theorems of the widest possible generality. The extension of the virial theorem to include the forces derived from the prevailing magnetic field provides such a starting point. We shall see that under conditions of equilibrium this extension of the virial theorem leads to the relation

$$2T + 3(\gamma - 1)\mathfrak{U} + \mathfrak{M} + \Omega = 0 \tag{1}$$

between the kinetic energy (T) of mass motion, the heat energy ($\mathfrak{U}$) of molecular motion, the magnetic energy ($\mathfrak{M}$) of the prevailing field, and the gravitational potential energy (Ω), where γ denotes the ratio of the specific heats. That a relation of the form (1) should exist is readily understood: For the balance between the pressures p_{kin}, p_{gas}, and p_{mag} due

[1] Cf. L. Biermann, *Annual Review of Nuclear Science*, **2** (Stanford: Annual Reviews, Inc., 1953), 349.

to the visible motions, the molecular motions, and the magnetic field, on the one hand, and the gravitational pressure, p_{grav}, on the other, requires

$$p_{\text{kin}} + p_{\text{gas}} + p_{\text{mag}} = p_{\text{grav}} , \tag{2}$$

while the order of magnitudes of these pressures are given by

$$p_{\text{kin}} = c_1 \frac{T}{V} , \qquad p_{\text{gas}} = c_2 \frac{\mathfrak{U}}{V} , \qquad p_{\text{mag}} = \frac{H^2}{8\pi} = c_3 \frac{\mathfrak{M}}{V} , \tag{3}$$

and

$$p_{\text{grav}} = \text{Density} \times \text{gravity} \times \text{linear dimension} = - c_4 \frac{\Omega}{V} , \tag{3a}$$

where V denotes the volume of the configuration and c_1, c_2, c_3, and c_4 are numerical constants. A relation of the form (1) is therefore clearly implied. We now proceed to establish the exact relation (1).

With the usual assumptions of hydromagnetics, the equations of motion governing an inviscid fluid can be written in the form

$$\rho \frac{du_i}{dt} = - \frac{\partial}{\partial x_i} \left(p + \frac{|\boldsymbol{H}|^2}{8\pi} \right) + \rho \frac{\partial V}{\partial x_i} + \frac{1}{4\pi} \frac{\partial}{\partial x_j} H_i H_j , \tag{4}$$

where ρ denotes the density, p the pressure, V the gravitational potential, and $\boldsymbol{H}$ the intensity of the magnetic field. (In eq. [1] and in the sequel, summation over repeated indices is to be understood.)

Multiply equation (4) by x_i and integrate over the volume of the configuration. Reducing the left-hand side of the equation in the usual manner, we find

$$\begin{aligned} \iiint \rho x_i \frac{du_i}{dt} dx_1 dx_2 dx_3 &= \int_0^M x_i \frac{d^2 x_i}{dt^2} dm \\ &= \frac{1}{2} \frac{d^2}{dt^2} \int_0^M r^2 dm - \int_0^M |\boldsymbol{u}|^2 dm , \end{aligned} \tag{5}$$

where $dm = \rho dx_1 dx_2 dx_3$ and the integration is effected over the entire mass, M, of the configuration. Letting

$$I = \int_0^M r^2 dm \qquad \text{and} \qquad T = \frac{1}{2} \int |\boldsymbol{u}|^2 dm \tag{6}$$

denote the moment of inertia and the kinetic energy of mass motion, respectively, we have

$$\begin{aligned} \frac{1}{2} \frac{d^2 I}{dt^2} - 2T = &- \iiint x_i \frac{\partial}{\partial x_i} \left(p + \frac{|\boldsymbol{H}|^2}{8\pi} \right) dx_1 dx_2 dx_3 \\ &+ \frac{1}{4\pi} \iiint x_i \frac{\partial}{\partial x_j} H_i H_j dx_1 dx_2 dx_3 + \int_0^M x_i \frac{\partial V}{\partial x_i} dm . \end{aligned} \tag{7}$$

The last of the three integrals on the right-hand side of this equation represents the gravitational potential energy, Ω, of the configuration. The remaining two volume integrals can be reduced by integration by parts. Thus the first of the two integrals gives

$$\begin{aligned} &- \iiint x_i \frac{\partial}{\partial x_i} \left(p + \frac{|\boldsymbol{H}|^2}{8\pi} \right) dx_1 dx_2 dx_3 \\ &\quad = - \int \left(p + \frac{|\boldsymbol{H}|^2}{8\pi} \right) \boldsymbol{r} \cdot d\boldsymbol{S} + 3 \iiint \left(p + \frac{|\boldsymbol{H}|^2}{8\pi} \right) dx_1 dx_2 dx_3 . \end{aligned} \tag{8}$$

The surface integral (over $d\boldsymbol{S}$) vanishes, since the pressure (including the magnetic pressure $|\boldsymbol{H}|^2/8\pi$) must vanish on the boundary of the configuration; and the volume integral over p and $|\boldsymbol{H}|^2/8\pi$ is readily expressible in terms of the internal energy ($\mathfrak{U}$) and the magnetic energy ($\mathfrak{M}$) of the configuration. Thus we have

$$-\iiint x_i \frac{\partial}{\partial x_i}\left(p+\frac{|\boldsymbol{H}|^2}{8\pi}\right) dx_1 dx_2 dx_3 = 3(\gamma-1)\mathfrak{U}+3\mathfrak{M}\,, \tag{9}$$

where γ denotes the ratio of the specific heats. In the same way the second volume integral in equation (7) gives

$$\frac{1}{4\pi}\iiint x_i \frac{\partial}{\partial x_j} H_i H_j dx_1 dx_2 dx_3 = -2\mathfrak{M}\,. \tag{10}$$

Now, combining equations (7), (9), and (10), we have

$$\frac{1}{2}\frac{d^2 I}{dt^2} = 2T + 3(\gamma-1)\mathfrak{U}+\mathfrak{M}+\Omega\,. \tag{11}$$

This is the required generalization of the virial theorem; it differs from the usual one only in the appearance of $\mathfrak{M}+\Omega$ in place of Ω.

3. *The condition for dynamical stability.*—If the configuration is one of equilibrium, then it follows from the virial theorem that

$$3(\gamma-1)\mathfrak{U}+\mathfrak{M}+\Omega = 0\,. \tag{12}$$

On the other hand, the total energy, $\mathfrak{E}$, of the configuration is given by

$$\mathfrak{E} = \mathfrak{U}+\mathfrak{M}+\Omega\,. \tag{13}$$

Eliminating $\mathfrak{U}$ between equations (12) and (13), we obtain

$$\mathfrak{E} = -\frac{3\gamma-4}{3(\gamma-1)}(|\Omega|-\mathfrak{M})\,. \tag{14}$$

From this equation for the total energy it follows that *a necessary condition for the dynamical stability of an equilibrium configuration is*

$$(3\gamma-4)(|\Omega|-\mathfrak{M}) > 0\,. \tag{15}$$

Thus, even when $\gamma > \frac{4}{3}$ (the condition for dynamical stability in the absence of a magnetic field) a sufficiently strong internal magnetic field can induce dynamical instability in the configuration. In fact, according to formula (15), the condition for dynamical stability, when $\gamma > \frac{4}{3}$, is

$$\mathfrak{M} = \frac{1}{8\pi}\iiint |\boldsymbol{H}|^2 dx_1 dx_2 dx_3 = \tfrac{1}{6}R^3 (H^2)_{\text{av}} < |\Omega|\,, \tag{16}$$

where $(H^2)_{\text{av}}$ denotes the mean square magnetic field.

For a spherical configuration of uniform density,

$$\Omega = -\frac{3}{5}\frac{GM^2}{R}\,, \tag{17}$$

where M is its mass, R is its radius, and G is the constant of gravitation. We can use this expression for Ω to estimate the limit imposed by the virial theorem on the magnetic fields which can prevail. On expressing M and R in units of the solar mass and the solar radius, we find from equations (16) and (17) that

$$\sqrt{(H^2)_{\mathrm{av}}} < 2.0 \times 10^8 \frac{M}{R^2} \text{ gauss}. \tag{18}$$

For the peculiar A stars for which Babcock has found magnetic fields of the order of 10^4 gauss, we may estimate that

$$M \simeq 4\odot \quad \text{and} \quad R \simeq 5R_\odot \quad \text{(A star)}; \tag{19}$$

and expression (18) gives

$$\sqrt{(H^2)_{\mathrm{av}}} < 3 \times 10^7 \text{ gauss} \quad \text{(A stars)}. \tag{19a}$$

Of greater interest is the limit set by expression (18) for an S-type star for which Babcock has found a variable magnetic field of the order of 1000 gauss. For an S-type star we can estimate that

$$M \simeq 15\odot \quad \text{and} \quad \mathrm{R} \simeq 300R_\odot \quad \text{(S star)}; \tag{20}$$

and (18) now gives

$$\sqrt{(H^2)_{\mathrm{av}}} < 3 \times 10^4 \text{ gauss} \quad \text{(S star)}. \tag{20a}$$

Thus the limit set by (18) is seen to be two to three orders of magnitude larger than the surface fields observed by Babcock. However, the fields in the interior may be much stronger than the surface fields; and it is even possible that the actual root-mean-square fields in these stars are near their maximum values. Indeed, from the fact that the periods of the magnetic variables are long compared with the adiabatic pulsation periods they would have if they were nonmagnetic, we may surmise that $\sqrt{(H^2)_{\mathrm{av}}}$ is near the limit set by (18); for, as is well known, we may lengthen the period of the lowest mode of oscillation of a system by approaching the limit of dynamical stability; and we can accomplish this by letting $\mathfrak{M} \rightarrow |\Omega|$.

Note added June 17.—Since we wrote this paper, Dr. Babcock has informed us that he has measured a variable magnetic field (+2000 to −1200 gauss) for the star VV Cephei. It has been estimated that for this star $M = 100\odot$ and $R = 2600R_\odot$. With these values, inequality (18) gives $\sqrt{(H^2)_{\mathrm{av}}} < 3000$ gauss. We may conclude that this star must be on the verge of dynamical stability and, anticipating the result established in Section IV, probably highly oblate.

4. *The virial theorem for an infinite cylindrical distribution of matter.*—Some care is needed in applying the results of §§ 2 and 3 to a distribution of matter which can be idealized as an infinite cylinder (such as, for example, a spiral arm); for the potential energy per unit length of an infinite cylinder is infinite. For this reason it is perhaps best that we consider the problem *de novo.*

We shall consider, then, an infinite cylinder in which the prevailing magnetic field is in the direction of the axis of the cylinder; and we shall suppose that all the variables are functions only of the distance r from the axis of the cylinder. Under these conditions the equations of motion reduce to the single one

$$\rho \frac{du_r}{dt} = -\frac{\partial}{\partial r}\left(p + \frac{H^2}{8\pi}\right) - \frac{2Gm(r)}{r} \rho, \tag{21}$$

where $m(r)$ is the mass per unit length interior to r.

Multiplying equation (21) by $2\pi r^2 dr$ and integrating over the entire range of r, we find in the usual manner that

$$\frac{1}{2}\frac{d^2}{dt^2}\int_0^M r^2 dm - \int_0^M \left(\frac{dr}{dt}\right)^2 dm = 2(\gamma - 1)\mathfrak{U} + 2\mathfrak{M} - GM^2, \tag{22}$$

where M denotes the mass per unit length of the cylinder and $\mathfrak{U}$ and $\mathfrak{M}$ are the kinetic and the magnetic energies per unit length of the cylinder, respectively.

From equation (22) it follows that, under equilibrium conditions, we should have

$$2(\gamma - 1)\mathfrak{U} + 2\mathfrak{M} - GM^2 = 0 \,. \tag{23}$$

A necessary condition for equilibrium to obtain is, therefore,

$$\mathfrak{M} < \tfrac{1}{2}GM^2 \,. \tag{24}$$

We can rewrite the condition (24) alternatively in the form

$$\mathfrak{M} = \frac{1}{4}\int_0^R H^2 r\,dr = \tfrac{1}{8}(H^2)_{\text{av}}R^2 < \tfrac{1}{2}\pi^2 R^4 \bar{\rho}^2 G \,, \tag{25}$$

or

$$\sqrt{(H^2)_{\text{av}}} < 2\pi R\bar{\rho}G \,. \tag{25a}$$

This last condition on the root-mean-square field is essentially equivalent to one of the formulae used in the preceding paper[2] (eq. [13]) for estimating the magnetic field in the spiral arm; the difference between the two formulae arises from the fact that in that paper the gravitational attraction was not limited to the interstellar gas only; allowance was also made for the stars contributing to the gravitational force acting on the gas.

II. THE RADIAL PULSATIONS OF AN INFINITE CYLINDER

5. *The pulsation equation.*—In view of the inconclusive nature of the current treatments[3] of the adiabatic pulsations of magnetic stars, it is perhaps of interest to see how the corresponding problem in infinite cylinders can be fully solved. We consider, then, the radial pulsation of an infinite cylinder, along the axis of which there is a prevailing magnetic field.

Choosing the time t and the mass per unit length, $m(r)$, interior to r, as the independent variables, we can write the equations of continuity and motion in the forms

$$\frac{\partial}{\partial m}(\pi r^2) = \frac{1}{\rho} \tag{26}$$

and

$$\frac{\partial^2 r}{\partial t^2} = -2\pi r\frac{\partial P}{\partial m} - \frac{2Gm(r)}{r} \,, \tag{27}$$

where

$$P = p + \frac{H^2}{8\pi} \tag{28}$$

denotes the total pressure. Distinguishing the values of the various parameters for the equilibrium configuration by a subscript zero and writing

$$r = r_0 + \delta r, \qquad P = P_0 + \delta P, \qquad \rho = \rho_0 + \delta\rho, \text{ etc.} \tag{29}$$

we find that the equations governing radial oscillations of small amplitudes are

$$\frac{\partial}{\partial m}(2\pi r_0 \delta r) = -\frac{\delta\rho}{\rho_0^2} \tag{30}$$

and

$$\frac{\partial^2}{\partial t^2}\delta r = -2\pi\delta r\frac{\partial P_0}{\partial m} - 2\pi r_0\frac{\partial}{\partial m}\delta P + \frac{2Gm}{r_0^2}\delta r \,. \tag{31}$$

[2] *Ap. J.*, **118**, 113, 1953.

[3] M. Schwarzschild, *Ann. d'ap.*, **12**, 148, 1949; G. Gjellestad, *Rep. No. 1, Inst. Ther. Ap.* (Oslo, 1950), and *Ann. d'ap.*, **15**, 276, 1952; V. C. A. Ferraro and D. J. Memory, *M.N.*, **112**, 361, 1952; T. G. Cowling, *M.N.*, **112**, 527, 1952.

Using the equation

$$2\pi r_0 \frac{\partial P_0}{\partial m} = -\frac{2Gm}{r_0}, \tag{32}$$

which must obtain in equilibrium, we can rewrite equation (31) in the form

$$\frac{\partial^2}{\partial t^2}\,\delta r = -2\pi r_0 \frac{\partial}{\partial m}\,\delta P + \frac{4Gm}{r_0^2}\,\delta r\,. \tag{33}$$

We shall now evaluate δP. For an adiabatic pulsation,

$$\delta P = \delta p + \frac{\boldsymbol{H}_0 \cdot \delta \boldsymbol{H}}{4\pi} = \gamma \frac{p_0}{\rho_0}\,\delta\rho + \frac{\boldsymbol{H}_0 \cdot \delta \boldsymbol{H}}{4\pi}. \tag{34}$$

Now when the medium is of infinite electrical conductivity, the change, $\Delta\boldsymbol{H}$, *at a given point* in a prevailing magnetic field, $\boldsymbol{H}_0$, caused by a displacement $\delta\boldsymbol{r}$, is given quite generally by

$$\Delta\boldsymbol{H} = \operatorname{curl}\,(\delta\boldsymbol{r} \times \boldsymbol{H}_0)\,. \tag{35}$$

This relation is derived in § 14 below (see eq. [130]); but we may note here that it merely expresses the fact that the changes in the magnetic field are simply a consequence of the lines of force being pushed aside. According to equation (35), the change in the magnetic field, $\delta\boldsymbol{H}$, *as we follow the motion*, is given by

$$\delta\boldsymbol{H} = \operatorname{curl}\,(\delta\boldsymbol{r} \times \boldsymbol{H}_0) + (\delta\boldsymbol{r}\cdot\operatorname{grad})\,\boldsymbol{H}_0\,. \tag{36}$$

When $\boldsymbol{H}_0$ is in the z-direction and $\delta\boldsymbol{r}$ is radial, the only nonvanishing component of $\delta\boldsymbol{H}$ is

$$\delta H_z = -\frac{1}{r}\frac{\partial}{\partial r}(H_0 r\,\delta r) + \delta r\,\frac{\partial H_0}{\partial r} = -\frac{H_0}{r}\frac{\partial}{\partial r}(r\,\delta r)\,, \tag{37}$$

in the z-direction. Hence in the case under consideration

$$\frac{\boldsymbol{H}_0 \cdot \delta\boldsymbol{H}}{4\pi} = -\frac{H_0^2 \rho_0}{4\pi}\frac{\partial}{\partial m}(2\pi r_0\,\delta r)\,, \tag{38}$$

and the expression for δP becomes

$$\delta P = -\left(\gamma p_0 + \frac{H_0^2}{4\pi}\right)\rho_0 \frac{\partial}{\partial m}(2\pi r_0\,\delta r)\,, \tag{39}$$

where we have substituted for $\delta\rho$ in equation (34) in accordance with equation (30).

With δP given by equation (38), equation (33) takes the form

$$\frac{\partial^2}{\partial t^2}\,\delta r = 4\pi^2 r\,\frac{\partial}{\partial m}\left\{\left(\gamma p + \frac{H^2}{4\pi}\right)\rho\,\frac{\partial}{\partial m}(r\,\delta r)\right\} + \frac{4Gm}{r^2}\,\delta r\,. \tag{40}$$

In writing equation (39), we have suppressed the subscripts zero distinguishing the equilibrium configuration, since there is no longer any cause for ambiguity.

When all the physical variables vary with time like $e^{i\sigma t}$, equation (39) reduces to

$$\left(\sigma^2 + \frac{4Gm}{r^2}\right)\delta r = -4\pi^2 r\,\frac{d}{dm}\left\{\left(\gamma p + \frac{H^2}{4\pi}\right)\rho\,\frac{d}{dm}(r\,\delta r)\right\}, \tag{41}$$

where δr has now the meaning of an amplitude.

The boundary conditions,

$$\delta r = 0 \quad \text{at} \quad m = 0 \quad \text{and} \quad \delta P = 0 \quad \text{at} \quad m = M\,, \tag{42}$$

in conjunction with the pulsation equation (40) will determine for σ^2 a sequence of possible characteristic values, σ_k^2. And it can be readily shown that the solutions, δr_k, belonging to the different characteristic values, are orthogonal:

$$\int_0^M \delta r_k \delta r_l dm = 0 \qquad (k \neq l)\,. \tag{43}$$

In view of this orthogonality of the functions δr_k, we should expect that the characteristic values themselves could be determined by a variational method. The basis for this method is developed in the following section.

6. *An integral formula for σ^2 and a variational method for determining it.*—Multiply equation (41) by δr and integrate over the range of m, i.e., from 0 to M. We obtain

$$\begin{aligned}\sigma^2 \int_0^M (\delta r)^2 dm + 4G \int_0^M \left(\frac{\delta r}{r}\right)^2 dm& \\ = -4\pi^2 \int_0^M r\,\delta r \frac{d}{dm}&\left\{\left(\gamma p + \frac{H^2}{4\pi}\right) \rho \frac{d}{dm}(r\,\delta r)\right\} dm\,.\end{aligned} \tag{44}$$

By integrating by parts the integral on the right-hand side, we obtain

$$\sigma^2 \int_0^M (\delta r)^2 dm + 4G \int_0^M \left(\frac{\delta r}{r}\right)^2 dm = 4\pi^2 \int_0^M \left(\gamma p + \frac{H^2}{4\pi}\right) \rho \left[\frac{d}{dm}(r\,\delta r)\right]^2 dm\,. \tag{45}$$

Writing $p = (P - H^2/8\pi)$ in equation (45), we obtain, after some elementary reductions,

$$\begin{aligned}\sigma^2 \int_0^M (\delta r)^2 dm = \gamma \int_0^M \frac{P}{\rho}\left[\frac{1}{r}\frac{d}{dr}(r\,\delta r)\right]^2 dm - 4G \int_0^M \left(\frac{\delta r}{r}\right)^2 m\,dm& \\ + \frac{1}{8\pi}(2-\gamma) \int_0^M \frac{H^2}{\rho}\left[\frac{1}{r}\frac{d}{dr}(r\,\delta r)\right]^2 dm\,.&\end{aligned} \tag{46}$$

It can be shown that the foregoing equations give a minimum value for σ^2 when the true solution δr belonging to the lowest characteristic value of the pulsation equation is substituted; and any other function δr (satisfying the boundary conditions) will give a larger value for σ^2. These facts can clearly be made the basis of a variational procedure for determining σ^2.

In the theory of the adiabatic pulsations of ordinary stars, it is known[4] that we get a very good estimate of σ^2 (for the fundamental mode) by setting

$$\delta r = \text{Constant } r, \tag{47}$$

in an integral formula for σ^2 similar to equation (46). We shall assume that this will continue to be the case in our present problem. Therefore, making the substitution (47) in equation (46), we obtain

$$\sigma^2 \int_0^M r^2 dm = 4\gamma \int_0^M \frac{P}{\rho} dm - 2GM^2 + (2-\gamma) \int_0^R H^2 r\,dr\,. \tag{48}$$

[4] P. Ledoux and C. L. Pekeris, *Ap. J.*, **94**, 124 1941.

On the other hand,

$$\int_0^M \frac{P}{\rho}\, d m = 2\pi \int_0^R P r d r = -\pi \int_0^R r^2 \frac{dP}{d r}\, d r = G \int_0^M m d m = \tfrac{1}{2} G M^2 . \tag{49}$$

Hence

$$\sigma^2 \int_0^M r^2 d m = 2\,(\gamma - 1) G M^2 + (2 - \gamma) \int_0^R H^2 r d r . \tag{50}$$

An alternative form of this equation is (cf. eq. [23])

$$\begin{aligned} \sigma^2 \int_0^M r^2 d m &= 4\,(\gamma - 1)\,[\tfrac{1}{2} G M^2 - \mathfrak{M}] + 4\,\mathfrak{M} \\ &= 4\,[\,(\gamma - 1)^2 \mathfrak{U} + \mathfrak{M}\,] . \end{aligned} \tag{51}$$

III. THE GRAVITATIONAL INSTABILITY OF AN INFINITELY LONG CYLINDER WHEN A CONSTANT MAGNETIC FIELD IS ACTING IN THE DIRECTION OF THE AXIS

7. *The formulation of the problem.*—In Section II we have seen that an infinitely long cylinder in which there is a prevalent magnetic field in the direction of the axis is stable for radial oscillations. But the question was left open as to whether the cylinder may not be unstable for transverse or for longitudinal oscillations. In Section III we shall take up the discussion of the transverse oscillations; however, in order not to complicate an already difficult problem, we shall restrict ourselves to the case when the medium is incompressible in addition to being an infinitely good electrical conductor.

We picture to ourselves, then, an infinite cylinder of uniform circular cross-section of radius R_0, along the axis of which a constant magnetic field of intensity H_0 is acting. Since any transverse perturbation can be expressed as a superposition of waves of different wave lengths, the question of stability can be investigated by considering, individually, perturbations of different wave lengths. We suppose, then, that the cylinder is subject to a perturbation, the result of which is to deform the boundary into

$$r = R + a \cos kz . \tag{52}$$

Since the fluid is assumed to be incompressible, the mass per unit length must be the same before and after the deformation; this, clearly, requires that

$$R_0^2 = R^2 + \tfrac{1}{2} a^2 . \tag{53}$$

We shall see that, as a result of the deformation, the mean field in the z-direction is also changed by an amount of order a^2 (see eq. [87] below).

The investigation of the stability of the cylinder consists of two parts. First, we must calculate the change in the potential energy, $\Delta\Omega$, and the magnetic energy, $\Delta\mathfrak{M}$, per unit length resulting from the perturbation. Then, depending on whether $\Delta\Omega + \Delta\mathfrak{M}$ is positive or negative, we shall have stability or instability. We shall see presently that $\Delta\Omega + \Delta\mathfrak{M} < 0$ for all k less than a certain determinate value depending on H_0. In other words, the cylinder is unstable for all wave lengths exceeding a certain critical value λ_*. The determination of λ_* is the first problem in the investigation of stability. The second problem concerns the specification of the wave number k_m (say) for which the instability will develop at the maximum rate. We can determine this mode of maximum instability by considering the amplitude of the deformation (cf. eq. [52]) as a function of time, constructing a Lagrangian for the cylinder and determining the manner of increase of the amplitudes of the unstable modes. We shall find that whenever $\lambda > \lambda_*$ (or $k < k_*$), the amplitude increases like e^{qt}, where q is a function of k (and H_0). The mode of maximum instability is clearly the one which makes q (for a given H) a maximum.

Before proceeding to the details of the calculations, we may state that the method we have described derives from an early investigation of Rayleigh's[5] on the stability of liquid jets.

8. *The change in the potential energy per unit length caused by the deformation.*—Following the outline given in § 7, we shall first calculate the change in the potential energy, $\Delta\Omega$, per unit length caused by the deformation which makes the cross-section change from one of a constant radius R_0 to one whose boundary is given by equation (52). Since the potential energy per unit length of an infinite cylinder is infinite, the evaluation of $\Delta\Omega$ requires some care. We proceed as follows:

Let U and V denote the external and the internal gravitational potentials of the deformed cylinder. They satisfy the equations

$$\nabla^2 U = 0 \qquad \text{and} \qquad \nabla^2 V = -4\pi G\rho \, . \tag{54}$$

We shall first solve these equations to the first order in the amplitude a appropriately for the problem on hand. The solutions must clearly be of the forms

$$U = -2\pi G\rho R^2 \log r + aAK_0(kr) \cos kz + c_0 \tag{55}$$

and

$$V = -\pi G\rho r^2 + aBI_0(kr) \cos kz \, , \tag{55a}$$

where c_0 is an additive constant (with which we need not further concern ourselves), A and B are constants to be determined, and I_n and K_n are the Bessel functions of order n for a purely imaginary argument, which have no singularity at the origin and at infinity, respectively.

The constants A and B in solutions (55) are to be determined by the condition that U and V and $\partial U/\partial r$ and $\partial V/\partial r$ must be continuous on the boundary (52). Carrying out the calculations consistently to the first order in a, we find that the continuity conditions require

$$A\, K_0\,(kR) = B\, I_0\,(kR) \tag{56}$$

and

$$A\, K_1\,(kR) + B\, I_1\,(kR) = \frac{4\pi G\rho}{k} \, . \tag{56a}$$

Solving these equations, we find

$$A = 4\pi G\rho R I_0(kR) \qquad \text{and} \qquad B = 4\pi G\rho R K_0(kR) \, . \tag{57}$$

The required solution for V is, therefore,

$$V = -\pi G\rho r^2 + 4\pi G\rho R \; aK_0(kR) I_0(kr) \cos kz + O(a^2). \tag{58}$$

Now suppose that the amplitude of the deformation is increased by an *infinitesimal* amount from a to $a + \delta a$. The change in the potential energy, $\delta\Delta\Omega$, consequent to this infinitesimal increase in the amplitude, can be determined by evaluating the work done in the redistribution of the matter required to increase the amplitude. For evaluating this latter work, it is necessary to specify in a quantitative manner the redistribution which takes place; and we shall now do this.

An arbitrary deformation of an incompressible fluid can be thought of as resulting from a displacement $\boldsymbol{\xi}$ applied to each point of the fluid. The assumed incompressibility of the medium requires that div $\boldsymbol{\xi} = 0$; and, since no loss of generality is implied by supposing that the displacement is irrotational, we shall write

$$\boldsymbol{\xi} = \text{grad}\, \psi, \tag{59}$$

[5] Lord Rayleigh, *Scientific Papers* (Cambridge: At the University Press, 1900), **2**, 361; also *Theory of Sound* ("Dover Reprints" [New York, 1945]), **2**, 350–362.

and require that

$$\nabla^2\psi = 0. \tag{60}$$

A solution of equation (60) which is suitable for considering the deformation of a uniform cylinder into one whose boundary is given by (52) is

$$\psi = AI_0(kr)\cos kz, \tag{61}$$

where A is a constant. The corresponding radial and z-components of $\boldsymbol{\xi}$ are

$$\xi_r = AkI_1(kr)\cos kz \quad \text{and} \quad \xi_z = -AkI_0(kr)\sin kz. \tag{62}$$

Since at $r = R$, ξ_r must reduce to $a\cos kz$ (cf. eq. [52]), we must have

$$A = \frac{a}{kI_1(kR)}. \tag{63}$$

The displacements,

$$\xi_r = a\frac{I_1(kr)}{I_1(kR)}\cos kz \quad \text{and} \quad \xi_z = -a\frac{I_0(kr)}{I_1(kR)}\sin kz, \tag{64}$$

applied to each point of the cylinder will deform it into the required shape. The displacement $\delta\boldsymbol{\xi}$, which must be applied to increase the amplitude from a to $a + \delta a$, is therefore,

$$\delta\xi_r = \delta a\frac{I_1(kr)}{I_1(kR)}\cos kz \quad \text{and} \quad \delta\xi_z = -\delta a\frac{I_0(kr)}{I_1(kR)}\sin kz. \tag{65}$$

The change in the potential energy, $\delta\Delta\Omega$, per unit length involved in the infinitesimal deformation (65) can be obtained by integrating over the whole cylinder the work done by the displacement $\delta\boldsymbol{\xi}$ in the force field specified by the gravitational potential (58). It is therefore given by

$$\delta\Delta\Omega = -2\pi\rho\left\{\int_0^{R+a\cos kz}\delta\boldsymbol{\xi}\cdot\text{grad}\,V\,r\,dr\right\}_{\text{av}}, \tag{66}$$

where the averaging is to be done with respect to z. Substituting for V and $\delta\boldsymbol{\xi}$ from equations (58) and (65), we obtain

$$\begin{aligned}\delta\Delta\Omega = -2\pi\rho\delta a\Bigg\{&\int_0^{R+a\cos kz}\cos kz\frac{I_1(kr)}{I_1(kR)}[-2\pi G\rho r\\ &+4\pi\rho GRakK_0(kR)I_1(kr)\cos kz]\,r\,dr\\ &+\int_0^{R+a\cos kz}\sin kz\frac{I_0(kr)}{I_1(kR)}[4\pi\rho GRakK_0(kR)I_0(kr)\sin kz]\,r\,dr\Bigg\}_{\text{av}}.\end{aligned} \tag{67}$$

Evaluating the foregoing expression consistently to the first order in a, we find

$$\delta\Delta\Omega = 2\pi^2\rho^2GR^2a\delta a - 4\pi^2\rho^2GRa\delta a\frac{kK_0(kR)}{I_1(kR)}\int_0^R[I_1^2(kr) + I_0^2(kr)]\,r\,dr, \tag{68}$$

or, using the readily verifiable result,

$$\int_0^R[I_1^2(kr) + I_0^2(kr)]\,r\,dr = \frac{R}{k}I_0(kR)I_1(kR), \tag{69}$$

we have

$$\delta\Delta\Omega = 4\pi^2\rho^2GR^2[\tfrac{1}{2} - I_0(x)K_0(x)]a\delta a, \tag{70}$$

where for the sake of brevity we have written

$$x = kR\,. \tag{71}$$

Finally, integrating equation (70) over a from 0 to a, we obtain

$$\Delta\Omega = 2\pi^2\rho^2GR^2[\tfrac{1}{2} - I_0(x)K_0(x)]a^2\,. \tag{72}$$

This is the required expression for the change in the potential energy per unit length caused by the deformation.

9. *The change in the magnetic energy per unit length caused by the deformation.*—The changes in the magnetic field inside the cylinder can best be determined from the condition that the magnetic induction across any section normal to the axis of the cylinder must remain unaffected by the deformation. This condition follows from the assumed infinite electrical conductivity of the medium. Thus, if

$$H\mathbf{1}_z + \boldsymbol{h} \tag{73}$$

represents the magnetic field inside the cylinder (where $\mathbf{1}_z$ is a unit vector in the z-direction, $\boldsymbol{h}$ is a field, of order a, varying periodically with z, and H is the mean field), we must require that

$$N = \int_0^R H_0\,r\,d\,r = \int_0^{R+a\cos kz} (H + h_z)\,r\,d\,r = \text{Constant}\,. \tag{74}$$

Turning to the determination of H and $\boldsymbol{h}$, we may first observe that $\boldsymbol{h}$ can be derived from a magnetostatic potential ϕ satisfying the equation $\nabla^2\phi = 0$. For the problem on hand we can represent ϕ as a series in powers of a of the form

$$\phi = \sum_{n=1}^{\infty} \frac{a^n A_n}{nk} I_0\,(n\,k\,r)\sin\,n\,k\,z\,, \tag{75}$$

where the A_n's are constants to be determined. Retaining terms up to the second order in a, we have

$$h_r = aA_1I_1(kr)\,\sin\,kz + a^2A_2I_1(2kr)\,\sin\,2kz \tag{76}$$

and

$$h_z = aA_1I_0(kr)\,\cos\,kz + a^2A_2I_0(2kr)\,\cos\,2kz\,, \tag{77}$$

for the components of $\boldsymbol{h}$.

With h_z given by equation (77), the magnetic induction across a normal section of the cylinder is given by

$$N = \int_0^{R+a\cos kz} \{H + a\,A_1 I_0\,(k\,r)\cos\,k\,z + a^2 A_2 I_0\,(2\,k\,r)\cos\,2\,k\,z\}\,r\,d\,r\,. \tag{78}$$

Evaluating N correct to the second order in a, we obtain

$$\begin{aligned} N = \tfrac{1}{2}H\,(R^2 + \tfrac{1}{2}a^2) + \tfrac{1}{2}a^2A_1RI_0\,(kR) + a\Big[HR + \frac{A_1}{k}\,I_1\,(kR)R\Big]\cos\,k\,z \\ + a^2\Big[\tfrac{1}{4}H + \tfrac{1}{2}A_1RI_0\,(kR) + \frac{A_2}{2k}RI_1\,(2kR)\,\Big]\cos\,2\,k\,z\,; \end{aligned} \tag{79}$$

and according to equation (74) this must be identically equal to (cf. eq. [53])

$$\tfrac{1}{2}H_0R_0^2 = \tfrac{1}{2}H_0(R^2 + \tfrac{1}{2}a^2)\,. \tag{80}$$

Hence we must require that

$$\tfrac{1}{2}H(R^2+\tfrac{1}{2}a^2)+\tfrac{1}{2}a^2A_1RI_0(kR)=\tfrac{1}{2}H_0(R^2+\tfrac{1}{2}a^2)\,, \tag{81}$$

$$HR+\frac{A_1}{k}I_1(kR)R=0\,, \tag{82}$$

and

$$\tfrac{1}{4}H+\tfrac{1}{2}A_1RI_0(kR)+\frac{A_2}{2k}RI_1(2kR)=0\,. \tag{83}$$

From equations (82) and (83) we find:

$$A_1=-\frac{H}{R}\frac{x}{I_1(x)} \tag{84}$$

and

$$A_2=\frac{H}{R^2}\frac{x}{I_1(2x)}\left\{\frac{xI_0(x)}{I_1(x)}-\frac{1}{2}\right\}, \tag{85$$

where $x=kR$ (cf. eq. [71]).

With A_1 given by equation (84), equation (81) gives (correct to order a^2)

$$H_0=H\left\{1-\frac{a^2}{R^2}\frac{xI_0(x)}{I_1(x)}\right\}, \tag{86}$$

or, equivalently,

$$H=H_0\left\{1+\frac{a^2}{R^2}\frac{xI_0(x)}{I_1(x)}\right\}. \tag{87}$$

This equation shows that the mean field inside the deformed cylinder is larger than that in the undeformed cylinder; the difference is of order a^2 and depends on the wave number of the deformation.

Equations (76), (77), (84), (85), and (87) determine the field inside the cylinder correct to the second order in a. It may be noted here that the same solution can also be derived from the alternative (but equivalent) condition that the magnetic lines of force follow the boundary of the cylinder (52).

With the field inside the cylinder determined, we can now evaluate the magnetic energy, $\mathfrak{M}$, per unit length. We have

$$\begin{aligned}\mathfrak{M}&=\frac{1}{4}\left\{\int_0^{R+a\cos kz}|\boldsymbol{H}|^2r\,dr\right\}_{\text{av}}\\&\doteq\frac{1}{4}\left\{\int_0^{R+a\cos kz}(H^2+2Hh_z+h_z^2+h_r^2)\,r\,dr\right\}_{\text{av}}\,,\end{aligned} \tag{88}$$

where the averaging is to be done with respect to z. Substituting for h_r and h_z from equations (76) and (77) and evaluating $\mathfrak{M}$ correct to the second order in a, we obtain (cf. eq. [69])

$$\begin{aligned}\mathfrak{M}=\tfrac{1}{8}H^2(R^2+\tfrac{1}{2}a^2)&+\tfrac{1}{2}aH\left\{\cos kz\int_0^{R+a\cos kz}A_1I_0(kr)\,r\,dr\right\}_{\text{av}}\\&+\tfrac{1}{2}a^2HA_2\left\{\cos 2kz\int_0^{R+a\cos kz}I_0(2kr)\,r\,dr\right\}_{\text{av}}\\&+\tfrac{1}{8}a^2A_1^2\int_0^R[I_0^2(kr)+I_1^2(kr)]\,r\,dr\\&=\tfrac{1}{8}H^2(R^2+\tfrac{1}{2}a^2)+\tfrac{1}{4}a^2HA_1RI_0(kR)+\tfrac{1}{8}a^2A_1^2\frac{R}{k}I_0(kR)\,I_1(kR)\,.\end{aligned} \tag{89}$$

On making further use of equations (53), (84), and (87), we can reduce this last expression for $\mathfrak{M}$ to the form

$$\mathfrak{M} = \tfrac{1}{8} H_0^2 R_0^2 + \tfrac{1}{8} a^2 H^2 \frac{x I_0(x)}{I_1(x)}. \tag{90}$$

But the magnetic energy per unit length of the undeformed cylinder is $\frac{1}{8} H_0^2 R_0^2$. Hence

$$\Delta \mathfrak{M} = \tfrac{1}{8} a^2 H^2 \frac{x I_0(x)}{I_1(x)}. \tag{91}$$

10. *The modes of deformation which are unstable.*—Combining the results of §§ 8 and 9, we have

$$\Delta \Omega + \Delta \mathfrak{M} = \left\{ 2\pi^2 \rho^2 G R^2 \left[\tfrac{1}{2} - I_0(x) K_0(x) \right] + \tfrac{1}{8} H^2 \frac{x I_0(x)}{I_1(x)} \right\} a^2. \tag{92}$$

Letting

$$H_s^2 = 16\pi^2 \rho^2 R^2 G \qquad \text{or} \qquad H_s = 4\pi\rho R \sqrt{G}, \tag{93}$$

we can rewrite equation (92) more conveniently in the form

$$\Delta \Omega + \Delta \mathfrak{M} = 2\pi^2 \rho^2 R^2 G \left\{ \left[\tfrac{1}{2} - I_0(x) K_0(x) \right] + \frac{x I_0(x)}{I_1(x)} \left(\frac{H}{H_s} \right)^2 \right\} a^2. \tag{94}$$

Whether the mode of deformation considered is stable or unstable will depend upon the sign of the quantity in braces in the foregoing expression.

Now the asymptotic behaviors of the Bessel functions which appear in equation (94) are:

$$I_0(x) \to 1, \quad I_1(x) \to \tfrac{1}{2} x, \quad \text{and} \quad K_0(x) \to -(\gamma + \log \tfrac{1}{2} x) \quad (x \to 0), \tag{95}$$

where γ (not to be confused with the ratio of the specific heats) is Euler's constant 0.5772 . . . , and

$$I_0(x) \to \frac{e^x}{(2\pi x)^{1/2}}, \quad I_1(x) \to \frac{e^x}{(2\pi x)^{1/2}}, \quad \text{and} \quad K_0(x) \to \left(\frac{\pi}{2x} \right)^{1/2} e^{-x} \quad (x \to \infty). \tag{96}$$

Hence $\Delta \Omega$ (cf. eq. [72]) tends to minus infinity logarithmically as $x \to 0$ and tends monotonically to the positive limit $\pi^2 \rho^2 R^2 G$ as $x \to \infty$, while $\Delta \mathfrak{M}$ (cf. eq. [91]) tends to the positive limit $\frac{1}{4} a^2 H^2$ as $x \to 0$ and increases monotonically to infinity (linearly) as $x \to \infty$. These behaviors of $\Delta \Omega$ and $\Delta \mathfrak{M}$ are illustrated in Figure 1, in which the functions $[\frac{1}{2} - I_0(x) K_0(x)]$ and $x I_0(x)/I_1(x)$ are plotted.

From the asymptotic behaviors of $\Delta \Omega$ and $\Delta \mathfrak{M}$ it follows that the equation

$$\tfrac{1}{2} - I_0(x) K_0(x) + \frac{x I_0(x)}{I_1(x)} \left(\frac{H}{H_s} \right)^2 = 0 \tag{97}$$

allows a single positive root. Let $x = x_*$ denote this root. Then

$$\Delta \Omega + \Delta \mathfrak{M} > 0 \qquad \text{for} \qquad x > x_*, \tag{98}$$

and

$$\Delta \Omega + \Delta \mathfrak{M} < 0 \qquad \text{for} \qquad x < x_*. \tag{98a}$$

Hence *all modes of deformation with $x < x_*$ are unstable.* Since $x = kR$, x_* specifies the minimum wave number (in units of $1/R$) for a stable deformation; alternatively, we could also say that *all modes of deformation with wave lengths exceeding*

$$\lambda_* = \frac{2\pi R}{x_*} \tag{99}$$

are unstable.

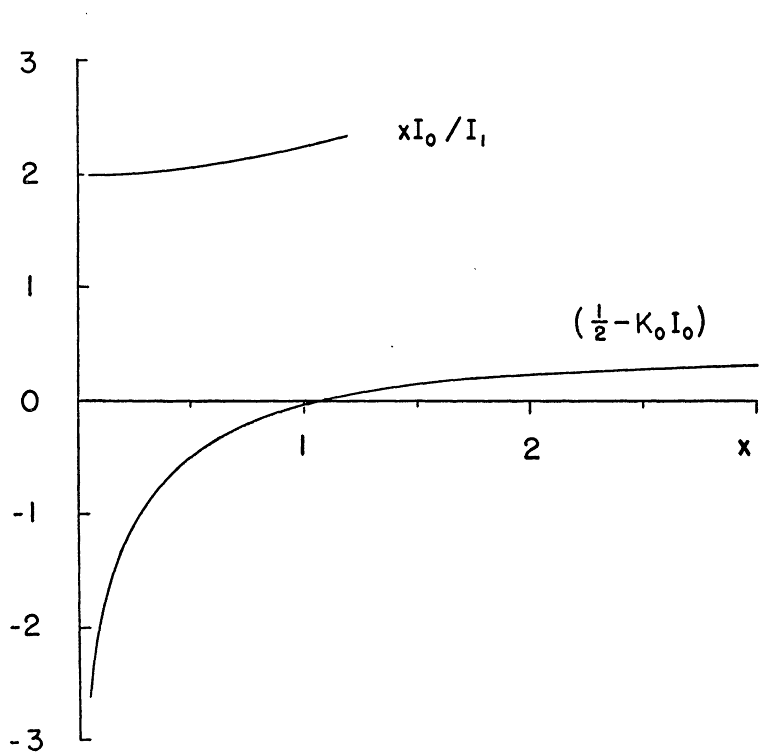


FIG. 1.—The dependence of the changes in the potential energy, $\Delta\Omega$, and magnetic energy, $\Delta\mathfrak{M}$, per unit length of an infinitely long cylinder on the wave number of the deformation; $\Delta\Omega$ is proportional to $[\frac{1}{2} - I_0(x)K_0(x)]$, while $\Delta\mathfrak{M}$ is proportional to $xI_0(x)/I_1(x)$, where x is the wave number measured in the unit $1/R$.

TABLE 1

DEPENDENCE OF WAVE NUMBERS x_* AND x_m AT WHICH INSTABILITY FIRST SETS IN AND AT WHICH IT IS MAXIMUM, ON PREVAILING MAGNETIC FIELD

H/H_s	x_*	x_m	$q_m/(4\pi G\rho)^{1/2}$
0.........	1.067	0.58	0.246
0.25.........	0.832	.47	.208
0.50.........	0.480	.28	.133
0.75.........	0.232	.14	.0685
1.00.........	0.092	.057	.0281
1.25.........	0.0299	.0182	.0091
1.50.........	0.00757	.00459	.00229
2.00.........	0.000228	0.000139	0.0000693

In Table 1 we have listed x_* for a few values of H/H_s. This table exhibits the strong stabilizing effect of the magnetic field: this is shown in the present connection by the very rapid increase, with increasing H, of the wave length at which instability sets in. In fact, for $H > H_s$ this increase becomes exponential; this can be shown in the following way:

Since $x_* = 0.092$ already for $H = H_s$, for $H > H_s$ we may replace the Bessel functions which occur in equation (97) by their dominant terms for $x \rightarrow 0$; thus,

$$\tfrac{1}{2} + \gamma + \log \tfrac{1}{2} x_* + 2 \left(\frac{H}{H_s}\right)^2 = 0 \qquad (H > H_s) \; . \tag{100}$$

Hence

$$x_* = 2 \exp \left\{ - \left[\gamma + \tfrac{1}{2} + 2 \left(\frac{H}{H_s}\right)^2 \right] \right\} \qquad (H > H_s) \; , \tag{101}$$

or, numerically,

$$x_* = 0.6811 \, e^{-2(H/H_s)^2} \qquad (H > H_s) \, . \tag{102}$$

11. *The mode of maximum instability.*—In the preceding section we have seen that an infinite cylinder is gravitationally unstable for all modes of deformation with wave lengths exceeding a certain critical value. We shall now show that there exists a wave length for which the instability is a maximum. For this purpose we shall suppose that the amplitude, a, of the deformation is a function of time and seek an equation of motion for it.

We have already seen that the potential energy (gravitational plus magnetic) per unit length of the cylinder measured from the equilibrium state is

$$\mathfrak{V} = \Delta\mathfrak{M} + \Delta\Omega = -2\pi^2\rho^2R^2GF(x)a^2 \, , \tag{103}$$

where

$$F(x) = I_0(x) K_0(x) - \tfrac{1}{2} - \left(\frac{H}{H_s}\right)^2 \frac{x I_0(x)}{I_1(x)} \, . \tag{104}$$

Defined in this manner, $F(x) > 0$ for $x < x_*$, i.e., it is positive for all unstable modes and negative for all stable modes.

To obtain the Lagrangian function for the cylinder, we must find the kinetic energy of the motion resulting from the varying amplitude. Since we have assumed that the fluid is incompressible, a velocity potential, ψ, exists which satisfies Laplace's equation. And the solution for the velocity potential appropriate to the problem on hand is

$$\psi = BI_0(kr) \cos kz \, , \tag{105}$$

where B is a constant to be determined. The components of the velocity derived from the foregoing potential are

$$u_r = \frac{\partial \psi}{\partial r} = + BkI_1(kr) \cos kz \tag{106}$$

and

$$u_z = \frac{\partial \psi}{\partial z} = - BkI_0(kr) \sin kz \, . \tag{106a}$$

The constant of proportionality, B, in the foregoing equations must be determined from the condition that the radial velocity, u_r, at $r = R$ must agree with that implied by equation (52); i.e., we should have

$$BkI_1(kR) \cos kz = \frac{da}{dt} \cos kz \, . \tag{107}$$

Hence

$$B = \frac{1}{kI_1(x)} \frac{da}{dt} \, . \tag{108}$$

From equations (106) we obtain, for the kinetic energy per unit length, the expression (cf. eq. [69])

$$\begin{aligned}\mathfrak{T} &= \tfrac{1}{2}\pi\rho B^2k^2\int_0^R [I_0^2(kr)+I_1^2(kr)]\,r\,dr \\ &= \tfrac{1}{2}\pi\rho B^2k^2\frac{R}{k}\,I_0(x)\,I_1(x)\ ,\end{aligned} \tag{109}$$

or, substituting for B from equation (108), we have

$$\mathfrak{T} = \tfrac{1}{2}\pi\rho R^2\,\frac{I_0(x)}{xI_1(x)}\left(\frac{da}{dt}\right)^2 . \tag{110}$$

The Lagrangian function (per unit length) for the infinite cylinder is therefore given by

$$\mathfrak{L} = \mathfrak{T} - \mathfrak{V} = \tfrac{1}{2}\pi\rho R^2\,\frac{I_0(x)}{xI_1(x)}\left(\frac{da}{dt}\right)^2 + 2\pi^2\rho^2R^2GF(x)\,a^2 . \tag{111}$$

The equation of motion for a derived from the Lagrangian (111) is

$$\pi\rho R^2\,\frac{I_0(x)}{xI_1(x)}\,\frac{d^2a}{dt^2} = 4\pi^2\rho^2R^2GF(x)\,a\ , \tag{112}$$

or, alternatively,

$$\frac{d^2a}{dt^2} = 4\pi G\rho\left\{\frac{xI_1(x)}{I_0(x)}\,[I_0(x)\,K_0(x)-\tfrac{1}{2}] - \left(\frac{H}{H_s}\right)^2x^2\right\}a\ , \tag{113}$$

where we have substituted for $F(x)$ in accordance with equation (104). The solution for a is therefore of the form

$$a = \text{Constant } e^{\pm qt}\ , \tag{114}$$

where

$$q^2 = 4\pi G\rho\left\{\frac{xI_1(x)}{I_0(x)}\,[I_0(x)\,K_0(x)-\tfrac{1}{2}] - \left(\frac{H}{H_s}\right)^2x^2\right\} . \tag{115}$$

Accordingly, q is purely imaginary for $x > x_*$ and is real for $x < x_*$; this is in agreement with the fact that all modes with $x > x_*$ are stable, while all modes with $x < x_*$ are unstable.

As defined by equation (115), $q = 0$ both for $x = 0$ and for $x = x_*$. There is, therefore, a determinate intermediate value of x—say, x_m—for which q attains a maximum—say, q_m. *The wave number x_m clearly represents the mode of maximum instability;* for it is the mode for which the amplitude of the deformation increases most rapidly. The wave length

$$\lambda_m = \frac{2\pi R}{x_m}\ , \tag{116}$$

corresponding to the wave number x_m, gives approximately the length of the "pieces" into which the cylinder will ultimately break up: for the component with the wave length λ_m, in the Fourier analysis of an arbitrary perturbation, is the one whose amplitude will increase most rapidly with time and, therefore, represents the mode in which the instability will first assert itself. Finally, it is clear that $1/q_m$ gives a measure of the time needed for the instability to make itself manifest.

In Table 1 the values of x_m and $q_m/(4\pi G\rho)^{1/2}$ are also listed. As in the case of x_* (§ 10), we can give explicit formulae for x_m and q_m for $H > H_s$. Since for $H > H_s$ we are concerned only with values of $x \ll 1$, we may replace the Bessel functions which occur in the expression for q^2 by their dominant terms for $x \rightarrow 0$. Thus

$$q^2 = 4\pi G\rho \left\{ -\tfrac{1}{2}x^2(\gamma + \tfrac{1}{2} + \log \tfrac{1}{2}x) - x^2 \left(\frac{H}{H_s}\right)^2 \right\} \qquad (H > H_s) . \quad (117)$$

The expression on the right-hand side attains its maximum when

$$(\gamma + \tfrac{1}{2} + \log \tfrac{1}{2}x) + \tfrac{1}{2} + 2\left(\frac{H}{H_s}\right)^2 = 0 . \quad (118)$$

Hence

$$x_m = 2 \exp \left\{ -(\gamma + 1) - 2\left(\frac{H}{H_s}\right)^2 \right\} = 0.4131\, e^{-2(H/H_s)^2} \qquad (H > H_s) . \quad (119)$$

The corresponding expression for q_m is

$$q_m = \tfrac{1}{2}x_m(4\pi G\rho)^{1/2} \qquad (H > H_s) . \quad (120)$$

These formulae emphasize the fact, apparent from an examination of Table 1, that, as the strength of the magnetic field increases, not only does the wave length of the mode of maximum instability increase exponentially, but the time needed for the instability to manifest itself also increases exponentially.

TABLE 2

WAVE LENGTHS λ_* AND λ_m AT WHICH INSTABILITY SETS IN AND AT WHICH IT IS MAXIMUM AND CHARACTERISTIC TIME, q_m^{-1}, NEEDED FOR INSTABILITY TO MANIFEST ITSELF FOR CASE $R = 250$ PARSECS AND $\rho = 2 \times 10^{-24}$ GM/CM^3

H (Gauss)	λ_* (Parsecs)	λ_m (Parsecs)	q_m^{-1} (Years)
0.	1.5×10^3	2.7×10^3	1.0×10^8
1.25×10^{-6}.	1.9×10^3	3.3×10^3	1.2×10^8
$2.5\ \times10^{-6}$.	3.3×10^3	5.6×10^3	1.8×10^8
3.75×10^{-6}.	6.8×10^3	1.1×10^4	3.6×10^8
$5.0\ \times10^{-6}$.	1.7×10^4	2.8×10^4	8.7×10^8
6.25×10^{-6}.	5.2×10^4	8.6×10^4	2.7×10^9
$7.5\ \times10^{-6}$.	2.1×10^5	3.4×10^5	1.1×10^{10}
$10.0\ \times10^{-6}$.	6.9×10^6	1.1×10^7	3.5×10^{11}

12. *Numerical illustrations.*—To illustrate the theory developed in the preceding sections we shall take, as typical of a spiral arm of a galaxy,

$$R = 250 \text{ parsecs} \quad \text{and} \quad \rho = 2 \times 10^{-24} \text{ gm/cm}^3 . \quad (121)$$

The corresponding value of H_s is (cf. eq. [93])

$$H_s = 5.0 \times 10^{-6} \text{ gauss} . \quad (122)$$

For these values of the physical parameters, the nondimensional results given in Table 1 can be converted into astronomical measures; they are given in Table 2. From the values given in this table it follows that between $H = H_s$ and $H = 2H_s$ the characteristic time of the instability becomes so long that, for all practical purposes, the instability is effectively removed by the presence of the magnetic field.

IV. THE FLATTENING OF A GRAVITATING FLUID SPHERE UNDER THE INFLUENCE OF A MAGNETIC FIELD

13. *The formulation of the problem.*—In this section we shall consider the gravitational equilibrium of an incompressible fluid sphere with a uniform magnetic field inside and a dipole field outside. We shall show that under these circumstances the sphere is not a configuration of equilibrium and that it will become oblate by contracting along the axis of symmetry.

We suppose, then, that initially the magnetic field in the interior of the sphere is uniform and of intensity $\boldsymbol{H}$ in the z-direction. In spherical polar co-ordinates (r, θ, φ) the components of $\boldsymbol{H}$ in the radial (r) and the transverse (θ) directions are

$$H_r^{(i)} = H\mu \quad \text{and} \quad H_\theta^{(i)} = -H(1-\mu^2)^{1/2} \qquad (r < R)\,, \tag{123}$$

where $\mu = \cos\theta$ and the superscript i indicates that these are the components of the field *inside* the sphere.

When the field inside the sphere is uniform, that outside the sphere must be a dipole field given by

$$H_r^{(e)} = H\left(\frac{R}{r}\right)^3 \mu \quad \text{and} \quad H_\theta^{(e)} = \tfrac{1}{2}H\left(\frac{R}{r}\right)^3 (1-\mu^2)^{1/2}\,, \tag{124}$$

where R denotes the radius of the sphere.

The energy, $\mathfrak{M}$, of the magnetic field specified by equations (123) and (124) is given by

$$\begin{aligned}\mathfrak{M} &= \frac{H^2}{8\pi}\left(\tfrac{4}{3}\pi R^3\right) + \tfrac{1}{4}H^2\int_R^\infty\int_{-1}^{+1}\left(\frac{R}{r}\right)^6\left\{\mu^2 + \tfrac{1}{4}(1-\mu^2)\right\}r^2 dr\, d\mu \\ &= \tfrac{1}{4}H^2R^3\,.\end{aligned} \tag{125}$$

Let the sphere be now deformed in such a way that the equation of the bounding surface is

$$r(\mu) = R + \epsilon P_l(\mu)\,, \tag{126}$$

where $\epsilon \ll R$, $\mu = \cos\theta$ (θ being the polar angle), and $P_l(\mu)$ denotes, as usual, the Legendre polynomial of order l. We shall call such a deformation of the sphere a "P_l-deformation." We shall investigate the stability of the sphere by examining whether or not it is stable to a P_l-deformation.

14. *The change in the magnetic energy of the sphere due to a P_l-deformation.*—As we have already pointed out in § 8, an arbitrary deformation of an incompressible body can be thought of as the result of applying to each point of the body a displacement $\boldsymbol{\xi}$. And if, as in § 8 (eqs. [59] and [60]), we express $\boldsymbol{\xi}$ as the gradient of a scalar function, ψ, the solution of Laplace's equation satisfied by ψ appropriate to a P_l-deformation of a sphere is

$$\psi = A r^l P_l(\mu)\,, \tag{127}$$

where A is a constant. The corresponding expressions for the components of $\boldsymbol{\xi}$ are

$$\xi_r = \frac{\partial\psi}{\partial r} = Alr^{l-1}P_l(\mu) \tag{128}$$

and

$$\xi_\theta = \frac{1}{r}\frac{\partial\psi}{\partial\theta} = -Ar^{l-1}(1-\mu^2)^{1/2}P_l'(\mu)\,, \tag{128a}$$

where a prime is used to denote differentiation with respect to μ. According to equation (126), at $r = R$, $\xi_r = \epsilon P_l(\mu)$; this determines A, and we have

$$\xi_r = \epsilon \left(\frac{r}{R}\right)^{l-1} P_l(\mu) \text{ and } \xi_\theta = -\frac{\epsilon}{l}\left(\frac{r}{R}\right)^{l-1} (1-\mu^2)^{1/2} P_l'(\mu) \; . \tag{129}$$

Now the deformation of a body will alter the prevailing magnetic field; and, since in a medium of infinite electrical conductivity a change in the existing magnetic field can be effected only by bodily pushing aside the lines of force, it follows that

$$\delta \boldsymbol{H} = \text{curl}\,(\boldsymbol{\xi} \times \boldsymbol{H}) \; . \tag{130}$$

[The truth of this last relation can be established in the following way: Suppose that the displacement $\boldsymbol{\xi}$ takes place as a slow continuous movement so that if $\boldsymbol{u}$ denotes the velocity, $\boldsymbol{u} = \partial\boldsymbol{\xi}/\partial t$ (i.e., if quantities of the second order of smallness are neglected). On the other hand, when the electrical conductivity is infinite,

$$\delta \boldsymbol{E} = -\boldsymbol{u} \times \boldsymbol{H} \, ,$$

where $\delta\boldsymbol{E}$ is the electrical field resulting from the changing magnetic field $\delta\boldsymbol{H}$ in accordance with Maxwell's equation,

$$\text{curl}\,\delta\boldsymbol{E} = -\frac{\partial}{\partial t}\,\delta\boldsymbol{H} \, .$$

Combining the last two equations, we have

$$\text{curl}\left(\frac{\partial\boldsymbol{\xi}}{\partial t} \times \boldsymbol{H}\right) = \frac{\partial}{\partial t}(\delta\boldsymbol{H}) \; .$$

The relation (130) is simply the integrated form of this equation.]

When the fluid is incompressible (i.e., when div $\boldsymbol{\xi} = 0$ in addition to div $\boldsymbol{H} = 0$), equation (130) can be written alternatively in the form

$$\delta\boldsymbol{H} = (\boldsymbol{H}\cdot\text{grad})\boldsymbol{\xi} - (\boldsymbol{\xi}\cdot\text{grad})\boldsymbol{H} \, . \tag{131}$$

And when the initial field is homogeneous, equation (131) simplifies still further to

$$\delta\boldsymbol{H} = (\boldsymbol{H}\cdot\text{grad})\boldsymbol{\xi} \, . \tag{132}$$

In spherical polar co-ordinates the foregoing equation is equivalent to

$$\delta H_r = \left(H_r \frac{\partial}{\partial r} + \frac{H_\theta}{r}\frac{\partial}{\partial \theta}\right)\xi_r - \frac{H_\theta \xi_\theta}{r} \tag{133}$$

and

$$\delta H_\theta = \left(H_r \frac{\partial}{\partial r} + \frac{H_\theta}{r}\frac{\partial}{\partial \theta}\right)\xi_\theta + \frac{H_\theta \xi_r}{r} \, . \tag{133a}$$

These equations in conjunction with equations (123) and (129) give

$$\delta H_r^{(i)} = \epsilon H\,(l-1)\frac{r^{l-2}}{R^{l-1}} P_{l-1}(\mu) \tag{134}$$

and

$$\delta H_\theta^{(i)} = -\epsilon H \frac{r^{l-2}}{R^{l-1}} (1-\mu^2)^{1/2} P'_{l-1}(\mu) \; . \tag{134a}$$

The corresponding change in the internal magnetic energy density is given by

$$\delta\left(\frac{|\boldsymbol{H}|^2}{8\pi}\right)=\frac{1}{4\pi}\,\boldsymbol{H}^{(i)}\cdot\delta\boldsymbol{H}^{(i)}$$
$$=\epsilon\,\frac{H^2}{4\pi}\frac{r^{\,l-2}}{R^{l-1}}\{\,(l-1)\,\mu P_{l-1}(\mu)+(1-\mu^2)P'_{l-1}(\mu)\,\}\,. \tag{135}$$

On further simplification this reduces to

$$\delta\left(\frac{|\boldsymbol{H}|^2}{8\pi}\right)=\epsilon\,(l-1)\,\frac{H^2}{4\pi}\frac{r^{\,l-2}}{R^{l-1}}P_{l-2}(\mu)\;. \tag{136}$$

Hence, when averaged over all directions, this is zero except when $l = 2$, in which case

$$\delta\left(\frac{|\boldsymbol{H}|^2}{8\pi}\right)=\frac{\epsilon}{4\pi}\frac{H^2}{R}\qquad\qquad(l=2)\;; \tag{137}$$

the corresponding change in the internal magnetic energy, $\Delta\mathfrak{M}^{(i)}$, is given by

$$\Delta\mathfrak{M}^{(i)}=\tfrac{1}{3}\epsilon H^2R^2\,. \tag{138}$$

15. *The change in the external magnetic energy of the sphere due to a P_l-deformation.*— Writing the magnetic field outside the deformed sphere in the form

$$H_r^{(e)}=H\left(\frac{R}{r}\right)^3\mu+\delta H_r^{(e)} \tag{139}$$

and

$$H_\theta^{(e)}=\tfrac{1}{2}H\left(\frac{R}{r}\right)^3(1-\mu^2)^{1/2}+\delta H_\theta^{(e)}\,, \tag{139a}$$

we shall suppose that $\delta H_r^{(e)}$ and $\delta H_\theta^{(e)}$ are derivable from a magnetic potential $\delta\phi^{(e)}$. Since the magnetic potential satisfies Laplace's equation, the solution for $\delta\phi^{(e)}$ must be expressible as a linear combination of the fundamental solutions $P_j(\mu)/r^{j+1}$, which vanish at infinity.

We shall find it convenient to write the solution for $\delta\phi^{(e)}$ in the form

$$\delta\phi^{(e)}=-\epsilon H\left\{\frac{l-1}{l}\left(\frac{R}{r}\right)^lP_{l-1}(\mu)+\Sigma A_j\left(\frac{R}{r}\right)^{j+1}P_j(\mu)\right\}, \tag{140}$$

where the A_j's are coefficients to be determined. The expressions for $\delta H_r^{(e)}$ and $\delta H_\theta^{(e)}$ derived from this potential are

$$\delta H_r^{(e)}=\epsilon H\left\{(l-1)\frac{R^l}{r^{\,l+1}}P_{l-1}(\mu)+\Sigma A_j(j+1)\frac{R^{j+1}}{r^{j+2}}P_j(\mu)\right\} \tag{141}$$

and

$$\delta H_\theta^{(e)}=\epsilon H\left\{\frac{l-1}{l}\frac{R^l}{r^{\,l+1}}P^1_{l-1}(\mu)+\Sigma A_j\frac{R^{j+1}}{r^{j+2}}P^1_j(\mu)\right\}. \tag{141a}$$

The coefficients A_j in equations (141) and (141a) can be determined from the condition that the component of the magnetic field normal to a bounding surface must be continuous. To the first order in ϵ this condition requires that

$$\{H_r^{(e)}\}_{R+\epsilon P_l}+\{H_\theta^{(e)}\}_R\,\frac{\epsilon}{R}(1-\mu^2)^{1/2}\frac{\partial P_l}{\partial\mu}$$
$$=\{H_r^{(i)}\}_{R+\epsilon P_l}+\{H_\theta^{(i)}\}_R\,\frac{\epsilon}{R}(1-\mu^2)^{1/2}\frac{\partial P_l}{\partial\mu}\,, \tag{142}$$

where $-(\epsilon/R)(1-\mu^2)^{1/2}\partial P_l/\partial\mu$ is the angle (to the first order in ϵ) which the deformed boundary makes with the θ-direction; the terms in H_θ in the foregoing equation arise from this latter circumstance. Now, according to equations (124), (139), and (140),

$$\begin{aligned}\{H_r^{(e)}\}_{R+\epsilon P_l}+\{H_\theta^{(e)}\}_R\,\frac{\epsilon}{R}(1-\mu^2)^{1/2}\frac{\partial P_l}{\partial\mu} &= H\mu\left(1-3\frac{\epsilon}{R}P_l\right)\\ &+\tfrac{1}{2}H\frac{\epsilon}{R}(1-\mu^2)\frac{\partial P_l}{\partial\mu}+\frac{\epsilon}{R}H\{(l-1)P_{l-1}+\Sigma A_j(j+1)P_j\}\,,\end{aligned} \tag{143}$$

while, according to equations (123) and (134),

$$\begin{aligned}\{H_r^{(i)}\}_{R+\epsilon P_l}+\{H_\theta^{(i)}\}_R\,\frac{\epsilon}{R}(1-\mu^2)^{1/2}\frac{\partial P_l}{\partial\mu} &= H\mu\\ &+\frac{\epsilon}{R}H(l-1)P_{l-1}-\frac{\epsilon}{R}H(1-\mu^2)\frac{\partial P_l}{\partial\mu}\,;\end{aligned} \tag{143a}$$

and the equality of the expressions on the right-hand sides of equations (143) and (143a) requires

$$\begin{aligned}\Sigma A_j(j+1)P_j &= 3\mu P_l-\tfrac{3}{2}(1-\mu^2)\frac{\partial P_l}{\partial\mu}\\ &=\frac{3}{2(2l+1)}\{(l+1)(l+2)P_{l+1}-l(l-1)P_{l-1}\}\,.\end{aligned} \tag{144}$$

Hence

$$A_{l-1}=-\frac{3(l-1)}{2(2l+1)}\,,\qquad A_{l+1}=\frac{3(l+1)}{2(2l+1)}\,, \tag{145}$$

and

$$A_j=0\qquad\text{for}\qquad j\neq l-1\qquad\text{or}\qquad l+1\,. \qquad\text{(145a}$$

Inserting these values of A in equations (141) and (141a), we obtain

$$\delta H_r^{(e)}=\epsilon H\left\{\frac{(l-1)(l+2)}{2(2l+1)}\frac{R^l}{r^{l+1}}P_{l-1}(\mu)+\frac{3(l+1)(l+2)}{2(2l+1)}\frac{R^{l+2}}{r^{l+3}}P_{l+1}(\mu)\right\} \tag{146}$$

and

$$\delta H_\theta^{(e)}=\epsilon H\left\{\frac{(l-1)(l+2)}{2l(2l+1)}\frac{R^l}{r^{l+1}}P_{l-1}^1(\mu)+\frac{3(l+1)}{2(2l+1)}\frac{R^{l+2}}{r^{l+3}}P_{l+1}^1(\mu)\right\}. \tag{146a}$$

Returning to equations (139) and (139a), we can write the change in the external magnetic energy, $\Delta\mathfrak{M}^{(e)}$, to the first order in ϵ, in the form

$$\begin{aligned}\Delta\mathfrak{M}^{(e)} &=\frac{H^2}{8\pi}\iiint_{R+\epsilon P_l\geq r\geq R}\left(\frac{R}{r}\right)^6\{\mu^2+\tfrac{1}{4}(1-\mu^2)\}\,r^2dr\,d\mu\,d\varphi\\ &+\frac{H}{8\pi}\iiint_{r>R}\left(\frac{R}{r}\right)^3\{2P_1(\mu)\,\delta H_r^{(e)}+P_1^1(\mu)\,\delta H_\theta^{(e)}\}\,r^2dr\,d\mu\,d\varphi\,.\end{aligned} \tag{147}$$

After some minor reductions we find

$$\begin{aligned}\Delta\mathfrak{M}^{(e)} &= \tfrac{1}{4}\epsilon H^2R^2\int_{-1}^{+1}\{\tfrac{1}{2}P_2(\mu)+\tfrac{1}{2}\}P_l(\mu)\,d\mu\\ &+\tfrac{1}{2}\epsilon H^2\int_R^\infty dr\,r^2\int_{-1}^{+1}d\mu\left(\frac{R}{r}\right)^3P_1(\mu)\left\{\frac{(l-1)\,(l+2)}{2\,(2l+1)}\,\frac{R^l}{r^{l+1}}P_{l-1}(\mu)\right.\\ &\qquad\left.+\frac{3\,(l+1)\,(l+2)}{2\,(2l+1)}\,\frac{R^{l+2}}{r^{l+3}}P_{l+1}(\mu)\right\} \qquad (148)\\ &+\tfrac{1}{4}\epsilon H^2\int_R^\infty dr\,r^2\int_{-1}^{+1}d\mu\left(\frac{R}{r}\right)^3P_1^1(\mu)\left\{\frac{(l-1)\,(l+2)}{2l\,(2l+1)}\,\frac{R^l}{r^{l+1}}P_{l-1}^1(\mu)\right.\\ &\qquad\left.+\frac{3\,(l+1)}{2\,(2l+1)}\,\frac{R^{l+2}}{r^{l+3}}P_{l+1}^1(\mu)\right\}.\end{aligned}$$

From this equation it is evident that $\Delta\mathfrak{M}^{(e)}$ vanishes (to the first order in ϵ) for all deformations except a P_2-deformation. And for a P_2-deformation we have

$$\Delta\mathfrak{M}^{(e)} = \tfrac{1}{8}\epsilon H^2R^2\int_{-1}^{+1}[P_2(\mu)]^2d\mu+\tfrac{1}{5}\epsilon H^2R^5\int_R^\infty\int_{-1}^{+1}\frac{dr}{r^4}\{\tfrac{1}{2}P_2(\mu)+\tfrac{1}{2}\}\,d\mu \qquad (149)$$

or

$$\Delta\mathfrak{M}^{(e)} = \tfrac{7}{60}\,\epsilon H^2R^2 \qquad (l=2)\,. \qquad (150)$$

Finally, combining equations (138) and (150), we obtain

$$\Delta\mathfrak{M} = \Delta\mathfrak{M}^{(i)}+\Delta\mathfrak{M}^{(e)} = \tfrac{9}{20}\,\epsilon H^2R^2\,, \qquad (151)$$

for the total change in the magnetic energy due to a P_2-deformation; it vanishes to this order for all higher deformations.

We have, therefore, shown that *the change in the magnetic energy is of the second order in ϵ for all deformations of the sphere except a P_2-deformation; and for a P_2-deformation it is of the first order in ϵ and is given by (151)*. Moreover, for a P_2-deformation $\Delta\mathfrak{M} > 0$ when the deformation is in the sense of making the sphere into a prolate spheroid; and $\Delta\mathfrak{M} < 0$ when the deformation is in the sense of making the sphere into an oblate spheroid.

16. *The change in the gravitational potential energy and the instability of the sphere to a P_2-deformation.*—The change in the potential energy, $\Delta\Omega$, due to a P_l-deformation can also be computed. The result is well known for a P_2-deformation. For a general P_l-deformation we can evaluate $\Delta\Omega$ by following the procedure used in § 8. We shall not give here the details of the calculations, which lead to the result

$$\Delta\Omega = \frac{3\,(l-1)}{(2l+1)^2}\left(\frac{\epsilon}{R}\right)^2\frac{GM^2}{R}\,. \qquad (152)$$

The change in the potential energy is therefore always positive and is of the second order in ϵ. This is in contrast to $\Delta\mathfrak{M}$, which, as we have seen, is of the first order in ϵ for a P_2-deformation and is negative for a deformation which tends to make it oblate. We can therefore conclude that the sphere is unstable and that it will tend to collapse toward an oblate spheroidal shape. To estimate the extent to which this collapse may proceed, let us consider $\Delta\Omega + \Delta\mathfrak{M}$ for a P_2-deformation. We have (cf. eqs. [151] and [152])

$$\Delta\Omega+\Delta\mathfrak{M} = \frac{3}{25}\,\frac{GM^2}{R^3}\,\epsilon^2+\tfrac{9}{20}H^2R^2\epsilon \qquad (l=2)\,. \qquad (153)$$

As a function of ϵ, $\Delta\Omega + \Delta\mathfrak{M}$ has a minimum which it takes when

$$\frac{6}{25}\frac{GM^2}{R^3}\epsilon + \tfrac{9}{20}H^2R^2 = 0\,, \tag{154}$$

or

$$\frac{\epsilon}{R} = -\frac{15}{8}\frac{H^2R^4}{GM^2}. \tag{155}$$

If H_* denotes the value of the constant magnetic field inside the sphere for which $\mathfrak{M}$ (given by eq. [125]) is equal to the numerical value of the gravitational potential energy Ω $(= -3GM^2/5R)$, then

$$\tfrac{1}{4}H_*^2R^3 = \frac{3}{5}\frac{GM^2}{R}. \tag{156}$$

In terms of H_* defined in this manner, we can rewrite equation (155) in the form

$$\frac{\epsilon}{R} = -4.5\left(\frac{H}{H_*}\right)^2. \tag{157}$$

We may interpret this relation by saying that when a star has a magnetic field approaching the limit set by the virial theorem (cf. Sec. I), then it tends to become highly oblate; in this respect the magnetic field has the same effect as a rotation.

V. THE GRAVITATIONAL INSTABILITY OF AN INFINITE HOMOGENEOUS MEDIUM IN THE PRESENCE OF A MAGNETIC FIELD

17. *The statement of the problem.*—It is well known that, by considering the propagation of a wave in an infinite homogeneous medium and allowing for the gravitational effects of the density fluctuations, Jeans[6] showed that the velocity of wave propagation is given by

$$V_J = \sqrt{(c^2 - 4\pi G\rho/k^2)}\,, \tag{158}$$

where $c = \sqrt{(\gamma p/\rho)}$ denotes the convectional velocity of sound and k is the wave number. Accordingly, when

$$k < c(4\pi\rho G)^{-1/2}\,, \tag{159}$$

the velocity of wave propagation becomes imaginary; and under these circumstances the amplitude of the wave will increase exponentially with time. The inequality (159) is therefore the condition for gravitational instability; this is Jeans's result. In Section V we shall show that Jeans's condition (159) is unaffected by the presence of a magnetic field. The physical reason for this is evident for a deformation in which the density waves are perpendicular to the lines of force because the motion of the particles in this case will be parallel to the lines of force and therefore will not be impeded by the magnetic field. But also a density wave forming an angle with the lines of force may be obtained by particle motions parallel to the lines of force, as shown in Figure 2.

18. *The three modes of wave propagation in the presence of a magnetic field and the condition for gravitational instability.*—Consider an extended homogeneous gaseous medium of infinite electrical conductivity, and suppose that there is present a uniform magnetic

[6] *Astronomy and Cosmogony* (Cambridge: At the University Press, 1929), pp. 345–347.

field of intensity $\boldsymbol{H}$. Then the fluctuations in density ($\delta\rho$), pressure (δp), magnetic field ($\boldsymbol{h}$), and gravitational potential (δV) are governed by the equations

$$\rho\frac{\partial \boldsymbol{u}}{\partial t}=\frac{1}{4\pi}(\text{curl }\boldsymbol{h}\times\boldsymbol{H})-\text{grad }\delta p+\rho\text{ grad }\delta V,$$

$$\frac{\partial \boldsymbol{h}}{\partial t}=\text{curl }(\boldsymbol{u}\times\boldsymbol{H}), \tag{160}$$

$$\frac{\partial}{\partial t}\delta\rho=-\rho\text{ div }\boldsymbol{u},$$

and

$$\nabla^2\delta V=-4\pi G\delta\rho.$$

If the changes in pressure and density are assumed to take place adiabatically, then

$$\delta p=c^2\delta\rho. \tag{161}$$

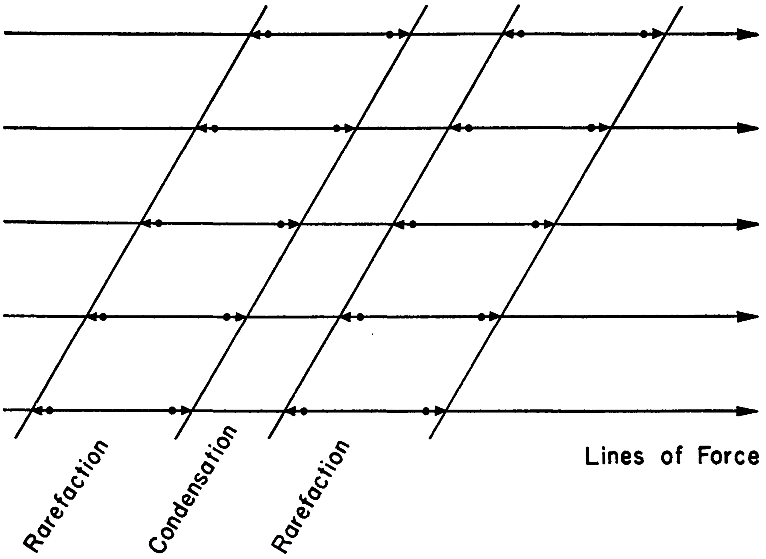


FIG. 2.—Illustrating why the presence of a magnetic field does not affect Jeans's condition for the gravitational instability of an infinite homogeneous medium.

We shall seek the solutions of equations (160) and (161) which correspond to the propagation of waves in the z-direction. Then $\partial/\partial z$ is the only nonvanishing component of the gradient. And if we further suppose that the orientation of the co-ordinate axes is so chosen that

$$\boldsymbol{H}=(0,H_y,H_z), \tag{162}$$

it readily follows that $h_z=0$; and we find that equations (160) and (161) break up into the two noncombining systems:

$$\rho\frac{\partial u_x}{\partial t}=\frac{H_z}{4\pi}\frac{\partial h_x}{\partial z},\qquad \frac{\partial h_x}{\partial t}=H_z\frac{\partial u_x}{\partial z}; \tag{163}$$

and

$$
\begin{aligned}
&\rho \frac{\partial u_y}{\partial t} - \frac{H_z}{4\pi} \frac{\partial h_y}{\partial z} = 0 \,, \\
&\rho \frac{\partial u_z}{\partial t} + \frac{H_y}{4\pi} \frac{\partial h_y}{\partial z} + c^2 \frac{\partial}{\partial z}\, \delta\rho - \rho \frac{\partial}{\partial z}\, \delta V = 0 \,, \\
&\frac{\partial h_y}{\partial t} + H_y \frac{\partial u_z}{\partial z} - H_z \frac{\partial u_y}{\partial z} = 0 \,, \\
&\frac{\partial}{\partial t}\, \delta\rho + \rho \frac{\partial u_z}{\partial z} = 0 \,, \\
&\frac{\partial^2}{\partial z^2}\, \delta V + 4\pi G \delta\rho = 0 \,.
\end{aligned}
\tag{164}
$$

Equations (163) can be combined to give

$$
\frac{\partial^2 h_x}{\partial t^2} = \frac{H_z^2}{4\pi\rho} \frac{\partial^2 h_x}{\partial z^2} \qquad \text{and} \qquad \frac{\partial^2 u_x}{\partial t^2} = \frac{H_z^2}{4\pi\rho} \frac{\partial^2 u_x}{\partial t^2} \,. \tag{165}
$$

These equations are the same as those leading to the ordinary hydromagnetic waves of Alfvén propagated with the velocity

$$
V_A = \frac{H_z}{\sqrt{(4\pi\rho)}} \,. \tag{166}
$$

This mode of wave propagation is therefore unaffected by gravitation and compressibility.

Turning next to solutions of equations (164), which also represent the propagation of waves in the z-direction, we can write

$$
\frac{\partial}{\partial t} = i\omega \qquad \text{and} \qquad \frac{\partial}{\partial z} = -ik \,, \tag{167}
$$

where ω denotes the frequency and k the wave number. Making the substitutions (167) in equations (164), we obtain a system of linear homogeneous equations which can be written in matrix notation in the following form:

$$
\begin{vmatrix}
\rho\omega & k\dfrac{H_z}{4\pi} & 0 & 0 & 0 \\
0 & -k\dfrac{H_y}{4\pi} & \rho\omega & -kc^2 & k\rho \\
kH_z & \omega & -kH_y & 0 & 0 \\
0 & 0 & -k\rho & \omega & 0 \\
0 & 0 & 0 & 4\pi G & -k^2
\end{vmatrix}
\begin{vmatrix}
u_y \\
h_y \\
u_z \\
\delta\rho \\
\delta V
\end{vmatrix}
= 0 \,. \tag{168}
$$

The condition that equation (168) has a nontrivial solution is that the determinant of the matrix on the left-hand side should vanish. Expanding the determinant, we find that it can be reduced to the form

$$\left(\frac{\omega}{k}\right)^4-\left\{\frac{H^2}{4\pi\rho}+\left(c^2-\frac{4\pi G\rho}{k^2}\right)\right\}\left(\frac{\omega}{k}\right)^2+\frac{H_z^2}{4\pi\rho}\left(c^2-\frac{4\pi G\rho}{k^2}\right)=0\,. \tag{169}$$

In terms of the velocity of wave propagation, $V = \omega/k$, we can rewrite equation (169) in the form

$$V^4-(V_A^2\sec^2\vartheta+V_J^2)V^2+V_A^2V_J^2=0\,, \tag{170}$$

where ϑ denotes the angle between the directions of $\boldsymbol{H}$ and of wave propagation and V_J and V_A have the same meanings as in equations (158) and (166).

It is seen that equation (170) allows two modes of wave propagation. If V_1 and V_2 denote the velocities of propagation of these two modes, we conclude from equation (170) that

$$V_1V_2 = V_AV_J$$

and

$$V_1^2+V_2^2=V_A^2\sec^2\vartheta+V_J^2\,. \tag{171}$$

Accordingly, *if V_J is imaginary, then either V_1 or V_2 must be imaginary*. In other words, one of the two modes of wave propagation will be unstable if Jeans's condition (159) is satisfied. The condition for gravitational instability is therefore unaffected by the presence of the magnetic field. However, as to which of the two modes will become unstable will depend on the strength of the magnetic field. Thus for $H\rightarrow 0$, the two modes given by equation (170) approach, respectively, Jeans's mode and Alfvén's mode. And if we suppose that

$$V_1\rightarrow V_J \quad\text{and}\quad V_2\rightarrow V_A \quad\text{as}\quad H\rightarrow 0\,, \tag{172}$$

then it follows from equation (170) that so long as $V_J^2 > 0$,

$$V_1\rightarrow V_A\sec\vartheta \quad\text{and}\quad V_2\rightarrow V_J\cos\vartheta \quad\text{as}\quad H\rightarrow\infty\,. \tag{173}$$

Hence, for $H\rightarrow\infty$, the mode which will become unstable when Jeans's condition is satisfied will be the mode which for $H\rightarrow 0$ is Alfvén's mode; and the mode which for $H\rightarrow 0$ is Jeans's mode becomes a hydromagnetic wave for $H\rightarrow\infty$ and is unaffected by gravitation. This "crossing-over" of the two modes with increasing strength of the magnetic field is in agreement with what is known[7] from the theory of wave propagation in a compressible medium in the absence of gravitation.

[7] Cf. H. van de Hulst, *Symposium: Problems of Cosmical Aerodynamics* (Dayton, Ohio: Central Air Documents Office, 1951), chap. vi; also N. Herlofson, *Nature*, **165**, 1020, 1950.

Studies of nonlinear problems (266)

Document LA-1940 (May 1955).

N° 266.

After the war, during one of his frequent summer visits to Los Alamos, Fermi became interested in the development and potentialities of the electronic computing machines. He held many discussions with me on the kind of future problems which could be studied through the use of such machines. We decided to try a selection of problems for heuristic work where in absence of closed analytic solutions experimental work on a computing machine would perhaps contribute to the understanding of properties of solutions. This could be particularly fruitful for problems involving the asymptotic—long time or "in the large" behavior of non-linear physical systems. In addition, such experiments on computing machines would have at least the virtue of having the postulates clearly stated. This is not always the case in an actual physical object or model where all the assumptions are not perhaps explicitly recognized.

Fermi expressed often a belief that future fundamental theories in physics may involve non-linear operators and equations, and that it would be useful to attempt practice in the mathematics needed for the understanding of non-linear systems. The plan was then to start with the possibly simplest such physical model and to study the results of the calculation of its long-time behavior. Then one would gradually increase the generality and the complexity of the problem calculated on the machine. The Los Alamos report LA–1940 (paper N° 266) presents the results of the very first such attempt. We had planned the work in the summer of 1952 and performed the calculations the following summer. In the discussions preceding the setting up and running of the problem on the machine we had envisaged as the next problem a two-dimensional version of the first one. Then perhaps problems of pure kinematics e.g., the motion of a chain of points subject only to constraints but no external forces, moving on a smooth plane convoluting and knotting itself indefinitely. These were to be studied preliminary to setting up ultimate models for motions of system where "mixing" and "turbulence" would be observed. The motivation then was to observe the *rates* of mixing and "thermalization" with the hope that the calculational results would provide hints for a future theory. One could venture a guess that one motive in the selection of problems could be traced to Fermi's early interest in the ergodic theory. In fact, his early paper (N° 11 *a*) presents an important contribution to this theory.

It should be stated here that during one summer Fermi learned very rapidly how to *program* problems for the electronic computers and he not only could plan the general outline and construct the so-called flow diagram but would work out himself the actual *coding* of the whole problem in detail.

The results of the calculations (performed on the old MANIAC machine) were interesting and quite surprising to Fermi. He expressed to me the opinion that they really constituted a little discovery in providing intimations that the prevalent beliefs in the universality of "mixing and thermalization" in non-linear systems may not be always justified.

A few words about the subsequent history of this non-linear problem. A number of other examples of such physical systems were examined by calculations on the electronic computing machines in 1956 and 1957. I presented the results of the original paper on several occasions at scientific meetings; they seemed to have aroused considerable interest among mathematicians and physicists and there is by now a small literature dealing with this problem. The most recent results are due to N. J. Zabusky.[1] His analytical work shows, by the way, a good agreement of the numerical computations with the continuous solution up to a point where a discontinuity developed in the derivatives and the analytical work had to be modified. One obtains from it another indication that the phenomenon discovered

[1] Exact Solutions for the Vibrations of a non-linear continuous string. A. E. C. Research and Development Report. MATT–102, Plasma Physics Laboratory, Princeton University, October 1961.

is not due to numerical accidents of the algorithm of the computing machine, but seems to constitute a real property of the dynamical system.

In 1961, on more modern and faster machines, the original problem was considered for still longer periods of time. It was found by J. Tuck and M. Menzel that after one continues the calculations from the first "return" of the system to its original condition the return is not complete. The total energy is concentrated again essentially in the first Fourier mode, but the remaining one or two percent of the total energy is in higher modes. If one continues the calculation, at the end of the next great cycle the error (deviation from the original initial condition) is greater and amounts to perhaps three percent. Continuing again one finds the deviation increasing—after eight great cycles the deviation amounts to some eight percent; but from that time on an opposite development takes place! After eight more i.e., sixteen great cycles altogether, the system gets very close—better than within one percent to the original state! This supercycle constitutes another surprising property of our non-linear system.

Paper N° 266 is not the only work that Fermi and I did together. In the summer of 1950 we made a study of the behavior of the thermonuclear reaction in a mass of deuterium and wrote a report, LA-1158, which is still classified. The problem is of enormous mathematical complexity, involving the hydrodynamics of the motion of the material, the hydrodynamics of radiation energy, all interwoven with the processes of the various reactions between the nuclei whose probabilities and properties depend i.a., on temperature, density, and the changing geometry of the materials. The aim of this work was to obtain, by a schematized but still elaborate picture of the evolution of all these physical processes, an idea of the propagation of such a reaction. This was to complement a previous work by Everett and myself, dealing with the problem of ignition of a mass deuterium. Assuming an ignition somehow started in a large volume, one wanted to evaluate the prospects of propagation of the reactions already started. Many ingenious schematizations and simplifications had to be introduced in order to describe the process, without the possibility of calculating in exact detail the innumerable geometrical and thermodynamical factors. The results of our computations on the chances of propagation were negative and the report played an important role in channeling imagination and energies towards a search for a different scheme for a successful hydrogen reaction. This was indeed found later on on a different basis. All the calculations on which the work of the report is based were performed on desk computers and slide rules. The subsequent massive and lengthy work on the electronic computer machines (organized and performed by von Neumann, F. and C. Evans and others) confirmed in large lines, qualitatively and to a good degree quantitatively the behavior of the system as estimated and predicted in our report—with its combination of intuitive evaluations, schematized equations and hand calculations.

S. M. Ulam.

266.

STUDIES OF NON LINEAR PROBLEMS

E. Fermi, J. Pasta, and S. Ulam

Document LA-1940 (May 1955).

Abstract.

A one-dimensional dynamical system of 64 particles with forces between neighbors containing nonlinear terms has been studied on the Los Alamos computer Maniac I. The nonlinear terms considered are quadratic, cubic, and broken linear types. The results are analyzed into Fourier components and plotted as a function of time.

The results show very little, if any, tendency toward equipartition of energy among the degrees of freedom.

The last few examples were calculated in 1955. After the untimely death of Professor E. Fermi in November, 1954, the calculations were continued in Los Alamos.

This report is intended to be the first one of a series dealing with the behavior of certain nonlinear physical systems where the non-linearity is introduced as a perturbation to a primarily linear problem. The behavior of the systems is to be studied for times which are long compared to the characteristic periods of the corresponding linear problems.

The problems in question do not seem to admit of analytic solutions in closed form, and heuristic work was performed numerically on a fast electronic computing machine (MANIAC I at Los Alamos).[1] The ergodic behavior of such systems was studied with the primary aim of establishing, experimentally, the rate of approach to the equipartition of energy among the various degrees of freedom of the system. Several problems will be considered in order of increasing complexity. This paper is devoted to the first one only.

We imagine a one-dimensional continuum with the ends kept fixed and with forces acting on the elements of this string. In addition to the usual linear term expressing the dependence of the force on the displacement of the element, this force contains higher order terms. For the purposes of numerical work this continuum is replaced by a finite number of points (at most 64 in our actual computation) so that the partial differential equation defining the motion of this string is replaced by a finite number of total differential equations. We have, therefore, a dynamical system of 64 particles with forces acting between neighbors with fixed end points. If x_i denotes the displacement of the i–th point from its original position, and α denotes the coefficient of the quadratic term in the force between the neighboring mass points and β that of the cubic term, the equations were either

$$(1) \qquad x_i = (x_{i+1} + x_{i-1} - 2\,x_i) + \alpha\,[(x_{i+1} - x_i)^2 - (x_i - x_{i-1})^2]$$
$$(i = 1\,, 2\,, \cdots, 64),$$

or

$$(2) \qquad x_i = (x_{i+1} + x_{i-1} - 2\,x_i) + \beta\,[(x_{i+1} - x_i)^3 - (x_i - x_{i-1})^3]$$
$$(i = 1\,, 2\,, \cdots, 64).$$

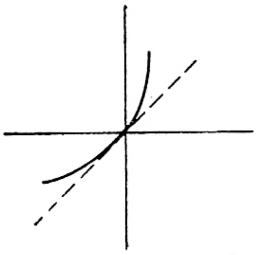
α and β were chosen so that at the maximum displacement the nonlinear term was small, e.g., of the order of one-tenth of the linear term. The corresponding partial differential equation obtained by letting the number of particles become infinite is the usual wave equation plus non-linear terms of a complicated nature.

Another case studied recently was

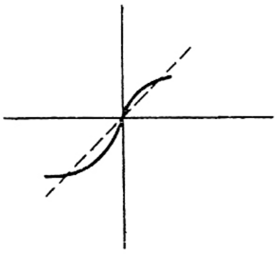
$$(3) \qquad \ddot{x}_i = \delta_1\,(x_{i+1} - x_i) - \delta_2\,(x_i - x_{i-1}) + c$$

[1] We thank Miss Mary Tsingou for efficient coding of the problems and for running the computations on the Los Alamos MANIAC machine.

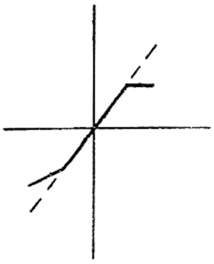
where the parameters δ_1, δ_2, c were not constant but assumed different values depending on whether or not the quantities in parentheses were less than or greater than a certain value fixed in advance. This prescription amounts to assuming the force as a broken linear function of the displacement. This broken linear function imitates to some extent a cubic dependence. We show the graphs representing the force as a function of displacement in three cases.



Quadratic



Cubic



Broken Linear

The solution to the corresponding linear problem is a periodic vibration of the string. If the initial position of the string is, say, a single sine wave, the string will oscillate in this mode indefinitely. Starting with the string in a simple configuration, for example in the first mode (or in other problems, starting with a combination of a few low modes), the purpose of our computations was to see how, due to nonlinear forces perturbing the periodic linear solution, the string would assume more and more complicated shapes, and, for t tending to infinity, would get into states where all the Fourier modes acquire increasing importance. In order to see this, the shape of the string, that is to say, x as a function of i and the kinetic energy as a function i were analyzed periodically in Fourier series. Since the problem can be considered one of dynamics, this analysis amounts to a Lagrangian change of variables: instead of the original $\dot{x}_i$ and x_i, $i = 1, 2, \cdots, 64$, we may introduce a and $\dot{a}_k$, $k = 1, 2, \cdots, 64$, where

$$(4) \qquad a_k = \Sigma x_i \sin \frac{ik\pi}{64} \cdot$$

The sum of kinetic and potential energies in the problem with a quadratic force is

$$(5\text{ a}) \qquad \mathrm{E}^{\mathrm{kin}}_{x_i} + \mathrm{E}^{\mathrm{pot}}_{x_i} = \frac{1}{2}\dot{x}_i^2 + \frac{(x_{i+1}-x_i)^2 + (x_i - x_{i-1})^2}{2}$$

$$(5\text{ a}) \qquad \mathrm{E}^{\mathrm{kin}}_{a_k} + \mathrm{E}^{\mathrm{pot}}_{a_k} = \frac{1}{2}\dot{a}_k^2 + 2\,a_k^2 \sin^2 \frac{\pi k}{128}$$

if we neglect the contributions to potential energy from the quadratic or higher terms in the force. This amounts in our case to at most a few percent.

The calculation of the motion was performed in the x variables, and every few hundred cycles the quantities referring to the a variables were computed by the above formulas. It should be noted here that the calculation of the motion could be performed directly in a_k and $\dot{a}_k$. The formulas, however become unwieldy and the computation, even on an electronic computer, would take a long time. The computation in the a_k variables could have been more instructive for the purpose of observing directly the interaction between the a_k's. It is proposed to do a few such calculations in the near future to observe more directly the properties of the equations for $\ddot{a}_k$.

Let us say here that the results of our computations show features which were, from the beginning, surprising to us. Instead of a gradual, continuous flow of energy from the first mode to the higher modes, all of the problems show an entirely different behavior. Starting in one problem with a quadratic force and a pure sine wave as the initial position of the string, we indeed observe initially a gradual increase of energy in the higher modes as predicted (e.g., by Rayleigh in an infinitesimal analysis). Mode 2 starts increasing first, followed by mode 3, and so on. Later on, however, this gradual sharing of energy among successive modes ceases. Instead, it is one or the other mode that predominates. For example, mode 2 decides, as it were, to increase rather rapidly at the cost of all other modes and becomes predominant. At one time, it has more energy than all the others put together! Then mode 3 undertakes this role. It is only the first few modes which exchange energy among themselves and they do this in a rather regular fashion. Finally, at a later time mode 1 comes back to within one percent of its initial value so that the system seems to be almost periodic. All our problems have at least this one feature in common. Instead of gradual increase of all the higher modes, the energy is exchanged, essentially, among only a certain few. It is, therefore, very hard to observe the rate of " thermalization " or mixing in our problem, and this was the initial purpose of the calculation.

If one should look at the problem from the point of view of statistical mechanics, the situation could be described as follows: the phase space of a point representing our entire system has a great number of dimensions. Only a very small part of its volume is represented by the regions where only one or a few out of all possible Fourier modes have divided among themselves almost all the available energy. If our system with nonlinear forces acting between the neighboring points should serve as a good example of a transformation of the phase space which is ergodic or metrically transitive, then the trajectory of almost every point should be everywhere dense in the whole phase space. With overwhelming probability this should also be true of the point which at time $t = 0$ represents our initial configuration, and this point should spend most of its time in regions corresponding to the equipartition of energy among various degrees of freedom. As will be seen from the results this seems hardly the case. We have plotted (figs. 1 to 7) the ergodic sojourn times in certain subsets of our phase space. These may show a tendency to approach limits as guaranteed by the ergodic theorem. These limits, however, do not seem to correspond to equipartition even in the time average. Certainly, there seems to be very little, if any, tendency towards equipartition

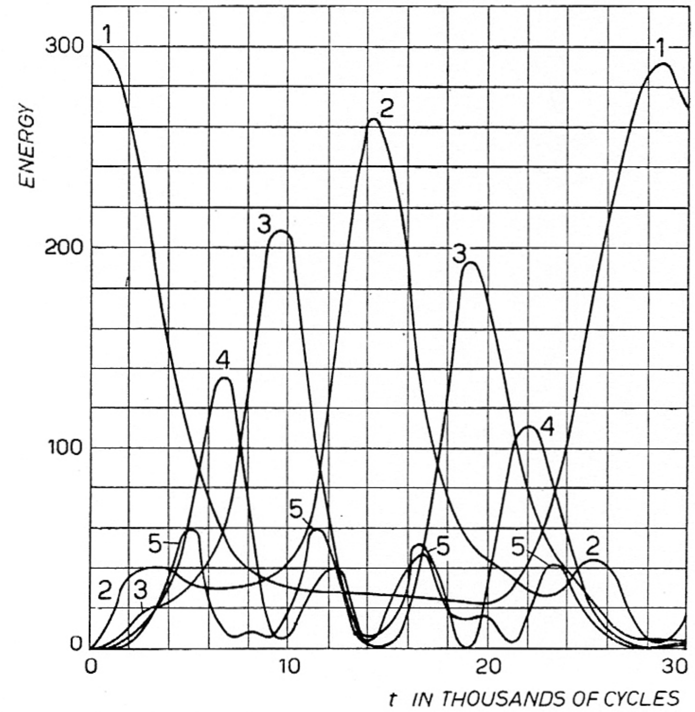


Fig. 1. – The quantity plotted is the energy (kinetic plus potential in each of the first five modes). The units for energy are arbitrary. N = 32 ; $\alpha = 1/4$; $\delta t^2 = 1/8$. The initial form of the string was a single sine wave. The higher modes never exceeded in energy 20 of our units. About 30,000 computation cycles were calculated.

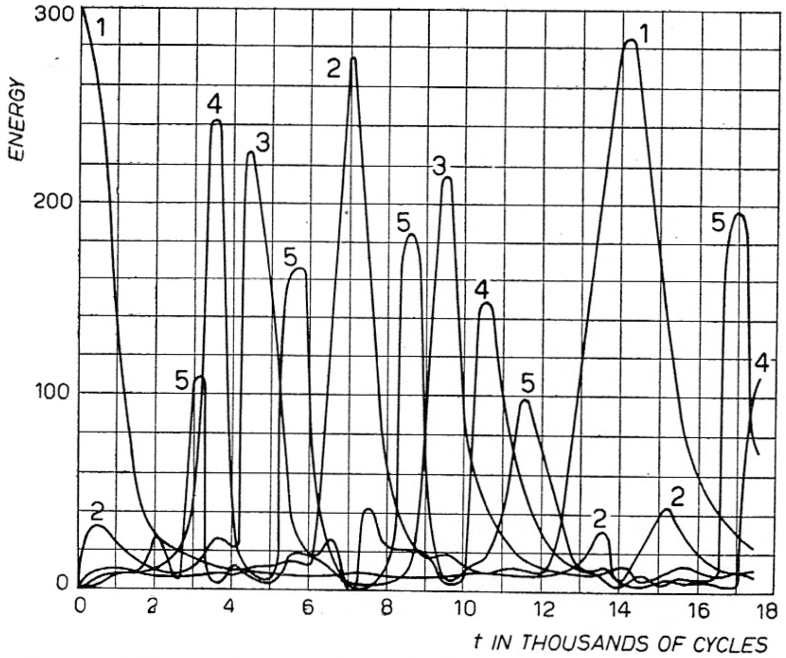


Fig. 2. – Same conditions ad fig. 1 but the quadratic term in the force was stronger. $\alpha = 1$. About 14,000 cycles were computed.

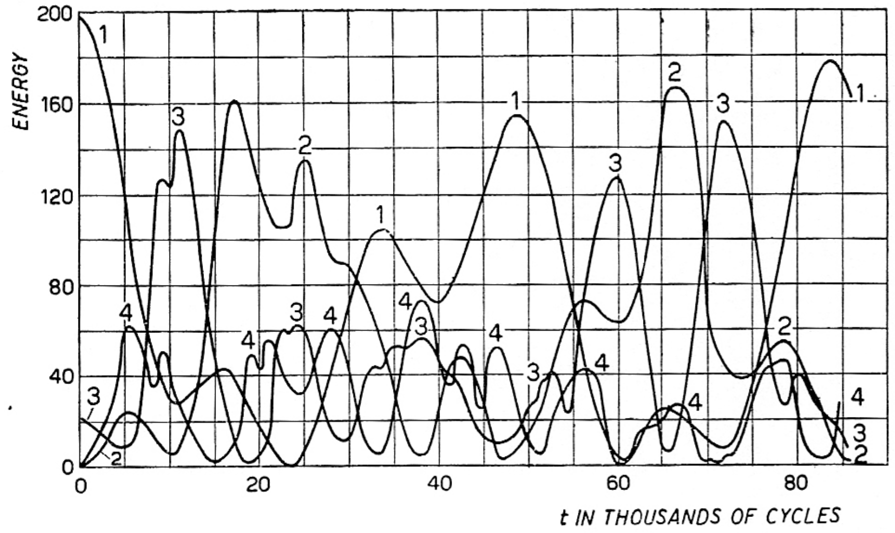


Fig. 3. – Same conditions as in fig. 1, but the initial configuration of the string was a " saw-tooth " triangular-shaped wave. Already at $t = 0$, therefore, energy was present in some modes other than 1. However, modes 5 and higher never exceeded 40 of our units.

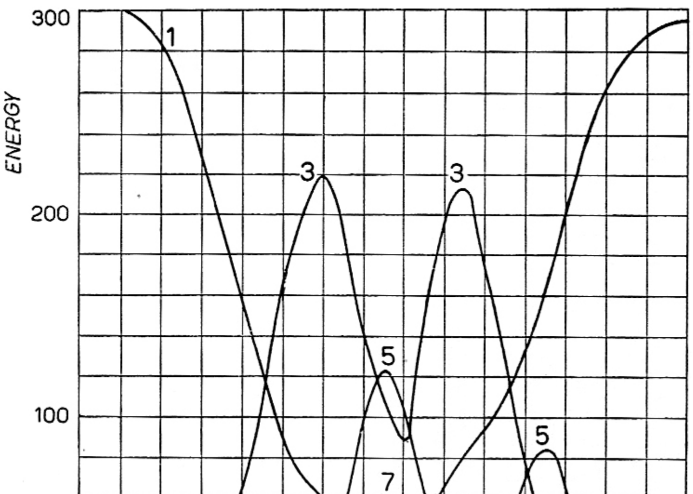


Fig. 4. – The initial configuration assumed was a single sine wave; the force had a cubic term with $\beta = 8$ and $\delta t^2 = 1/8$. Since a cubic force acts symmetrically (in contrast to a quadratic force), the string will forever keep its symmetry and the effective number of particles for the computation is $N = 16$. The even modes will have energy 0.

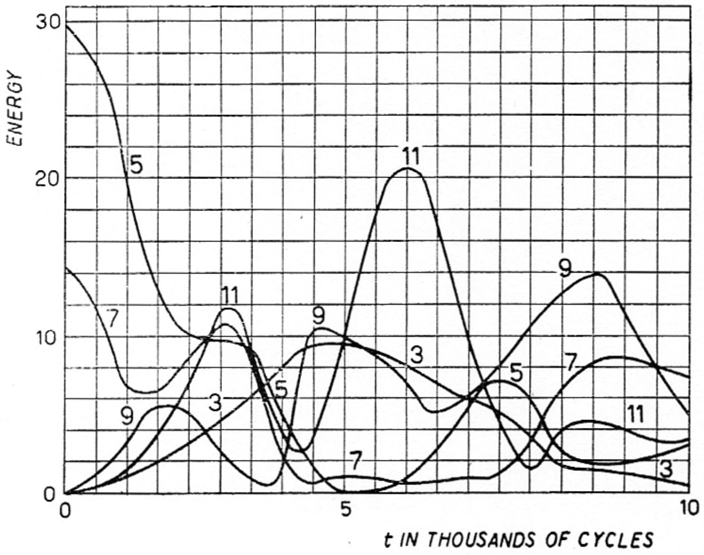


Fig. 5. – $N = 32$; $\delta t^2 = 1/64$; $\beta = 1/16$. The initial configuration was a combination of 2 modes. The initial energy was chosen to be 2/3 in mode 5 and 1/3 in mode 7.

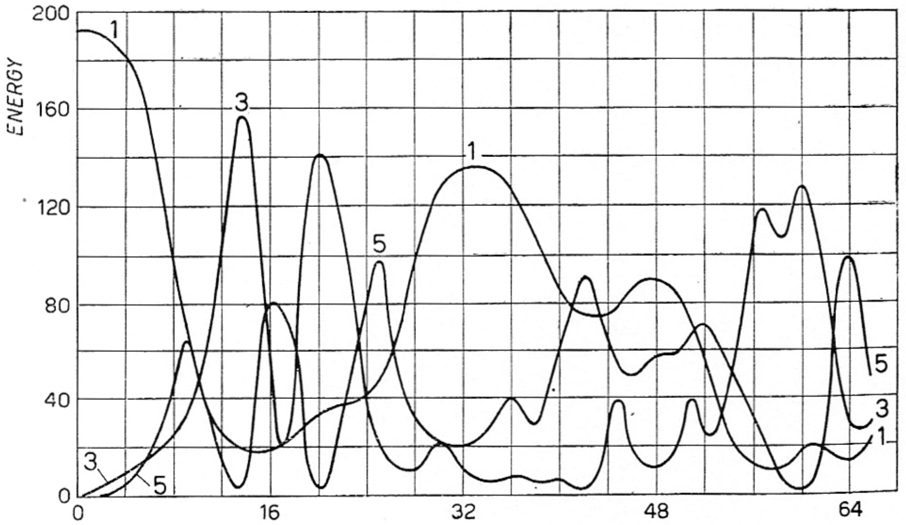


Fig. 6. – $\delta t^2 = 2^{-6}$. The force was taken as a broken linear function of displacement. The amplitude at which the slope changes was taken as $2^{-5} + 2^{-7}$ of the maximum amplitude. After this cut-off value, the force was assumed still linear but the slope increased by 25 percent. The effective $N = 16$.

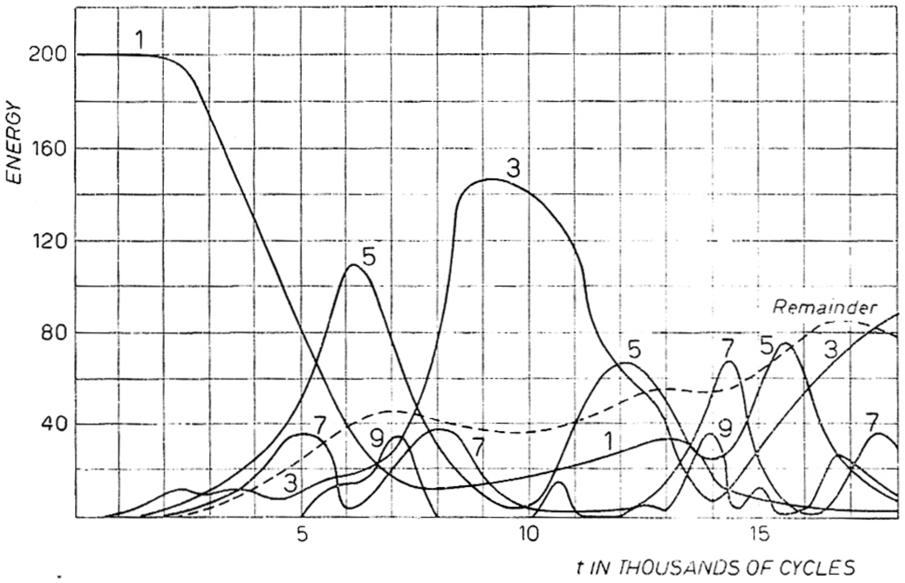


Fig. 7. – $\delta t^2 = 2^{-6}$. Force is again broken linear function with the same cut-off, but the slopes after that increased by 50 percent instead of the 25 percent charge as in problem 6. The effective N = 16.

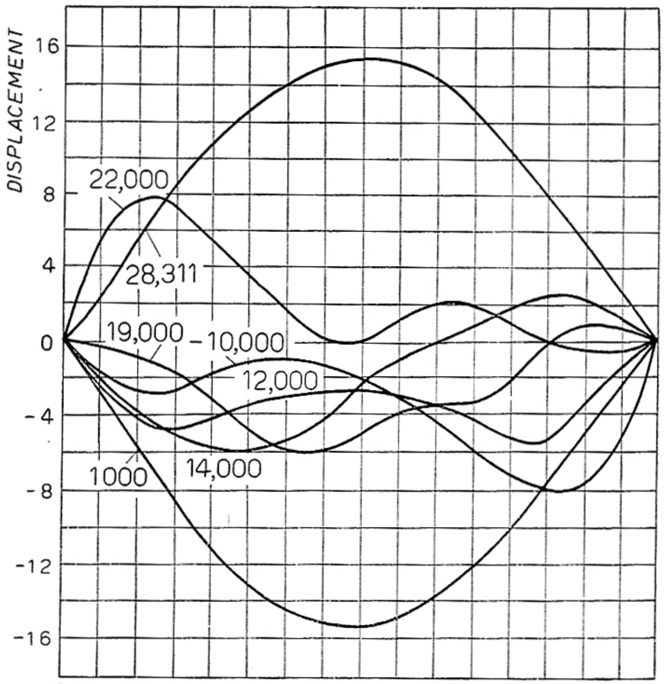


Fig. 8. – This drawing shows not the energy but the actual *shapes*, i.e., the displacement of the string at various times (in cycles) indicated on each curve. The problem is that of fig. 1.

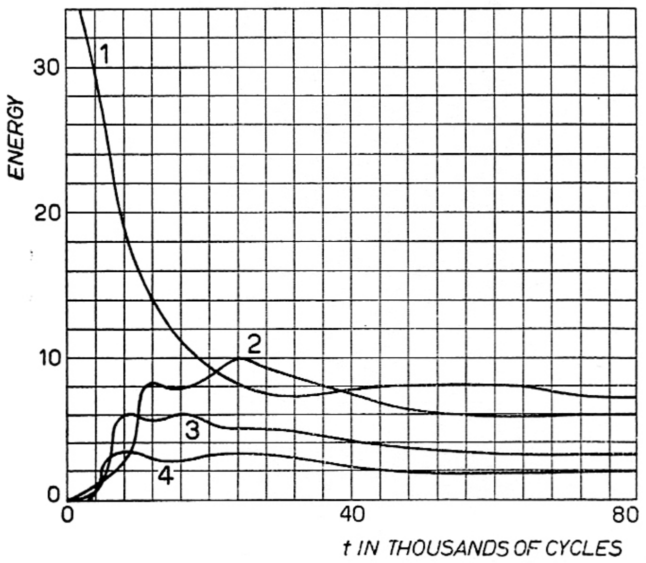


Fig. 9. This graph refers to the problem of fig. 6. The curves, numbered 1, 2, 3, 4, show the time averages of the kinetic energy contained in the first 4 modes as a function of time. In other words, the quantity is $\frac{1}{\nu}\sum_{i=1}^{\nu} T^{i}_{a_k}$. ν is the cycle no., $k = 1, 3, 5, 7$.

of energy among all degrees of freedom at a given time. In other words, the systems certainly do not show mixing.[2]

The general features of our computation are these: in each problem, the system was started from rest at time $t = 0$. The derivatives in time, of course, were replaced for the purpose of numerical work by difference expressions. The length of time cycle used varied somewhat from problem to problem. What corresponded in the linear problem to a full period of the motion was divided into a large number of time cycles (up to 500) in the computation. Each problem ran through many " would-be periods " of the linear problem, so the number of time cycles in each computation ran to many thousands. That is to say, the number of swings of the string was of the order of several hundred, if by a swing we understand the period of the initial configuration in the corresponding linear problem. The distribution of energy in the Fourier modes was noted every few hundred of the computation cycles. The accuracy of the numerical work was checked by the constancy of the quantity representing the total energy. In some cases, for checking purposes, the corresponding linear problems were run and these behaved correctly within one percent or so, even after 10,000 or more cycles.

It is not easy to summarize the results of the various special cases. One feature which they have in common is familiar from certain problems in me-

[2] One should distinguish between metric transitivity or ergodic behavior and the stronger property of mixing.

chanics of systems with a few degrees of freedom. In the compound pendulum problem one has a transformation of energy from one degree of freedom to another and back again, and not a continually increasing sharing of energy between the two. What is perhaps surprising in our problem is that this kind of behavior still appears in systems with, say, 16 or more degrees of freedom.

What is suggested by these special results is that in certain problems which are approximately linear, the existence of quasi-states may be conjectured.

In a linear problem the tendency of the system to approach a fixed "state" amounts, mathematically, to convergence of iterates of a transformation in accordance with an algebraic theorem due to Frobenius and Perron. This theorem may be stated roughly in the following way. Let A be a matrix with positive elements. Consider the linear transformation of the n–dimensional space defined by this matrix. One can assert that if $\bar{x}$ is any vector with all of its components positive, and if A is applied repeatedly to this vector, the directions of the vectors $\bar{x}$, $A(\bar{x}), \cdots, A^i(\bar{x}), \cdots$, will approach that of a fixed vector $\bar{x}_0$ in such a way that $A(\bar{x}_0) = \lambda(\bar{x}_0)$. This eigenvector is unique among all vectors with all their components non-negative. If we consider a linear problem and apply this theorem, we shall expect the system to approach a steady state described by the invariant vector. Such behavior is in a sense diametrically opposite to an ergodic motion and is due to a very special character, linearity of the transformations of the phase space. The results of our calculation on the nonlinear vibrating string suggest that in the case of transformations which are approximately linear, differing from linear ones by terms which are very simple in the algebraic sense (quadratic or cubic in our case), something analogous to the convergence to eigenstates may obtain.

One could perhaps conjecture a corresponding theorem. Let Q be a transformation of a n–dimensional space which is nonlinear but is still rather simple algebraically (let us say, quadratic in all the coordinates). Consider any vector $\bar{x}$ and the iterates of the transformation Q acting on the vector $\bar{x}$. In general, there will be no question of convergence of these vectors $Q^n(\bar{x})$ to a fixed direction.

But a weaker statement is perhaps true. The directions of the vectors $Q^n(\bar{x})$ sweep out certain cones C_α or solid angles in space in such a fashion that the time averages, i.e., the time spent by $Q^n(\bar{x})$ in C_α, exist for $n \rightarrow \infty$. These time averages may depend on the initial $\bar{x}$ but are able to assume only a finite number of different values, given C_α. In other words, the space of all direction divides into a finite number of regions R_i, $i = 1, \cdots, k$, such that for vectors $\bar{x}$ taken from any one of these regions the percentage of time spent by images of $\bar{x}$ under the Q^n are the same in any C_α.

The graphs fig. 1–9 show the behavior of the energy residing in various modes as a function of time; for example, in fig. 1 the energy content of each of the first 5 modes is plotted. The abscissa is time measured in computational cycles, δt, although figure captions give δt^2 since this is the term involved directly in the computation of the acceleration of each point.

In all problems the mass of each point is assumed to be unity; the amplitude of the displacement of each point is normalized to a maximum of 1. N denotes the number of points and therefore the number of modes present in the calculation. α denotes the coefficient of the quadratic term and β that of the cubic term in the force between neighboring mass points.

We repeat that in all our problems we started the calculation from the string at rest at $t = 0$. The ends of the string are kept fixed.

E. Fermi: Theories on the origins of the elements (240.3)

"Teoria sulle origini degli elementi,"
translated from Conferenze di Fisica Atomica (Fondazione Donegani) - Terza Conferenza,
Roma, Accademia Nazionale dei Lincei, pp 31-45 (1950),
compiled by Prof. E. Pancini

All the known matter is made up of various chemical elements each present with a different abundance, so the problem arises, first experimentally and then theoretically, of understanding for what reason some elements are abundant, others rare.

The problem is first of all an experimental one and, not wishing to discuss the question in detail here, a few general considerations are enough to understand it. What we are trying to establish are the amounts of the various chemical elements which are, so to say, in the whole Universe or, at least, in a large part of it and, obviously, the result which we may expect to obtain depends to a large extent on the samples taken for the analysis. For instance, if it is possible to determine the relative amounts of oxygen, iron, hydrogen and the other elements present in the part of the terrestrial crust which is approachable by our direct observations, one will get for each of them a definite relative abundance. But if, on the contrary, one determines for instance, the percentages of the same elements by analysing the meteorites, a different distribution of the elements with respect to the one found in surface rocks on the Earth will be discovered.

Therefore the problem, rather than a problem of chemical analysis, is essentially a problem of the selection of the samples to analyse.

Obviously, the question is not a new one; the data which will be presented here have been obtained in rather recent research by Harrison S. Brown,[*] of the University of Chicago, who has extended, enlarged and perfected the results of Goldschmidt.[†] The data have been obtained by analysing a large quantity of samples and this assures their reliability because data obtained from a particular sample display the special characteristics of it instead of what can be considered the cosmic distribution of the elements.

It is noteworthy the fact that, in spite of this observation, the conclusion of these analysis is that, if the selection of the samples is made with suitable attention, the results are highly uniform even if derived from materials of very disparate origin. For instance, in some favorable cases, it is possible to assign the ratio of the cosmic abundance of two elements with a precision of the order of the 1 or 2%.

Note that the measures of the abundances of the elements performed on the terrestrial crust, even if of utmost practical importance, have a rather limited the-

[*] *Rev. Mod. Phys.* **21**, 625 (1949).

[†] V.M. Goldschmidt, *Geochemische Verteilungsgesetze der Elemente und der Atomarten*, IX, Oslo, 1938.

oretical importance. In fact the terrestrial crust, indeed all the material which constitutes the Earth, during the geological ages, has been subjected to a deep chemical separation so that one could obtain significant results only through an analysis of samples taken in zones which go from the Earth's surface to its center and this is, obviously, impossible.

Luckily, this impossibility of taking samples from the interior of the Earth can be circumvented by studying the composition of the meteorites which, in the opinion of the experts, turn out to be samples taken from various zones of missing planets. Thus, if the Earth, due to a cosmic catastrophe, were to break up, the meteorites coming from its crust would be essentially made of iron or, more precisely, of an alloy consisting mostly of iron and then nickel and then, to an ever lesser extent, other elements.

In effect a statistical analysis of the meteorites which arrive on the Earth (and the meteorites which arrive on the Earth are really of these two kinds) indicates that the ratio between the amount of matter from stone meteorites and iron meteorites is not very idfferent from the ratio between the stone part and the iron part of the Earth which results from the investigation in depth through seismic waves.

The research of which we are speaking has been carried out mostly by a painstaking collection of a large number of meteorites and performing extremely accurate quantitative analyses of them. It's important to remark that the problem of making these analyses is much less simple than it might seem because most of the elements, indeed, as we shall see, almost all, are so rare as to be present in amounts of few parts per million or even less. Thus one of the greater difficulties of the problem is that of finding the means of performing quantitative chemical analyses of extreme sophistication. To overcome this difficulty one has been even obliged to use (at least in the recent studies performed at the Chicago University) nuclear reactors for irradiating the material under examination with the aim of observing the resultant characteristic activities of the elements of interest. Therefore the identification of the elements, in this way, is realized essentially through radioactive rather than chemical methods. By those means one has succeeded in analysing the material arriving on the Earth in the shape of iron and stone meteorites and, taking suitable averages of the data so obtained, a table has been constructed which for most of elements will coincide with other data of completely different origin as, for instance, those obtained through the spectroscopic observation of the stellar atmosphere. In this way one has an indication that the matter which constitutes the meteorites is not substantially different from that which constitutes stellar atmospheres. As a matter of fact there are some remarkable exceptions, but easily understood: for instance there are some elements in the meteorites which are practically missing, or present in an amount smaller than the one expected. This is the case of the noble gases which are present both in the Earth and in the meteorites in an amount largely smaller than that corresponding to their cosmic abundance because in the process of formation of the planets they have not been kept inside. Another exception even more notable is that of hydrogen, but also this exception is accounted for by arguments of the same kind of those put forward for the rare gases.

Facts of this kind make it clear how the results obtained from the analysis of the meteorites (from which we take most of the quantitative data on the cosmic abundances of the elements) should be revised through a careful chemical discussion which, on the other hand, is unfortunately almost completely arbitrary since it involves assumptions about the process of formation of the meteorites themselves and the characteristics both chemical and physical of the environment where they have been formed. In short the analysis of meteorites, element by element, must be integrated together through chemical considerations of a theoretical nature which allow us to decide if the element in question has retained its cosmic proportions in the meteorites.

Besides meteorites also the stellar atmospheres have been investigated (through spectroscopic analysis) and, in part, the matter clouds of the interstellar space through the analysis of their absorption spectrum. These data are, nevertheless, extremely limited and must be used only as additional ones. But the fact that is anyway remarkable is that all these data (taking into account the quoted exceptions which, in any case, can be justified by very convincing arguments) coming from the analysis of quite different celestial objects as the stellar atmosphere, the interstellar matter and the meteorites all match very well. On the contrary, the data which come from the Earth's crust vary significantly because, as we have said, the Earth's crust is not a faithful specimen of what can be considered to be the material which constitutes the Earth.

That said it is interesting to consider the table reported here (Table 1) in which the numbers represent the relative abundances of the various elements.

These data are extracted from the papers of Harrison S. Brown which can be considered the most up-to-date; the numbers listed in the table refer to some of the most significant elements and are sufficient to point out some strange features of the behavior of the relative abundance of the various elements as a function of the atomic number. They represent the number of atoms of each element present, on the average, in the cosmic matter for every 10^4 atoms of silicon 14.

When analysing this table it is convenient to begin at hydrogen, which is not only the simplest one of the elements, but also the most abundant: the number of its atoms present in the cosmic matter for every 10^4 atoms of silicon amounts to three or four hundred million. After the hydrogen, both in the periodic system of the elements and in the scale of the abundances, there is helium whose relative abundance is, on our scale, 35 million. For the elements which follow helium, the relative abundance decreases very rapidly to extremely low values: lithium, beryllium and boron are extremely rare: for instance, the relative abundance of beryllium is two tenths (that is, there are 50,000 atoms of silicon for every atom of beryllium). As one can see, between helium and beryllium there is a jump on the order of one hundred million.

Table 1 Relative abundance of the elements

	A	Z	Atoms per 10^4 atoms of silicon
H	1.01	1	3.5×10^8
He	4	2	3.5×10^7
Be	9.02	4	2×10^{-1}
C	12.01	6	8×10^4
O	16.00	8	1×10^5
Si	28.06	14	10^4
Cl	35.46	17	2.5×10^2
Mn	54.93	25	1×10^2
Fe	55.85	26	2.6×10^4
Co	58.94	27	1.6×10^2
Ni	58.69	28	2.0×10^3
Cu	63.57	29	7
Ga	69.72	31	5×10^{-3}
Sr	87.63	38	10^{-1}
Cd	112.41	48	2×10^{-2}
Cs	132.91	55	10^{-2}
Pt	195.23	78	10^{-1}
Pb	207.21	82	4×10^{-3}
Th	231.12	90	10^{-2}
U	238.07	92	3×10^{-3}

The other light elements which follow the three quoted above in the periodic table have abundances slightly different from that of silicon: to carbon, for instance, an abundance of 2×10^4 must be attributed. Immediately after oxygen heads upwards: 10^5. It is, after hydrogen and helium, the most abundant element as regards the number of atoms (not the weight).

Proceeding on this scale one finds abundances on the order of few units until one arrives at iron which has a considerably high abundance: 2.6×10^4. Then cobalt: 2.6×10^2, nickel: 2×10^3, and continuing on in the order of the periodic table, at this point the abundance begins to decrease rapidly and does not rise any more. From gallium on until uranium the abundances oscillate more or less irregularly between one tenth and one hundredth. A slight exception is lead which has a little higher abundance, but one might think that the amount of lead is increased due to the decay of the radioactive substances which are located immediately over it. Another exception, in the opposite sense, is uranium but one can think it became impoverished owing to its radioactive decay.

All these arguments will probably assume more clarity if we represent in a diagram (Fig. 1) the relative abundance of the elements as a function of the atomic number. From this diagram one can see that immediately after the peak represented by hydrogen and helium there is a tendency to exhibit, though with high regularity, a decreasing feature of the relative abundance of the elements. So that one who wanted to draw a curve through these points, neglecting the irregularities, could draw the curve shown in Fig. 1. And, if one wanted to trust that, having accounted for the exceptions, this curve represents with good approximation the relative abundance of the elements, one should also conclude that the relative abundance of each element is one of its essential characteristics like, for instance, its atomic number, its energy of formation or its mass. Then one forms the impression that the relative abundance of each element is really a property of its own, connected, as is obvious, both with the other properties of the element and with the mechanism, quite unknown, through which the element has been formed.

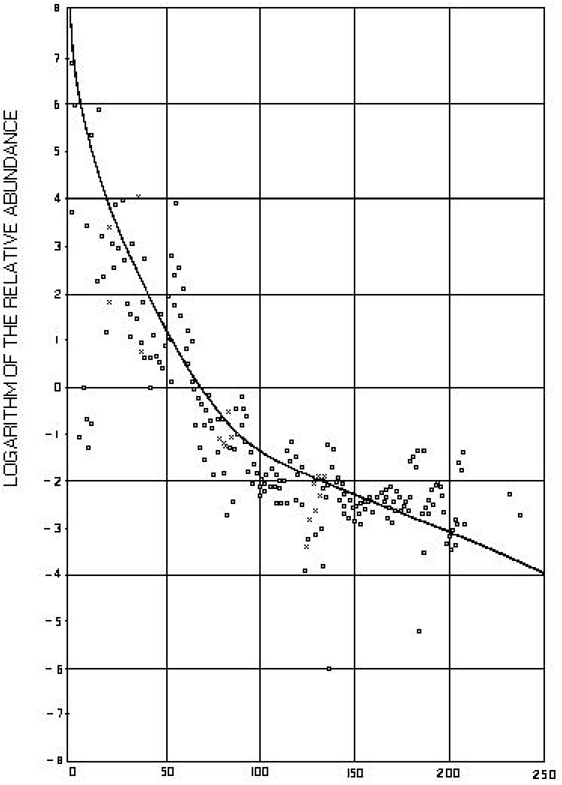


Fig. 1

Obviously, in a discussion of this kind one must take into account the abundance of the different isotopes of each element, but this is not a complication of the problem since the relative abundances of the isotopes of each element are known and rather constant: so if we know the abundance of each element it is a question of trivial arithmetic to calculate the abundance of the isotopes.

Also when studying the relative abundance of the isotopes one will notice some regularities that are worthwhile to call attention to because we shall come back to them below.

In Fig. 2 a diagram of isotopes is shown with the number of protons which constitute each nucleus given along the abscissa and the number of neutrons along the ordinate: as one can see the various elements are distributed in a region which initially has the direction of the bisector of the axes and then turns upwards (and this means that in the nuclei of low atomic number the number of protons equals that of neutrons whereas for the high atomic numbers the percentage of neutrons is, more and more, higher than 50%). Almost always one finds that in the lower part of the periodic system — that is for the light nuclei — the more abundant isotopes are those richer in protons or, what amounts to the same thing, poorer in neutrons; then there is a transition zone and finally in the higher part one observes quite the opposite tendency: the more abundant isotopes tend to have more neutrons than protons.

Obviously the idea of justifying all these facts, that is to justify the abundance of every single element and, for each element, the abundance of its isotopes, is, certainly, an extremely ambitious program and constitutes a problem whose solution is assuredly very far off. Nevertheless recently there have been attempts in this direction but with quite unsatisfactory results. This fact does not exclude that they are extremely interesting in the sense that they represent an attempt at research in the direction which most probably will be one of the most important in the future. On the other hand it is obvious that if the solutions obtained to date are not satisfactory one cannot exclude that in the future one cannot make conclusive steps along this road.

One of the most natural hypotheses that has been formulated, from long time and by various people, is that the elements we find in nature are the result of a process which bases itself on a kind of a chemical or, as one says, superchemical equilibrium. In other words one can ask if it permissible to imagine that if we put the constitutive elements of the chemical elements in a cauldron, that is protons and neutrons, and then heating it all to a suitable temperature and finally, when this matter is, so to say, well cooked, cool it suddenly, one can obtain a mixture of elements which looks like that which seems to us to constitute the Universe.

Many attempts in this sense have been made but the results obtained are indeed not very convincing. Obviously the temperatures and the pressures in this kind of cauldron must be thought to have rather amazing values if one wants to obtain results which do not disagree with the experimental data: for instance, the temperature should be about 10 billion degrees and the pressure about one million grams per cm^2. The necessity of such temperatures and pressures can be understood without difficulty if one considers that the temperature must be very high to bring forth, in a conspicuous manner, nuclear reactions, and that the pressures must be also be very high to have the possibility of forming very heavy nuclei. In fact if, at such high temperatures, the pressure were not proportionally high, all the nuclei consisting of many particles would disintegrate and the possibility of the existence of heavy nuclei present in nature would not exist.

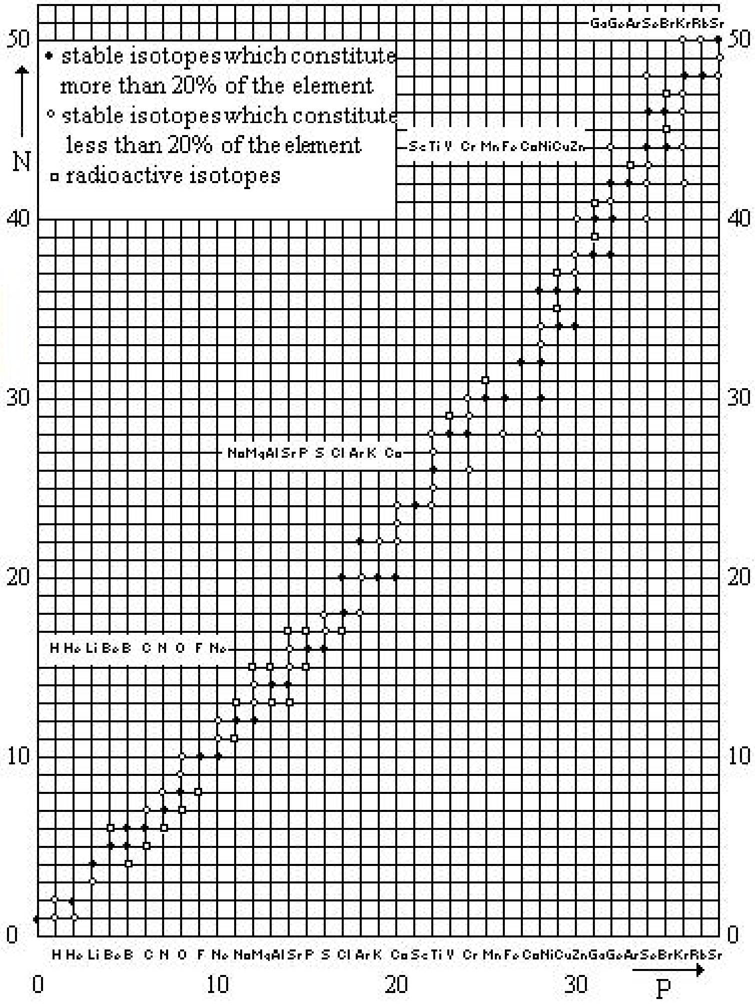


Fig. 2

On the other hand, the fact of the matter is that starting from such hypotheses one does not succeed in obtaining a distribution of elements which looks very much like the real one: for instance, the relative abundances of various isotopes turn out to have a random distribution absolutely different from that observed experimentally.

The more recent theories are based, instead, on a rather different scheme: we shall limit ourselves to speak here about only one of them, that which, in our opinion, is the most interesting one even if it cannot be considered in any way to give a satisfactory explanation of the facts. This theory is due chiefly to Gamow who, being a joker as everybody knows, joined with two other physicists, Alpher and Bethe, with the aim, perhaps, of playing with the fact that the three names, read in the American way mangling the words, sound like the first three letters of the Greek alphabet. As a matter of fact, the essential contribution of the theory of which we are speaking was given by Gamow and in part by Alpher: Bethe, instead, appears to have been associated only to complete the play on words.

Anyway, the theory can essentially be divided into two parts. The first is based on the observation, in reality not new, that there is the possibility of forming elements, even when the temperature and pressure do not assume such amazing values as those quoted above, provided that one conjectures forming the elements through successive additions of neutrons. Without dwelling at the moment on an explanation of the origin of these neutrons, let us try to give an idea of the way in which this formation can take place.

Let us still refer to a P-N diagram (Fig. 2) in which each element is represented by a point whose abscissa is equal to the number of its protons and whose ordinate is equal to the number of its neutrons. As we have already said, all stable elements are located in a well defined zone.

If we now assume that we submit a certain element to a "bath" of neutrons it may happen that its nuclei capture one of these neutrons. Thus, if the composition of this nucleus is represented by the point A of Fig. 3, after the capture the new formed nucleus will have a composition represented by the point B which is obtained by taking from A a step upwards (in fact N is increased by one and Z remains constant).

The new nucleus will be able, in turn, to absorb another neutron and produce an element represented by the point C and so on, until one ends up ging out of the zone of stable elements. The newly formed unstable element will evidently be beta radioactive and then will disintegrate through a beta process which is a change of a neutron into a proton: the new representative point will therefore be obtained taking a step downwards (decrease of a neutron) and a step rightwards (increase of a proton).

If now there are still neutrons present, the nucleus so formed will be able to absorb another neutron, then another neutron and after it will emit a beta ray; and in this way little by little we will climb up the slope of the stable elements. Thus little by little very heavy elements are formed through a mechanism of successive additions of neutrons to light nuclei assumed to be pre-existing.

At this point, if one wants to be ambitious (and, as we shall see, Gamow puts forward demands still more ambitious than these), one can even intend to explain

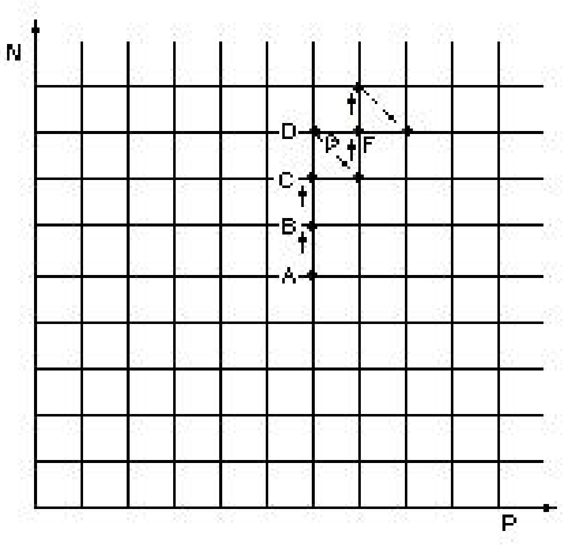


Fig. 3

the formation of all the elements starting from only neutrons.

Let us assume, in fact, that in a region of space, at a certain instant, are contained some neutrons. As is known, the neutron is not a stable particle, on the contrary its life time is quite short (it has not yet been measured very well but it cannot be appreciably different from 15') and therefore after about ten minutes, half of the neutrons will be decayed producing as many protons. But neutrons and protons have a certain affinity and the neutrons tend to latch onto the protons in this way forming nuclei of deuterium. In this way, starting initially from only neutrons, through their decay and association with the generated protons, it will be possible to form the first light nuclei and then, from them, with a process of the kind described above, one will arrive presumably at the formation of the heavy elements.

Gamow has made an attempt at investigating this model (or better put, a model which looks like this) from a quantitative point of view. Of course for a quantitative investigation it is necessary to introduce data on the probability of capture of neutrons by a given element, because it is this probability which essentially determines the speed of the process of this phenomenon. Now, much data is available on the capture of slow neutrons, but presumably phenomena of this type happened at a temperature high enough to advise taking data regarding the capture of rather fast neutrons — and Gamow has taken data of this type. In Fig. 4 we report the cross sections for the capture of fast neutrons as functions of the atomic weight.

Gamow, simplifying (maybe too much) that what the experimental results really give, assumes that these cross sections for the capture of neutrons increase linearly for values of the atomic weight between 0 and 100 and then remain constant as indicated by the broken line depicted in Fig. 4. If we observe the figure and take into account that the scale is logarithmic, we can judge how much Gamow's schematization is strong: anyway it is convenient to follow Gamow's reasoning till to the bottom before criticizing it.

Therefore let us assume, for the time being, that the cross sections are really the ones Gamow claims. In this case we can plainly write down the differential equations describing how heavier elements are successively formed. Let us call N_a the number of atoms with atomic weight a; the derivative of this number with respect to the time will depend on two terms: one which represents the increase in the number of atoms of weight a due to the aggregation of atoms of weight $a-1$ (and this will be a positive term proportional to N_{a-1} , to the cross section σ_{a-1} of the element $a-1$ and to the flux $\Phi(t)$ of the neutrons). Then there will be a negative term which in the same way represents the decrease of N_a due to the absorption of neutrons which changes the atoms a into atoms $a+1$. In a formula

$$\frac{dN_a}{dt} = \Phi(t)\,(\sigma_{a-1}N_{a-1} - \sigma_a N_a) \qquad (a = 1, 2, ...238) \ . \tag{1}$$

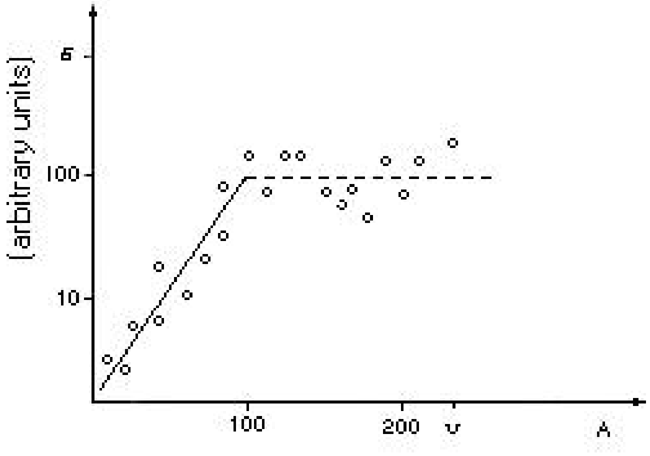


Fig. 4

Equations like this must be written for every value of a and a system will be obtained which we can solve, at a fixed neutron flux, deriving the way in which the abundances of the single elements evolve in time.

Now, the most significant result (which could be even more significant if the curve of the capture cross sections assumed by Gamow were a more faithful representation of the experimental facts) is that, owing to the peculiarity of this curve — namely the fact that for a certain atomic weight the tendency of the cross sections to increase stops suddenly — by assuming conveniently the time and the flux of neutrons, one finds a distribution of the abundances of the elements of the type represented in Fig. 5 and which is not very different from the experimental one (Fig. 1).

Of course the result one obtains depends on the time interval we choose in the sense that, if we fix a certain flux of neutrons, the material must be exposed to its action for a suitable time: in fact, if the time is too long, too many heavy elements are formed, if it is too short, too few elements are formed. But, by "cooking" so to say the material to the right point one succeeds in obtaining something which has a certain resemblance with the experimental data. This resemblance arrives at such a point as to give, in the case of elements with high atomic number, a distribution of isotopes resembling the real one. In the region of light elements the result is, instead, contrary to the experimental one but one can think that a successive heat treatment, even at not an exceedingly high temperature, might have modified the situation.

As we have said, Gamow was not content with these results and has taken a further step, a very risky and almost certainly wrong step. Almost certainly wrong since the step one takes when, to explain the facts, one assumes very precise hypotheses. In that case, as is obvious, the more precise the hypotheses are, the more easily one demonstrates that they are wrong. At any way, Gamow resolved to determine the time in which the formation of elements described above has happened by resorting to the theory of the expansion of the Universe. This theory is connected with the theory of general relativity and we attempt to give a short account of it.

Unfortunately also for general relativity, as for other physical theories, there does not exist a single theory and this entails a certain freedom of choice, but at present this choice cannot be made on the reliable basis of experimental results. But if we base ourselves on the simplest one of the theories of relativity, that one without a cosmological term, we can construct as has been done, a theory of the expansion of the Universe according to the following general lines.

One starts from the hypothesis, which has at least the merit of being very simple, that the energy density (matter and radiation) is uniform in the entire Universe, at least when one averages over very large regions of it.

Furthermore one assumes that the space has a constant curvature; this means that the Universe is homogeneous not only with respect to the energy density but also with respect to its geometrical properties. From this hypothesis one can infer that the Universe at a certainly well-defined time has the shape of a sphere or that of a pseudo-sphere; for particular reasons, connected with the present matter density, one must choose the pseudo-sphere, which is a sphere with an imaginary radius and obviously it is not possible to represent it by a figure. But if for the moment we leave out of consideration the fact that the object of which we want to speak is a pseudo-sphere and not a sphere and furthermore if we limit ourselves to represent

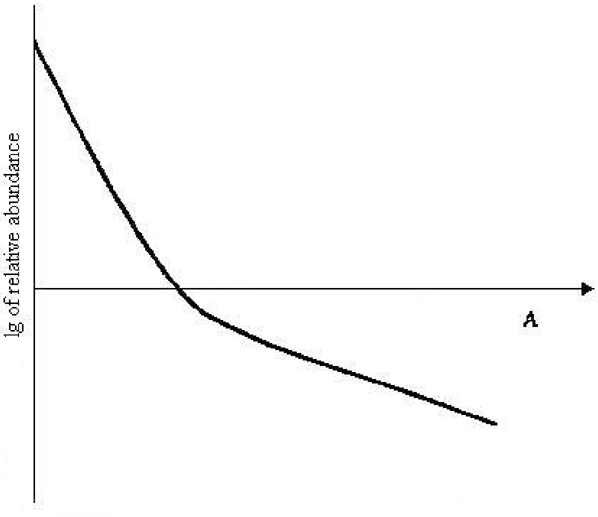


Fig. 5

only three of its four dimensions (the time and two spatial coordinates) it will be possible (Fig. 6) to give an idea of how the Universe evolves expanding itself. In Fig. 6 the Universe is represented by circles whose radius is increasing with time.

If now, by using the formulas of general relativity, one makes calculations, one can find a relation which connects the velocity of expansion of the radius, $r = iu$, of the pseudo-sphere with the energy density w:

$$\left(\frac{du}{dt}\right)^2 = c^2 + \frac{Kc^2}{3}u^2 w \ . \tag{2}$$

This formula says that the square of the time derivative of the modulus of the radius u of the pseudo-sphere equals the square of the light velocity plus a term which contains the radius itself, the energy density w and a constant K related to the gravitational constant G through

$$K = 8\pi \frac{G}{c^4}$$

and having the value of about $2 \times 10^{-48} dy\, n^{-1}$. If we now want to use this formula to describe the expansion of the Universe when its radius is very small, we can see that the first term of the right hand side becomes negligible since the energy density increases much faster than the squared radius decreases. Then formula (2) can be simplified in the following way:

$$\left(\frac{du}{dt}\right)^2 = \frac{Kc^2}{3}u^2 w \ . \tag{3}$$

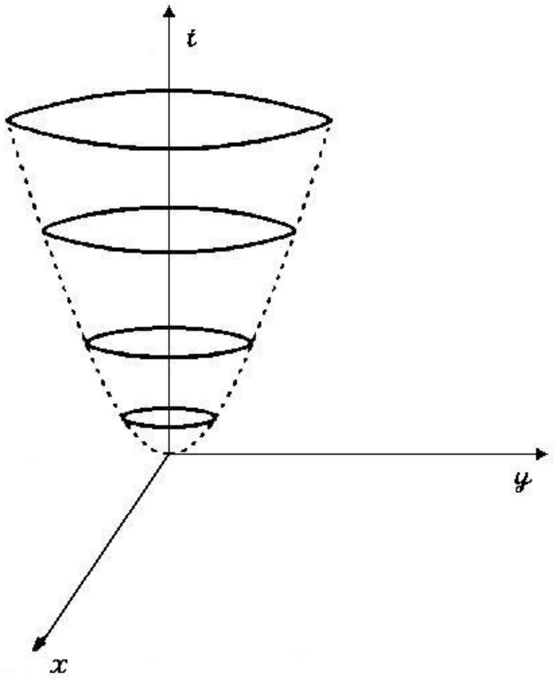


Fig. 6

At this point Gamow made an interesting remark: if one admits that, when formula (3) holds, the energy is essentially radiation energy, one can, using the formula, obtain a relation between temperature and time which contains only universal constants. In this way, one arrives at eliminating the arbitrariness due to the value of temperature at which nuclear reactions had taken place and given origin to the elements. This arbitrariness might have allowed one to obtain, to a certain extent, any result whatsoever. We give here, without proof, the formula which links temperature to the time elapsed since the instant when the Universe had infinitesimal size

$$T = \left(\frac{3}{4Kc^2\sigma}\right)^{1/4} \frac{1}{\sqrt{t}} \text{ degrees} . \tag{4}$$

In this formula, σ represents the Stefan's law constant, T the absolute temperature and t the time. As we have already pointed out, this is an approximate formula and holds for values of t not too large, for instance not exceeding a few million years.

If we substitute the constants appearing in formula (4) by their values, we obtain

$$T = \frac{1.52 \cdot 10^{10}}{\sqrt{t}} \text{ degrees} \tag{5}$$

from which one sees that when t is, for instance, equal to one second, the temperature is, as one could have imagined, enormously high: of the order of 10^{10} degrees. But it decreases very fast so that, after one thousand seconds, it is already reduced to the order of magnitude of one hundred million degrees and this value is low enough to stop strong nuclear phenomena.

Then the temperature varies with time following a well-defined law and the pseudo-radius of the Universe vaires with an equally determined law (it is proportional to the squared time) and therefore only one parameter is left undetermined: the density of neutrons; Gamow intends, by working on this single arbitrary parameter, to succeed in predicting the distribution of the abundances of the elements and indeed in a certain sense he succeeds. As a matter of fact, he succeeds as long as one is content with a very rough analysis of his results, but as one tries to enter into details, immediately one runs into trouble and probably troubles would increase if it was possible to carry out this analysis which is extremely complicated to do.

The first difficulties are met already in the lower region of the periodic system as soon as one asks oneself a little in detail in what manner the elements are gradually forming themselves. As we have already said, the first nucleus to be formed will be that of hydrogen, then through the merging of a proton with a neutron deuterium will be formed and then with the addition of another neutron tritium, which will decay through a beta decay into helium three. By addition of a new neutron helium three changes into helium four. Already here one meets a little difficulty because helium three capturing neutrons tends to break rather than to form helium four, nevertheless one can still think that at least a small fraction of helium three changes by capture of a neutron into helium four. But at this point the difficulty one meets is much more important because the nucleus of mass five does not exist: if one would try to form it by addition of a neutron to helium four it would break to peaces creating an insurmountable barrier which prevents the successive formation of elements through the addition of neutrons.

In reality one can find some way to jump or even better to avoid this barrier, and it is the following: according to the formulas written above, in the time in which these phenomena should take place, the temperature, though already much decreased, is still on the order of 10^8-10^9 degrees and at such high temperatures nuclear reactions can still take place in a conspicuous way and are produced by the collisions among the nuclei which move under the effect of thermal agitation. Thus it is not impossible to think that a nucleus of mass six can be formed, without the preliminary existence of a nucleus of mass five, by making a nucleus of deuterium react directly with a nucleus of helium four. As a matter of fact, this reaction is extremely unlikely, but not impossible, therefore it is not excluded that a small amount of lithium six is formed allowing, through successive additions of neutrons the formation of heavier nuclei. And many other difficulties of this type are met; for instance, also the nucleus eight does not exist in any stable form and this missing step will be jumped by a device of the kind already described.

But the difficulties do not end here, another one which cannot be passed over in silence is that, if one assumes an initial density of neutrons large enough so that they can form heavy elements in nonnegligible quantities, one finds that the ratio between helium and hydrogen has nothing to do with the actual one: one will have more abundant helium than hydrogen contrary to experimental evidence.

So it only remains sadly to conclude that this theory is unable to explain the way in which the elements have been formed in time, and this after all is what one should have expected.

However, we must recognize the courage with which Gamow has set about constructing an attempt at a theory based on extremely determined hypotheses: the theory has failed and this means that some of his hypotheses are wrong, but the result he has obtained in this way (to be at least certain of having made a mistake) is certainly more remarkable than one that which could have been obtained from a theory so indefinite as to be able to explain a lot of experimental facts, exactly because of the great deal of arbitrariness contained in it, but that would not have made evident what are its incorrect points allowing them to be corrected and to proceed to the construction of new and more satisfactory theories.

FA 9 - E. Fermi and A. Turkevich (Fermi-Turkevich): An excerpt from "Theory of the origin and relative abundance distribution of the elements," by Ralph A. Alpher and Robert C. Herman, pp. 193–197

Rev. Mod. Phys. ***22****, 153–212 (1950).*

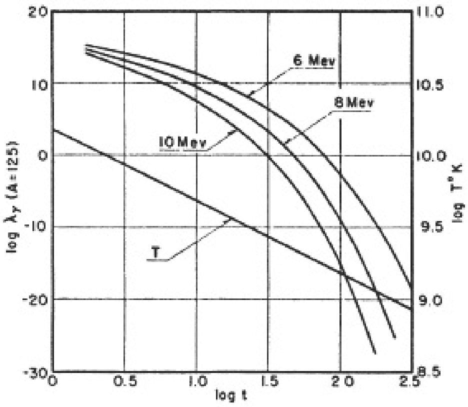


FIG. 18. Logarithm of the (γ, n) reaction probability, λ_γ, *versus* $\log t$ for the nucleus of $A=125$ for various binding energies of the last neutron, according to Smart (unpublished). The graph shows the decrease in λ_γ as the temperature, T, decreases in an expanding universe controlled by radiation [Eq. (100)].

equation for neutrons. Recognizing that this case including the universal expansion would require a higher initial density of matter than in the static case, five starting densities were considered. The solution which appears to be most promising is that for which the coefficients P_j in Eqs. (108) were taken as 100 times those given in Table XI. This corresponds to an initial density $\rho_m \cong 5\times 10^{-7}$ or 4×10^{-6} g/cm^3 depending on whether one takes $\lambda^{-1}=1800$ or 1000 sec., with corresponding starting times $t_0 \cong 230$ or 130 sec. The solutions of these five simultaneous equations for five initial densities were obtained on an IBM relay calculator in the range $0.128 \leqq \tau \leqq 1.17$. In Fig. 19 the results for the case $100P_j$ discussed are plotted as $\log\xi_j$ *versus* $\log\tau$. A comparison of these solutions with those of the static case given in Fig. 16 shows the marked effect of the expansion in that the relative concentrations quickly go through a peak and by $\tau \cong 1.2$ all the concentrations are decreasing as $\tau^{-\frac{3}{2}}$, i.e., the universal expansion is by then the only term of any importance. It will be seen that the curves for $j=2-4$ are essentially parallel by this time and the relative concentrations cannot change thereafter. The neutron concentration, on the other hand, continues to decrease mainly according to both decay and expansion and it is clear that the sum $\xi_n+\xi_p$ is approximately parallel to the other curves. The relative concentrations of species $j=1$, 2, 3, and 4 calculated in this approximation, with coefficients in Eq. (108) taken as 100 times those in Table XI, are in good agreement with those obtained in the static case which led to the fit over the entire range of atomic weight shown in Fig. 17. Calculations are in progress for species of higher atomic weights and it may be expected that a satisfactory fit to the observed data might result.

To be sure, calculating the relative abundances of the very light elements in the manner described possibly represents a poor approximation to the state of affairs for these light elements. The specific neutron-capture cross sections have not been used for these very light elements although the smoothed equation for σ is not too far in general from the observed values. More important is the fact that there are reactions other than neutron capture among the light elements which should be taken into account. Recently, Fermi and Turkevich have considered in detail all the nuclear reactions involved in the formation of the elements through helium in an expanding universe. This work is described in detail in Section IV(d)2.

2. *The Formation of Light Nuclei*

A preliminary study of the non-equilibrium formation of light nuclei, taking into account specific reactions and the actual cross sections, was first carried out by Gamow [59, 61]. He considered the building up of deuterons by neutron-proton capture. This process is described by Eqs. (108a) and (108b), taking only the first term in the summation in the former equation. Replacing p_1 by $\sigma_1 v$, where σ_1 is the absolute cross section for deuteron formation [20], setting $\rho_m=\rho_0 t^{-\frac{3}{2}}$ [see Eq. (101)], and replacing all temperature dependent quantities by their equivalent time-dependent forms, one obtains for the concentrations by weight of neutrons and protons [7, 9],

$$dx_n/d\tau = -x_n - \alpha_G x_n x_H/\tau, \qquad (128a)$$

$$dx_H/d\tau = +x_n - \alpha_G x_n x_H/\tau, \qquad (128b)$$

where,

$$\alpha_G = [(2^{9/4}\pi^{5/4}G^{1/4}a_r{}^{1/4}e^2h)/(3^{1/4}m_0{}^{9/2}c^{11/2}k)] \times(|\mu_H|+|\mu_n|)^2(\epsilon^{1/2}+\epsilon_0{}^{1/2})\epsilon^{3/2}\rho_0. \qquad (128c)$$

In Eq. (128c), G is the gravitational constant, a_r the radiation density constant, m_0 the unit of atomic mass,

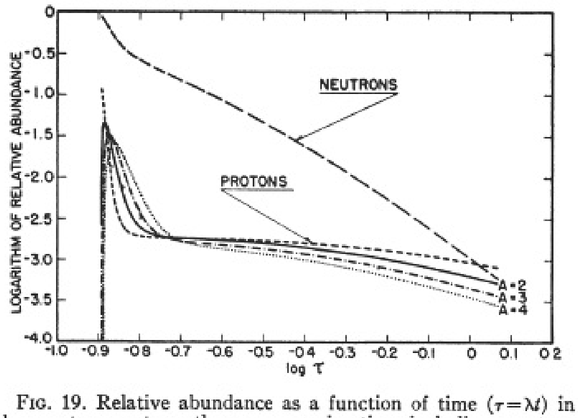


FIG. 19. Relative abundance as a function of time $(\tau=\lambda t)$ in the neutron-capture theory approximation including neutron decay and the universal expansion, according to Alpher and Herman (unpublished). These curves are solutions of Eqs. (108), $J=4$, with the density of matter at the start of the element-forming process $\rho_m=4\times 10^{-6}$ g/cm^3. The neutron decay constant is taken as $\lambda=10^{-3}$ sec.$^{-1}$. The ordinate is actually $\log\xi_i$, where according to Eq. (108d) the ξ_i are concentrations normalized with respect to the concentration of nucleons at the start of the element-forming process. Hence the effect of the universal expansion is evident in the solution.

μ_n and μ_H the magnetic moments of neutron and proton in nuclear magnetons, and ϵ and ϵ_0 are the binding energies of the singlet and virtual triplet states of the deuteron. Equations (128) have been integrated with $x_n(\tau=0)=1$, $x_H(\tau=0)=0$, and with the condition that the final concentration by weight of protons be 0.5, since hydrogen constitutes about 50 percent by weight of all matter. To obtain this final condition [7] one must take $\alpha_G=1$. This integration yields for the density of matter at 1 sec., 4.8×10^{-4} g/cm^3, which is in moderately good agreement with the matter densities obtained in other non-equilibrium calculations.

The non-equilibrium formation of the very light elements in an expanding universe has been examined in greater detail by Fermi and Turkevich.‡ All thermonuclear reactions which are less endothermic than the disintegration of the deuteron and which can go on between neutrons (N), protons (H), deuterons (D), tritons (T), He^3, and He^4 were considered, as well as the radioactive decay of the neutron and triton. The cosmological model chosen was that of a radiation universe containing a relatively small quantity of matter, for which the dependence of temperature on time is given by Eq. (100), namely,

$$T=1.52\times10^{10}t^{-\frac{1}{2}}\ {}^\circ\text{K}.$$

In this model, density or particle concentration varies as $t^{-\frac{3}{2}}$, as shown by Eq. (101). Fermi and Turkevich have assumed that the nucleon concentration was 10^{21} cm^{-3} at 1 sec., so that the nucleon concentration at any t is

$$C_{\text{nuc}}=10^{21}t^{-\frac{3}{2}}\ \text{cm}^{-3}. \tag{129}$$

This corresponds to an assumed matter density of $\sim1.7\times10^{-3}$ g/cm^3 at $t=1$ sec., or $\sim5\times10^{-7}$ g/cm^3 at $t=230$ sec. [Compare Section IV(d)1.]

The 28 reactions considered in detail are listed in Table XII. Examination of the reaction rates for the nuclear processes listed confirmed that until $t\cong300$ sec. the only event of any importance was neutron decay. The high temperature prevents the formation of an appreciable concentration of deuterons, and the nuclei past the deuteron must form through the deuteron, since at the density and temperature under consideration, many-body processes should not be important. A starting time of 300 sec. was therefore selected for the calculation. The initial relative concentrations of

TABLE XII. Reaction rates. [The quantities a_1 and a_2 are defined in Eq. (130), T_8 is the temperature in units of 10^8 °K, $T_8=152t^{-1/2}$ from Eq. (100), and $q_0=10^{21}$ sec.$^{3/2}$ cm^{-3}.]

No.	Reaction	Specific reaction rates	Term in rate equations, $\mathfrak{R}'$ [See Eq. (132)]
1	$\text{N}=\text{H}+e^-$	10^{-3} sec.$^{-1}$	$10^{-3}\mathbf{x}_\text{N}$
2	$\text{N}+\text{H}=\text{D}+h\nu$	6.6×10^{-20} sec.$^{-1}$	$6.6\times10^{-20}q_0\mathbf{x}_\text{N}\mathbf{x}_\text{H}t^{-3/2}$
3	$\text{N}+\text{D}=\text{T}+h\nu$	2.0×10^{-22} sec.$^{-1}$	$2.0\times10^{-22}q_0\mathbf{x}_\text{N}\mathbf{x}_\text{D}t^{-3/2}$
4	$\text{N}+\text{D}=\text{N}+\text{N}+\text{H}$	Negligible (see reaction 18)	0
5	$\text{N}+\text{He}^3=\text{He}^4+h\nu$	10^{-21} sec.$^{-1}$ (estimated)	$10^{-21}q_0\mathbf{x}_\text{N}\mathbf{x}_{\text{He}3}t^{-3/2}$
6	$\text{N}+\text{He}^3=\text{T}+\text{H}$	1.5×10^{-15} sec.$^{-1}$	$1.5\times10^{-15}q_0\mathbf{x}_\text{N}\mathbf{x}_{\text{He}3}t^{-3/2}$
7	$\text{H}+\text{H}=\text{D}+e^+$	$a_1=2\times10^{-39}$; $a_2=3.16$	$7.0\times10^{-41}q_0(\mathbf{x}_\text{H})^2t^{-7/6}10^{-0.592t^{1/6}}$
8	$\text{H}+\text{D}=\text{He}^3+h\nu$	$a_1=8.6\times10^{-21}$; $a_2=3.48$	$3.0\times10^{-22}q_0\mathbf{x}_\text{H}\mathbf{x}_\text{D}t^{-7/6}10^{-0.652t^{1/6}}$
9	$\text{H}+\text{D}=\text{H}+\text{H}+\text{N}$	Negligible (see reaction 18)	0
10	$\text{H}+\text{T}=\text{He}^4+h\nu$	$a_1=1.5\times10^{-19}$; $a_2=3.62$	$5.3\times10^{-21}q_0\mathbf{x}_\text{H}\mathbf{x}_\text{T}t^{-7/6}10^{-0.678t^{1/6}}$
11	$\text{H}+\text{T}=\text{He}^3+\text{N}$	$1.5\times10^{-15}\times10^{-36.8/T_8}$ sec.$^{-1}$	$1.5\times10^{-15}q_0\mathbf{x}_\text{H}\mathbf{x}_\text{T}t^{-3/2}10^{-0.242t^{1/2}}$
12	$\text{D}+\text{D}=\text{He}^4+h\nu$	$a_1=3.07\times10^{-19}$; $a_2=3.99$	$1.08\times10^{-20}q_0(\mathbf{x}_\text{D})^2t^{-7/6}10^{-0.747t^{1/6}}$
13	$\text{D}+\text{D}=\text{He}^3+\text{N}$	$a_1=3.0\times10^{-15}$; $a_2=3.99$	$1.1\times10^{-16}q_0(\mathbf{x}_\text{D})^2t^{-7/6}10^{-0.747t^{1/6}}$
14	$\text{D}+\text{D}=\text{H}+\text{T}$	$a_1=3.0\times10^{-15}$; $a_2=3.99$	$1.1\times10^{-16}q_0(\mathbf{x}_\text{D})^2t^{-7/6}10^{-0.747t^{1/6}}$
15	$\text{D}+\text{T}=\text{He}^4+\text{N}$	$a_1=5.0\times10^{-13}$; $a_2=4.24$	$1.8\times10^{-14}q_0\mathbf{x}_\text{D}\mathbf{x}_\text{T}t^{-7/6}10^{-0.794t^{1/6}}$
16	$\text{D}+\text{He}^3=\text{He}^4+\text{H}$	$a_1=1.5\times10^{-12}$; $a_2=6.72$	$5.3\times10^{-14}q_0\mathbf{x}_\text{D}\mathbf{x}_{\text{He}3}t^{-7/6}10^{-1.259t^{1/6}}$
17	$\text{D}+\text{He}^4=\text{Li}^6+h\nu$	$a_1=1.4\times10^{-21}$; $a_2=6.96$	$4.9\times10^{-23}q_0\mathbf{x}_\text{D}\mathbf{x}_{\text{He}4}t^{-7/6}10^{-1.304t^{1/6}}$
18[a]	$\text{D}+h\nu=\text{H}+\text{N}$	$5.9\times10^{12}T_8^{3/2}10^{-110/T_8}$ sec.$^{-1}$	$1.1\times10^{+16}\mathbf{x}_\text{D}t^{-3/4}10^{-0.723t^{1/2}}$
19	$\text{T}=\text{He}^3+e^-$	1.8×10^{-9} sec.$^{-1}$	$1.8\times10^{-9}\mathbf{x}_\text{T}$
20	$\text{T}+\text{T}=\text{He}^4+\text{N}+\text{N}$	$a_1=2.6\times10^{-13}$; $a_2=4.57$	$9.1\times10^{-15}q_0(\mathbf{x}_\text{T})^2t^{-7/6}10^{-0.856t^{1/6}}$
21	$\text{T}+\text{T}=\text{He}^6+h\nu$	$a_1=2.6\times10^{-19}$; $a_2=4.57$	$9.1\times10^{-21}q_0(\mathbf{x}_\text{T})^2t^{-7/6}10^{-0.856t^{1/6}}$
22	$\text{T}+\text{He}^3=\text{He}^4+\text{N}+\text{H}$	$a_1=1.5\times10^{-12}$; $a_2=7.24$	$5.3\times10^{-14}q_0\mathbf{x}_\text{T}\mathbf{x}_{\text{He}3}t^{-7/6}10^{-1.356t^{1/6}}$
23	$\text{T}+\text{He}^3=\text{He}^4+\text{D}$	$a_1=1.0\times10^{-13}$; $a_2=7.24$	$3.5\times10^{-15}q_0\mathbf{x}_\text{T}\mathbf{x}_{\text{He}3}t^{-7/6}10^{-1.356t^{1/6}}$
24	$\text{T}+\text{He}^3=\text{Li}^6+h\nu$	$a_1=3.1\times10^{-18}$; $a_2=7.24$	$1.1\times10^{-19}q_0\mathbf{x}_\text{T}\mathbf{x}_{\text{He}3}t^{-7/6}10^{-1.356t^{1/6}}$
25	$\text{T}+\text{He}^4=\text{Li}^7+h\nu$	$a_1=5.5\times10^{-19}$; $a_2=7.56$	$1.9\times10^{-20}q_0\mathbf{x}_\text{T}\mathbf{x}_{\text{He}4}t^{-7/6}10^{-1.416t^{1/6}}$
26	$\text{He}^3+\text{He}^3=\text{Be}^6+h\nu$	$a_1=1.4\times10^{-17}$; $a_0=11.49$	$4.9\times10^{-19}q_0(\mathbf{x}_{\text{He}3})^2t^{-7/6}10^{-2.151t^{1/6}}$
27	$\text{He}^3+\text{He}^3=\text{He}^4+\text{H}+\text{H}$	$a_1=1.4\times10^{-11}$; $a_2=11.49$	$4.9\times10^{-13}q_0(\mathbf{x}_{\text{He}3})^2t^{-7/6}10^{-2.151t^{1/6}}$
28	$\text{He}^3+\text{He}^4=\text{Be}^7+h\nu$	$a_1=1.7\times10^{-19}$; $a_2=12.01$	$6.0\times10^{-21}q_0\mathbf{x}_{\text{He}3}\mathbf{x}_{\text{He}4}t^{-7/6}10^{-2.250t^{1/6}}$

[a] The photon concentration is included in the constant.

‡ We are indebted to Drs. E. Fermi and A. Turkevich for their cooperation and communication of unpublished results. The authors take the responsibility for the correctness of this transcription and interpretation of their work.

neutrons and protons selected, namely, 0.70 and 0.30, respectively, correspond approximately to those resulting from free neutron decay starting at $t=0$ sec. One assumes neutrons only to begin with, and a neutron decay constant $\lambda=10^{-3}$ sec.$^{-1}$.

The specific rates taken for each of the reactions are also listed in Table XII, and were obtained as follows. The neutron and triton decay constants were taken as 10^{-3} sec.$^{-1}$ and 1.8×10^{-9} sec.$^{-1}$, respectively, in accordance with experiment. Of the three neutron-capture reactions, (2), (3), and (5), specific reaction rate constants consistent with experiment were assigned to reactions (2) and (3), while a constant was estimated for reaction (5). In all three cases the rate constants are independent of temperature and therefore of time. Many of the reactions in Table XII are of the form

$$X+X'\rightarrow Y+Y'+\text{energy},$$

where X and X' are heavy charged particles, while Y and Y' are either heavy charged particles or a gamma-ray plus a heavy charged particle. The specific rate constant for such thermonuclear reactions may be obtained from the usual expression for thermonuclear reaction rates [61], as¶

$$\mathcal{K}=a_1T_8^{-\frac{2}{3}}10^{-a_2T_8^{-\frac{1}{3}}}\ \text{cm}^3/\text{sec.},\tag{130}$$

where T_8 is the temperature in units of 10^8 °K, a_1, which depends mainly on the reaction probability after penetration, is given by

$$a_1=[(4h\Gamma r_0^2a_2^2\ \ln^2 10)/(3^{5/2}m_re^2Z_1Z_2)]\\ \times\exp(32m_re^2r_0Z_1Z_2/h^2)^{1/2},\tag{130a}$$

while a_2, which depends on the height of the potential barrier for the reaction may be written

$$a_2=3\times10^{-8/3}(\log e)[(\pi^2m_re^4Z_1^2Z_2^2)/(2h^2k)]^{1/3}.\tag{130b}$$

In the above, A_1, A_2, Z_1, Z_2, are the atomic weights and numbers of the reacting nuclei, m_r, the reduced mass, is given by $m_r=m_1m_2(m_1+m_2)^{-1}$ g, the combined radius $r_0\cong1.6\times10^{13}(A_1+A_2)^{\frac{1}{3}}$ cm, and Γ/h is the probability per second for the reaction after penetration of the barrier. Fermi and Turkevich have used for Γ the values given by Bethe [19, 61] or values obtained by procedures consistent with those used by Bethe. As is well known, in the absence of resonances, Γ is about 10^6 times smaller for radiative-capture reactions than for reactions with particle emission only.

TABLE XIII. Relative abundances of the light nuclei.

	Computed	Observed
H^1	1.00	1.00
H^2	1.3×10^{-2}	2×10^{-4}
He^3	2.6×10^{-4}	10^{-7}
He^4	1.5×10^{-1}	10^{-1}

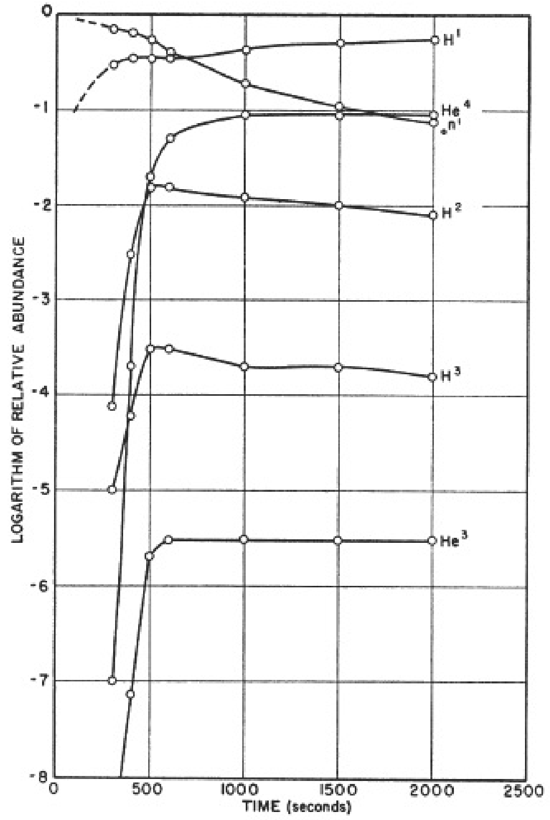


FIG. 20. Relative abundance of the very light elements as a function of time according to the non-equilibrium formulation of Fermi and Turkevich (unpublished). The nucleon concentration was taken to be 10^{21} cm^{-3} at $t=1$ sec. in an expanding universe controlled by radiation. The relative abundances are the ratios of the number of nuclei of a given species in a volume V to the total number of nucleons in that volume. Since both quantities vary with the universal expansion in the same way the effect of the expansion is not evident.

The specific reaction rates can be written in terms of the time as variable instead of the temperature since, from Eq. (100), $T_8=152t^{-\frac{1}{2}}$. In the last column in Table XII terms are given corresponding to the specific reactions as they would appear in rate equations involving concentrations of nuclei by weight, x_j. From Eqs. (110) it is clear that these terms have the form

$$\mathcal{R}=\mathcal{K}m_j(m_{j'}m_{j''})^{-1}\rho_mx_{j'}x_{j''}\ \text{sec.}^{-1},\tag{131}$$

or, using Eq. (101),

$$\mathcal{R}=\mathcal{K}m_j(m_{j'}m_{j''})^{-1}q_0m_0t^{-\frac{3}{2}}x_{j'}x_{j''}\ \text{sec.}^{-1},\tag{131a}$$

where m_0 is the mass of a nucleon, $q_0=10^{21}$ sec.$^{\frac{3}{2}}$ cm^{-3}, and the $\mathcal{K}$ correspond to the p_j previously discussed. This form denotes the contribution to dx_j/dt arising from the reaction between species j' and j'' leading to species j. For example, in the case of reaction (10) in Table XII, namely, $H+T=He^4+h\nu$,

$$\mathcal{K}_{10}=1.5\times10^{-19}T_8^{-\frac{2}{3}}10^{-3.62T_8^{-\frac{1}{3}}},$$

¶ In reference [61], p. 266, the quantity N there is incorrectly given insofar as stated dimensions are concerned. It is (g sec.)$^{-1}$ rather than (cm^3 sec.)$^{-1}$ as stated.

and the term denoted by Eq. (131a) is as given in the third column of Table XII for this reaction. Fermi and Turkevich have found that many of the reactions listed in Table XII can be neglected to a sufficient approximation. The reactions retained are evident from the following set of equations whose simultaneous solutions were obtained by Fermi and Turkevich through numerical integration. Denoting the term in Table XII corresponding to a given reaction by $\mathcal{R}'$, one has

$$d\mathbf{x}_N/dt = -\mathcal{R}_1' - \mathcal{R}_2' + \mathcal{R}_{15}', \qquad (132a)$$

$$d\mathbf{x}_H/dt = +\mathcal{R}_1' + \mathcal{R}_{14}' + \mathcal{R}_{18}' - \mathcal{R}_2', \qquad (132b)$$

$$d\mathbf{x}_D/dt = +\mathcal{R}_2' - \mathcal{R}_{12}' - \mathcal{R}_{13}' - \mathcal{R}_{14}' - \mathcal{R}_{15}', \qquad (132c)$$

$$d\mathbf{x}_T/dt = +\mathcal{R}_{14}' - \mathcal{R}_{15}', \qquad (132d)$$

and

$$d\mathbf{x}_{He^4}/dt = +\mathcal{R}_{15}', \qquad (132e)$$

where $\mathbf{x}_j = x_j/j$ and j is the atomic weight. The quantities $\mathcal{R}'$ are related to the $\mathcal{R}$ in Eq. (131a) by

$$\mathcal{R}' = [j/(j'j'')]\mathcal{R}, \qquad (132f)$$

in which j is the atomic weight of the particular nuclear species whose rate of change is being considered and where j' and j'' are the atomic weights of the two reacting species, respectively. An equation for He^3 is not included for the reason that to a sufficient approximation the concentration $\mathbf{x}_{He^3}$ is maintained at a steady state by the fast reactions (6) and (13), the other reactions involving this species being unimportant. From $d\mathbf{x}_{He^3}/dt = 0$, Fermi and Turkevich have found that to within a factor of ~ 3,

$$\mathbf{x}_{He^3}\mathbf{x}_D^{-2} \cong 2\times 10^{-2},$$

in which it is assumed that $\mathbf{x}_n$ is constant. The factor of 3 arises from the rather small time dependence of the rates in the steady state equations. While tritium balance is maintained almost as well yielding, $\mathbf{x}_T\mathbf{x}_D^{-1} \cong 10^{-2}$, the tritium reactions were considered in detail. The results of the integration of Eqs. (132), subject to the initial conditions already discussed, are given in Fig. 20 where the relative concentrations particle-wise are plotted *versus* the time for the various nuclear species. As can be seen in Table XIII, the computed relative abundances (at $t = 2000$ sec.) may be considered as in agreement with those observed for these species, in view of the fact that very light element abundances are not well known and would vary in different locales because of the participation of these species in thermonuclear reactions after the element-forming epoch. The computed relative abundances of H^3 and He^3 have been added together because of the radioactivity of the former. The observed relative abundances are obtained from Brown (Table III of reference [30]) and corrected for isotopic abundance ratios [129].

Fermi and Turkevich have also examined the problem of forming the light elements heavier than He^4. The only reactions involved are capture reactions since there are no exothermic reactions giving heavy particles. Reactions (17), (21), and (25) are of this kind, with (17) and (21) giving Li^6, the latter by β-decay of He^6, while (25) yields Li^7. Under the conditions discussed, and assuming no resonances, these reactions are very slow and lead to an insufficient amount of material, about 10^{-7} by weight, past He^4. The existence of a resonance in reaction (25) might repair this difficulty. In this case Fermi and Turkevich have considered how close a resonance would have to be in order to get an appreciable conversion of He^4 and H^3 to Li^7. A detailed examination of this reaction under the conditions discussed indicates that a resonance would have to be at about 400 kev or closer in order to convert any appreciable amount of the material into Li^7. At the present time the first observed level is at about 4 Mev. Turkevich has also considered this reaction (25) under a different set of initial conditions, namely, a nucleon concentration of 10^{23} instead of 10^{21} cm^{-3}, and a neutron-proton ratio of 6 to 1, both at $t=1$ sec. Non-resonance processes with these initial conditions lead to 3×10^{-4} by weight of Li^7, which is much closer to what is required and it is concluded that even a resonance closer than 1 Mev would be interesting. To the best of the authors' knowledge this reaction has not been studied directly. Another possibility first proposed by Wigner‖ for building appreciable amounts of elements past He^4 involves the idea of a "seed" nucleus. An example of an exothermic chain reaction involving a seed nucleus, studied by Turkevich, is

$${}_6C^{10} + {}_1H^3 \rightarrow {}_3Li^6 + {}_4Be^7 + 2\text{ Mev}.$$

If the two product nuclei would again build up to C^{10}, then one has a method of forming appreciable amounts of nuclei past the gap. However, since C^{10} is neutron deficient it is difficult to see, in this particular case, how it could be re-formed from the product nuclei. As pointed out by Gamow [60] there may exist other possible reactions of this type in which the "seed" nucleus would have a neutron rather than a proton excess. The difficulty of finding a scheme to bridge the non-existent nuclei at $A=5$ and 8 is discussed again later. In any event, neutron capture should be the predominant reaction at the temperatures considered and should certainly be responsible for building the elements past the first eight or ten.

Another question of interest in connection with the non-equilibrium formation of elements is the possible participation of some of the light elements in thermonuclear reactions in the expanding universe after the completion of the initial element forming process. As has been seen in previous discussion, the temperature after the element forming process was still quite high so that thermonuclear reactions between light nuclei and protons could go on. This problem has been studied

‖ See reference [60].

by Alpher, Herman, and Gamow [10] in the following manner. An examination of the relative abundance data suggests that the abundances of certain of the light elements such as Li, Be, and B have been markedly decreased since the "original" formation process. It is to be noted that these elements have relatively large cross sections for proton reactions [19].

It can be shown that for certain of the light elements the difference between the present relative abundance and the abundance computed according to the neutron-capture theory [10] is consistent with known thermonuclear reaction rates and with cosmological information furnished by the theory. Making use of Eq. (130) one may write for the number of thermonuclear reactions per gram of matter per second,

$$r_{\text{th}}=\mathcal{K}x_jx_p(m_jm_p)^{-1}\rho_m(t), \qquad (133)$$

where ρ_m is the density of matter, x_j and x_p the concentrations by weight of the reacting species j and protons, of mass m_j and m_p, respectively. In a manner analogous to that used in obtaining Eq. (110c) one finds

$$d[\ln x_j]/dt=-r_{\text{th}}m_j/x_j. \qquad (134)$$

In this equation ρ_m and T are taken to be defined as functions of time according to Eqs. (101) and (100). Assuming the concentration by weight of protons to be a constant, $x_p\cong 0.5$, one obtains for the ratio, α_R, of the observed relative abundance to that computed according to the neutron-capture theory,

$$\ln\alpha_R=B_1[I(t_0)-I(t_P)], \qquad (135)$$

in which

$$B_1=(152)^{-\frac{1}{3}}a_1(m_jm_p)^{-1}\rho_0, \qquad (135a)$$

$$I(t)=t^{-1/6}\exp(-B_2t^{1/6})+B_2Ei(-B_2t^{1/6}) \qquad (135b)$$

and

$$B_2=(152)^{-\frac{1}{3}}(\ln 10)a_2. \qquad (135c)$$

The quantity ρ_0 is the matter density at $t=1$ sec. ($\rho_m=\rho_0t^{-\frac{3}{2}}$) and a_1 and a_2 are as given in Eqs. (130). One can find $\log\alpha_R$ as the difference between the logarithms of the observed and the computed relative abundances directly from the datum points and curve in Fig. 17. The time t_0 represents the time at which the proton reactions became important while t_P refers to the present epoch. The reaction probabilities Γ in a_1 are tabulated by Bethe [19], and Gamow and Critchfield [61] for the reactions of interest.

Applying Eqs. (135) to the reactions of Li, Be, and B with protons, one finds that if t_0 is of the order of 10^3 sec. then the present relative scarcity of these elements can be explained. This value of t_0 is consistent with the time estimated for the cessation of neutron-capture processes. The same analysis applied to other light elements such as F^{19}, which are now scarce, yields similar results since these elements have high cross sections for proton reactions. On the other hand, those light elements which have relatively low cross sections for proton reactions are now found to lie approximately on the computed abundance curve. For example, in the cases of C and N there is no appreciable depletion. In fact, taking $t_0\cong 10^3$ sec. for N^{14} leads to depletion due to thermonuclear reactions of only one part per million up to the present epoch. In the foregoing treatment it has been assumed that the time dependences of ρ_m and T given by Eqs. (101) and (100) are a sufficient approximation for the problem since the main part of the depletion occurs during the period when these expressions are valid.

3. *Effects of Nuclear Stability*

In the theory of the neutron-capture process presented thus far, it has been assumed that the time between successive neutron captures was sufficiently long to allow any necessary adjustment of charge by β-decay. Clearly the validity of this assumption depends upon the density of the reacting material and upon the β-decay rates of the nuclei participating in the process. Smart [134, 136] has recently examined this problem in detail. One may consider the effect of β-decay on the neutron-capture process in three situations differentiated according to the density of matter. In the case of very low density, the time between successive neutron captures will in general be long enough to allow β-decay between captures for all species, and the capture reaction rates will control the formation processes. The nuclei involved will be stable or have one excess neutron. For a very high density, on the other hand, the reacting, nuclei would probably have as high a neutron excess as is consistent with neutron binding, and the rate of growth of particular species would be principally determined by their β-decay rates in these neutron-rich states. Finally, at intermediate densities one would expect a competition between neutron capture and β-decay processes.

The case of low density can be dismissed readily. Clearly it is meaningless to speak of a non-equilibrium process of successive neutron captures if the mean time between successive captures is appreciable as compared with the neutron lifetime. It can be shown that the upper limit to this low density case is of the order of 10^{-11} g/cm³. If one equates the neutron lifetime, $\sim$1000 sec., to the mean time between captures, then one has $1000=m_0(\rho_l\sigma v)^{-1}$, where m_0 is the nucleon mass, ρ_l the limiting density, σ the neutron-capture cross section for a particular species, and v the relative velocity, $\sim 4\times 10^8$ cm/sec. for 0.1 Mev. Taking $\sigma\cong 10^{-24}$ cm²

TABLE XIV. Decay constants for neutron-saturated nuclei.

A	$(T_{\frac{1}{2}})_m$	W_β (in m_ec^2)	$\gamma(Z,W_\beta)$	λ_β (sec.$^{-1}$)
50	10.8	17.3	1.37	6.9
64	13.3	16.5	1.55	7.4
100	19.9	16.3	1.93	13.5
125	24.5	14.7	2.32	11.8
180	34.5	11.6	3.48	7.6

Appendix A

Papers relevant to Fermi's Italian period

A.1 D. Bini, A. Geralico, R.T. Jantzen and R. Ruffini: On Fermi's resolution of the "4/3 problem" in the classical theory of the electron and its logical conclusion: hidden in plain sight

Abstract. We discuss the solution proposed by Fermi to the so called "4/3 problem" in the classical theory of the electron, a problem which puzzled the physics community for many decades before and after his contribution to the discussion. Unfortunately his early resolution of the problem in 1922–1923 published in three versions in Italian and German journals (after three preliminary articles on the topic) went largely unnoticed, and even recent texts devoted to classical electron theory still do not present his argument or acknowledge the actual content of those articles. The calculations initiated by Fermi at the time are finally brought to their logical physical conclusion here.

Introduction

The simplest classical model of the nonrotating electron in special relativity consists of a static spherically symmetric distribution of total electric charge e over the surface of a rigid sphere of radius r_0, as measured by an observer at rest with respect to the sphere. This model was first developed and studied during the first decade of the 1900s by Abraham [1], Lorentz [2] and Poincaré [3], based entirely on Maxwell's theory of electromagnetism. For an unaccelerated electron, the rest frame integral of the local energy density of the Coulomb field over the exterior of the electron sphere representing the total energy $W = e^2/(2r_0)$ stored in that field, is equal to the self-energy of the charge distribution. For any static isolated configuration of charge, this self-energy is equal to the work needed to assemble it by slowly bringing the charge elements in from spatial infinity.

The factor of 1/2 in the energy formula is a geometric factor which is replaced by 3/5 if the model of the electron is a constant charge density solid sphere rather than a constant density spherical surface charge distribution and one also considers the additional contribution to the electromagnetic field energy inside the sphere (zero in the surface distribution case by spherical symmetry): $1/2 + 1/10 = 3/5$. Dropping these factors and converting the Coulomb energy to the entire observed mass m_e of the electron by Einstein's famous mass-energy relation $E = mc^2$ defines a corresponding radius $r_{\rm e} = e^2/(m_e c^2)$ that pure dimensional analysis would lead to, called the classical radius of the electron.

With the birth of special relativity occurring during the same time years as the Abraham-Lorentz model development, there was the expectation that apart from any additional "bare mass" m_0 that the electron might have, the electromagnetic energy W should contribute to the inertial mass of the electron an amount $m_{\rm em} = W/c^2$ satisfying Einstein's mass-energy relation, leading to a total mass $m_e = m_0 + m_{\rm em}$. Instead they had found $m_{\rm em} = \frac{4}{3}W/c^2$ in the limit of nonrelativistic accelerated motion of the electron. This became the famous "4/3" problem.

After three preliminary papers on the inertial and gravitational mass of electromagnetic fields in 1921–1922 (Fermi 1–3, [4–6]), in 1923 Fermi (Fermi 4c [7]) reconsidered this problem for any regular spherically symmetric distribution of charge in motion that satisfies Born's definition of relativistic rigidity [8], namely that this distribution is time-independent in its instantaneous rest frame. Re-examining the Abraham-Lorentz derivation of the inertial mass of such a distribution of charge due entirely to its self-field, Fermi managed to correct the troublesome factor of 4/3 in their result which he showed is entirely due to their imposition of conventional rigidity with respect to a single inertial frame instead of the sequence of instantaneous rest frames following Born's criterion. The former is not a Lorentz invariant condition like Born's and so is in direct conflict with special relativity.

By an unfortunate coincidence the same numerical factor of 4/3 appears in the integral definition of the total 4-momentum observed by any inertial observer moving relative to an (unaccelerated) spherically symmetric charge distribution which is time-translation symmetric in its own inertial rest frame. Contracting the stress-energy tensor of the electromagnetic field due to such a distribution with the 4-velocity of any inertial observer gives the local 4-momentum distribution as seen by that observer, and integrating it over a time slice in that observer's reference frame gives the total 4-momentum seen by that observer at that moment of inertial time. Since it arises as the hypersurface integral of a second rank tensor field (invariant under translation in the rest frame of the charge distribution), this quantity is a linear 4-vector-valued function of the 4-velocity of the inertial observer and indeed the 4/3 factor enters because of the tensor transformation law. In the absence of sources, the 4-momentum of the electromagnetic field is actually independent of the observer, as shown by textbook applications of Gauss's law to the divergence-free stress-energy tensor, but in the presence of sources, this divergence is nonzero and leads to the complications encountered in this problem that of course were not understood in the early days of special relativity.

Kwal in 1949 [9] and later independently Rohrlich [10] in 1960 made the observation that by fixing the 4-velocity of the inertial observer in this calculation to be the one associated with the rest frame of the unaccelerated charge distribution, one obtains a fixed 4-momentum independent of time which equals the rest frame 4-momentum by definition and again the troublesome factor of 4/3 disappears. Unfortunately this is not the end of the story: the classical theory of charge distributions and electromagnetic self-forces and radiation reaction forces is a complicated and controversial subject into which many have entered the discussion over the past century since it began, and Fermi's own contribution has been largely ignored.

Indeed the Fermi coordinates and Fermi-Walker transport for which Fermi is well known in relativity were developed specifically in 1922 to treat this problem (Fermi 3, [6]) while he was a university student already knowledgeable in general relativity only a few years after its birth in 1916. In that very paper in its final section he considers the Lagrangian for an extended charged body with a given charge and mass distribution moving in an external electromagnetic field, where the distributions are confined to a length scale that in his subsequent paper will be assumed to be small compared to the variation of the external field. In that next paper he focuses only on the contribution of the charge distribution to its equation of motion in the external field, but one can easily retain the mass contribution as well, as in many discussions of this problem, where this mass is referred to as the bare mass or mechanical mass of the object. The result is the Lorentz force law with the inertial rest mass contribution to that equation consisting of the sum of the bare mass and the electromagnetic energy in the self-field of the charge distribution, the latter energy not preceded by the famous 4/3 factor of Abraham and Lorentz. Fermi's derivation in this larger context is discussed in detail in the textbook on special relativity by Aharoni [11] who came out with his second edition in 1965 specifically to include this part as explained in his preface, after attention had been called to the problem by Rohrlich in 1960.

Following the analysis by Abraham and then Lorentz of the accelerated version of their model for the electron, Fermi considered a regular spherically symmetric distribution of accelerated charges held in a rigid configuration by some external force and applied the Lagrangian variational principle to compute the time rate of change of the momentum in the force law, without specializing to a particular charge density profile. In order to show exactly how the mistaken 4/3 factor in the inertial mass due to the energy of the self-field arises, Fermi contrasted the Born rigid calculation of the Lagrangian variation (variation B) with that assuming rigidity with respect to a particular inertial observer (variation A), which is not relativistically invariant and hence not to be trusted. He showed that the latter assumption in deriving the equations of motion leads to the Abraham-Lorentz result with the mistaken 4/3 factor multiplying the electromagnetic energy of the charge distribution in its contribution to the inertial mass, but that the Born assumption gives the correct factor as expected by the equivalence of mass and energy through the famous equation $E = mc^2$.

Operationally, a congruence of timelike world lines is said to be Born-rigid if it has vanishing expansion. As discussed in detail by Salzman and Taub [12] in 1954, any timelike curve determines a family of orthogonal hyperplanes in special relativity and their orthogonal trajectories define the world lines of a body in Born-rigid motion (referred to as planar motion by Herglotz and Noether). The remaining class of motions are called group motions, and consist of curve segments from a continuous 1-parameter subgroup of Lorentz transformations of Minkowski spacetime into itself. The best example of these are the Rindler observers whose world lines are the integral curves of a single generator of Lorentz transformations, each world line with a unique constant acceleration. For the electron model, a time-independent spherically symmetric distribution of charge in the Fermi coordinate system adapted to the central world line is Born rigid.

In order to fix the 4/3 problem Poincaré [3] (followed up by von Laue [13]) seriously confused the issue by mixing it with the question of explaining the rigid configuration of charge through internal stresses. Long after Fermi's resolution of the 4/3 problem, even in the commentary by his friend Persico on Fermi's paper in the collected work of Fermi, it was thought that Poincaré stresses were necessary to explain this discrepancy. In fact the stability of the electron is an entirely different matter from the correct relation of the inertial mass to the electromagnetic energy as explained by Fermi.

Although Wilson [14] discussed the problem of the proper definition of the 4-momentum of the electromagnetic field in 1936 with no citations, he did not succeed in clarifying matters. In 1949 Kwal [9] showed that a slight modification of Abraham's original integral definitions for the unaccelerated electron leads to an electromagnetic 4-momentum endowed with the correct Lorentz transformation properties. Even later Rohrlich [10] in 1960 came to the same conclusion without being aware of previous work. They both explained that the correct result can only be obtained from the usual special relativistic integrals over a hypersurface of constant inertial time if that hypersurface represents a time slice in the rest frame of the electron, although Kwal only discussed changing the element of hypersurface volume without relating the region of integration to that rest frame. The classical electron model has continued to intrigue people ever since, see for example, Feynman [15], Teitelboim [16–17], Boyer [18], Rohrlich [19], Nodvik [20], Schwinger [21], Campos and Jimenéz [22], Cohen and Mustafa [23], Comay [24], Moylan [25], Kolbenstvedt [26], Rohrlich [27], Appel and Kiessling [28], de Leon [29], Harte [30], Pinto [31], Bettini [32], Galley et al [33], and more. At least three entire books are devoted to the topic of the classical theory of the charge distributions, those by Rohrlich [34], Yaghjian [35], and Spohn [36], and the model is described in detail by Jackson [37], the universally accepted reference textbook on classical electrodynamics (see also Chapter 8 of Anderson [38]). Some interesting historical details may be found in the recent article of Janssen and Mecklenburg [39]. This whole problem is not without explicit controversy, as detailed by Parrott in his archived exchange with *Physical Review* which would not publish his criticism of Rohrlich's recent work [40].

Except for Aharoni [11] and much later Kolbenstvedt [26] in 1997, and for Nodvik [20] and Appel and Kiessling [28] who consider a spinning generalization of the relativistically rigidly rotating electron model reviewed by Spohn [36], none of these references seem to take into account Fermi's actual argument nor connect it to that of Kwal and Rohrlich even though most of them cite Fermi's original article. Kolbenstvedt [26] called attention to Fermi's argument with a slightly different but equivalent explanation of his own, and not in an obscure physics journal, and yet the latest edition of the books of Jackson, Rohrlich, and Yaghjian, all published after that year still do not reflect this news. Jackson does explain that his nonrelativistic treatment can be relativistically corrected, referring to Fermi, and to be fair, the stated purpose of Yaghjian was to update the Abraham-Lorentz model which he did, apparently unaware of the content of Fermi's articles. Misner, Thorne and Wheeler's tome *Gravitation* [41], affectionately known as MTW, really raised the level of mathematical discussion of special and general relativity after 1973, and allowed Spohn to more cleanly and covariantly discuss the relativisitically rigid electron model to include spin, but without discussing the observer-dependent 4-momentum integral for the electromagnetic self-field.

An important element of this discussion is the conserved nature of the integrals of the local densities of energy and momentum associated with the divergence-free stress-energy tensor of the sourcefree electromagnetic field when integrated over an entire spacelike hyperplane of Minkowski spacetime due to Gauss's law. Such a conservation law fails to exist when the divergence is instead nonzero in the presence of sources or if a world tube containing sources is excluded from the integral, leading either to a spacetime volume divergence integral or (equivalently) to an internal boundary integral that must be taken into account in Gauss's law. This is an important discussion since none of the textbooks on special or general relativity describe this more general situation, while textbooks on classical electrodynamics typically only use local such integrals within bounded regions of space.

Since this discussion is crucial in understanding the present problem, it is included in the section following this introduction where the preliminary details about the electromagnetic field needed to consider the spherical model of an unaccelerated electron are introduced together with the definitions of the 4-momentum in the field as observed by any inertial observer, and the role played by Gauss's law in conservation laws is then explained, leaving the details of more exotic regions of spacetime integration to the appendix. The calculation of the 4-momentum integrals for the Abraham-Lorentz model of the unaccelerated electron is then reproduced in the subsequent section to explain the role played by Kwal and Rohrlich in this matter. Next we present Fermi's re-analysis of the Abraham-Lorentz calculation of the inertial mass for their model of the accelerated electron taking into account Born's rigidity condition. Finally the Kwal-Rohrlich definition of 4-momentum is related directly back to this correction using Gauss's law.

One finds that the Kwal-Rohrlich restriction of the observer-dependent electromagnetic field 4-momentum integrals to the electron rest frame time hyperplanes associates a unique 4-momentum with the unaccelerated electron which is the one special relativity assigns, which has long been known. However, for a single static electron configuration in the absence of interaction, the 4-momentum is not so interesting since there is no way even of revealing its inertial mass from at most uniform translational motion in flat spacetime. To get information about the inertial mass and 4-momentum, the electron must be accelerated and if we limit our attention to electromagnetic interactions, it will be accelerated by an external electromagnetic field through the Lorentz force law. We expect that the total momentum of the electron and the electromagnetic field (for a closed system) should be conserved.

We will show here for the first time that indeed the logical conclusion of Fermi's calculation of the lowest order contributions to the equations of motion of the electron is that the total 4-momentum as observed in the time slices in the sequence of instantaneous rest frames along its path is conserved, i.e., is independent of time, and is the usual one we associate with the system. The key idea of Fermi of the importance of this sequence of hyperplanes orthogonal to the path of a given world line in spacetime was imbedded in his Fermi coordinate system adapted to that world line and which outlived the purpose for which he introduced it in those initial days of the theory of general relativity.

Electrodynamic preliminaries

Although Fermi does not specify the density profile of the spherically symmetric charge distribution that he analyzes in his re-examination of the earliest classical electron theory proposed by Abraham [1] and improved by Lorentz [2], he refers specifically to their spherical shell model of the electron in his introduction. Without acceleration of the electron this model cannot help identify the inertial mass which arises as the proportionality constant between the applied force and the resulting acceleration. However, it was the interest in their unaccelerated model which helped push towards the understanding of the 4-momentum hypersurface integrals for the electromagnetic field so it is useful to review this case first. We re-examine their work in light of modern notation and perspective.

The model for the electron first proposed by Abraham [1] and improved by Lorentz [2] consisted of a uniform spherically symmetric distribution of total electric charge e over the surface of a rigid sphere of radius r_0 in its rest frame. This was called the contractile electron since it would then undergo Lorentz contraction with respect to an inertial frame in relative motion, while Abraham had assumed that the electron remained a rigid sphere with respect to all inertial observers. Einstein's understanding of special relativity only came after this model had been developed, and Lorentz had interpreted the Lorentz contraction as a dynamical effect rather than as a universal property of spacetime itself. They attempted to explain the mass-energy of the electron as due wholy to the electromagnetic field of the electron, equating the electron's energy and momentum to the energy and momentum of its electromagnetic field, which can be evaluated by suitably integrating the normal components of the stress-energy tensor of the electromagnetic field over a spacelike hyperplane representing a moment of time in an inertial reference frame. This is a useful example to keep in mind.

In an inertial system of Cartesian coordinates $(x^\mu) = (t = x^0, x^1, x^2, x^3)$ associated with an inertial reference frame in Minkowski spacetime with signature $(-{+}{+}{+})$ following the conventions of Misner, Thorne and Wheeler [41] with $c = 1$, Maxwell's equations for the electromagnetic field tensor $F_{\alpha\beta}$ due to the 4-current density J^α are

$$F^{\alpha\beta}{}_{,\beta} = 4\pi J^\alpha\,, \quad F_{\alpha\beta,\gamma} + F_{\beta\gamma,\alpha} + F_{\gamma\alpha,\beta} = 0\,, \tag{1}$$

where Greek indices assume the values $0, 1, 2, 3$, and Latin indices instead $1, 2, 3$. Indices may be raised and lowered with the flat Minkowski spacetime metric whose inertial coordinate components are $(g_{\alpha\beta}) = \mathrm{diag}(-1, 1, 1, 1) = (g^{\alpha\beta})$.

Of course when these equations are expressed in noninertial coordinate systems the comma here signifying partial coordinate derivatives $f_{,\alpha} = \partial_\alpha f = \partial/\partial x^\alpha$ must be replaced by the semicolon indicating the components of the covariant derivative. We will have need later for an arbitrary covariant constant covector field Q_α of vanishing covariant derivative $Q_{\alpha;\beta} = 0$, the components of which reduce to $Q_{\alpha,\beta} = 0$ in an inertial coordinate system where the components Q_α (and Q^α) are actual constants. In fact such covariant constant vector fields Q^α correspond to the translational Killing vector fields of Minkowski spacetime, which are special solutions of the general Killing equations that the symmetrized covariant derivative $Q_{(\alpha;\beta)} = 0$ vanish. The noncovariant constant Killing vectors generate the rotations and boost symmetries of Minkowski spacetime.

The stress-energy tensor of the electromagnetic field

$$T^{\mu\nu}_{\mathrm{em}} = \frac{1}{4\pi}\left(F^{\mu\alpha}F^\nu{}_\alpha - \frac{1}{4}g^{\mu\nu}F^{\alpha\beta}F_{\alpha\beta}\right) \tag{2}$$

has the following explicit inertial coordinate components

$$\begin{aligned}
T^{00}_{\mathrm{em}} &= \frac{1}{8\pi}(E^2 + B^2) = U_{\mathrm{em}}\,,\\
T^{0i}_{\mathrm{em}} &= \frac{1}{4\pi}(E \times B)^i = S^i\,,\\
T^{ij}_{\mathrm{em}} &= \frac{1}{4\pi}[-E^iE^j - B^iB^j + \frac{1}{2}g^{ij}(E^2 + B^2)]\,,
\end{aligned} \tag{3}$$

where U_{em} and S are the electromagnetic energy density and the Poynting vector respectively, and of course E and B are the usual electric and magnetic fields observed in the associated reference frame in index-free notation, with nontrivial inertial coordinate components $E^i = F^{0i} = F_{i0}$ and $B^1 = F_{23}$ etc. In general if u^α is the 4-velocity of an observer at a point of spacetime, the electric field as seen by that observer there is $E(u)^\alpha = F^\alpha{}_\beta u^\beta$. In a system of inertial coordinates adapted to that observer, then $u^\alpha = \delta^\alpha{}_0$, so that one has $E(u)^\alpha = F^\alpha{}_\beta\delta^\beta{}_0 = F^\alpha{}_0 = \delta^\alpha{}_i F^i{}_0$ since due to the change of sign under index raising and the antisymmetry of the field tensor $F^0{}_0 = -F^{00} = 0$. Note that in inertial coordinates associated with a second inertial observer in relative motion to a given 4-velocity u^α, its components are given by $(u^\alpha) = (\gamma, \gamma v^i)$, where v^i are the components of the relative velocity of the first observer and $\gamma = (1 - v^i v_i)^{-1/2}$ is the associated gamma factor.

The divergence of this stress-energy tensor in inertial coordinates is easily calculated using Maxwell's equations

$$T^{\mu\nu}_{\mathrm{em}\,,\nu} = -F^\mu{}_\nu J^\nu\,, \tag{4}$$

as shown by Exercise 3.18 of Misner, Thorne and Wheeler [41], for example. Thus in source-free regions where the 4-current $J^\mu = 0$ vanishes, this divergence is zero, which is the condition needed to obtain a conserved 4-momentum for the free electromagnetic field in textbook discussions using Gauss's law. When the the 4-current density $J^\alpha = \rho U^\alpha$ is due to the motion of a distribution of charge moving with 4-velocity field U^α and rest frame charge density ρ, then this divergence has the value

$$T_{\mathrm{em}}^{\mu\nu}{}_{,\nu} = -\rho F^\mu{}_\nu U^\nu = -\rho E(U)^\mu\,, \tag{5}$$

which apart from the sign is the 4-force density exerted by the electromagnetic field on the charge distribution, expressable as the product of the charge density and the electric field in the rest frame of the moving charge. This divergence plays a crucial role in the Lagrangian equations of motion of the electron and in the conservation or not of the 4-momentum of the electromagnetic field. Unlike the 4-momentum of a particle which is locally defined and independent of the observer (but whose components depend on the choice of coordinates of course), the 4-momentum of the electromagnetic field is nonlocal and can only be defined at a momentum of time with respect to some inertial observer through an integral over an entire hyperplane Σ of spacetime corresponding to the extension of the local rest space of that observer at that moment. In the presence of sources $J^\alpha \neq 0$, this 4-momentum not only generally depends on the time for nonstationary sources, but also on the choice of observer, since there is no a priori reason to expect integrals over different regions of spacetime to agree. When instead $J^\alpha = 0$ as is the case for a free electromagnetic field, a conservation law applies due to the vanishing divergence and if those integrals are finite, they in fact all define the same 4-momentum vector on Minkowski spacetime.

The components of the 4-momentum of the electromagnetic field as seen by an inertial observer with 4-velocity u^α at a moment of time t in the observer rest frame represented by a time coordinate hyperplane Σ (for which u^α is in fact the future-pointing unit normal vector field) is given by the integral formula

$$P(\Sigma)^\alpha = \int_\Sigma T_{\mathrm{em}}^{\alpha\beta} d\Sigma_\beta\,, \tag{6}$$

where one can integrate over an object with a free index only if that index is expressed in some inertial coordinate system where it makes sense to compare 4-vectors at different spacetime points in the flat spacetime due to the path independence of parallel transport. The contracted pair of indices can be evaluated in any coordinates. For any spacelike hyperplane Σ with future-pointing timelike unit normal u^α, the hyperplane volume element

$$d\Sigma_\alpha = -u_\alpha dV_\Sigma \tag{7}$$

and induced volume element dV_Σ are most easily evaluated in inertial coordinates (t, x^i) adapted to the observer with 4-velocity u^α, where $u^\alpha = \delta^\alpha{}_0$ while $u_\alpha =$

$-\delta^0{}_\alpha$ and $dV_\Sigma = dx^1 dx^2 dx^3$, while the spacetime volume element is simply $d^4V = dt\, dV_\Sigma$. The minus sign $-1 = u^\alpha u_\alpha$ in $d\Sigma_\alpha$ is needed to pick out the future normal component $X_u = -X^\alpha u_\alpha$ of a vector field in its integral

$$\int_\Sigma X^\alpha d\Sigma_\alpha = -\int_\Sigma X^\alpha u_\alpha dV_\Sigma = \int_\Sigma X_u dV_\Sigma\,. \tag{8}$$

In an inertial system of coordinates the above integral then has the components

$$P(\Sigma)^0 = \int_\Sigma T^{00}_{\rm em} dV_\Sigma\,, \quad P(\Sigma)^i = \int_\Sigma T^{0i}_{\rm em} dV_\Sigma\,, \tag{9}$$

which represents the integral of the local density of energy and momentum in the field as seen by the associated inertial observer.

For a given fixed choice of hyperplane Σ, the above integral formula (6) for the 4-momentum defines a unique 4-vector whose components can be evaluated in (Cartesian) inertial coordinates with respect to any other inertial observer, resulting in a Lorentz transformation of those components. However, if the hypersurface is changed, the result is a different 4-vector, unrelated to the original one by any simple transformation.

Only in the special case of a divergence-free stress energy tensor is the result actually independent of the hypersurface because of Gauss's law, and so defines a single 4-vector no matter what time slice or what inertial observer is chosen. When the components of this single 4-vector are transformed from one system of inertial (Cartesian) coordinates to another, they then transform according to the associated Lorentz transformation. Perhaps influenced by this atypical special case, early on there was the expectation that this should be the situation in general when sources are present which make the divergence of the electromagnetic stress-energy tensor nonzero, but this was a completely unjustified expectation.

Since Gauss's law is so essential to this question, it is crucial to have its application understood before embarking on the details of the classical model of the electron. We consider the 4-dimensional spacetime region R bounded by two hyperplanes Σ_1 and Σ_2 each representing a moment of time with respect to some inertial observer and each oriented by its future-pointing unit normal vector field, a constant vector field which represents the 4-velocity of the observer. These hyperplanes are parallel for the same inertial observer and hence do not intersect, with one in the future of the other, but they do intersect for two observers in relative motion, in which case one has to be careful about the signs in the two disjoint contributions to the 4-dimensional integral relative to the future-pointing normals of the hyperplanes, since the future halves of each hyperplane switch passing from one to the other across the 2-plane of their intersection. The appendix discusses these details.

In its metric form rather than its metric-independent form involving only differential forms, Gauss's law in Minkowski spacetime only applies to the integral of a vector field over the bounding hypersurface of a region R of spacetime, equating the integral of its divergence over R with respect to the spacetime volume element to the hypersurface integral of the outward normal component of the vector field with respect to the induced or intrinsic volume element on the hypersurface. Suppose Σ_1 and Σ_2 do not intersect, and Σ_2 is to the future of Σ_1. Then provided that the integral of the boundary at spatial infinity which closes the boundary between these two hyperplanes can be neglected due to the fall-off properties of the vector field there, Gauss's law states that

$$\int_R \mathcal{J}^\beta{}_{;\beta} d^4V = \int_{\Sigma_2} \mathcal{J}^\beta d\Sigma_\beta - \int_{\Sigma_1} \mathcal{J}^\beta d\Sigma_\beta \,, \tag{10}$$

where the negative sign proceeds the second integral since its future pointing normal is not outward but inward.

If the divergence $\mathcal{J}^\beta{}_{;\beta} = 0$ vanishes, then

$$\int_{\Sigma_2} \mathcal{J}^\beta d\Sigma_\beta = \int_{\Sigma_1} \mathcal{J}^\beta d\Sigma_\beta \,, \tag{11}$$

so the integral is the same for these two parallel hyperplanes and so is independent of the moment of time for this single inertial observer. To extend this "conservation law" to any two inertial observers in relative motion, we just need to be careful about the signs of the orientations of the interior and bounding hyperplanes in the two disjoint regions and pairs of boundaries into which their intersection divides them. However, if the divergence is zero, this is all irrelevant and one again finds that the two integrals are the same, and hence the result is independent of the choice of spacelike hyperplane, giving the same result for all observers and all moments of their time. When the divergence is nonzero, the two integrals differ by a nonzero amount which depends on the region of integration and hence in general one finds a different result for every inertial observer and every moment of their time.

Gauss's law can be applied to a second rank symmetric tensor $T^{\alpha\beta}$ only by contracting it with a covector Q_α to form a vector field $\mathcal{J}^\beta = Q_\alpha T^{\alpha\beta}$, so introduce a covariant constant such covector, in terms of which the divergence becomes

$$\mathcal{J}^\beta{}_{;\beta} = Q_\alpha T^{\alpha\beta}{}_{;\beta} \,. \tag{12}$$

We then get the result

$$\int_R \mathcal{J}^\beta{}_{;\beta} d^4V = \int_{\Sigma_2} Q_\alpha T^{\alpha\beta} d\Sigma_\beta - \int_{\Sigma_1} Q_\alpha T^{\alpha\beta} d\Sigma_\beta \,. \tag{13}$$

If we agree to evaluate these expressions in inertial coordinates where Q_α are constants, then they can be factored out of the equation and one gets a relation involving the 4-momentum as seen by the corresponding inertial observers

$$\int_R T^{\alpha\beta}{}_{;\beta} d^4V = \int_{\Sigma_2} T^{\alpha\beta} d\Sigma_\beta - \int_{\Sigma_1} T^{\alpha\beta} d\Sigma_\beta = P(\Sigma_2)^\alpha - P(\Sigma_1)^\alpha \,, \tag{14}$$

or using Eq. (5) for the electromagnetic field we get

$$P(\Sigma_2)^\alpha - P(\Sigma_1)^\alpha = -\int_R \rho E^\alpha(U) d^4V \,. \tag{15}$$

Thus if the divergence is nonzero, as occurs for the electromagnetic field in the presence of sources, the two 4-momenta differ by a quantity that depends on the region of integration, so there is no common agreement among inertial observers about the 4-momentum in the field, nor is the result independent of time for a single inertial observer. This is the source of the complication for defining the 4-momentum of the electromagnetic field in the classical model of the electron.

A covariant constant vector field is a Killing vector generating translational symmetries of Minkowski spacetime from which the conservation of linear momentum follows for translation invariant Lagrangians according to Noether's theorem. The arbitrary translational Killing vector field Q^α allows us to pick out the components of linear momentum. A general Killing vector field satisfies the condition that its symmetrized covariant derivative vanish $Q_{(\alpha;\beta)} = 0$. If instead we use a nontranslational Killing vector field in the above argument, then since the stress-energy tensor is symmetric and only the symmetric part contributes to its contraction with the covariant derivative of Q_α, we get the same divergence formula as before

$$\mathcal{J}^\beta{}_{;\beta} = Q_\alpha T^{\alpha\beta}{}_{;\beta} + Q_{(\alpha;\beta)} T^{\alpha\beta} = Q_\alpha T^{\alpha\beta}{}_{;\beta}\,. \tag{16}$$

For the nontranslational Killing vector fields which generate rotations, for example, this process leads to picking out the components of the conserved angular momentum in the case of vanishing divergence. See Misner, Thorne and Wheeler [41], for example. However, we will not consider angular momentum here.

For a static electric field due to a static charge distribution ρ in its rest frame, when expressed in terms of inertial coordinates in that rest frame for a time slice Σ_{rest} in that frame, the quantity

$$P(\Sigma)^0 = \frac{1}{8\pi}\int_{\Sigma_{\text{rest}}} E^2(\mathbf{x})d^3\mathbf{x} = W \tag{17}$$

is just the self-energy of the charge configuration defined alternatively by

$$W = \frac{1}{2}\int\int d^3\mathbf{x}d^3\mathbf{x}'\,\frac{\rho(t,\mathbf{x})\rho(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\,, \tag{18}$$

using the vector notation $\mathbf{x} = (x^i)$, $d^3\mathbf{x} = dx^1dx^2dx^3 = dV$. Jackson (see p. 41 of the Third edition [37]) shows how the latter formula for the self-energy of such a static charge configuration is equivalent to the energy in its associated electric field using the integral formula for the potential

$$\phi(\mathbf{x}) = \int d^3\mathbf{x}'\,\frac{\rho(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|} \tag{19}$$

and the Poisson equation $\nabla^2\phi = -4\pi\rho$. Then replacing the primed factors in the double integral for W by this expression for the potential, and with a crucial integration by parts identity, we get

$$\begin{aligned}W &= \frac{1}{2}\int\int d^3\mathbf{x}d^3\mathbf{x}'\,\frac{\rho(\mathbf{x})\rho(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|} = \frac{1}{2}\int d^3\mathbf{x}\,\rho(\mathbf{x})\phi(\mathbf{x})\\ &= -\frac{1}{8\pi}\int d^3\mathbf{x}\,\phi(\mathbf{x})\nabla^2\phi(\mathbf{x}) = \frac{1}{8\pi}\int d^3\mathbf{x}\,\left[\nabla_i\phi(\mathbf{x})\nabla^i\phi(\mathbf{x}) - \nabla_i(\phi(\mathbf{x})\nabla^i\phi(\mathbf{x}))\right]\\ &= \frac{1}{8\pi}\int d^3\mathbf{x}\,\left[E(\mathbf{x})^2 - \nabla_i(\phi(\mathbf{x})\nabla^i\phi(\mathbf{x}))\right]\,. \end{aligned}\tag{20}$$

This integral is only over the charge distribution but one can extend it to over all space since the extra contribution is zero where the charge density is zero, but as an integral over all space, the divergence term by Gauss's law is equivalent to a surface integral at spatial infinity, where the integrand goes to zero fast enough in this static case so that the surface integral evaluates to zero in the limit. The result is just the first term representing the total energy in the electric field.

$$W = \frac{1}{8\pi} \int E(\mathbf{x})^2 \, d^3\mathbf{x} = \int T^{00} d^3\mathbf{x} \,. \tag{21}$$

This self-energy plays a key role in the lowest order approximation to the equations of motion of the charge distribution.

Returning now to the divergence $-\rho E(U)^i$ in inertial coordinates of the rest frame of a static distribution of charge, its spatial integral reversed in sign is just the total electric force on the charge distribution which of course must be zero for a static configuration of charge, assuming that the charge elements are held in place by forces that are not addressed yet in this model. Otherwise the situation would not remain static. However, if the charge distribution is accelerated, there is no a priori reason to expect that the total electric force in its instantaneous rest frame be zero, and this was the error made in the Abraham and Lorentz model. Fermi showed that by requiring that the rigidity of the model respect Born's special relativistic rigidity condition, this total force integral is modified by a simple factor that his Fermi coordinate system provided, and resolves the 4/3 problem. Gauss's law is then the key to picking out the correct conserved 4-momentum of the total system which remains ambiguous in the static unaccelerated case, as we will show in the final section.

The static electron model

Fermi considers an arbitrary spherically symmetric static distribution of total charge e with density ρ in the rest frame of the electron, while referring specifically to the Abraham-Lorentz model of a uniform surface distribution of charge on a sphere of radius r_0 as the motivation for his analysis. The latter is an instructive example to keep in mind. Let the spherically symmetric charge distribution remain at rest at the spatial origin of a system of inertial coordinates (t, x^i) associated with the inertial frame K in which it is at rest for all time. The inertial observer 4-velocity is $u = \partial_t$ (in index-free notation). Let (t, r, θ, ϕ) be a corresponding system of spherical coordinates in terms of which the sphere containing the charge has the equation $r = r_0$. The metric is

$$ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta \, d\phi^2) \,. \tag{22}$$

Then whatever the internal distribution of charge, the exterior field outside its outer surface at $r = r_0$ in index-free notation is

$$F = -\frac{e}{r^2} dt \wedge dr \,, \qquad E = \frac{e}{r^2} \, \partial_r \,, \qquad B = 0 \quad (r \geq r_0) \,. \tag{23}$$

so that the Poynting vector is also zero. Introducing orthonormal components with respect to the normalized spherical coordinate frame via

$$X^{\hat{0}} = X^0\,,\ X^{\hat{r}} = X^r\,,\ X^{\hat{\theta}} = r^{-1}X^\theta\,,\ X^{\hat{\phi}} = (r\sin\theta)^{-1}X^\phi\,, \tag{24}$$

the nonvanishing such components of the stress-energy tensor of the exterior field (for $r \geq r_0$) are

$$T^{00}_{\rm em} = -T^{rr}_{\rm em} = T^{\hat{\theta}\hat{\theta}}_{\rm em} = T^{\hat{\phi}\hat{\phi}}_{\rm em} = \frac{1}{8\pi}E^2 = \frac{e^2}{8\pi r^4} = U_{\rm em}\,. \tag{25}$$

Its divergence is zero in the exterior of the electron sphere. For the shell model, the interior electromagnetic field is zero by spherical symmetry, but if instead one assumes a constant density model inside a ball of radius r_0, the interior electric field has magnitude er/r_0^3 (for $r \leq r_0$). This interior field then contributes to the total energy of the field.

The inertial coordinate components of the 4-momentum (9) in this rest frame K for any rest frame time slice $\Sigma_{\rm rest}$ are

$$P(\Sigma_{\rm rest})^0 = \int_{\Sigma_{\rm rest}} U_{\rm em}dV = W\,,\quad P(\Sigma_{\rm rest})^k = \int_{\Sigma_{\rm rest}} S^k dV = 0\,, \tag{26}$$

where W is the self-energy of the static charge distribution due to its own electric field. These are time-independent because of the time-independence of the electric field in this frame, leading to the constant 4-momentum

$$P(\Sigma_{\rm rest})^\alpha = WU^\alpha\,. \tag{27}$$

For the Coulomb field of the spherical shell electron, evaluating this quantity in spherical coordinates gives the energy of the Coulomb field

$$W = \frac{e^2}{2r_0}\,. \tag{28}$$

For the constant density model of the electron, this integral over the internal field produces an additional contribution of $e^2/(10r_0)$ leading to the total energy $3e^2/(5r_0)$. If one assumes that this electromagnetic energy makes a contribution $m_{\rm em}$ to the inertial mass of the electron via Einstein's mass-energy relation $E = mc^2$, then $m_{\rm em} = W$ (in units where $c = 1$). However, the inertial mass can only be ascertained from the equation of motion of an accelerated electron, so this must be confirmed by the evaluation of the equation of motion.

Note that the tracefree condition $T^{00}_{\rm em} = T^{11}_{\rm em}+T^{22}_{\rm em}+T^{33}_{\rm em}$ in the Cartesian inertial coordinates when integrated over the same region yields the condition

$$\int_{\Sigma_{\rm rest}} T^{00}_{\rm em}dV = \int_{\Sigma_{\rm rest}} \left(T^{11}_{\rm em} + T^{22}_{\rm em} + T^{33}_{\rm em}\right) dV\,, \tag{29}$$

but by the spherical symmetry of the electric field in the rest frame each of the terms on the right hand side has the same value

$$\int_{\Sigma_{\rm rest}} T^{11}_{\rm em}dV = \int_{\Sigma_{\rm rest}} T^{22}_{\rm em}dV = \int_{\Sigma_{\rm rest}} T^{33}_{\rm em}dV = \frac{1}{3}\int_{\Sigma_{\rm rest}} T^{00}_{\rm em}dV\,. \tag{30}$$

Consider a second inertial system K' with inertial Cartesian coordinates $(t', x^{1'}, x^{2'}, x^{3'})$, such that the original rest system K of the electron is moving with velocity v along the $x^{1'}$-axis, and let $U' = \partial/\partial t'$ be the 4-velocity of the new time lines and let Σ' be a new time slice of constant time t'. These are related to each other by the Lorentz coordinate transformation $x^{\mu'} = L^{\mu}{}_{\nu} x^{\nu}$, namely

$$t' = \gamma(t + vx^1)\,, \; x^{1'} = \gamma(x^1 + vt)\,, \; \gamma = (1 - v^2)^{-1/2}\,. \tag{31}$$

If the 4-momentum (6) defined the same 4-vector for every inertial observer, then its inertial coordinate components would simply transform like those of a 4-vector should under this Lorentz transformation, namely the components $(W, 0, 0, 0)$ would transform to $(W', p^{1'}, p^{2'}, p^{3'})$ whose nonzero values would be

$$W' = \gamma W = \gamma m_{\rm em}\,, \qquad p^{1'} = \gamma W v = \gamma m_{\rm em} v\,, \tag{32}$$

which in any case represent the new coordinate components of the 4-vector representing the 4-momentum as seen in the rest frame. However, the 4-momentum as seen by the new observer is a different 4-vector, as Gauss's law requires, so this transformed 4-momentum is not the result of evaluating the 4-momentum formulas in the new frame, and it is senseless to actually compare the transformed components of the old 4-vector with the new components of the new 4-vector.

To instead evaluate the 4-momentum (6) as seen by the new inertial observer in the frame K' in the new inertial coordinates, we must first Lorentz transform the components of the electromagnetic energy-momentum tensor to the new frame and then perform the integration over the new time coordinate hyperplane Σ', and finally relate that integral to the integral over the original rest frame time coordinate hyperplane $\Sigma_{\rm rest}$ using the condition of time invariance in the rest frame. The stress-energy tensor transforms as follows

$$T_{\rm em}^{\alpha'\beta'} = L^{\alpha}{}_{\mu} L^{\beta}{}_{\nu} T_{\rm em}^{\mu\nu}\,. \tag{33}$$

Using $T_{\rm em}^{01} = 0$, the nontrivial part of this transformation in the t-x^1 components is explicitly

$$\begin{aligned} T_{\rm em}^{0'0'} &= \gamma^2[T_{\rm em}^{00} + 2vT_{\rm em}^{01} + v^2 T_{\rm em}^{11}] = \gamma^2[T_{\rm em}^{00} + v^2 T_{\rm em}^{11}]\,, \\ T_{\rm em}^{0'1'} &= \gamma^2[T_{\rm em}^{01} + v(T_{\rm em}^{00} + T_{\rm em}^{11}) + v^2 T_{\rm em}^{01}] = \gamma^2 v(T_{\rm em}^{00} + T_{\rm em}^{11})\,. \end{aligned} \tag{34}$$

The 3-volume element on the hyperplane Σ' transforms according to $dV' = dV/\gamma$ due to the Lorentz contraction of the differential dx^1. This follows from the relation $dx^1 = \gamma(dx'^1 - vdt')$ restricted to $dt' = 0$, while $dx^2 = dx'^2, dx^3 = dx'^3$, so that

$$dV' = dx^{1'} dx^{2'} dx^{3'} = \gamma^{-1} dx^1 dx^2 dx^3 = \gamma^{-1} dV\,. \tag{35}$$

Then taking the symmetry property (30) into account, one finds

$$\begin{aligned} P(\Sigma')^{0'} &= \int_{\Sigma'} T_{\rm em}^{0'0'} dV' = \gamma\left(1 + \frac{1}{3}v^2\right) \int_{\Sigma'} T_{\rm em}^{00} dV = \gamma\left(1 + \frac{1}{3}v^2\right) \int_{\Sigma_{\rm rest}} T_{\rm em}^{00} dV \\ &= \gamma\left(1 + \frac{v^2}{3}\right) m_{\rm em} = \left(\frac{4}{3}\gamma - \frac{1}{3\gamma}\right) m_{\rm em}\,, \\ P(\Sigma')^{1'} &= \int_{\Sigma'} T_{\rm em}^{0'1'} dV' = \gamma v\left(1 + \frac{1}{3}\right) \int_{\Sigma'} T_{\rm em}^{00} dV = \gamma v\left(1 + \frac{1}{3}\right) \int_{\Sigma_{\rm rest}} T_{\rm em}^{00} dV \\ &= \frac{4}{3}\gamma m_{\rm em} v\,. \end{aligned} \tag{36}$$

Here the integral over Σ' of the integrand with respect to dV in each case equals the integral over $\Sigma_{\rm rest}$ because its value at $x^{i'}$, re-expressed in terms of the old coordinates x^i of the same point, is independent of t because the charge configuration is static in the rest frame, and so has the same value at the corresponding point of $\Sigma_{\rm rest}$. For example, on the hyperplane Σ', when expressed in terms of the old coordinates, the old components $T_{\rm em}^{00}(t', x^{1'}, x^{2'}, x^{3'}) = T_{\rm em}^{00}(x^1, x^2, x^3)$ simply don't depend on t, and so the integral against dV on that hyperplane is equal to its integral against dV on the original rest frame hyperplane $\Sigma_{\rm rest}$. The appendix shows how to re-express the difference between the 4-vectors $P(\Sigma')$ and $P(\Sigma_{\rm rest})$ independent of inertial coordinates.

In the nonrelativistic limit $|v| \ll 1$ where $\gamma \to 1$, the energy is unchanged, but the momentum has an unwanted extra factor of 4/3. This is the famous 4/3 problem for the unaccelerated electron. Furthermore, at nonzero speeds for which v^2 becomes appreciable compared to 1, the ratio between the magnitude of the linear momentum and the energy is a complicated function of the speed $|v|$ rather than the simple result $|v|/c$ as in special relativity. However, this apparent problem is based on a misconception since as explained after Eq. (5) in the previous section, the 4-momentum of the electromagnetic field depends on the observer in the presence of sources, and each distinct inertial observer produces a different 4-vector from this process, so it makes no sense to compare the result (36) to the Lorentz transformation of the original 4-vector produced by the rest frame observer. For some reason this was never understood in the early days of relativity. Because historically people insisted on finding some conserved 4-momentum to assign to the electromagnetic field, they arbitrarily picked the only natural choice for an unaccelerated electron, the 4-momentum as seen in the rest frame of the electron, and in fact, this is the one we associate with a particle whose rest mass is $m_{\rm em}$. This was first proposed by Kwal in 1939 [9] although not stated so clearly and later independently by Rohrlich [10] in 1960. The real 4/3 problem is instead its unwanted appearance as a factor in the inertial mass evaluated for the accelerated electron model developed by Abraham and Lorentz. Unfortunately their calculation preceded the introduction by Born of a relativistically invariant notion of rigidity for that model, which Fermi eventually realized was the key to resolving that apparent conflict with the equivalence of mass and energy in special relativity.

For completeness we explain what Kwal and Rohrlich actually did. In the integral formulas in the primed inertial coordinates Kwal replaced the hypersurface volume element

$$d\Sigma'_{\beta'} = -u_{\beta'} dV' = \delta^0{}_\beta dV' \tag{37}$$

by the one corresponding to the rest frame hypersurface volume element at the same spacetime point but expressed in the new coordinates

$$d\Sigma_{\beta'} = -u^{\rm rest}_{\beta'} \gamma dV' = u^{\rm rest}_{\beta'} dV \,, \tag{38}$$

where $dV = \gamma dV'$ and $(-u^{\rm rest}_{\beta'}) = \gamma(1, -v_i)$. This changes the integral to a new one. In other words this substitution disconnects the hypersurface volume element

4-vector from the hypersurface of integration, changing both its direction and magnitude. See Fig. 1. Then with this substitution, we get

$$\begin{aligned}
P(\Sigma')^{\alpha'} &= \int_{\Sigma'} T_{\rm em}^{\alpha'\beta'} d\Sigma'_{\beta'} \rightarrow \\
P_{\rm KR}(\Sigma')^{\alpha'} &\equiv \int_{\Sigma'} T_{\rm em}^{\alpha'\beta'} (-u^{\rm rest}_{\beta'} \gamma dV') = \int_{\Sigma'} L^{\alpha}{}_{\delta} T_{\rm em}{}^{\delta\beta} (-u^{\rm rest}_{\beta}) dV \\
&= \int_{\Sigma'} L^{\alpha}{}_{\delta} T_{\rm em}{}^{\delta\beta} \delta^{0}{}_{\beta} dV \\
&= L^{\alpha}{}_{\delta} \int_{\Sigma'} T_{\rm em}{}^{\delta 0} dV \\
&= L^{\alpha}{}_{\delta} \int_{\Sigma_{\rm rest}} T_{\rm em}{}^{\delta 0} dV = L^{\alpha}{}_{\delta} P(\Sigma_{\rm rest})^{\delta} \,, \qquad (39)
\end{aligned}$$

using $-u^{\rm rest}_{\beta} = \delta^{0}{}_{\beta}$ for the rest frame inertial coordinate components. Again one must use the time invariance in the rest frame to conclude that the integral over Σ' when expressed in the rest frame inertial coordinates is independent of time and so agrees with the integral over $\Sigma_{\rm rest}$, allowing the components of the new 4-momentum to transform like those of a 4-vector from the components in the rest frame.

This redefinition of the momentum integral is perhaps more simply understood as the result of merely inserting the projection operator along the unit rest frame 4-velocity vector $-u^{{\rm rest}\,\alpha'} u^{\rm rest}_{\beta'}$ into the contracted pair of indices and using the relation $\gamma = -u^{{\rm rest}\,\delta'} u_{\delta'}$ for the relative gamma factor of the two 4-velocities to get the gamma factor in the integrand which undoes the Lorentz contraction to get the rest frame volume element $dV_{\Sigma_{\rm rest}} = \gamma dV_{\Sigma'}$ at the same point

$$\begin{aligned}
P_{\rm KR}(\Sigma')^{\alpha'} &= \int_{\Sigma'} T_{\rm em}^{\alpha'\beta'} (-u^{{\rm rest}\,\delta'} u^{\rm rest}_{\beta'}) d\Sigma'_{\delta'} \\
&= \int_{\Sigma'} T_{\rm em}^{\alpha'\beta'} (-u^{{\rm rest}\,\delta'} u^{\rm rest}_{\beta'}) (-u_{\delta'} dV_{\Sigma'}) \\
&= -\int_{\Sigma'} T_{\rm em}^{\alpha'\beta'} u^{\rm rest}_{\beta'} (-u^{{\rm rest}\,\delta'} u_{\delta'}) dV_{\Sigma'} \\
&= -\int_{\Sigma'} T_{\rm em}^{\alpha'\beta'} u^{\rm rest}_{\beta'} (\gamma) dV_{\Sigma'} \\
&= -\int_{\Sigma'} T_{\rm em}^{\alpha'\beta'} u^{\rm rest}_{\beta'} dV_{\Sigma_{\rm rest}} \,. \qquad (40)
\end{aligned}$$

Since there is only one free index here, if we re-express the integral in the rest frame inertial coordinates, then we get

$$P_{\rm KR}(\Sigma')^{\alpha'} = -L^{\alpha}{}_{\mu} \int_{\Sigma'} T_{\rm em}{}^{\mu\beta} u^{\rm rest}_{\beta} dV_{\Sigma_{\rm rest}} \,, \qquad (41)$$

but the integral is still over the new time hyperplane. However, the integrand is a static function independent of the rest frame time coordinate t, so it is equivalent

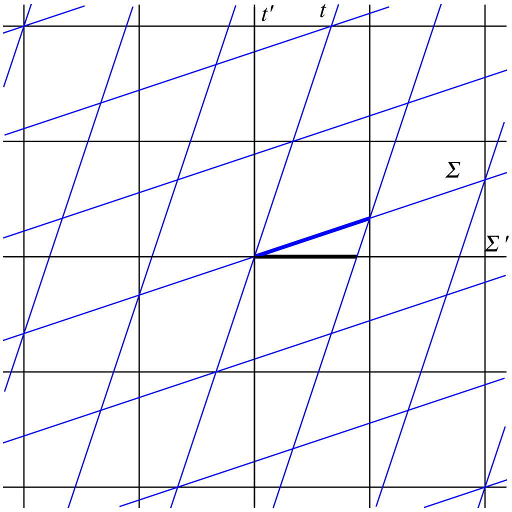


Fig. 1 A 2-dimensional diagram of the rest frame time coordinate line t (slanted forward) and a moment of rest frame coordinate time Σ (slanted upward) and the moving frame with time coordinate line t' (vertical) and a moment Σ' of its time (horizontal). For a differential region independent of time in the rest frame, like the strip between the t axis and the parallel line immediately to its right, the differential of volume dV' on Σ' as seen in the moving frame is Lorentz contracted with respect to the rest frame differential on Σ: $dV' = \gamma^{-1} dV$. Thus integrating on Σ' with respect to the differential $\gamma dV'$ is equivalent to integrating over the corresponding region of Σ (obtained by projection from Σ' to Σ along the t coordinate lines), provided that the integrand is independent of time in the rest frame.

to the integral over Σ_{rest} instead

$$\begin{aligned} P_{\text{KR}}(\Sigma')^{\alpha'} &= -L^{\alpha}{}_{\mu} \int_{\Sigma_{\text{rest}}} T_{\text{em}}{}^{\mu\beta} u_{\beta}^{\text{rest}} dV_{\Sigma_{\text{rest}}} \\ &= L^{\alpha}{}_{\mu} P(\Sigma_{\text{rest}})^{\mu} = P(\Sigma_{\text{rest}})^{\alpha'} . \end{aligned} \tag{42}$$

Kwal was not sophisticated enough to do more than examine the volume element without ever referring explicitly to the actual region of integration, where the staticity condition in the rest frame is essential to allow the integral to be done on any time hyperplane. Rorhlich simply demanded that the original integral for the 4-momentum only be performed on a time hyperplane in the rest frame of the electron, which eliminates the consideration of the integrals on other hyperplanes which yield results different from that evaluated in the rest frame. Thus one always evaluates the 4-momentum integral to the same 4-vector, whose components one can express in any inertial coordinate system, and which will then transform under the corresponding relative Lorentz transformation.

Fermi's contribution

Fermi's first paper in 1921 (Fermi 1: "On the dynamics of a rigid system of electric charges in translational motion," [4] studied a special relativistic system of electrons in rigid motion as then understood by Abraham and Lorentz and found the 4/3 factor in its inertial mass formula, while this factor was not present in the mass corresponding to the "weight" he calculated using general relativity in his second paper (Fermi 2: "On the electrostatics of a homogeneous gravitational field and on the weight of electromagnetic masses," [5]), referring to Levi-Civita's uniformly accelerated metric for the calculations [42]. This contradicted the assumed equivalence of these two masses in general relativity. These papers were both written within five years of the birth of Einstein's theory of general relativity in 1916, during which Fermi was first a high school student and then a university student writing his first two scientific papers. During the next year 1922 in preparation for his revisit to the problem, Fermi published his third paper on his famous Fermi comoving coordinate system adapted to the local rest spaces along the world line of a particle in motion (Fermi 3: "On phenomena occurring close to a world line, [6]), and calculated the variation of the action for a system of charges and masses interacting with an electromagnetic field in such a coordinate system. He then used this approach to resolve the 4/3 puzzle in his fourth paper (two versions Fermi 4a and 4c published in Italian and one Fermi 4b in German, the most complete of which is Fermi 4c: "Correction of a contradiction between electrodynamic theory and the relativistic theory of electromagnetic masses") without explicitly referring to the third paper. These were published in 1922–1923. Still in 1923 collaborating with A. Pontremoli (Fermi 10, [43]), Fermi applied his same argument to correct the calculation of the inertial mass of the radiation in a cavity with reflecting walls, where the same 4/3 factor had appeared when the cavity is in rigid motion not respecting the Born criterion; Boughn and Rothman provide a detailed alternative analysis which confirms Fermi's result in that case [44].

His approach was to use a variational principle in a region of spacetime containing the world tube of an accelerated electron charge distribution within which one has to make certain assumptions on how the relative motion of the individual charge elements in the distribution behaves. Following the Born notion of rigidity compatible with special relativity, the only way an electron can move rigidly so that its shape in its rest frame does not change is if the individual world lines of the charge distribution all cut the local rest frame time slices orthogonally, a Lorentz invariant geometrical condition which is equivalent to stating that their relative velocities are all zero at that moment. This condition must hold in a sequence of different inertial observers with respect to which the charge distribution is at rest. If instead one takes the family of time slices associated with a single inertial observer and require that the shape not change, i.e., that the relative velocities are all zero at each such time, this corresponds to the nonrelativistic notion of rigidity, and the world lines may be varied by arbitrary time-dependent translations, so that their

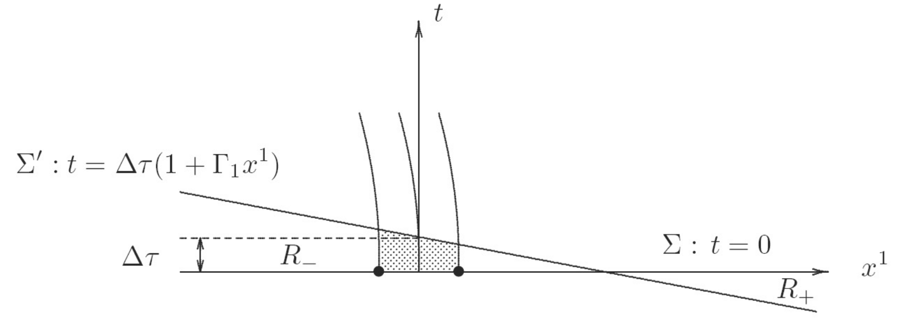


Fig. 2 A constant x^2, x^3 slice of inertial coordinates (t, x^i) showing the world tube of an electron sphere instantaneously at rest at $t = 0$ but accelerated in the negative x^1 direction ($\Gamma_1 < 0$) and two successive rest frame Fermi time coordinate slices separated by infinitesimal proper time $\Delta\tau$ at the center of the sphere, with the Fermi time slices intersecting to the right of the world tube (equivalent to the assumption $|\Gamma_1| r_0 < 1$). The spacetime region within the electron world tube between the two slices (shaded in this plane cross-section) occurs in the Gauss's law application to the wedge between the two time slices, namely $R_- \cup R_+$, two regions which are separated from each other by a plane of constant x^1 within the hypersurface $t = 0$ shown as the intersection point in this diagram.

variations of the spatial inertial coordinates from a given state can be arbitrary functions of time. However, such a conventional rigid motion with respect to that single observer will not be seen as rigid in that sense with respect to any other single inertial observer, so it is clearly incompatible with special relativity as emphasized by Fermi. This was perhaps obvious, but no one had examined the equations of motion starting from the Lagrangian to understand that the usual starting point for the Abraham-Lorentz evaluation of their assumed equations of motion was equivalent to this assumption. This was the insight that Fermi had had to resolve the problem. Assuming conventional rigidity, one finds the starting point equations of motion of the Abraham-Lorentz model whose analysis yields the incorrect inertial mass factor with the 4/3 factor, but with Born rigidity one instead finds the one expected from Einstein's mass-energy relation which removes this factor. The only difference in the two calculations is the resulting Fermi correction factor in the integral of the total force on the charge distribution, a factor arising from the spacetime volume element in Fermi coordinates due to the acceleration of its central world line.

Fermi considers a laboratory frame with inertial coordinates (t, x^1, x^2, x^3) in which at the end of his argument, the accelerated electron is momentarily at rest centered about the spatial origin at the initial coordinate time which we will assume for simplicity to be $t = 0$. Assuming that the Fermi coordinate system (T, X^1, X^2, X^3) is adapted to a world line in the electron charge distribution passing through the origin of these spatial coordinates at $t = 0$ when $v^i = 0$, its time hypersurface $T = 0$ can be chosen to coincide with $t = 0$, but after a small interval dt of laboratory time along the central world line, equal to the increment dT in the proper time along that world line to first order, the Fermi time slice is instead tilted slightly to remain orthogonal to that world line as shown in Fig. 1. The metric in the Fermi coordinate system is

$$ds^2 = -N^2 dT^2 + \delta_{ij} dX^i \, dX^j \,, \quad N = c(1 + \Gamma_i X^i/c^2)\,, \tag{43}$$

where $\Gamma_i = \dot{v}^i = dv^i/dT$ are the Cartesian components of the proper acceleration of the central world line (functions of T), and the speed of light c is not taken to be unity in this paragraph only in order to appreciate how factors of c enter the discussion. The proper time along the central Fermi coordinate time line is initially approximately $dT = dt$ at $t = 0 = T$, but away from the spatial origin at that world line there is a linear correction factor due to the lapse function N in the Fermi coordinate system. The proper time interval along the normal to the initial hypersurface (measured by the increment in t or T to first order) to a nearby Fermi time slice is the increment $c^{-1}N\,dT = (1 + \Gamma_i x^i/c^2)dT$, namely the proper time along the time lines in the Fermi coordinate system. Misner, Thorne and Wheeler discuss the Fermi coordinate system in detail [41]. Of course because the proper time of each charge element world line varies by the Fermi lapse function factor compared to the central world line, the accelerations of the actual charge elements away from the central world line differ slightly from that of the central world line.

If we imagine doing a variation of the action integral over a spacetime region in inertial coordinates between two slices of inertial time (his variation A), then if we use the same coordinate symbols (t, x^i) for the corresponding variation in Fermi coordinates between two slices of Fermi coordinate time (his variation B), the only formal difference in the action integrand is the additional Fermi lapse factor which enters through the spacetime volume element. This lapse correction factor is the entire basis for Fermi's correction, and multiplies the coordinate volume element to provide the covariant spacetime volume element in Minkowski spacetime: d^4V which is $d^4x = dt\,dV$ in inertial coordinates but $N dt\,dV$ in Fermi coordinates, where $Ndt = d\tau$ is the proper time along the time world lines orthogonal to the flat time slices and $dV = dx^1dx^2dx^3$ is the spatial volume element in both cases. Fermi does not mention his mathematical article on these coordinates, but just presents a short derivation of the correction factor based on the curvature of the world line. The extra acceleration term in the integral with coefficient $\Gamma_i x^i$ (with $c = 1$ again) provides exactly the necessary correction to produce the desired result in the inertial mass coefficient in the equations of motion for any smooth spherically symmetric model of the electron.

However, to justify this variation of the action yielding the Lagrange equations, the variations must vanish on the bounding time slices and be arbitrary functions of time for the intermediate times. For the variation A, Fermi explicitly states that the variations of the spatial coordinates are arbitrary functions of t which vanish at the end slices, but for the variation B he only examines an infinitesimal contribution of an interval of Fermi time to the whole 4-dimensional integral and he emphasizes that for that interval of time, the variations in the spatial coordinates of the world lines should be arbitrary constants to represent an overall translation of those world lines. However, in order to claim his resulting Lagrangian equation is valid, it has to be understood that as in the first case, the variations in the spatial coordinates must be arbitrary functions of the time coordinate which vanish at the end times. This implies that the Lagrangian variation extremizes the action among all those world lines which break the rigid Born symmetry assumed in the solution about which the variation takes place. It does not allow for a variation among the family of Born rigid motions of the electron nearby the given solution. None of this is made explicit in Fermi's article.

If the spatial variations were arbitrary constants in the Fermi coordinate system in order to preserve the rigidity in the variation, and if they were to vanish on the end time slices, they would vanish everywhere, so could one not conclude that at every time along the world tube of the electron that the spatial integral coefficients of the variation must vanish. On the other hand if they did not vanish at the end times, one could not ignore the boundary terms which result from the integration by parts along the time lines. Furthermore, without being independent variations at each time, one cannot conclude that their coefficients must vanish. This is a very tricky point since in general one cannot impose symmetries on a Lagrangian and be guaranteed to get the same equations of motion for the restricted variational principle as those that result from imposing the symmetries on the Lagrangian equations of motion derived from the general variational principle as discussed by Maccallum and Taub for the complementary problem of spatial rather than temporal symmetry imposed on a Lagrangian [45]. It is the boundary terms which play the key role in this discussion. By not requiring that the variations about a symmetric solution conform to the symmetry, Fermi appears to have avoided these difficulties.

Note that the model of the charge distribution as some kind of rigid body is necessary in order to assign some common acceleration to the system at each moment of time (that of the central world line) so that its coefficient in the equations of motion can be interpreted as the inertial mass. Consider therefore as Fermi does such an accelerated system of electric charge in special relativity held at rest relative to each other by some external forces (i.e., in conventional or relativistic rigid motion). The corresponding action is given in inertial coordinates by the usual Lagrangian integral in inertial coordinates with the additional term in the mechanical mass added back into the discussion representing a rest mass distribution with differential mass dm assumed to have the same rigidity properties as the charge distribution with differential charge de, i.e., they mass and charge elements share the same world lines

$$S = S(A_\mu, x^\alpha) = \int \left(-\frac{1}{16\pi} F^{\alpha\beta} F_{\alpha\beta} + A_\mu J^\mu \right) d^4x - \int d\tau\, dm\,. \qquad (44)$$

The region of integration is an arbitrary region of spacetime, and the 4-current

$J^\mu = \rho U^\mu$ depends on the parametrized world lines of the charged particles, whose unit 4-velocity is $U^\mu = dx^\mu/d\tau$ if $d\tau$ is the increment of proper time along them. The charge and mass terms are first integrated over the world lines of the charge and mass elements and then over the family of these world lines. Both the charge and mass profiles as a function of the family of world lines of the matter distribution are assumed to be given and fixed along those world lines. Fermi discusses and varies this action in his Fermi coordinate article (Fermi 3, [4]). The line integrals in the charge and mass distribution terms are parametrization independent, so the world lines can be parametrized by any parameter, including coordinate time.

Varying S with respect to the vector potential A_μ, fixing the world lines of the charge distribution, leads to the inhomogeneous Maxwell's equations. In fact

$$\begin{aligned}\delta S|_{x^\alpha=const.} &= \int d^4x \left(-\frac{1}{8\pi}F^{\alpha\beta}\delta F_{\alpha\beta} + J^\mu \delta A_\mu\right)\\ &= \int d^4x \left(-\frac{1}{4\pi}F^{\alpha\beta}\delta(\partial_\alpha A_\beta) + J^\mu \delta A_\mu\right)\\ &= \int d^4x \left(\frac{1}{4\pi}\partial_\alpha F^{\alpha\beta}\delta A_\beta + J^\mu \delta A_\mu\right) - \int d^4x\, \frac{1}{4\pi}\partial_\alpha(F^{\alpha\beta}\delta A_\beta)\\ &= \int d^4x \left(\frac{1}{4\pi}\partial_\alpha F^{\alpha\mu} + J^\mu\right)\delta A_\mu\,, \qquad (45)\end{aligned}$$

that is

$$\partial_\alpha F^{\mu\alpha} = 4\pi J^\mu\,. \qquad (46)$$

The next to last equality in this sequence follows from the usual Lagrangian variation integration by parts, resulting in the integral of a divergence which by Gauss's law is equivalent to a boundary integral where the variation is assumed to vanish and hence does not contribute to the final expression.

The variation of S with respect to the coordinates of the charge element world lines where the above variations A and B are relevant requires first reinterpreting the spacetime volume integral of the interaction term as the integral over a family of line integrals along those world lines. This is most easily done using the adapted Fermi coordinate system where the spatial coordinates parametrize the world lines of the charge elements, which are the time lines of the system. The spacetime volume element is $d^4x = N dt\, dV = d\tau\, dV$ with $dV = d^3x$ and $d\tau = N\, dt$. The 4-current is $J^\mu = \rho U^\mu$, where ρ is the rest frame charge density, which is a constant (along the world lines but zero everywhere else), and $U^\alpha = dx^\alpha/d\tau$ is the charge element 4-velocity. Then let $de = \rho dV$. The interaction term in the action can then be represented as the integral of the line integral alone the world line with respect to the rest frame charge density

$$\int J^\mu A_\mu d^4x = \int\int \rho A_\mu \frac{dx^\mu}{d\tau} d\tau\, dV = \int \left(\int A_\mu dx^\mu\right) de\,. \qquad (47)$$

Keeping in mind the geometrical origin of de, the line integral is coordinate independent and so one can use this expression also in inertial coordinates using any parametrization of the world lines.

Since variations of the electromagnetic field Lagrangian at constant A_α vanish, we only have to vary the source term, where A_α is instead evaluated along the charge element world lines so $\delta A_\mu = A_{\mu,\nu}\delta x^\nu$. Using the fact that $\delta(dx^\mu) = d(\delta x^\mu)$ as usual in the Lagrangian variation, we find step by step for the variation of the interaction term

$$\begin{aligned}
\delta S|_{A_\alpha=const.} &= \delta\left(\int de\, dx^\mu A_\mu\right)\\
&= \int de \int \left[A_{\mu,\sigma}dx^\mu \delta x^\sigma + A_\sigma \delta dx^\sigma\right]\\
&= \int de \int \left[A_{\mu,\sigma}dx^\mu \delta x^\sigma - dA_\sigma \delta x^\sigma + d(A_\sigma \delta x^\sigma)\right]\\
&= \int de \int \left[(A_{\mu,\sigma} - A_{\sigma,\mu})dx^\mu \delta x^\sigma + d(A_\sigma \delta x^\sigma)\right]\\
&= \int de \int F_{\sigma\mu} dx^\mu \delta x^\sigma + \int de \int d(A_\sigma \delta x^\sigma)\,. \qquad (48)
\end{aligned}$$

Ignoring the boundary term, the first integral (where the line integral part is independent of the parametrization of the world lines) can be expressed in terms of inertial coordinates or proper time in Fermi coordinates, where the Fermi lapse correction factor depends on the location of the charge element

$$\int \left(\int F_{\sigma\mu}\frac{dx^\mu}{dt}de\right)\delta x^\sigma\, dt = \int \left(\int F_{\sigma\mu}\frac{dx^\mu}{d\tau}N de\right)\delta x^\sigma\, dt\,. \qquad (49)$$

Both expressions are equivalent but the presence of a nonunit lapse function in the Fermi coordinate system is crucial.

If we consider the left expression in inertial coordinates in which the electron is momentarily at rest (so that $N = 1$, $dx^\mu/dt = \delta^\mu{}_0$ and dV agrees with the Fermi coordinate volume element), it reduces to

$$\int \left(\int E_i\, de\right)\delta x^i\, dt = \int \left(\int \rho E_i dV\right)\delta x^i\, dt \qquad (50)$$

since $F_{i0} = E_i$ is the electric field in inertial coordinates and $F_{00} = 0$. The factor in parentheses is just the total electric force on the distribution of electric charge at this moment. For the Fermi variation A in these inertial coordinates, one has $\delta x^\sigma = \delta^\sigma{}_i \delta x^i(t)$ and one can require that $\delta x^i(t_1) = 0 = \delta x^i(t_2)$ at the boundary inertial time hyperplanes of the region of integration, while leaving $\delta x^i(t)$ arbitrary in between. This allows one to ignore the boundary term which integrates to the end times where the variation vanishes, while forcing the expression in parentheses to zero if we ignore the mechanical mass term in the Lagrangian for the moment, leading to the condition

$$\int \rho E_i dV = 0\,. \qquad (51)$$

This is the starting point of the Abraham-Born derivation of the equations of motion in the model of the electron with zero mechanical mass, showing that it is equivalent to assuming the noncovariant rigidity condition, which Fermi concludes must obviously invalidate that model.

The only difference for his variation B in the Fermi coordinate system is the additional factor of the Fermi lapse in the differential of proper time needed to define the electric field in that coordinate system

$$0 = \int F_{\sigma\mu} \frac{dx^\mu}{d\tau} N de = \int F_{\sigma\mu} U^\mu N de = \int \rho E(U)_\mu N dV \,, \tag{52}$$

an expression which only has nonzero components $E(U)_\mu = \delta^i{}_\mu E(U)_i$ in either Fermi coordinates or in inertial coordinates in which the electron is momentarily at rest, where $E(U)_i = E_i$ then agree. Clearly when the acceleration is identically zero $\Gamma_i = 0$ and $N = 1$, the final conditions are the same for both cases A and B, so one must have nonzero acceleration to see a difference in these two cases. Of course without acceleration one cannot measure the inertial mass.

To finish the story we must analyze these conditions in terms of the internal forces exerted on the charge elements by other charge elements and the forces exerted by the external electromagnetic field responsible for the acceleration of the electron. It is the separation of the self-field and the external field that allows one to extract the Lorentz force law relation to the acceleration of the central world line (corrected by radiation reaction terms if one expands if far enough in the acceleration) and thus identify the inertial mass coefficient where the 4/3 problem is apparent, and Fermi's correction restores this factor to 1. The uncorrected Abraham-Lorentz condition is discussed in detail in Jackson [37] (although the Third Edition omits the final explicit evaluation of the famous 4/3 term), so we only summarize it here. We then follow Fermi in explicitly evaluating the correction term to see its effect in removing the unwanted 4/3 factor. Finally we will consider the additional mechanical mass term in the Lagrangian to follow Fermi's original Lagrangian discussion in his third paper. For the moment we set this term to zero as in Fermi's fourth paper.

- Field separation for variations of type A

Consider first the system of variations A.

$$0 = \int E_a \, de \,. \tag{53}$$

Let $E = E_{\text{self}} + E_{\text{ext}}$, where E_{self} and E_{ext} the contributions to the total field due to the self-interaction of the system and to the external electric field respectively, the latter of which is assumed to be sufficiently uniform over the small dimensions of the system that it can be pulled out of the integral, which results in the total charge multiplying the external electric field evaluated at the central world line. Eq. (53) thus becomes

$$F^a_{\text{ext}} \equiv \int E^a_{\text{ext}} \, de = E^a_{\text{ext}} \int de = -\int E^a_{\text{self}} \, de \equiv -F^a_{\text{self}} \,. \tag{54}$$

The self-force is the result of the interaction of each element of charge of the sphere with every other element. The explicit details of the calculation involving the retarded times can be found in Jackson's textbook [37]. The self-field can be expressed in terms of the self-potentials A and ϕ by

$$E_{\rm self} = -\nabla\phi - \frac{1}{c}\frac{\partial A}{\partial t}\,, \tag{55}$$

so that

$$F_{\rm ext} = \int \rho\left[\nabla\phi + \frac{1}{c}\frac{\partial A}{\partial t}\right]\, d^3\mathbf{x}\,, \tag{56}$$

since the charge element is $de = \rho\, d^3\mathbf{x}$. We now adopt the Jackson notation that $\mathbf{x}$ is the spatial position vector in the Cartesian coordinate system and $dV = d^3\mathbf{x}$ is the spatial volume element, and let $\mathbf{v}$ and $\mathbf{a} = \dot{\mathbf{v}} = \Gamma$ be the velocity and acceleration of the charge distribution, which at the initial time t of our calculation satisfies $\mathbf{v}(t) = 0$ (all elements of the charge distribution are simultaneously at rest) and $\mathbf{a} = \mathbf{a}(t)$ (the acceleration is the same for all elements of the charge distribution at that moment), expressing the nonrelativistic rigidity of the charge distribution. We also reintroduce factors of the speed of light c into the discussion.

By evaluating the potentials at the retarded time $t' = t - |\mathbf{x}-\mathbf{x}'|/c$, i.e.,

$$A = \frac{1}{c}\int \frac{[J(t',\mathbf{x}')]_{\rm ret}}{|\mathbf{x}-\mathbf{x}'|}\, d^3\mathbf{x}'\,, \qquad \phi = \int \frac{[\rho(t',\mathbf{x}')]_{\rm ret}}{|\mathbf{x}-\mathbf{x}'|}\, d^3\mathbf{x}'\,, \tag{57}$$

and using the rule (Taylor series expansion about the time $t' = t$)

$$[\ldots]_{\rm ret} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!}\left(\frac{|\mathbf{x}-\mathbf{x}'|}{c}\right)^n \frac{\partial^n}{\partial t^n}[\ldots]|_{t'=t}\,, \tag{58}$$

Eq. (56) becomes

$$F_{\rm ext} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\,c^n}\int d^3\mathbf{x}\int d^3\mathbf{x}'\,\rho(t,\mathbf{x})\frac{\partial^n}{\partial t^n}\Bigg[\rho(t,\mathbf{x}')\nabla(|\mathbf{x}-\mathbf{x}'|^{n-1}) + \frac{|\mathbf{x}-\mathbf{x}'|^{n-1}}{c^2}\frac{\partial J(t,\mathbf{x}')}{\partial t}\Bigg]\,. \tag{59}$$

Consider the first term in the brackets. The $n = 0$ term

$$\int d^3\mathbf{x}\int d^3\mathbf{x}'\,\rho(t,\mathbf{x})\rho(t,\mathbf{x}')\nabla|\mathbf{x}-\mathbf{x}'|^{-1} \tag{60}$$

vanishes in the case of a spherically symmetric charge distribution, whereas the $n = 1$ term is identically zero (gradient of a constant), implying that the first nonvanishing contribution comes from $n = 2$. Changing the summation indices thus leads to

$$F_{\rm ext} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\,c^{n+2}}\int d^3\mathbf{x}\int d^3\mathbf{x}'\,\rho(t,\mathbf{x})|\mathbf{x}-\mathbf{x}'|^{n-1}\frac{\partial^{n+1}}{\partial t^{n+1}}\Bigg[J(t,\mathbf{x}') + \frac{\partial\rho(t,\mathbf{x}')}{\partial t}\frac{\nabla(|\mathbf{x}-\mathbf{x}'|^{n+1})}{(n+1)(n+2)|\mathbf{x}-\mathbf{x}'|^{n-1}}\Bigg]\,. \tag{61}$$

The continuity equation, spherical symmetry and angular averaging can be used to simplify this expression, taking into account also that for a rigid charge distribution the current is $J(t, \mathbf{x}') = \rho(t, \mathbf{x}')\mathbf{v}(t)$, where $\mathbf{v}(t) = 0$ holds at the time t at which this calculation is carried out, so only its time derivatives contribute to the series expansion. The term in this expansion containing the first time derivative of the acceleration $\dot{\Gamma} = \ddot{\mathbf{v}}$ is associated with the radiation reaction, not discussed here.

The final result, obtained by neglecting all nonlinear powers of the acceleration and its derivatives (which appear for $n \geq 4$), at lowest order can be written as

$$F_{\text{ext}} = -F_{\text{self}} = \frac{2}{3}\sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\frac{I_n}{c^{n+2}}\frac{\partial^n}{\partial t^n}\dot{\mathbf{v}}\,, \tag{62}$$

where

$$I_n = \int\!\!\int d^3\mathbf{x}d^3\mathbf{x}'\,\rho(t,\mathbf{x})|\mathbf{x}-\mathbf{x}'|^{n-1}\rho(t,\mathbf{x}')\,. \tag{63}$$

The lowest order term is the only one considered by Fermi to make his point and is twice the self-energy of the charge distribution

$$I_0 = 2W = \int\!\!\int d^3\mathbf{x}d^3\mathbf{x}'\,\frac{\rho(t,\mathbf{x})\rho(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\,, \tag{64}$$

which for the spherical shell model of the electron is $2W = e^2/r_0$. In the point particle limit, I_0 diverges corresponding to the infinite self-energy of a point particle, $I_1 = e^2$, and $I_n = 0$ for $n > 1$. When the charge is uniformly distributed over the surface of the sphere one has $I_n = 2e^2(2r_0)^{n-1}/(n+1)$.

In the nonrelativistic limit for any smooth spherically symmetric distribution of charge (i.e., considering only the $n = 0$ term of the series) Eq. (62) becomes

$$F^{\text{NR}}_{\text{self}} = -\frac{4}{3}\frac{W}{c^2}\dot{\mathbf{v}}\,, \tag{65}$$

so that the Newton's equation of motion for the system takes the form

$$F^{\text{NR}}_{\text{ext}} = \frac{4}{3}m_{\text{em}}\dot{\mathbf{v}}\,, \qquad m_{\text{em}} = \frac{W}{c^2}\,. \tag{66}$$

This is 4/3 times the electromagnetic mass m_{em} defined by the Einstein mass-energy relation. Recall that this is understood to be expressed in an inertial frame in which the electron is momentarily at rest, ignoring higher order terms in the acceleration which include the famous radiation reaction terms.

- Field separation for variations of type B

The "correct" result in which the unwanted factor of 4/3 is removed is achieved starting instead with Fermi's corrected integral condition, so that in the previous calculation of Jackson we must replace the factor of $\rho(t, \mathbf{x})$ in the double spatial integral by $\rho(t, \mathbf{x})(1 + \dot{\mathbf{v}}(t) \cdot \mathbf{x})$, assuming that we are using a Fermi coordinate system at a time slice which coincides with the previous inertial coordinate slice of the preceding discussion when the electron is momentarily at rest. Thus the vanishing integral $n = 0$ term, namely Eq. (60), of the original expansion now becomes

$$\begin{aligned}&\int d^3x \int d^3x'\, \rho(t,\mathbf{x})[1+\dot{\mathbf{v}}(t)\cdot \mathbf{x}/c^2]\rho(t,\mathbf{x}')\nabla|\mathbf{x}-\mathbf{x}'|^{-1}\\ &= \int d^3x \int d^3x'\, \rho(t,\mathbf{x})[\dot{\mathbf{v}}(t)\cdot \mathbf{x}/c^2]\rho(t,\mathbf{x}')\nabla|\mathbf{x}-\mathbf{x}'|^{-1}\,. \qquad (67)\end{aligned}$$

Fermi noted that this double spatial integral will give the same value if the two dummy vector integration variables are switched, and hence can also be replaced by the average of these two ways of writing the same integral. Letting $\nabla|\mathbf{x}-\mathbf{x}'|^{-1} = -(\mathbf{x}-\mathbf{x}')/|\mathbf{x}-\mathbf{x}'|^3$

$$\begin{aligned}&c^{-2}\int d^3x \int d^3x'\, \rho(t,\mathbf{x})\rho(t,\mathbf{x}')[\dot{\mathbf{v}}(t)\cdot\mathbf{x}](\mathbf{x}'-\mathbf{x})/|\mathbf{x}-\mathbf{x}'|^3\\ &= c^{-2}\int d^3x \int d^3x'\, \rho(t,\mathbf{x}')\rho(t,\mathbf{x})[\dot{\mathbf{v}}(t)\cdot\mathbf{x}'](\mathbf{x}-\mathbf{x}')/|\mathbf{x}-\mathbf{x}'|^3\\ &= -c^{-2}\frac{1}{2}\int d^3x \int d^3x'\, \rho(t,\mathbf{x})\rho(t,\mathbf{x}')[\dot{\mathbf{v}}(t)\cdot(\mathbf{x}'-\mathbf{x})](\mathbf{x}'-\mathbf{x})/|\mathbf{x}-\mathbf{x}'|^3\,. \qquad (68)\end{aligned}$$

Now imposing spherical symmetry about the origin, the components of this vector integral are nonzero only along the acceleration vector, with a coefficient which can be replaced by the average value of the vector component integral

$$-[\dot{\mathbf{v}}(t)\cdot(\mathbf{x}'-\mathbf{x})](\mathbf{x}'-\mathbf{x}) \to -\dot{\mathbf{v}}(t)\frac{1}{3}(\mathbf{x}'-\mathbf{x})\cdot(\mathbf{x}'-\mathbf{x}) = -\dot{\mathbf{v}}(t)\frac{1}{3}|\mathbf{x}'-\mathbf{x}|^2 \qquad (69)$$

so it reduces to

$$-\frac{1}{3}\frac{\dot{\mathbf{v}}(t)}{c^2}\left[\frac{1}{2}\int d^3x \int d^3x'\, \rho(t,\mathbf{x})\rho(t,\mathbf{x}')/|\mathbf{x}-\mathbf{x}'|\right] = -\frac{1}{3}\frac{W}{c^2}\dot{\mathbf{v}}(t)\,, \qquad (70)$$

since the expression in square brackets is the self-energy of the charge distribution at the time t. This is the only additional term linear in the acceleration which contributes to the lowest terms of the previous calculation (so that the lowest order radiation reaction term is unchanged, although not shown here)

$$F^{\rm NR}_{\rm ext} = \frac{4}{3}\frac{W}{c^2}\dot{\mathbf{v}} - \frac{1}{3}\frac{W}{c^2}\dot{\mathbf{v}} = \frac{W}{c^2}\dot{\mathbf{v}}\,, \qquad (71)$$

which leads to the desired result

$$F^{\rm NR}_{\rm ext} = m_{\rm em}\dot{\mathbf{v}}\,, \qquad m_{\rm em} = \frac{W}{c^2}\,. \qquad (72)$$

in the nonrelativistic limit, according to Newton's law with the electromagnetic mass $m_{\rm em} = W/c^2$.

Finally to consider the contribution to the Lagrangian from a mechanical mass distribution, we must vary the final term in the Lagrangian which has been ignored until now. In the Fermi coordinate system this is trivial. The Lagrangian term is simply

$$-\int d\tau\, dm = -\int \left(\int N\, dt\right) dm = -\int \left(\int 1+\Gamma_i x^i\, dt\right) dm\,, \qquad (73)$$

and its variation is

$$-\delta\int\left(\int 1+\Gamma_i x^i\, dt\right) dm = -\int\left(\int \Gamma_i \delta x^i\, dt\right) dm$$
$$= -\left(\int\, dm\right)\int \Gamma_i \delta x^i dt = -\int (m_0\Gamma_i)\delta x^i dt\,, \tag{74}$$

where m_0 is the total mechanical mass. The contribution to the above Fermi condition are the coefficients of the arbitrary variations $\delta x^i = \delta x^i(t)$, namely just the term $-\int(m_0\Gamma_i) = -m_0\dot{v}^i$. The complete equation of motion is then first

$$\int \rho E(U)_i(1+\Gamma_j x^j)\, dV - m_0\Gamma_i = 0\,, \tag{75}$$

and then after splitting off the self-force and passing to the lowest order approximation

$$(m_0 + m_{\rm em})\dot{\mathbf{v}} = F^{\rm NR}_{\rm ext}\,. \tag{76}$$

Thus mechanical mass and the electromagnetic mass contribute in the same way to the total inertial rest mass of the spherical distribution of charged matter.

Relating Kwal-Rohrich back to Fermi through Gauss

Given the Kwal-Rohrlich 4-momentum evaluated for an unaccelerated electron and the inertial mass contribution from the electromagnetic field found by Fermi for the accelerated electron, it is natural to look for a relation between them. In the unaccelerated case, one has an entire family of distinct 4-momenta which depend on the inertial observer, but the one we usually associate with the electron of a certain rest energy is the one defined by the rest frame observer. Although Fermi stopped his analysis once he achieved his limited goal, in light of the 4-momentum integral situation in which interest later arose, it is natural to continue his line of thought to its logical conclusion. We do this here and find that Fermi's corrected condition which generates the correct equations of motion guarantees the conservation of the total 4-momentum as seen in the instantaneous rest frame of the accelerated electron at each point of its world line.

All we need do do is specialize the Gauss law discussion begun in Section 2 to the electromagnetic stress-energy tensor over the spacetime region R between two successive time hyperplanes Σ_t and $\Sigma_{t+\Delta t}$ associated with a Fermi coordinate system adapted to the central world line of the accelerated electron, as in Fig. A.1 of the Appendix where the case of 1-dimensional motion is illustated. Let $\Delta t > 0$ so $t+\Delta t$ is to the future of t along the central world line where t measures the elapsed proper time. Fig. A.1 shows the tilting of the Fermi time slices to remain orthogonal to the central world line of the electron and to the common local rest space of the elements of charge which make up the electron sphere. Then Eqs. (13) and (5) lead to the fundamental relation

$$-\int_R Q_\alpha \rho E(U)^\alpha d^4V = Q_\alpha\left[P(\Sigma_{t+\Delta t})^\alpha - P(\Sigma_t)^\alpha\right]\,, \tag{77}$$

where on the right hand side the components have to be expressed in inertial coordinates or the components Q_α are not constant and cannot be factored out of the integral. On the left hand side, if evaluated in the Fermi coordinate system, these components are functions of time to compensate for the time-dependent change of direction of the 4-velocity of the central world line, and so can only be pulled out of the spatial integral. Recall that $E(U)^\alpha$ is the electric field seen in the electron rest frame and ρ is the rest frame charge density.

Let R_- be the half-region for which the hyperplane $\Sigma_{t+\Delta t}$ is in the future of Σ_t, while R_+ has the reverse relationship, as in Fig. A.1, so that the world tube of the electron cuts through the region R_- as shown there. Splitting the integral into the spatial integral and then the temporal integral, using the spacetime volume element $d^4V = (1 + \Gamma_i x^i)dV\, dt$, one then has

$$-\int_t^{t+\Delta t} Q_\alpha \int_{\Sigma_\tau \cap R_-} \left[\rho E(U)^\alpha (1 + \Gamma_i x^i) dV\right] d\tau = Q_\alpha \left[P(\Sigma_{t+\Delta t})^\alpha - P(\Sigma_t)^\alpha\right] . \tag{78}$$

For the Born rigid distribution of charge according to the Fermi condition (75), the spatial integral in parentheses on the left hand side of Eq. (78) at each Fermi time (which the proper time parameter along the central world line) equals the mechanical mass times the proper time covariant derivative $D/d\tau$ of the 4-velocity of the central world line

$$m_0 \delta^\alpha{}_i \Gamma^i = m_0 \frac{Du^\alpha}{d\tau} = \frac{D}{d\tau}(m_0 u^\alpha) = \frac{D}{d\tau} p_0^\alpha \tag{79}$$

where $p_0^\alpha = DU^\alpha/d\tau$ is the mechanical momentum. Here we use the notation $D/d\tau$ to remind us that in noninertial coordinates like those of Fermi, the covariant derivative along the parametrized curve does not coincide with the action of the ordinary such derivative, but when we evaluate the expression in components with respect to a fixed inertial coordinate system, it does. The final integral with respect to the Fermi time coordinate, if performed with the components taken in an inertial coordinate system, is then just the difference of the mechanical momentum between the two Fermi times

$$\int_t^{t+\Delta t} Q_\alpha \left(\frac{dp_0^\alpha}{dt}\right) dt = Q_\alpha \left[p_0(t+\Delta t)^\alpha - p_0(t)^\alpha\right] , \tag{80}$$

so that

$$-Q_\alpha \left[p_0^\alpha(t+\Delta t) - p_0^\alpha(t)\right] = Q_\alpha \left[P(t+\Delta t)^\alpha - P(t)^\alpha\right] , \tag{81}$$

Expressing this in inertial coordinates, since Q_α are arbitrary constants, we find

$$p_0(t+\Delta t)^\alpha + P(t+\Delta t)^\alpha = p_0(t)^\alpha + P(t)^\alpha , \tag{82}$$

namely that the sum of the mechanical 4-momentum and the 4-momentum of the external electromagnetic field $p_0^\alpha + P^\alpha$ must be the same on the two Fermi time slices and hence on every Fermi time slice. In other words the Fermi condition is equivalent to the conservation of the Kwal-Rohrlich 4-momentum for the total system, a fact which no one seems to have realized until now. Thus Fermi also pointed the way towards selecting the only observer-defined total 4-momentum which is conserved and which corresponds to what we associate with this system. The proper time derivative of this relation gives its rate of change version

$$\frac{D}{dt}(p_0(t)^\alpha + P(t)^\alpha) = 0\,. \tag{83}$$

Thus the calculations initiated by Fermi nearly a century ago have finally reached their natural conclusion.

Apart from Kolbenstvedt [26] much later in 1997, only Aharoni [11] seems to have seen and understood Fermi's argument, explaining exactly what Fermi did in detail in his 1965 textbook revised because of the then recent Rohrlich work on this topic and re-interpreting it in his own way, explaining in detail how the 4-momentum integrals first explained by Kwal and later Rohrlich are connected to Fermi's approach to the problem. Anaroni's equations (6.5), (6.18) and (6.19) for the total self-force due to the electron charge distribution involve through his (6.18) the proper time rate of change of an integral over the spacetime region between two successive proper time hypersurfaces of the electron (his own reformulation of the self-force in view of the Kwal-Rohrlich integral definition as noted in a footnote). Aharoni considers the following equivalent reformulation of the previous equations valid for the total electromagnetic field, but restricted only to the self-field in order to define the self-force due only to the self-field of the charge distribution

$$\frac{dP^\mu}{d\tau} = -\frac{d}{d\tau}\int_{\tau_0}^{\tau}\int F_{\text{self}}{}^{\mu}{}_{\nu} J^\nu \, d\tau d\Sigma = -\delta^{\mu}{}_{i}\int (1+\Gamma_j x^j) E^i_{\text{self}}\rho \, d^3x\,. \tag{84}$$

However, Aharoni failed to relate his"postulated" self-force expression to Gauss's law to show that it actually is related to the proper time rate of change of the Kwal-Rohrlich 4-momentum integral restricted to the self-field. Spohn and Yaghjian both have long bibliographies in their textbooks, but neither mentions Aharoni, while Rohrlich has an author index indicating Aharoni's name on page 283 where no reference to anyone can be found. Only the much later work of Kolbenstvedt acknowledges Fermi's approach, rederiving it in a slightly different but equivalent form, also ignored by Rohrlich, Spohn and Yaghjian in their later editions.

Concluding Remarks

It is unfortunate that the first four papers by one of the leading physicists of the twentieth century were never translated from their original Italian. The fourth paper which concludes this series and which appeared in preliminary versions in both Italian and German, was the culmination of Fermi's early work in relativity only a few years after the birth of general relativity and written while he was a university student. Its actual contents seem to have remained a mystery to nearly all those who have cited it in discussions of the classical theory of the electron which still interests people even today, while the leading textbook on classical electrodynamics still

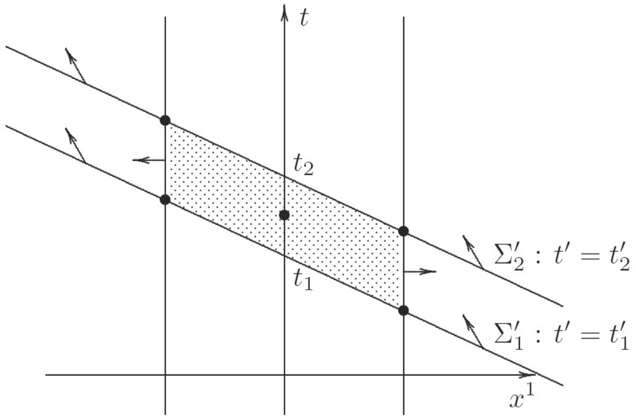


Fig. 3 Figure 5.3.b from Misner, Thorne and Wheeler redrawn with an inner cylindrical boundary which is the world tube of the electron sphere boundary. The arrows show the chosen unit normal direction for the orientation of each hypersurface, but in the single Gauss law relation for the region of spacetime between Σ_1 and Σ_2 excluding the shaded region inside the cylinder, the sum of the outward normally oriented integral contributions is zero for a divergence-free vector field. Here the boundary term due to the portion σ of the cylinder between the two parallel hyperplanes vanishes by spherical symmetry.

repeats the Abraham-Lorentz derivation of the equations of motion without Fermi's correction, although admitting that it can be relativistically corrected following Fermi. Ironically Fermi's third paper (see [46] for a historical discussion), which he considered only a tool for obtaining his result in that fourth paper, and which Fermi never even explicitly cited there, did make an indelible mark on relativity with the terms Fermi coordinates and Fermi-Walker transport, although even the much later paper by Walker that coupled together their names forever also ignores Fermi's original paper in Italian. Surprisingly even the text by Rohrlich updated only recently four decades after its original publication fails to connect his own adjustment of the definition of the 4-momentum of the electromagnetic field of the classical electron to Fermi's argument about the equations of motion, while the more recent books by Yaghjian and Spohn devoted to this area also show no sign that they have ever seen Fermi's argument. We hope the present work restores Fermi's message to its rightful place and perhaps provoke some thought about its meaning. A shorter version of this discussion has been published elsewhere [47] and reproduced in Appendix B.

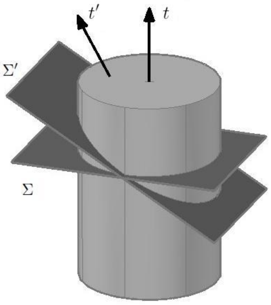


Fig. 4 The world tube of the electron sphere is a cylinder in spacetime about the t axis, shown here with one spatial dimension suppressed. The time slices $t = 0$ (Σ) and $t' = 0$ (Σ') cut this cylinder, intersecting in the spacelike 2-plane $x^1 = 0, t = 0$, which separates the spacetime region between these time slices into two disjoint subregions $x^1 > 0$ and $x^1 < 0$. Gauss's law applies separately to each of these two simply connected regions outside the electron sphere cylinder, but the signs of the outward normals of the time slices switch between these two regions, while remaining the same for the cylindrical portion of their boundaries.

Appendix. Gauss's theorem and "conservation laws"

For a divergence-free stress-energy tensor in all of Minkowski spacetime which falls off sufficiently fast at spatial infinity, its integral over any two parallel inertial time hyperplanes would be the same by Gauss's law, as explained in most standard textbooks in relativity, see Chapter 5 of Misner, Thorne and Wheeler [41], for example, or Appendix A1–5 of Rohrlich's Third Edition [34], or Anderson [38]. This gives the usual 4-momentum conservation law that the 4-momentum has the same value for different time slices for a given inertial observer. However, for two time slices associated with a pair of inertial observers in relative motion, the time slices necessarily intersect so one has to be more careful in applying Gauss's law to this more general situation, though again one finds that the 4-momentum is independent of the observer as well as the time slice. However, in the present case the nonzero divergence due to the source inside the timelike world tube of the electron sphere surface, or equivalently the boundary term on that world tube if one excludes the sources from Gauss's law, interferes with this more familiar picture, forcing the 4-momentum of the electromagnetic field to depend explicitly on the inertial observer. We consider these complications in detail in this appendix since they do not seem to be discussed in standard textbooks. The spherical shell model of the electron discussed in the first section is used to illustrate the evaluation of the Gauss law integrals.

Fig. 1 generalizes Fig. 5.3.b of Misner, Thorne and Wheeler: it represents a constant x^2, x^3 slice of the unaccelerated electron world tube centered at the origin of the unprimed spatial coordinates in spacetime. As in section 2, the unprimed coordinates are associated with the rest frame K of the electron, while the primed coordinates are associated with a frame K' in relative motion with respect to the unprimed frame is in the x^1 direction with velocity $-v < 0$ as shown in the figure. Consider the spacetime region devoid of electromagnetic sources between two spacelike hyperplanes Σ_1' and Σ_2' of constant inertial times t_1' and $t_2' > t_1'$ and outside of an internal lateral boundary σ between them which is a subset of the cylindrical timelike surface representing the world tube of the electron spherical surface ($r = r_0$ in its rest frame). Let $\overline{\Sigma}_1'$ and $\overline{\Sigma}_2'$ be the portions of those planes exterior to this cylinder. Suppose Σ_1' and Σ_2' are oriented by their future-pointing unit normal vector fields and σ by its inward unit normal $\partial/\partial r$ relative to the region of spacetime in question. Let Q be any covariant constant 4-vector so that $q^\mu = Q_\nu T_{\rm em}^{\nu\mu}$ is a divergence-free vector field in the spacetime region bounded by the three hypersurfaces $\overline{\Sigma}$, $\overline{\Sigma}'$ and σ, as well as by the lateral boundary at spacelike infinity, a region to which Gauss's law with zero volume integral and outward pointing normals applies. Taking the orientations into account relative to the outward normal on each boundary hypersurface, one then has

$$\int_{\overline{\Sigma}_2'} Q_\mu T_{\rm em}^{\mu\nu} d\Sigma_\nu - \int_{\overline{\Sigma}_1'} Q_\mu T_{\rm em}^{\mu\nu} d\Sigma_\nu = \int_\sigma Q_\mu T_{\rm em}^{\mu\nu} d\sigma_\nu \,. \tag{85}$$

If the lateral boundary term vanishes, then the integral is the same over each of the two time hypersurfaces outside the world tube of the electron sphere. Indeed for time slices in the rest frame of the electron, or in the moving frame, these integrals are time-independent, which corresponds exactly to the vanishing of the integral over the electron surface tube between the two slices. This follows for all possible projections Q_α in the explicit evaluation of the lateral integral from the vanishing of $T_{\rm em}^{0r}$ itself and of the surface integral of the spatial stress components

$$T_{\rm em}^{x^i r} = T_{\rm em}^{r x^i} = T_{\rm em}^{rr} \frac{\partial x^i}{\partial r} = -T_{\rm em}^{00} \frac{x^i}{r} \tag{86}$$

over the 2-sphere $r = r_0$, which follows from the spherical symmetry and the fact that the integral along the time direction on the cylinder is the constant rest frame time difference $t_2 - t_1 = \gamma(t_2' - t_1')$. However, even though for each such inertial coordinate system, the integral at constant time is time-independent, we must do a second calculation to relate the results of the integration with respect to inertial coordinate systems in relative motion.

In the usual textbook situation of a free electromagnetic field with no sources, one does not exclude any world tube from the Gauss law application so the internal boundary integral is not present and the divergence integral is zero. As a result the difference of the integrals over the two time parallel hyperplanes is zero. The same remarks will apply to the Gauss law application to two intersecting time hyperplanes, extending the equality of the 4-momentum integral to all inertial time slices.

The situation between the time hyperplanes of two different inertial frames is more complicated since the hyperplanes necessarily intersect, as shown in Fig. 2 with one spatial dimension suppressed, assuming that the relative velocity v along the direction x^1 of the electron rest frame relative to the moving primed frame is

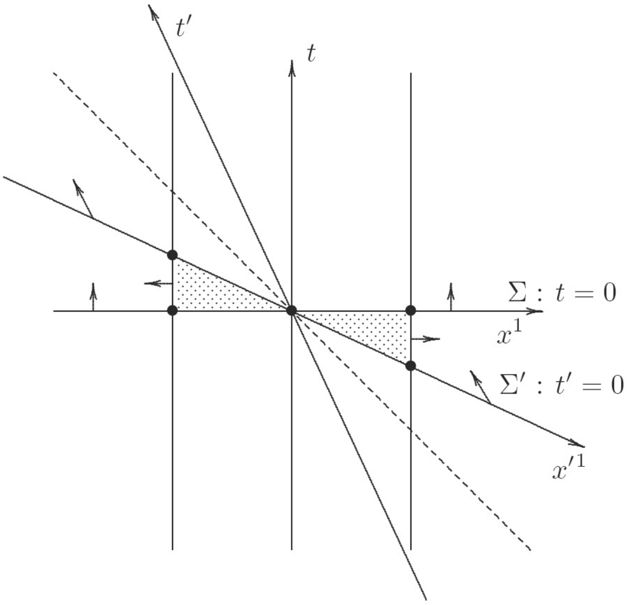


Fig. 5 Fig. 5.3.c from Misner, Thorne and Wheeler (or Fig. A1–3 from Rohrlich's Third Edition) redrawn with an inner cylindrical boundary which is the world tube of the electron sphere boundary, showing a constant x^2, x^3 slice of the previous figure. The arrows show the chosen unit normal direction for the orientation of each hypersurface, which changes sign relative to the unit outward normal of the exterior region outside the cylinder going from $x^1 > 0$ to $x^1 < 0$. Here the boundary term due to the portion σ of the cylinder between the two parallel hyperplanes is now nonvanishing. The two halves σ_+ ($x^1 > 0$) and σ_- ($x^1 < 0$) contribute terms with opposite signs to the two separate Gaussian integral relations because of the change in sign of the outward normals on Σ and Σ', and hence in the difference relation needed to reassemble the two halves of those time hypersurface integrals, they contribute a nonzero correction term.

positive, as in the previous figure. Fig. 3 shows a constant x^2, x^3 slice of Fig. 2 generalizing Fig. 5.3.c of Misner, Thorne and Wheeler [41] (or Fig. A1–3 from Rohrlich's Third Edition), but with an additional internal lateral boundary, here the portion σ of the cylinder representing the electron sphere centered around the t axis and extending between the two time slices. Consider the region of spacetime exterior to the electron sphere bounded by the time hypersurfaces $t = 0$ and $t' = 0$, with unit future-pointing normals $U = \partial/\partial t$ and $U' = \partial/\partial t'$. Let $\sigma = \sigma_- \cup \sigma_+$ be the portion of the cylindrical world tube of the electron sphere between these two time hyperplanes, divided into two disjoint parts σ_+ for $x^1 > 0$ and σ_- for $x^1 < 0$, each with the orientation induced by the outward radial normal $\partial/\partial r$ relative to the sphere. For each point on the electron sphere, σ consists of the region between $t = 0$ and $t = -vx^1$, so the integral on σ along t leads to a factor $\Delta t = 0 - (-vx^1) = vx^1 > 0$ for $x^1 > 0$ and a factor $\Delta t = -vx^1 - 0 = -vx^1 > 0$ for

$x^1 < 0$ since the integrand is independent of t along the cylinder.

Similarly let $\Sigma = \Sigma_- \cup \Sigma_+$ and $\Sigma' = \Sigma'_- \cup \Sigma'_+$, each with the future-pointing normal orientation, and let $\overline{\Sigma} = \overline{\Sigma}_- \cup \overline{\Sigma}_+$ and $\overline{\Sigma}' = \overline{\Sigma}'_- \cup \overline{\Sigma}'_+$ be the portions of those regions outside the world tube of the electron sphere. One can separately apply Gauss's law to the two disjoint regions with these boundaries and reassemble the pieces to get a relation between the integrals over $\overline{\Sigma}$, $\overline{\Sigma}'$ and σ. Since the outer normal directions switch directions for $\overline{\Sigma}$ and $\overline{\Sigma}'$ but not σ going from $x^1 > 0$ to $x^1 < 0$, one must take the difference of the two separate Gauss law relations to reassemble the total integrals over $\overline{\Sigma}$ and $\overline{\Sigma}'$, which leads to a net nonvanishing contribution from σ in spite of the spherical symmetry. One has

$$\begin{aligned}
\int_{\overline{\Sigma}_+} Q_\mu T^{\mu\nu}_{\rm em} d\Sigma_\nu - \int_{\overline{\Sigma}'_+} Q_\mu T^{\mu\nu}_{\rm em} d\Sigma_\nu &= \int_{\sigma_+} Q_\mu T^{\mu\nu}_{\rm em} d\sigma_\nu \,,\\
\int_{\overline{\Sigma}'_-} Q_\mu T^{\mu\nu}_{\rm em} d\Sigma_\nu - \int_{\overline{\Sigma}_-} Q_\mu T^{\mu\nu}_{\rm em} d\Sigma_\nu &= \int_{\sigma_-} Q_\mu T^{\mu\nu}_{\rm em} d\sigma_\nu \,,
\end{aligned} \tag{87}$$

and therefore taking the difference

$$\begin{aligned}
&\int_{\overline{\Sigma}'} Q_\mu T^{\mu\nu}_{\rm em} d\Sigma_\nu - \int_{\overline{\Sigma}} Q_\mu T^{\mu\nu}_{\rm em} d\Sigma_\nu \\
&\quad = \int_{\overline{\Sigma}'_+} Q_\mu T^{\mu\nu}_{\rm em} d\Sigma_\nu - \int_{\overline{\Sigma}_+} Q_\mu T^{\mu\nu}_{\rm em} d\Sigma_\nu \\
&\qquad\qquad - \left(\int_{\overline{\Sigma}_-} Q_\mu T^{\mu\nu}_{\rm em} d\Sigma_\nu - \int_{\overline{\Sigma}'_-} Q_\mu T^{\mu\nu}_{\rm em} d\sigma_\nu \right) \\
&\quad = - \int_{\sigma_+} Q_\mu T^{\mu\nu}_{\rm em} d\sigma_\nu + \int_{\sigma_-} Q_\mu T^{\mu\nu}_{\rm em} d\sigma_\nu \,.
\end{aligned} \tag{88}$$

Consider applying the above relation in this setting for $Q = -U'$, so that $q^\alpha = -U'_{\nu'} T^{\nu'\alpha}_{\rm em} = T^{t'\alpha}_{\rm em} = \gamma(T^{t\alpha}_{\rm em} + v T^{x^1\alpha}_{\rm em})$. Then

$$\int_{\overline{\Sigma}'} q^\alpha d\Sigma_\alpha = \int_{\overline{\Sigma}'} T^{t't'}_{\rm em}\, dV' = W' \tag{89}$$

while

$$\int_{\overline{\Sigma}} q^\alpha d\Sigma_\alpha = \int_{\overline{\Sigma}} T^{t't}_{\rm em}\, dV = \gamma W \,. \tag{90}$$

Using the exterior field in the source free region outside the electron spherical shell model as an example, one finds that the cylindrical world tube integrals, since the integrand is independent of t, are explicitly

$$\begin{aligned}
\int_{\sigma_+} q^\alpha d\sigma_\alpha &= \int_{\sigma_+} T^{t'r}_{\rm em}\, dt\, dS = \int_{\sigma_+} \gamma(T^{tr}_{\rm em} + v T^{x^1 r}_{\rm em})\, dt\, dS \\
&= \int_{-\pi/2}^{\pi/2} \int_0^\pi (0 - (-v x^1)) \gamma v \left(\frac{x^1}{r_0} T^{rr}_{\rm em} \right) r_0^2 \sin\theta\, d\theta d\phi \\
&= \gamma v^2 r_0^3 T^{rr}_{\rm em} \int_{-\pi/2}^{\pi/2} \int_0^\pi (\sin\theta \cos\phi)^2 \sin\theta\, d\theta d\phi \\
&= \frac{1}{6} \gamma v^2 (4\pi r_0^3 T^{rr}_{\rm em}) = -\frac{1}{6} \gamma v^2 W \,.
\end{aligned} \tag{91}$$

and

$$\begin{aligned}\int_{\sigma_-} q^\alpha d\sigma_\alpha &= \int_{\pi/2}^{3\pi/2}\int_0^\pi ((-vx^1)-0)\gamma v\left(\frac{x^1}{r_0}T^{rr}_{\rm em}\right) r_0^2\sin\theta\, d\theta d\phi \\ &= -\gamma v^2 r_0^3 T^{rr}_{\rm em}\int_{\pi/2}^{3\pi/2}\int_0^\pi (\sin\theta\cos\phi)^2\sin\theta\, d\theta d\phi \\ &= -\gamma v^2 r_0^3(4\pi T^{rr}_{\rm em})\frac{1}{6} = \frac{1}{6}\gamma v^2 W\,. \end{aligned} \tag{92}$$

Since the outward normals on Σ and Σ' reverse direction on the second set of integrals, but the outward normal on σ does not, the separate Gauss's law relations are

$$\begin{aligned}\int_{\overline{\Sigma}'_+} q^\alpha d\Sigma'_\alpha - \int_{\overline{\Sigma}_+} q^\alpha d\Sigma_\alpha &= -\int_{\sigma_+} q^\alpha d\sigma_\alpha \\ \int_{\overline{\Sigma}'_-} q^\alpha d\Sigma'_\alpha - \int_{\overline{\Sigma}_-} q^\alpha d\Sigma_\alpha &= \int_{\sigma_-} q^\alpha d\sigma_\alpha \end{aligned} \tag{93}$$

and their sum is

$$\begin{aligned} W' - \gamma W &= \int_{\overline{\Sigma}'} q^\alpha d\Sigma'_\alpha - \int_{\overline{\Sigma}} q^\alpha d\Sigma_\alpha \\ &= -\int_{\sigma_+} q^\alpha d\sigma_\alpha + \int_{\sigma_-} q^\alpha d\sigma_\alpha = \frac{1}{3}v^2\gamma W\,. \end{aligned} \tag{94}$$

Thus the unwanted correction factor is exactly the integral over the cylindrical boundary over the electron sphere of the moving frame 4-velocity component of the stress-energy tensor, with the factor of 1/3 equal to

$$\begin{aligned} &\frac{1}{4\pi}\int_0^{2\pi}\int_0^\pi (\sin\theta\cos\phi)^2\sin\theta\, d\theta d\phi \\ &\quad = \frac{1}{4\pi}\int_{S_2}\frac{(x^1)^2}{r_0^2}d\Omega = \frac{1}{3}\frac{1}{4\pi}\int_{S_2}\frac{r_0^2}{r_0^2}\,d\Omega = \frac{1}{3}\,, \end{aligned} \tag{95}$$

whose value follows from spherical symmetry as expressed in Eq. (30). This term which causes the result to differ from the 4-momentum as seen in the rest system is exactly due to the unbalanced outward radial stress on the charge distribution at the surface of the electron sphere.

One can repeat this calculation for $Q = \partial/\partial x'^1$ in order to express the momentum correction factor as an integral over this boundary, with one less factor of v in the correction term since

$$T_{\rm em}^{x^{1'} r} = \gamma(T_{\rm em}^{x^1 r} + vT^{tr}_{\rm em}) = \gamma T_{\rm em}^{x^1 r} \tag{96}$$

compared to the previous calculation where

$$T^{t'r}_{\rm em} = \gamma(T^{tr}_{\rm em} + vT_{\rm em}^{x^1 r}) = \gamma v T_{\rm em}^{x^1 r}\,. \tag{97}$$

With this corresponding correction term the integral relationship now becomes

$$p^{1'} - \gamma v W = \frac{1}{3}\gamma v W \to p^{1'} = \frac{4}{3}\gamma v W\,. \tag{98}$$

explaining the famous factor of 4/3.

On the other hand for the model with a uniform distribution of charge within the electron sphere, one must extend the hypersurface integrals over the interior region to evaluate the total 4-momentum in the electromagnetic field since the field is no longer zero there. Only by doing this does the self-energy integral of the static charge configuration agree with the energy in the electric field it generates. This forces one instead to consider the spacetime volume divergence integral over that region in applying Gauss's law, rather than the spherical boundary hypersurface integral. One could do the same for the spherical shell model, where the divergence integral would yield the same result as the spherical boundary integral evaluated above when excluding the region containing the charge.

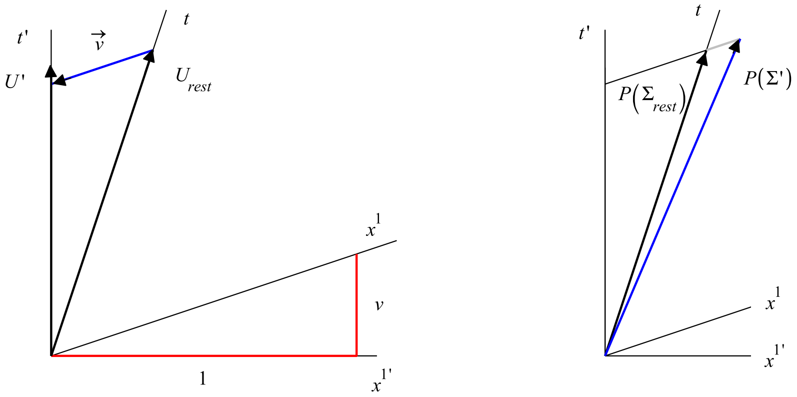


Fig. 6 Left: The plane of the two inertial observer 4-velocities for motion along the x^1-axis. The rest frame axis x^1 has slope ν. A unit vector along this axis has primed $0'$ and $1'$ components $\langle\gamma\nu,\gamma\rangle$. The relative velocity of U' as seen by the rest frame observer with 4-velocity $U_{\rm rest}$ is $\vec{\nu}(U',U_{\rm rest})$, which extends from the tip of $U_{\rm rest}$ to the vertical axis along U', and whose $0'$ and $1'$ components are $-\nu\langle\gamma\nu,\gamma\rangle$. Right: The rest frame 4-momentum and the moving frame 4-momentum.

We can easily re-express the above component relationships (94) and (98) in 4-vector form. The subtracted terms on the left hand side are exactly the moving frame inertial coordinate components of the 4-momentum as seen by the rest frame

$$\langle P(\overline{\Sigma}_{\rm rest})^{0'}, P(\overline{\Sigma}_{\rm rest})^{1'}\rangle = \langle \gamma W, \gamma v W\rangle\,. \tag{99}$$

The right hand sides instead have corresponding primed components $\frac{1}{3}\gamma v\langle v,1\rangle$, which can be re-expressed as follows. The 4-vector with its first two primed components equal to $v\gamma\langle v,1\rangle$ is just the sign-reversed relative velocity of the moving frame compared to the rest frame as seen in the rest frame, call its components $-\nu(U',U_{\rm rest})^{\alpha'}$. See Fig. 5. The rest energy is just $W = P(\overline{\Sigma}_{\rm rest})^0 = -P(\overline{\Sigma}_{\rm rest})_\beta U^\beta_{\rm rest}$, where for emphasis we include the subscript notation for the rest frame quantities. Thus we get in index-free notation

$$\begin{aligned} P(\overline{\Sigma}') - P(\overline{\Sigma}_{\text{rest}}) &= -\frac{1}{3} W \nu(U', U_{\text{rest}}) \\ &= \frac{1}{3} P(\overline{\Sigma}_{\text{rest}})_\beta U^\beta_{\text{rest}} \nu(U', U_{\text{rest}})\,, \end{aligned} \tag{100}$$

so that we get the following orthogonal decomposition of the general 4-momentum

$$\begin{aligned} P(\overline{\Sigma}') &= P(\overline{\Sigma}_{\text{rest}}) + \frac{1}{3} P(\overline{\Sigma}_{\text{rest}})_\beta U^\beta_{\text{rest}} \nu(U', U_{\text{rest}}) \\ &= W \left(U_{\text{rest}} - \frac{1}{3} \nu(U', U_{\text{rest}}) \right) . \end{aligned} \tag{101}$$

This extra 4-vector piece aligned with the relative velocity of the moving frame with respect to the rest frame is what causes the 4-momentum to depend on the observer 4-velocity relative to the rest system, causing it to deviate from the desired 4-vector momentum. Its scalar coefficient is directly related to the unbalanced radial stress at the surface of the electron sphere.

Poincaré stresses are introduced within the electron sphere so that they exactly compensate for this radial stress, but then they add their own contribution to the total conserved 4-momentum, which is aligned with the 4-velocity of the electron sphere. Schwinger has a detailed discussion of these additional stresses [21]. The best choice to fix the arbitrariness of his family of models simply eliminates the extra unwanted term along the relative velocity to make the total 4-momentum equal to the rest frame value for the electromagnetic field alone ($h = -1$ in the notation of Jackson [37]). For this choice the stress-energy tensor of the Poincaré stresses is proportional to the projection $g_{\alpha\beta} + U_{\text{rest}\,\alpha} U_{\text{rest}\,\beta}$ into the local rest space of the rest frame, whose contraction with the volume element of the rest frame time hyperplane therefore vanishes, so in that frame the total 4-momentum integral reduces to the integral of the electromagnetic stress-energy tensor alone. Since the total stress-energy tensor is divergence-free, observers in relative motion therefore measure the same 4-vector 4-momentum $W U_{\text{rest}\,\alpha}$ as in the rest frame.

One can directly evaluate the difference in the 4-momentum 4-vector observed by the rest and moving frame observers in a few lines using Gauss's law Eq. (15) expressed in rest frame inertial coordinates applied to the entire spacetime region between the rest frame time slice Σ_{rest}: $t = 0$ and the moving frame time slice Σ': $t' = \gamma(t + vx^1) = 0$, or $t = -vx^1$. See Fig. 6. For the shell model of the electron the spherical surface density limited to the sphere $r = r_0$ is $\rho = e/(4\pi r_0^2)\delta(r - r_0)$. Because of the delta function, the spacetime volume integral reduces to a hypersurface integral over the spherical cylinder with volume element $dt\, r_0^2 d\Omega$, but one has to take into account the fact that the corresponding radial electric field has a Heaviside function factor: the inertial components of the rest frame Coulomb field are those of the radial inverse square field $H(r - r_0)e/r^3 \langle 0, x^1, x^2, x^3 \rangle$, using the rest frame inertial coordinate component notation: $z = \langle z^0, z^1, z^2, z^3 \rangle$, where

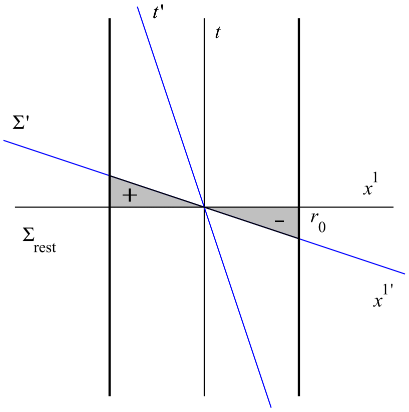


Fig. 7 A 2-dimensional cross-section of the region between the rest and moving frame inertial time hypersurfaces $t = 0$ and $t' = 0$ for relative motion along the x^1 direction. Applying Gauss's law Eq. (15) to this region requires opposite signed orientations for the spacetime regions on opposite sides of the plane of intersection $x^1 = 0 = t$.

$H(r)$ is the Heaviside function:

$$H(x) = \begin{cases} 0 & x < 0 \\ 1/2 & x = 0 \\ 1 & x > 0 \end{cases} . \tag{102}$$

Recalling that the two regions into which the plane of intersection of these time hyperplanes are divided have opposite orientation for the spacetime region integral, we get for the integral over the spherical shell between the hyperplanes

$$\begin{aligned} P(\Sigma') - P(\Sigma_{\text{rest}}) &= -\int_R \rho E(U) d^4V \\ &= -\int_{x^1<0}\int_0^{-vx^1} \frac{e}{4\pi r_0^2}\delta(r-r_0)H(r-r_0)\frac{e}{r^3}\langle 0, x^1, x^2, x^3\rangle \, dt\, r^2 dr\, d\Omega \\ &\quad + \int_{x^1>0}\int_{-vx^1}^0 \frac{e}{4\pi r_0^2}\delta(r-r_0)H(r-r_0)\frac{e}{r^3}\langle 0, x^1, x^2, x^3\rangle \, dt\, r^2 dr\, d\Omega \\ &= v\frac{e^2}{r_0}H(0)\left(\int_{S_2,x^1<0} \frac{1}{4\pi r_0^2} x^1\langle 0, x^1, x^2, x^3\rangle \, d\Omega \right. \\ &\quad \left. + \int_{S_2,x^1>0} \frac{1}{4\pi r_0^2} x^1\langle 0, x^1, x^2, x^3\rangle \, d\Omega\right) \\ &= v\frac{e^2}{r_0}H(0)\langle 0,1,0,0\rangle \int_{S_2} \frac{1}{4\pi}\left(\frac{x^1}{r_0}\right)^2 d\Omega \\ &= \frac{1}{3}v\frac{e^2}{2r_0}\langle 0,1,0,0\rangle = \frac{1}{3}Wv\langle 0,1,0,0\rangle = -\frac{1}{3}W\nu(U', U)\,, \end{aligned} \tag{103}$$

where the integrals over the half spheres of the products x^1x^2 and x^1x^3 vanish by symmetry, and the one remaining integral by symmetry is 1/3 the integral with $(x^1)^2$ replaced by r_0^2. The factor $H(0) = 1/2$ results from the limiting situation of integration over a thin shell of finite thickness where the radial electric field rises from 0 to its value at the outer edge of the shell, so its integral over the shell against the constant charge density function leads to the average value of the electric field, which has an additional factor of 1/2.

Just for fun, suppose we evaluate the spatial momentum in the moving frame in terms of the rest frame inertial coordinates, where the future-pointing normal to the primed inertial time hypersurface $t' = \gamma(t + vx^1) = 0$ is $n = \gamma\langle 1, -v, 0, 0\rangle$, while $\gamma dV' = dV$, so that on Σ', the volume element is $\langle d\Sigma_\alpha\rangle = -\langle n_\alpha dV'\rangle = dV\langle 1, v, 0, 0\rangle$. Then $-T^{1\alpha}n_\alpha dV' = T^{1j}n_j = T^{11}(\gamma v)dV' = T^{11}v$, so that

$$\begin{aligned} P(\Sigma')^1 &= \int_{\Sigma'} T^{1\alpha} d\Sigma_\alpha = \int_{r_0}^{\infty}\int_{S_2} T^{11} v r^2 dr\, d\Omega = -\frac{1}{3}\int_{r_0}^{\infty}\int_{S_2} T^{00} v r^2 dr\, d\Omega \\ &= \frac{vW}{3}\,. \end{aligned} \tag{104}$$

Noting that $\nu(U', U) = -\nu$ and $P(\Sigma_{\text{rest}})^1 = 0$, we thus recover exactly the previous result (103).

References

[1] M. Abraham, Ann. Phys. (Leipzig) **10**, 105 (1903); Phys. Z. **5**, 576 (1904).

[2] H.A. Lorentz, *The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat* (Dover, New York, 1952); First Edition 1909, from lectures of 1906.

[3] H. Poincaré, C. R. Acad. Sci. **140**, 1504 (1905); Rend. Circ. Mat. Palermo **21**, 129 (1906): English translation by H.M. Schwartz, "Poincaré's Rendiconti Paper on Relativity. Part I, II, III," Am. J. Phys. 39: 1287 (1971), 40: 862872 (1972), 40: 1282 (1972).

[4] E. Fermi, Nuovo Cim. **22**, 199 (1921).

[5] E. Fermi, Nuovo Cim. **22**, 176 (1921).

[6] E. Fermi, Rend. Lincei, **31** (I), 21–23, 51–52, 101–103 (1922).

[7] E. Fermi, Phys. Z. **23**, 340 (1922); Nuovo Cim. **25**, 159 (1923); original Italian article at http://www.archive.org/details/collectedpapersn007155mbp.

[8] M. Born, Ann. Physik **30** 1, (1909); Phys. Z. **11**, 233 (1910).

[9] B. Kwal, J. Phys. Radium **10**, 103 (1949).

[10] F. Rohrlich, Am. J. Phys. **28**, 639 (1960).

[11] J. Aharoni, *The Special Theory of Relativity*, Oxford University Press, First Edition: 1959, Second Edition: 1965, see Chapter 5.

[12] G. Salzman and A.H. Taub, Phys. Rev. **95**, 1659 (1954)

[13] M. von Laue, *Das Relativitätsprinzip* (Vieweg: Braunschweig, 1911).

[14] W. Wilson, Proc. Phys. Soc (London) **48**, 736 (1936).

[15] R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman Lectures on Physics* (Addison-Wesley, Reading, MA, 1964): Volume II , Chapter 28.

[16] C. Teitelboim, Phys. Rev. **D1**, 1572, 1970; **D3**, 297, 1971; **D4**, 345, 1971.

[17] C. Teitelboim, D. Villarroel and Ch. Van Weert, Revista Nuovo Cim. **3**, 1, 1980.

[18] T.H. Boyer, Phys. Rev. **D25**, 3246 (1982).

[19] F. Rohrlich, Phys. Rev. **D25**, 3251 (1982).

[20] Nodvik, J.S.: Ann. Phys. **28**, 225 (1964).

[21] J. Schwinger, Found. Phys. **13**, 373 (1983).

[22] I. Campos and J.L. Jimnez, Phys. Rev. **D33**, 607 (1986); Eur. J. Phys.]bf 13, 117 (1992).

[23] J.M. Cohen and E. Mustafa, Int. J. Theor. Phys., **25**, 717 (1986).

[24] E. Comay, Int. J. Theor. Phys. **30**, 1473 (1991).

[25] P. Moylan, Amer. J. Phys. **63**, 818 (1995).

[26] H. Kolbenstvedt, Phys. Lett. **A234**, 319 (1997).

[27] F. Rohrlich, Am. J. Phys. **65**, 1051 (1997).

[28] W. Appel and M. Kiessling, Ann. Phys. **289**, 24 (2001).

[29] J.P. de Leon, Gen. Relativ. Grav. **36**, 1453 (2004).

[30] A.I. Harte, Phys. Rev. **D73**, 065006 (2006).

[31] F. Pinto, Phys. Rev. **D73**, 104020 (2006).

[32] A. Bettini, *Riv. del N. Cimento* **32**, 295–337 (2009).

[33] C.R. Galley, A.K. Leibovich, and I.Z. Rothstein, Phys. Rev. Lett. **105**, 094802 (2010).

[34] F. Rohrlich, *Classical Charged Particles* (Addison-Wesley, Reading, 1965); updated Third Edition (World Scientific, Singapore, 2007).

[35] A.D. Yaghjian, *Relativistic Dynamics of a Charged Sphere: Updating the Lorentz-Abraham Model*, (Berlin: Springer, Berlin, 1992; Second Edition 2006); Phys. Rev. **E78**, 046606 (2008).

[36] H. Spohn, *Dynamics of Charged Particles and their Radiation Field* (Cambridge University Press, Cambridge, 2004).

[37] J. Jackson, *Classical Electrodynamics* (Wiley, New York, Second Edition: 1975, Third Edition: 1999); see respectively Chapters 17, 16.

[38] J.L. Anderson, *Principles of Relativity Physics*, (Academic Press, New York, 1967).

[39] M. Janssen and M. Mecklenburg, pp. 65-134 in V.F. Hendricks, K.F. Jørgensen, J. Lützen, and S.A. Pedersen (Eds.), *Interactions: Mathematics, Physics and Philosophy, 1860-1930* (Springer, Dordrecht 2007); available at http://www.tc.umn.edu/∼janss011/; see also M. Janssen, PhD thesis "A comparison between Lorentz's ether theory and special relativity in the light of the experiments of Trouton and Noble," see Chapter 2, http://www.mpiwg-berlin.mpg.de/litserv/diss/janssen_diss/ and M. Janssen, Studies in History and

Philosophy of Modern Physics **40**, 26 (2009).

[40] S. Parrott, Comment on Phys. Rev. D 60 084017 "Classical self-force" by F. Rohrlich, arXiv: gr-qc/0502029.

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

[42] T. Levi-Civita, Rendiconti della Accademia dei Lincei **26** (1917); **27** (1918); **28** (1918); Fermi refers to Note II of this series in 1917, p. 3.

[43] E. Fermi and A. Pontremoli, Rend. Lincei **32**, 162 (1923).

[44] S. Boughn and T. Rothman, 2011, e-print: http://arxiv.org/abs/1108.2250.

[45] M.A.H. MacCallum and A.H. Taub, Commun. Math. Phys. **25**, 173 (1972).

[46] D. Bini and R.T. Jantzen, Proceedings of the Ninth ICRA Network Workshop on Fermi and Astrophysics, edited by V. Gurzadyan and R. Ruffini (World Scientific, Singapore, 2003): Nuovo Cim. **117B**, 983 (2002) [http://arXiv.org/abs/gr-qc/0202085]

[47] R.T. Jantzen and R. Ruffini, Gen. Relativ. Grav. **44**, 2063 (2012).

A.2 R.T. Jantzen and R. Ruffini: Fermi and electromagnetic mass

R.T. Jantzen and R. Ruffini: “Fermi and electromagnetic mass,” *Gen. Rel. Grav.* **44**, 2063 (2012).

Gen Relativ Gravit (2012) 44:2063–2076
DOI 10.1007/s10714-012-1381-0

RESEARCH ARTICLE

Fermi and electromagnetic mass

Robert T. Jantzen · Remo Ruffini

Received: 18 December 2011 / Accepted: 7 May 2012 / Published online: 23 May 2012

Abstract Fermi's analysis of the contribution of the electromagnetic field to the inertial mass of the classical electron within special relativity is brought to its logical conclusion, leading to the conservation of the total 4-momentum of the field plus mechanical mass system as seen by the sequence of inertial observers in terms of which the accelerated electron is momentarily at rest.

Keywords Special relativity · Electromagnetic mass · Conservation laws

1 Introduction

In 1921–1923 Enrico Fermi [1–7] wrote his first four scientific papers in a series addressing the question of the contribution of the energy in the Coulomb field of a classical model of the electron to its inertial mass within special relativity. This model had been developed in the first decade of the 1900s by Abraham [8,9] and Lorentz [10] during the same period in which special relativity was being born. Fermi's second paper [2] studied this question within general relativity using a metric introduced by Levi–Civita representing a spacetime reference frame accelerated along one spatial direction. Fermi's third paper [3,11,12] addressed a side issue in this series—the mathematical theory of his Fermi coordinate system and Fermi–Walker transport (both

R. T. Jantzen
Department of Mathematics and Statistics, Villanova University, Villanova, PA 19085, USA

R. T. Jantzen (✉) · R. Ruffini
ICRA, University of Rome "La Sapienza", P.le Aldo Moro 5, 00185 Rome, Italy
e-mail: robert.jantzen@villanova.edu

R. Ruffini
Physics Department, ICRANet, P.zza della Repubblica 10, 65122 Pescara, Italy

extensively employed by Synge in his early textbook on general relativity [13]), the latter of which became a key tool in the theory of general relativity—while the culminating fourth paper written in three versions in Italian and German but never available in English until now [14], though often quoted, has rarely been appreciated nor understood for its actual content. Fermi himself stopped short of considering his result in the fourth paper in the context of his third, namely by considering the electromagnetic contribution to the inertial mass together with a contribution from an additional mass source (mechanical or bare mass). We finish his calculation here. Furthermore although the topic continues to interest people even today as an interesting physics question, the natural completion of his work by applying it to the controversial question of the nature of the 4-momentum integrals for the electromagnetic field has never been correctly considered. We do so here. Details may be found in [15].

An unfortunate complication in this story was the confusion of the entirely separate issue of the stability of the electron with the issue of attributing a unique 4-momentum to its electromagnetic field. Unlike the 4-momentum of a point particle which is a uniquely defined 4-vector at a spacetime point, the 4-momentum of the electromagnetic field in the presence of sources is a nonlocal measurement by an inertial observer which is represented mathematically by an integral over a spacelike hyperplane of constant inertial coordinate time in the observer's associated inertial reference system, and whose result depends on the entire field at such a moment of time. In general this produces a different 4-vector for every inertial observer and for every choice of time in that observer's system of reference. This is a consequence of the nonvanishing divergence of the stress-energy tensor of the electromagnetic field when sources are present, in contrast to the situation for divergence-free such tensors where Gauss's law guarantees that the 4-momentum is independent of the inertial observer and choice of inertial time. Historically the Lorentz transformed components of the rest frame 4-momentum were compared to the components of the distinct 4-momentum seen by an observer in relative motion in the associated inertial coordinate system, but since these are components of two distinct 4-vectors, they cannot agree. It should have been expected that this comparison would fail, but instead this was seen as an apparent problem.

Poincaré [16–20] attempted to restore a unique total 4-momentum result by considering the combined system of the electromagnetic extended charge model with stabilizing stresses that would yield a divergence-free total energy-momentum tensor, thus "closing the system." However, in so doing, he obscured the fact that the electromagnetic field, which gave birth to special relativity through its Lorentz invariance, should make a contribution to the total mass-energy of the electron which is by itself relativistically correct. This perpetuated a basic error with the Abraham–Lorentz model rather than correcting it.

The key to resolving these complications with the model was the notion of rigidity later introduced by Born in 1909 [21,22], the only notion of rigidity that is compatible with special relativity. Fermi understood how to use this condition to invalidate the starting point of the Abraham–Lorentz calculation of the equation of motion for a rigid extended spherically symmetric electron accelerated by an external electromagnetic field—that the total electromagnetic force on the electron at a moment of inertial time in which it is instantaneously at rest be zero—and correct it using his Fermi coordinate system which inserts a Fermi coordinate lapse function factor into the integration of the differential forces over the corresponding time hyperplane. This led to the "correct" mass-energy relationship between the energy of the self-field of the electron and its inertial mass. However, as we will see, it also leads to a conserved total 4-momentum that is naturally associated with the 4-velocity of the electron in the expected way.

2 Electromagnetic preliminaries

We follow the conventions of Misner, Thorne and Wheeler [23] for the $-+++$ signature metric $g_{\alpha\beta}$ of Minkowski spacetime, which in inertial coordinates $(x^\alpha) = (t, x^i)$ with the identification $x^0 = t$ has nonzero components $-g_{00} = g_{ii} = 1$ (Greek and Latin indices run from 0 to 3 and 1 to 3 respectively, units are chosen so that $c = 1$) and for the electromagnetic field tensor $F_{\alpha\beta}$, whose stress-energy tensor

$$T^{\mu\nu} = \frac{1}{4\pi}\left(F^{\mu\alpha}F^{\nu}{}_{\alpha} - \frac{1}{4}g^{\mu\nu}F^{\alpha\beta}F_{\alpha\beta}\right) \tag{1}$$

has nonzero divergence

$$T^{\mu\nu}{}_{;\nu} = -F^{\mu}{}_{\nu}J^{\nu}, \tag{2}$$

as a result of the Maxwell equation $F^{\alpha\beta}{}_{;\beta} = 4\pi J^\alpha$, where J^α is the 4-current.

Gauss's law can only be applied to a 4-vector field on Minkowski spacetime, so introduce a covariant constant vector field Q^α, $Q^\alpha{}_{;\beta} = 0$ representing a translation Killing vector field and let $\mathcal{J}^\alpha = Q_\beta T^{\beta\alpha}$, so that $\mathcal{J}^\beta{}_{;\beta} = Q_\alpha T^{\alpha\beta}{}_{;\beta}$. Let R be the spacetime region between two spacelike hyperplanes Σ_1 and Σ_2 oriented by their future-pointing unit normals $u^\alpha_{(1)}$ and $u^\alpha_{(2)}$ which are the 4-velocities of the corresponding inertial observers. Provided that the fields fall off sufficiently fast at spatial infinity so that the closing timelike boundary integral there between the two hyperplanes vanishes, Gauss's law states

$$\int_R \mathcal{J}^\beta{}_{;\beta}\, d^4V = \int_{\Sigma_2} \mathcal{J}^\beta d\Sigma_\beta - \int_{\Sigma_1} \mathcal{J}^\beta d\Sigma_\beta, \tag{3}$$

where for a single hypersurface Σ, the hypersurface volume element is $d\Sigma_\beta = -u_\beta d\Sigma$, so that

$$\int_\Sigma \mathcal{J}^\beta d\Sigma_\beta = \int_\Sigma (-u_\beta \mathcal{J}^\beta)\, d\Sigma \tag{4}$$

is the integral of the future-normal component of the vector field with respect to the intrinsic volume element $d\Sigma = dV_\Sigma$. In inertial coordinates adapted to the 4-velocity u^α so that Σ coincides with a hyperplane of constant inertial time t, this is just $dV_\Sigma = dx^1dx^2dx^3$, and the hyperplane integral is just a triple integral with respect to these spatial coordinates, while the spacetime volume element is then $d^4V = dt\,dx^1dx^2dx^3$. For intersecting such hyperplanes Σ_1 and Σ_2 associated with observers in relative motion, R must be oriented oppositely on the two disjoint pieces into which the intersection divides it, with the half for which Σ_2 is the future boundary oriented positively, and the other half oriented negatively (see Fig. 5.3.c of Misner, Thorne and Wheeler [23]). Thus

$$\int_R Q_\alpha T^{\alpha\beta}{}_{;\beta} d^4V = \int_{\Sigma_2} Q_\alpha T^{\alpha\beta} d\Sigma_\beta - \int_{\Sigma_1} Q_\alpha T^{\alpha\beta} d\Sigma_\beta, \tag{5}$$

where if we agree to evaluate these expressions in inertial coordinates where Q_α are constants, then they can be factored out of the equation.

The inertial coordinate components of the 4-momentum of the electromagnetic field as seen by an inertial observer with 4-velocity u^α at a moment of time t in the observer rest frame represented by a spacelike time coordinate hyperplane Σ (for which u^α is in fact the future-pointing timelike unit normal vector field) is given by the integral formula

$$P(\Sigma)^\alpha = \int_\Sigma T^{\alpha\beta} d\Sigma_\beta. \tag{6}$$

In inertial coordinates where $u^\alpha = \delta^\alpha{}_0$ this gives the energy and momentum as the integral of the local energy density and the Poynting vector respectively

$$P(\Sigma)^0 = \int_\Sigma T^{00} dV_\Sigma,\, P(\Sigma)^i = \int_\Sigma T^{0i} dV_\Sigma. \tag{7}$$

While the contracted pair of indices in the integral (6) can be evaluated in any coordinates, one can integrate over an object with a free index only if that index is expressed in some inertial coordinate system where it makes sense to compare 4-vectors at different spacetime points in the flat spacetime due to the path independence of parallel transport. In such coordinates we then have from Eqs. (2), (5) and the definition (6)

$$\int_R -F^\alpha{}_\beta J^\beta d^4V = P(\Sigma_2)^\alpha - P(\Sigma_1)^\alpha. \tag{8}$$

When $J^\alpha = 0$, the left hand side is zero, showing that the 4-momentum vector functional is independent of the hyperplane and defines a single 4-vector which represents the conserved 4-momentum of the free electromagnetic field.

3 Lagrangian equations in Fermi coordinates

The Born rigidity condition requires that the charge and mass density profiles of an electron model be time-independent in the Fermi coordinate system adapted to a world

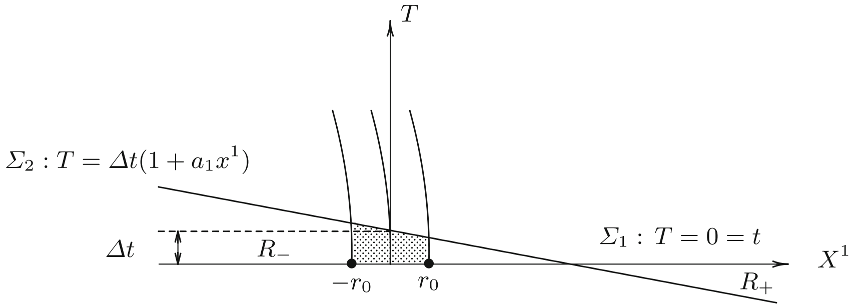


Fig. 1 Inertial Cartesian coordinates (T, X^1) with Fermi coordinates (t, x^1) such that $t = 0 = T$ coincide, showing an $X^2 = 0 = X^3$ cross-section of the world tube of an electron sphere instantaneously at rest at the origin at $T = 0$ but accelerated in the negative $x^1 = X^1$ direction ($a_1 < 0$). At a successive Fermi time Δt later, the Fermi time hyperplanes intersect to the right of the world tube (equivalent to the assumption $|a_1|r_0 < 1$). The spacetime region within the electron world tube between the two slices (*shaded* in this plane cross-section) occurs in the Gauss's law application to the wedge between the two time slices, namely $R = R_- \cup R_+$, two regions which are separated from each other by a plane of constant x^1 within the hypersurface $t = 0$ shown as the intersection point in this diagram; R_- must be positively oriented, but R_+ negatively oriented for Eq. (3)

line within the localized matter distribution. The constant Fermi time hyperplanes in such a coordinate system are orthogonal to this world line at their point of intersection, representing the local rest space of the associated comoving observer at that point of the world line. In fact the Fermi time coordinate lines are always orthogonal to the Fermi time coordinate hyperplanes; for this reason these coordinates are often known as Fermi normal coordinates.

The classical model of the nonrotating electron assumes a spherically symmetric distribution of mass and charge within a sphere of radius r_0 of the central world line in such a coordinate system, where the metric line element has the form

$$ds^2 = -N_F^2 dt^2 + \delta_{ij} dx^i dx^j , \quad N_F = 1 + a_i x^i \tag{9}$$

and a_i are the Fermi coordinate components of the 4-acceleration a^α of the central world line $x^i = 0$, where the proper time derivative along the time lines $d/d\tau = N_F^{-1} d/dt$ reduces to the Fermi coordinate time derivative; in the Fermi coordinates one has $a^\alpha = \delta^\alpha{}_i a^i$. The spacetime volume element is $d^4 V = N_F \, dt \, dV$, where $dV = dx^1 dx^2 dx^3$ is the spatial volume element. See Misner, Thorne and Wheeler [23] for details of this coordinate system. Figure 1 shows a 2-dimensional cross-section of two successive Fermi time hyperplanes for a central world line decelerating along the x^1 direction, and the interpretation of the Fermi lapse function for an infinitesimal increment Δt of Fermi time.

The time lines are the world lines of the elements of the charged matter distribution, having Fermi coordinate 4-velocity components

$$U^\alpha = \frac{dx^\alpha}{d\tau} = \frac{1}{N_F} \frac{dx^\alpha}{dt} = \frac{1}{N_F} \delta^\alpha{}_0 , \tag{10}$$

from which one obtains the acceleration $a^\alpha = DU^\alpha / dt|_{x^i = 0}$ of the central world line. Let ρ and $\rho_{(\mathrm{me})}$ be the spherically symmetric charge and mechanical or bare mass densities, which are functions only of the radius $r = (\delta_{ij} x^i x^j)^{1/2}$ and which vanish outside $r = r_0$. The Abraham–Lorentz spherical shell model assumes a delta function distribution at $r = r_0$: $\rho_{\text{shell}} = \delta(r - r_0)e/(4\pi r_0^2)$, where e is the total charge of the electron; one may also easily consider a uniform density distribution within the sphere of radius r_0. Let $de = \rho\, dV$ and $dm_{(\text{me})} = \rho_{(\text{me})} dV$ be the elements of the charge and mechanical mass distributions, so that $e = \int \rho\, dV$ is the total charge and $m_{(\text{me})} = \int \rho_{(\text{me})}\, dV$ is the total mechanical mass (also called bare mass). The 4-current is then $J^\alpha = \rho\, U^\alpha$.

The action for the electromagnetic field together with the matter distribution considered by Fermi in his third paper [3] is

$$S = \int\limits_R \left(-\frac{1}{16\pi} F^{\alpha\beta} F_{\alpha\beta} + A_\alpha J^\alpha \right) d^4x - \int d\tau\, dm_{(\text{me})}, \tag{11}$$

where the second integral in the Lagrangian is the integral with respect to the differential of mechanical mass of the line integral over the world line of the matter element, while the second term in the first integral here can be similarly expressed as

$$\begin{aligned} \int \rho A_\alpha U^\alpha N_F\, dt\, dV &= \int \rho A_\alpha \frac{dx^\alpha}{d\tau} N_F\, dt\, dV \\ &= \int \rho A_\alpha \frac{dx^\alpha}{dt}\, dt\, dV = \int A_\alpha dx^\alpha\, de, \end{aligned} \tag{12}$$

showing that it is a parametrization-independent line integral integrated over the charge distribution. The region R of integration is assumed to be a cylindrical region with respect to the Fermi coordinate system between two fixed Fermi times, over an arbitrary time-independent spatial region B in the Fermi coordinate system. The action is a function of the 4-potential of the electromagnetic field, in terms of which $F_{\alpha\beta} = dA_{\alpha\beta} = A_{\beta,\alpha} - A_{\alpha,\beta}$, and of the world lines of the matter distribution, which are the time lines of the Fermi coordinate system. Varying the action with respect to A_α yields the remaining Maxwell's equations.

The first term in the action is independent of the world lines. Varying the world lines such that $\delta x^\alpha = \delta^\alpha{}_i \delta x^i$ leads to the Lagrangian equations of motion for the central world line of the rigid charged matter distribution. Varying the 4-current term with respect to the world lines, as shown by Fermi [3–6], ignoring a boundary term which arises from an integration by parts in time, leads to

$$\begin{aligned} &\int\limits_{t_1}^{t_2} \left(\int\limits_B F_{\alpha\beta} \frac{dx^\alpha}{dt} de \right) \delta x^\beta\, dt \\ &\quad = \int\limits_{t_1}^{t_2} \left(\int\limits_B F_{\alpha\beta} U^\alpha N_F\, de \right) \delta x^\beta\, dt \\ &\quad = \int\limits_{t_1}^{t_2} \left(\int\limits_B E(U)_i N_F\, de \right) \delta x^i\, dt. \end{aligned} \tag{13}$$

where $E(U)^\alpha = F^\alpha{}_\beta U^\beta$ is the electric field as seen by the Fermi coordinate observer with 4-velocity U^α. The variation of the mechanical mass term yields $-\int_{t_1}^{t_2} m_{(\mathrm{me})} a_i\, \delta x^i\, dt$.

If the variations δx^i are arbitrary functions of the Fermi time, vanishing at the endpoint Fermi times to justify ignoring the integration by parts boundary term, then one obtains the Fermi condition, now amended by the re-insertion of the mechanical mass term

$$\int\limits_B \rho E(U)_i N_F\, dV - m_{(\mathrm{me})} a_i = 0. \tag{14}$$

This condition with $m_{(\mathrm{me})} = 0$, as assumed by Fermi in his fourth paper and by Abraham and Lorentz in their purely electromagnetic model of the electron, differs from the Abraham–Lorentz starting condition for their derivation of the equations of motion only by the additional factor of the Fermi coordinate lapse function in the integral, see Jackson [27] who reproduces their calculation. However, the first term in (14) reversed in sign is exactly the Gauss integral integrand for the integral over t in Eq. (8), namely

$$\int\limits_R -F^\alpha{}_\beta J^\beta d^4V = -\int\limits_{t_1}^{t_2} \left(\int\limits_B \rho \delta^\alpha{}_i E(U)^i N_F\, dV \right) dt, \tag{15}$$

so that using the Fermi condition to replace the expression in parentheses, the latter becomes

$$\int\limits_R -F^\alpha{}_\beta J^\beta d^4V = -\int\limits_{t_1}^{t_2} m_{(\mathrm{me})} \delta^\alpha{}_i a^i\, dt. \tag{16}$$

However, to evaluate this vector integral we need to express its components in an inertial coordinate system where we can utilize the relation $a^\alpha = DU^\alpha/d\tau = dU^\alpha/d\tau$, remembering that the Fermi coordinate time is the proper time along the central world line

$$\begin{aligned} m_{(\mathrm{me})} \int\limits_{t_1}^{t_2} a^\alpha\, d\tau &= m_{(\mathrm{me})} \int\limits_{t_1}^{t_2} \frac{dU^\alpha}{d\tau}\, d\tau \\ &= m_{(\mathrm{me})} U^\alpha |_{t_1}^{t_2} = p^\alpha_{(\mathrm{me})} |_{t_1}^{t_2}, \end{aligned} \tag{17}$$

which are the inertial coordinate components of the mechanical 4-momentum of the rigid matter distribution. Gauss's law (8) using (15)–(17) then becomes

$$-p^\alpha_{(\mathrm{me})} |_{t_1}^{t_2} = P(\Sigma_{t_2})^\alpha - P(\Sigma_{t_1})^\alpha, \tag{18}$$

or

$$p^{\alpha}_{(\mathrm{me})}(t_1) + P(\Sigma_{t_1})^{\alpha} = p^{\alpha}_{(\mathrm{me})}(t_2) + P(\Sigma_{t_2})^{\alpha}, \tag{19}$$

showing that the total 4-momentum of the system as seen by the Fermi coordinate comoving observer is independent of the Fermi time, a result apparently overlooked until now. Aharoni [24], who put out a new edition of his textbook on special relativity in 1965 in order to explain Fermi's work on this particular problem after its rediscovery, got very close to this result with his postulated self-force introduced in his reinterpretation of Fermi's results—but he missed it by neglecting to consider Gauss's law for the electromagnetic field with sources. Attention had been brought to Fermi's work in 1960 by Rohrlich's discussion of the 4-momentum integral for the electromagnetic field of the unaccelerated spinless classical electron [25] without knowledge of Fermi's work or of the same conclusions reached earlier in 1949 by Kwal [26], who was also unaware of Fermi's work.

It should also be noted that Nodvik generalized this model to include spin by adding Euler angles describing the orientation of the electron spin axis with respect to a Fermi–Walker propagated orthonormal frame along the central world line [28], as more recently updated and extended by Appel and Kiessling [29,30] and reviewed by Spohn [31]. The bare mass contribution to the Lagrangian is then modified by the Lorentz gamma factor of the motion of the elementary elements of the mass distribution with respect to the central world line, described by the intrinsic "gyration" angular velocity of the electron. However, the resulting discussion becomes extremely complicated and very difficult to follow for those of us who are not experts in advanced classical electrodynamics.

The advantage of our presentation for the spinless model is that it retains the elegance and simplicity of the work initiated by Fermi himself while remaining at the comprehension level of the standard reference text for classical electrodynamics by Jackson [27]. This allows the central idea of much more sophisticated analyses to be accessible to the general audience, an idea which is not explicitly described in the leading books on this subject [31–35].

4 The unaccelerated electron

For the case of an unaccelerated distribution of charge $a_i = 0$ when the exterior electromagnetic field vanishes and $N_F = 1$, Fermi's condition reduces to equating to zero the total electric force on the electron from its own Coulomb field

$$\int_B \rho E(U)_i \, dV = 0, \tag{20}$$

in which case the volume integral in Gauss's law vanishes and the total 4-momentum $P(\Sigma)^{\alpha}$ in the electromagnetic field is independent of time in the Fermi coordinate system, as expected since the state of the system is static in that inertial reference frame. However, although the 4-momentum of the Coulomb field is time-independent for any inertial observer, different inertial observers in relative motion measure different 4-vectors for this 4-momentum. Because this was misunderstood, and it is natural to want to associate a 4-vector representing the 4-momentum of the Coulomb field of the electron which is aligned with the 4-velocity of its central world line, people looked for solutions.

Poincaré [16–19] introduced stresses needed to balance the electromagnetic stresses in the charge distribution for the case of zero mechanical mass soon after the Abraham–Lorentz model was developed, but unfortunately retained their mistaken nonrelativistic notion of rigidity for the accelerated electron, mixing up the separate issue of the stability of this model with the lack of consistency within special relativity. By adding a nonunique ad hoc stress-energy tensor to cancel out the divergence of the electromagnetic one, he re-established the existence of a conserved 4-momentum at the cost of being inconsistent with special relativity, deriving an inertial mass for the electromagnetic field related to the energy W by the relation $\frac{4}{3}W/c^2$ instead of W/c^2.

Decades later in 1949 Kwal [26] essentially realized that in order to have a unique 4-momentum associated with the Coloumb field of the unaccelerated electron, one simply had to restrict the time hyperplane in the 4-momentum integral to one associated with the electron's inertial rest frame, although he was not sophisticated enough to actually talk about the region of integration and only examined the volume element for the hyperplane integration. In fact one can simply insert a projection along the rest frame 4-velocity into the contracted pair of indices in the 4-momentum integral definition to enforce this result for any time hyperplane. In inertial coordinates the components of this adjusted 4-momentum are

$$P_{\text{Kwal}}(\Sigma)^\alpha = \int_\Sigma T^{\alpha\beta}(-U_\beta U^\delta)\,d\Sigma_\delta, \tag{21}$$

where as above U^α is the 4-velocity of the rest frame of the electron and u^α is the 4-velocity of the inertial observer associated with the time slice Σ. However,

$$-U^\delta\,d\Sigma_\delta = -U^\delta u_\delta\,d\Sigma = \gamma(U,u)\,d\Sigma = d\Sigma_{\text{rest}}, \tag{22}$$

leads to the differential of volume $d\Sigma_{\text{rest}}$ on the tilted hyperplanes associated with a different rest frame time hypersurface at each point of Σ, a differential whose Lorentz contraction $d\Sigma_\delta = \gamma(U,u)^{-1}d\Sigma_{\text{rest}}$ by the relative gamma factor $\gamma(U,u) = -u_\delta U^\delta$ yields the original differential. This corresponds to integrating over the the corresponding region of Σ_{rest} related by moving to it from Σ along the rest frame time lines. Thus we have

$$P_{\text{Kwal}}(\Sigma)^\alpha = \int_\Sigma T^{\alpha\beta}U_\beta\,d\Sigma_{\text{rest}}. \tag{23}$$

However, if we express the components of this equation in rest frame inertial coordinates where the system is static, the components of the 4-vector integrand $T^{\alpha\beta}U_\beta$ are independent of the rest frame time coordinate and so have the same values along each rest frame time line, so the integral is equivalent to integrating over any time hyperplane Σ_{rest} in the rest frame with respect to the actual differential of volume on that rest frame time hyperplane and the result is the unique 4-vector $P(\Sigma_{\text{rest}})^\alpha$. Kwal essentially only describes replacing the factor $-u_\beta\, d\Sigma$ by $-U_\beta\, d\Sigma_{\text{rest}}$ in the integral, without using the time translation invariance in the rest frame to relate the different time hyperplane regions of integration. A decade later in 1960 Rohrlich [25] came to the same conclusion without being aware of the work of Kwal or Fermi or being explicit about the time translation invariance needed to evaluate the rest frame 4-momentum on time slices not associated with the rest frame. Jackson describes in detail Rohrlich's discussion in the Second and Third Editions [27], and provides an alternative explanation for getting the same result on other time slices using the invariants of the electromagnetic field tensor. Unfortunately Rohrlich claims the Abraham–Lorentz definition of the observer-dependent 4-momentum of the electromagnetic field is wrong in the case of the very special case of an unaccelerated electron, which is simply not the case. When the integral is restricted to a bounded region of a constant inertial time hyperplane, this integral is essential in describing the transport of energy and momentum in and out of the region for any configuration of charges, currents and electromagnetic fields. See Sect. 6.7 on Poynting's Theorem in Jackson's Third Edition [27].

Explicit evaluation of the electromagnetic 4-momentum with respect to an inertial observer with 4-velocity u^α on a constant inertial time hyperplane Σ in the shell model of the electron shows that it can be expressed in the form [15]

$$\begin{aligned} P(\Sigma)^\alpha &= P(\Sigma_{\text{rest}})^\alpha + \frac{1}{3} P(\Sigma_{\text{rest}})_\beta U^\beta \nu(u,U)^\alpha \\ &= W\left(U^\alpha - \frac{1}{3}\nu(u,U)^\alpha\right), \end{aligned} \tag{24}$$

where $\nu(u,U)^\alpha$ is the relative velocity of the moving frame compared to the rest frame as seen in the rest frame and $P(\Sigma_{\text{rest}})^\alpha = WU^\alpha$, and W is the rest frame energy of the Coulomb field defined explicitly in the next section. The second term on the right hand side of this equation (orthogonal to the first term) shows the explicit dependence of the 4-momentum on the observer 4-velocity u^α. See Fig. 2.

Schwinger [40] has considered a special 1-parameter family of internal stress-energy tensors compatible with this shell model, resulting in a total stress-energy tensor which is divergence-free and hence the total 4-momentum is a single conserved 4-vector. Among these is the choice corresponding to $h = -1$ in the notation of the Third Edition of Jackson [27] where this tensor is proportional to the orthogonal projection $g_{\alpha\beta} + U_\alpha U_\beta$ into the local rest spaces of the electron sphere and hence does not contribute at all to the rest frame evaluation of the total 4-momentum, which therefore equals the 4-momentum of the electromagnetic field alone, the first term on the right hand side of Eq. (24). In any other inertial frame, the integral of the internal stress-energy tensor inside the electron sphere therefore exactly cancels the extra velocity-dependent term in that equation to yield the same total 4-momentum 4-vector.

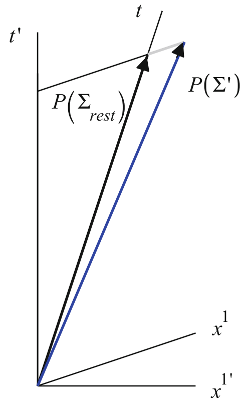


Fig. 2 The relationship between the rest frame 4-momentum $P(\Sigma_{\text{rest}})$ and the 4-momentum $P(\Sigma')$ observed in an inertial frame in relative motion, referred to the corresponding inertial coordinate axes for the case of motion along the x^1 direction

5 Equations of motion for the rigid charge distribution

The actual equations of motion for the central world line of the rigid charge distribution can be evaluated in the quasi-stationary limit of small enough and sufficiently slowly changing acceleration that one can linearize Eq. (17) with respect to the acceleration and ignore its time derivatives (thus neglecting radiation reaction terms) as described in detail in Jackson [27] for the case of zero mechanical mass, without the Fermi lapse factor. First one must separate out the self-field due to the charge distribution from the external field in which the electron is moving, assuming that the latter is essentially constant over the charge distribution so that it may be factored out of the integral. The self-field is defined through the retarded time integrals of the 4-potential over the charge distribution in Lorentz gauge. One then has

$$\int E^{(\text{self})}(U)_i N_F \, de + \int E^{(\text{ext})}(U)_i N_F \, de - m_{(\text{me})} a_i = 0. \tag{25}$$

The lowest order contribution to the self-force in this approximation as shown by Fermi is $-m_{(\text{em})} a_i$, where the inertial mass coefficient $m_{(\text{em})}$ is a constant equal to the total energy W of the Coulomb field of the charge distribution, defined by

$$W = \frac{1}{2} \int \int d^3\mathbf{x} d^3\mathbf{x}' \, \frac{\rho(\mathbf{x})\rho(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} \tag{26}$$

in the notation of Jackson [27], expressed as seen by an inertial observer at a time when the electron is momentarily at rest. The Fermi coordinate lapse factor in the integrand of the self-field integral in (25) corrects the Abraham-Lorentz result $m_{(\text{em})} = \frac{4}{3} W$ to conform with the Einstein mass-energy relation $E = mc^2$ (with $c = 1$), as we show next.

Then since the first term on the left hand side and the right hand side of Eq. (25) are proportional to the acceleration, the second term must be first order in the acceleration, so keeping only first order terms, one can ignore the Fermi lapse factor in the second term which becomes

$$\int E^{(\text{ext})}(U)_i N_F \, de = E^{(\text{ext})}(U)_i \int de = e E^{(\text{ext})}, \tag{27}$$

leading to the Lorentz force law in the Fermi frame for which the spatial velocity is zero

$$(m_{(\text{me})} + m_{(\text{em})}) a_i = e E_i^{(\text{ext})}. \tag{28}$$

In other words the mass formula for the electromagnetic contribution to the inertial mass is

$$m = m_{(\text{me})} + m_{(\text{em})}. \tag{29}$$

For the spherical shell model, one easily finds $m_{(\text{em})} = e^2/(2r_0)$, which compares very nicely with the Reissner–Nordstrom irreducible mass formula [36]

$$m = m_{(\text{irred})} + e^2/(2r_+) \tag{30}$$

for the gravitational mass m of a static spherically symmetric charge distribution of total charge e and outer horizon radius r_+ within general relativity.

6 Conclusions

The classical theory of the electron and related issues has attracted the attention of many of the great physicists of the past century, and has been the subject of many articles and a few books that continue to appear, most of which seem not to reflect Fermi's simple argument, although a relatively recent article by Kolbenstved [37] offers an alternative explanation of that argument. For a complete list of such references see [15], as well as the recent analysis by Boughn and Rothman [38] of a related problem considered by Fermi in his fifth paper [39]. Ultimately the problematic issues of a finite-sized classical electron were sidestepped by the point particle model and renormalization techniques introduced in the quantum theory. However, as recently as the past decade, new results in the classical theory have appeared [29–31,33–35], but which still leave this loose end of Fermi's work unaddressed.

Gauss's law for stress-energy tensors with nonzero divergence is straightforward to consider yet, until now no one has connected up Fermi's results with the question of the 4-momentum in the electromagnetic field of the classical electron model, an issue which arose after he lost interest in the problem. Doing so has provided a useful pedagogical example omitted in all textbooks on general relativity or electrodynamics and has led to the following satisfying result. While in the case of the unaccelerated electron, there is no selection mechanism to pick out the obvious candidate for the 4-momentum aligned with the 4-velocity of the rigid electron other than the alignment itself, for the accelerated electron it is only the instantaneous rest frame observer which leads not only to aligning the 4-momentum of the electromagnetic self-field with the 4-velocity, but also to a 4-momentum conservation law for the total 4-momentum.

Acknowledgments Andrea Geralico and Donato Bini are thanked for collaborating in laying down the foundation from which our new results were obtained.

References

1. Fermi, E.: Nuovo Cim. **22**, 199 (1921)
2. Fermi, E.: Nuovo Cim. **22**, 176 (1921)
3. Fermi, E.: Rend. Lincei, **31**, 21–23, 51–52, 101–103 (1922)
4. Fermi, E.: Phys. Z. **23**, 340 (1922)
5. Fermi, E.: Rendi. Lincei **31**, 184–187, 306–309 (1922)
6. Fermi, E.: Nuovo Cim. **25**, 159 (1923)
7. Fermi, E.: Note e Memorie (Collected Papers), Accademia Nazionale dei Lincei and The University of Chicago Press, (Vol. 1, 1961, Vol. 2, 1965) [original Italian articles available at English translations of early articles in: Fermi and Astrophysics, Ruffini, R., Boccaletti, D (Eds.), World Scientific, Singapore (2012)] http://www.archive.org/details/collectedpapersn007155mbp
8. Abraham, M.: Ann. Phys. (Leipzig) **10**, 105 (1903)
9. Abraham, M.: Phys. Z. **5**, 576 (1904)
10. Lorentz, H.A.: The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat, Dover, New York. (First Edition 1909, from lectures of 1906) (1952)
11. Bini, D., Jantzen, R.T.: In: Gurzadyan, V., Ruffini, R. (Eds.) Proceedings of the Ninth ICRA Network Workshop on Fermi and Astrophysics, World Scientific, Singapore, (2003) [Nuovo Cim. vol. 117B, p. 983 (2002)] http://arXiv.org/abs/gr-qc/0202085
12. Bini, D., Jantzen, R.T.: In: Ruffini, R., Boccaletti, D. (Eds.) Fermi and Astrophysics, World Scientific, Singapore (2012)
13. Synge, J.L.: Relativity, The General Theory. North Holland, Amsterdam (1960)
14. Ruffini, R., Boccaletti, D.: Fermi and Astrophysics. World Scientific, Singapore (2011)
15. Bini, D., Geralico, A., Jantzen, R.T., Ruffini, R.: On Fermi's resolution of the 4/3 problem in the classical theory of the electron and its logical conclusion: hidden in plain sight. In: Fermi and Astrophysics, World Scientific, Singapore (2012)
16. Poincaré, H.: C. R. Acad. Sci. **140**, 1504 (1905)
17. Poincaré, H.: Rend. Circ. Mat. Palermo **21**, 129. [English translation by Schwartz, H.M.: Poincaré's Rendiconti Paper on Relativity. Part I, II, III. Am. J. Phys. **39**, 1287 (1971)] (1906)
18. Poincaré, H.: Am. J. Phys. **40**, 862 (1972)
19. Poincaré, H.: Am. J. Phys. **40**, 1282 (1972)
20. von Laue, M.: Das Relativitätsprinzip. Vieweg, Braunschweig (1911)
21. Born, M.: Ann. Physik **30**, 1 (1909)
22. Born, M.: Phys. Z. **11**, 233 (1910)
23. Misner, C.W., Wheeler, J.A., Thorne, K.S.: Gravitation. Freeman, San Francisco (1973)
24. Aharoni, J.: The Special Theory of Relativity, Oxford University Press, Oxford [First Edition: 1959, Second Edition: 1965, see Chapter 5]
25. Rohrlich, F.: Am. J. Phys. **28**, 639 (1960)
26. Kwal, B.: J. Phys. Radium **10**, 103 (1949)
27. Jackson, J.: Classical Electrodynamics, Wiley, New York, [Second Edition: 1975, Third Edition: 1999; see respectively Chapters 17, 16]
28. Nodvik, J.S.: Ann. Phys. **28**, 225 (1964)
29. Kiessling, M.: Phys. Lett. A **258**, 197 (1999)
30. Appel, W., Kiessling, M.: Ann. Phys. **289**, 24 (2001)
31. Spohn, H.: Dynamics of Charged Particles and Their Radiation Field. Cambridge University Press, Cambridge (2004)
32. Rohrlich, F.: Classical Charged Particles, Addison-Wesley, Reading (1965)
33. Rohrlich, F.: Classical Charged Particles, (updated 3rd Edn) World Scientific, Singapore (2007)
34. Yaghjian, A.D.: Relativistic Dynamics of a Charged Sphere: Updating the Lorentz-Abraham Model, Springer, Berlin (Second Edition 2006) (1992)
35. Yaghjian, A.D.: Phys. Rev. **E78**, 046606 (2008)
36. Christodoulou, D., Ruffini, R.: Phys. Rev. D **4**, 3552 (1971)
37. Kolbenstvedt, H.: Phys. Lett. A **234**, 319 (1997)

38. Boughn, S., Rothman, T.: Hasenöhrl and the Equivalence of Mass and Energy (2011) e-print: http://arxiv.org/abs/1108.2250
39. Fermi, E., Pontremoli, A.: Rend. Lincei **32**, 162 (1923) [English translation in [14]]
40. Schwinger, J.: Found. Phys. **13**, 373 (1983)

A.3 D. Boccaletti: When a problem is solved too early. Enrico Fermi and the infamous 4/3 problem

Introduction

It has often happened, particularly in the past few centuries, that some scientific results had been reobtained more than once, each time ignoring the authors of the preceding discoveries. In the case of mechanics this happened many times, as recalled by A. Wintner in the preface to his famous book:[1] "... even the classical literature of the great century of celestial mechanics appears to be saturated with rediscoveries (sometimes *bona fide* and sometimes not assuredly so) ...". In times closer to us, this has happened again for "the infamous 4/3 problem."[2] It took thirty years for the result obtained by Fermi to have its "consecration" in an authoritative book (see below) and ten more to begin circulating among the community of experts. In the next pages we shall first try to historically contextualize Fermi's paper in an extremely concise way and then to bring into question the procedures through which the paper itself has been interpreted. At the end we shall advance a conjecture which, as with all conjectures, is based on circumstantial but not incontrovertible evidence.

The story in short

In the early twenties, when Fermi was concluding his studies at the University of Pisa, in Italy the problems related to the rising quantum mechanics had not yet filtered into academic circles. Instead the electromagnetic theory and the theory of relativity (special and general) were well-known and studied (even if in restricted circles) through the works of Abraham, Lorentz, Poincaré, Richardson . . . for electromagnetic theory and the papers of Einstein, Levi-Civita, and the book of Weyl for the theory of relativity. Fermi, still a student, had a deep knowledge of these theories and of classical analytical mechanics. Besides being testified to in Fermi's biography written by Emilio Segrè,[3] this appears clearly in the first papers he published.[4] Paper 1) is substantially a generalization of a result which, at that time,

[1] A. Wintner: *The Analytical Foundations of Celestial Mechanics*, Princeton University Press, 1947, p. IX.

[2] The expression is due to J. D. Jackson in his textbook *Classical Electrodynamics*, Third Edition, Wiley 1998, p. 755.

[3] E. Segrè: *Enrico Fermi Physicist*, The University of Chicago Press, 1970

[4] For the first few Fermi's papers also see, in Italian,

C. Tarsitani: *I lavori di Fermi sulla relatività nei commenti di Persico e Segrè*, Atti del IV congresso nazionale di storia della fisica, Como, 1983,

F. Cordella, F. Sebastiani: *Il debutto di Enrico Fermi come fisico teorico: I primi lavori sulla relatività (1921–1922–23)*, Quaderno di Storia della Fisica N. 5, 1999 and

F. Cordella, A. De Gregorio, F. Sebastiani: *Enrico Fermi: Gli anni italiani*, Editori Riuniti, 2001. To avoid possible misunderstandings, we follow the convention of the present volume and refer to Fermi's papers making use of the numbered classification scheme given in *Enrico Fermi: Note e Memorie (Collected Papers)*, Accademia Nazionale dei Lincei and University of Chicago Press, was quoted in the circulating textbooks on electrodynamics.

Besides the various editions of the Abraham's *Theorie der Elektrizität* (which originated as a second part of the treatise of Föppl published for the first time in 1894) and Lorentz's *The theory of electrons* (1909, second edition 1915), the textbook having a larger circulation was Richardson's *The electron theory of matter* (1914). Fermi refers to this latter textbook. At that time (1921–23) it was generally accepted that, for a charged particle moving with variable velocity, the electromagnetic mass was 4/3 times the inertial mass.[5] The whole theoretical work done in the last two decades, mainly by Abraham and Lorentz, had led to considering the electron (discovered by J. J. Thomson in 1897) to be a rigid sphere with a uniform charge distribution on its surface. In particular, Abraham was convinced that the electron's entire mass was of electromagnetic origin and in 1902 announced the realization of an "electromagnetic mechanics." He also called "longitudinal mass" the mass associated only with a force oriented along the electron's trajectory and called "transverse mass" that associated with a force oriented perpendicular to the electron's trajectory.[6] (These terms had a long life since were used in various papers on the special theory of relativity, including the fundamental Einstein paper of 1905). Since $E_0^e = \frac{e^2}{2R}$ (R radius of the sphere) is the electrostatic energy, the current theory drove to evaluate the electromagnetic contribution to the electron's mass as $m_e = \frac{2}{3}\frac{e^2}{c^2 R}$. As a consequence this made the electromagnetic mass equal to 4/3 times the mass entering into Einstein's equation $E = mc^2$.

Fermi demonstrates that, in the context of the then current theory, one obtains the same result for any system of moving charges, i.e., the factor 4/3. Therefore inertial mass and electromagnetic mass do not match. He also announces that in a forthcoming paper he will consider electromagnetic masses as masses endowed with weight from the point of view of the general theory of relativity. In point of fact, in paper 2), Fermi obtains the result that the electromagnetic mass and the passive gravitational mass (the weight of the charged particle) do match. This is a blatant contradiction: either this result disproves the equivalence principle (largely accepted by that time) or a new problem arises on the possible electromagnetic nature of mass (remember Abraham's ideas on "electromagnetic mechanics"!). In paper 4c), Fermi first solves the problem. He is well aware of the importance of the result obtained. In fact he writes and publishes three equivalent versions of his work (in *Il nuovo Cimento*, in the *Rendiconti dell'Accademia Nazionale dei Lincei*

Vol. 1, 1961, Vol. 2, 1965. The relevant papers 1), 2), 3), 4c) of which we will be concerned in the following are given in English translation in Chapter 2 of this volume.

[5] See, for instance O. W. Richardson: *The electron theory of matter*, Cambridge University Press, 1914, Chapters XI, XII.

[6] For a historical analysis of the problem of the electromagnetic mass see
A. I. Miller: *Albert Einstein's Special Theory of Relativity, Emergence (1905) and Early Interpretation (1905–1911)*, Springer, 1998 and
E. T. Whittaker: *A History of the Theories of the Aether and Electricity*, Thomas Nelson & Sons, London 1951 and 1953.

and in *Physikalische Zeitschrift*).[7] He also confides to his friend Enrico Persico that there will be some troubles to obtain an agreement with his ideas: "*... I am trying with great effort to launch the business of the 4/3. The main difficulty derives from the fact that they have a hard time understanding—in part because the thing is not easy to understand, in part because I express myself too concisely—but little by little they begin to understand what it is all about ...*".[8]

But, as the saying goes, no man is a prophet in his own country and the three versions (even the German one) went unnoticed. Thus, as Rohrlich said,[9] the result was bound to be rediscovered. It did not find its way into the standard references or textbooks until 1953 when E. T. Whittaker, in the second volume of his *History* (on p. 51, see footnote 6) quoted Fermi's Lincei communications saying "It was shown long afterwards by E. Fermi that the transport of the stress system set up in the material of the sphere should be taken into account, and that when this is done, Thomson's result becomes

$$Additional\ mass = \frac{1}{c^2}\ Energy\ of\ the\ field\text{ ''}$$

In the meantime two papers had appeared. W. Wilson obtained the same result of Fermi in a different way[10] and analogously B. Kwal 13 years later in a short note arrived at the same conclusions exploiting the relativistic transformation of the electromagnetic energy-momentum tensor.[11] Finally, the result was discovered for a fourth time by F. Rohrlich,[12] again (apparently) without the knowledge of any of the previous papers. Fundamentally, Fermi showed that factor 4/3 was produced by an incorrect application of (or more precisely by failing to apply) the theory of relativity. The circumstance which, at first sight, might appear rather strange is that Fermi, in his teaching activity of those years continued teaching the old result. Only in his textbook *Introduzione alla Fisica Atomica*[13] he introduced a short sentence mentioning relativistic corrections (without demonstration). In this connection, W. Joffrain[14] put forward the hypothesis of a sort of deontological scruple: not to teach, in an institutional course, results which are not yet universally accepted. Subsequently, in collaboration with A. Pontremoli,[15] Fermi applied successfully

[7] Besides the *Nuovo Cimento* version 4c), Fermi published the two Lincei communications XXI, 1922, pp. 184–187 and 306–309 (4a) and the paper *Über eine Widerspruch zwischen der elektrodynamischen und relativistischen theorie der elektromagnetischen Masse* in Physikalische Zeitschrift XXIII, 340-344, 1922 (4b).

[8] E. Segrè, op. cit., p. 197.

[9] F. Rohrlich: *Charged Classical Particles*, Addison-Wesley, 1965, p. 17.

[10] W. Wilson: The mass of a convected field and Einstein's mass-energy law, Proc. Phys. Soc. (London) **48**, 736–740 (1936). This paper is also mentioned in Whittaker's book.

[11] B. Kwal: Les expressions de l'énergie et de l'impulsion du champ électromagntique propre de l'électron en mouvement, J. Phys. Radium **10**, 103–104 (1949).

[12] F. Rohrlich: Self-energy and stability of the classical electron Am. J. Phys. **28**, 639–643 (1960).

[13] E. Fermi: *Introduzione alla Fisica Atomica*, Zanichelli, 1928, p. 66.

[14] W. Joffrain: *Un inedito di Enrico Fermi — Elettrodinamica*, Atti del XVIII Congresso di Storia della Fisica e dell'Astronomia, Como (Italy), May 15–16, 1998.

[15] E. Fermi, A. Pontremoli: *Sulla massa della radiazione in uno spazio vuoto*, Rend. Lincei, **32** (1), 162–164 (1923).

the same method to the calculation of the mass of the radiation contained in a cavity with reflecting walls, for which the standard textbooks of the time had an expression containing the same factor 4/3. Anyway, the problem of the nature of the electromagnetic mass was been dragging on for various decades through the contributions, after that of Fermi, of Rohlrich, Dirac, etc.. However, almost always, the successive results—at least apparently—went unnoticed.

The resistible path of Fermi's paper

At this point, in retrospect, if we look at the whole story some circumstances appear at the very least to be strange. Let us start from the beginning of the sequence. Fermi obtained the result published in 4c) in January 1922.[16] It is clear that he feels proud of the conclusions obtained. This turns out clearly in the letter to Persico in which he already announces his intent of publishing the paper also on a German review (which will result to be *Physikalische Zeitschrift*), to make it known outside of Italy.

At this point we can notice that, completely immersed in the academic context of those times, Fermi thought that the paper concerning the factor 4/3 was much more important, since it was solving a problem already several decades old, than paper 3), only published in Italian in *Rendiconti dell'Accademia Nazionale dei Lincei* presented by G. Armellini in January 1922. As we know, paper 3), after the generalization due to Walker (1932), spread far and wide and still is considered of lasting importance. The German version of 4c), i.e., 4b), sent to *Physikalische Zeitschrift*, was received by the journal the ninth of May 1922. The paper was immediately published and also reviewed in *Physikalische Berichte* by Erich Kretschmann in the issue of December 15.[17] We point out that Erich Kretschmann, who was a habitual reviewer of the journal for at least three sections regarding the foundations of physics (in German: *Allgemeines, Allgemeine Grundlagen der Physik, Mechanik*, respectively), was not an obscure physicist, but a quite well known expert in the theory of relativity. In fact a paper published by him in 1917 on the physical meaning of the postulates of the theory of relativity[18] had caused a lot of talk and even aroused a reply by Einstein himself.

Then Fermi's paper, which clarified how one can correctly apply the principles of the (special) theory of relativity to solve the problem of the factor 4/3, outwardly

[16] This date can be fixed at a sufficiently good approximation by comparing the Fermi's letter to Enrico Persico (see note 8), which is of January twenty-five, 1922, with what Persico writes in *Note e Memorie* Vol. 1, p. 24, introducing the paper. Persico also reports a discussion of Fermi with Luigi Puccianti and Giovanni Polvani regarding the factor 4/3 which seems to coincide with what Fermi writes in the letter (where, however, Fermi does not name the names). The strange thing is that Persico does not quote the letter here. Moreover Fermi dates "January 1922" the German version of the paper.

[17] *Physikalische Berichte, Dritter Jahrgang* 1922, N. 24, p. 1293.

[18] E. Kretschmann: Über den physikalischen Sinn der Relativitätstheorie, Annalen der Physik **53**, 575–614 (1917).

appears to have fallen into the hands of the right person. Unfortunately, this was not the case. In fact instead of investigating the method used by Fermi to applying correctly the relativistic concepts, Kretschmann limited himself to repeating Fermi's words describing the two possible ways of performing the variation for applying Hamilton principle and then to conclude that the solution of the problem of the factor 4/3 given by von Laue in his book was "*much more transparent*".[19] We know from Fermi's biography written by Segrè and from the reminiscences published by Persico that Fermi had studied Weyl's textbook[20] thoroughly, which moreover is quoted in the paper itself when Fermi follows Weyl in applying Hamilton's principle. It is enough to make a comparison of Fermi's paper with the page where Weyl says "*This theory does not, of course, explain the existence of the electron, since cohesive forces are lacking in it*"[21] for understanding that Fermi, following Weyl, only means to deal with a charged sphere (with a surface distribution of charge) without tackling the problem of its internal structure and stability. Then the comparison that Kretschmann makes with von Laue's solution, which involves the introduction of the so called "Poincaré stresses" which turn out to be necessary for ensuring the stability of the electron, is completely misleading. Fermi, as those who will find the solution of the factor 4/3 after him, considers this problem as having nothing to do with the problem of stability. It is curious that even Enrico Persico, who in January 1922 received the letter in which Fermi mentioned the subject, in 1961 writes "*It is now well known that the factor 4/3 can be interpreted as due to the part of the energetic tensor contributed by the internal non-electromagnetic stresses, whose existence must be assumed to assure the equilibrium of the charges. However, in the books known to Fermi, this discrepancy was not explained (he had evidently overlooked the explanation contained in M. von Laue, Die Relativittstheorie, 1, third edition, 1929, p. 218) and so he found for it an explanation of his own, essentially equivalent to the former but obtained through Weyl's variational method*".[22] At that date Rohrlich's paper had already been published, but perhaps Persico had not had enough time to see it. However, a good eight years before (1953) the second volume of Whittaker's book[23] had been published in which Fermi's paper (the Lincei version) was mentioned with the explanation reported above. We point out that Whittaker's book did not go unnoticed, both for the reputation of the author and for the *vexata quaestio* of the authorship of the special theory of relativity. As is known, Whittaker ascribed to Poincaré the authorship of the special theory of relativity and was also charged with ahistoricisms concerning the theory of relativ-

[19] See 17 Kretschmann quotes the 1919 third edition of von Laue's book, but the author continued to maintain the same conclusions in the subsequent fourth edition (see Die Relativitätstheorie on Dr. M. von Laue, Braunschweig, 1929, pp. 224–227 and also its French translation).

[20] Fermi always quoted the fourth 1921 edition of H. Weyl: Raum. Zeit. Materie, Springer, Berlin.

[21] This excerpt is from the English translation of the 1921 German edition republished by Dover in 1952 with the title "*Space-Time-Matter*".

[22] See *Note e Memorie* Vol. 1, p. 24. This strange and uncorrected (for what regards "it is now well known...") sentence has been also remarked by Tarsitani, loc. cit. in 4.

[23] See 6.

ity.[24] Then it is strange that even Rohrlich did not know about the quotation of Fermi by Whittaker, particularly if we bear in mind that the subject of Whittaker's book was the origin and the development of the e. m. theory (Abraham, Poincaré, Lorentz, ...). In the 1960 paper (see 12), which is the first Rohrlich dedicated to the problem of the electromagnetic mass of the electron and related questions, Fermi's paper is not mentioned and the same for Wilson's and Kwal's papers.

Apparently Rohrlich solved independently the "4/3 problem" without knowing the contributions of his predecessors. Two years later, in a lecture given before the Joseph Henry Society[25] he said "*. . . For a finite electron this was first pointed out by Fermi in 1922. It is closely related to the definition of rigidity in special relativity where the difference in the simultaneity of relatively moving observers plays an essential role. Unfortunately, Fermi's paper was either never understood or soon forgotten*". Rohrlich, at this point, quotes as a reference the German version (see 7) of Fermi's paper but there is no mention of the papers of Wilson and Kwal. In his 1965 book (see 9), he mentions all three authors (Fermi, Wilson, Kwal) who had preceded him. As a matter of fact this is the last time Rohrlich mentions Fermi's contribution. On this subject he has published papers for about forty years but, as one can check considering the most important journals, Fermi's name is no longer mentioned. The odd thing is that, even in the last paper known to us[26] which contains an appendix with the title "*The history and eventual solution of the stability problem (the 4/3 problem)*", Fermi's name does not appear. A prospective reader could only find the reference to Fermi in the bibliographies of the books and the papers quoted.

In the same year (1965) in which Rohrlich's book appeared, the second revised edition of *The Special Theory of Relativity* by J. Aharoni also came out.[27] In the preface the author says that Rohrlich's 1960 paper "*...initiated new interest in the problem and it turned out that actually a similar solution had already been proposed by B. Kwal in 1949 and the same result obtained as for back as 1922 by E. Fermi who used a different method. It can now be stated that the abolition of the 4/3 factor is also implicit in Dirac's paper on the classical theory of the electron (1938). It is difficult to explain why all the earlier papers passed unnoticed. Possibly this was due to Poincaré's idea to link the 4/3 factor with the instability of an electric charge on purely electrostatic forces*". It should be noted that Aharoni cannot have had

[24] See G. Holton: On the Origins of the Special Theory of Relativity (1960) in
G. Holton: *Thematic Origins of Scientific Thought. Kepler to Einstein*, Harvard University Press, 1977 and
A.J. Miller: A study of Henri Poincaré's "Sur la Dynamique de l'Electron," Arch. Hist. Exact. Scis. **10**, 207–328 (1973).

[25] The theory of the electron, Thirty-first Joseph Henry Lecture (read before the Society May 11, 1962).

[26] F. Rohrlich: The dynamics of a charged sphere and the electron, Am. J. Phys. **65**, 1051-1056, (1997).

[27] J. Aharoni: The special Theory of Relativity, Second revised edition, Oxford University Press, 1965 (reprinted by Dover, 1985).

knowledge from Rohrlich's paper, which he quotes, of the name of Fermi, Wilson and Kwal since in that paper they are not mentioned. Before the two 1965 books, Kwal's name does not appear, while Wilson's name only appears in Whittaker's book together with Fermi's. Evidently, there is a missing link in the chain!

Let us turn again to Aharoni's book which, from our point of view, assumes a particular importance. In fact, Aharoni is the only one, among all who spoke about Fermi's paper, who devoted himself to effectively understand the method Fermi used for applying his relativistic concept of rigidity. He spends about two pages (170–171) to explain and explicitly reconstruct Fermi's calculations omitted by the author who sums up "*...we have manifestly...*". Therefore, on the part of Aharoni there is a true appreciation for the work done by the man who first solved the problem. Retrospectively, it comes to mind that to the early readers those calculations might not be so transparent (see the above letter to Persico where Fermi admits to expressing himself "*too concisely*") and also that the course of differential geometry given by Luigi Bianchi about which Fermi speaks in a letter to Persico, came in very useful to him so to consider obvious the calculations and then to omit them.[28] All the known biographies of Fermi report that he went to Göttingen with a fellowship from the Italian Ministry of Public Instruction in the winter 1922–23 to study with the group headed by Max Born and he remained there seven months. "...when Fermi arrived at Göttingen, he found several brilliant contemporaries there, among them Werner Heisenberg and Pascual Jordan, two of the brightest luminaries of theoretical physics. Indeed the two had already been recognized for their exceptional abilities, and Born was writing papers in collaboration with them at about the time of Fermi's residence in Göttingen. Unfortunately it seems that Fermi did not become a member of that extraordinary group or interact with them. I do not know the reason for this ...".[29] "...Born himself was kind and hospitable. But he did not guess that the young man from Rome, for all his apparent self-reliance, was at the very moment going through that stage of life which most young people cannot avoid. Fermi was groping with uncertainty and seeking reassurance. He was hoping for a pat on the back from Professor Max Born ...".[30] Both the biographers (Emilio Segrè and Fermi's wife) agree in maintaining that Fermi came back to Rome not satisfied with the German experience, somehow disappointed. It is known from other sources that Born and his collaborators thought the best of Fermi and this is born out from the fact that subsequently he was on friendly terms with them. Moreover, the biographers confirm that Fermi's German was certainly good enough to allow easy communication and then not to be excluded. Why then, as far as we know, he did not join the Born's group and went back to Rome disappointed?

[28] In the letter Fermi writes "I will pass the examination in higher analysis (differential geometry) which is a terrific bore, in which the problem studied are chosen by the sole criterion that they should lack all interest," see Segrè, op. cit. p. 201–202.

[29] See Segrè, op. cit. p. 33.

[30] Laura Fermi: *Atoms in the Family. My life with Enrico Fermi*, The University of Chicago Press, 1954, p. 31

Both Emilio Segrè and Laura Fermi put forward the hypothesis that the ideas of Born's group at that time appeared to be very concrete, even philosophical, and then not able to catch the interest of Fermi or Fermi himself was not mature enough to get himself to be appreciated in that environment.

Our conjecture is that Fermi had effectively taken his paper to Göttingen to be appreciated, but he did not achieve his aim. When Fermi arrived in Göttingen, the paper on the "4/3 problem" had already been published in German and so readable by Born and the others. It is unthinkable that Fermi, who was so proud of his result, had not exhibited it and asked Born for his opinion. We recall that, what's more, in his paper Fermi quotes a relativistic definition of rigidity due to Born (in a paper of 1909).[31] The most obvious thing to do for a brilliant young physicist, as Fermi was, would have been to display the paper he was proud of to the authoritative professor. To the best of our knowledge, no proof exists even if it is reasonable to suppose that this had happened. The only thing we can say for certain is that Born's book on Relativity theory,[32] which in its second edition of 1921 held the "traditional" point of view of the "4/3 problem", continued to give the same version till to the last edition.[33] The same thing happened for Pauli's famous lectures[34] as if Fermi's paper had never existed. Born and Pauli were not alone in ignoring Fermi's paper and related conclusions; to the list we can add even Feynman.[35] Coming back to Born, from his authobiography[36] it turns out that over the years he continued to think about the problem of the electromagnetic mass of the electron, but there is no connection with Fermi's conclusions which are never mentioned. Our conjecture, for all its worth, is that the disappointment for having not received appreciation embittered Fermi and also deterred him from the subject. Moreover, the problems raised by the new quantum mechanics and statistical theories definitively averted his interest from classical electrodynamics.

[31] Max Born: *Die Theorie des starren Elektrons in der Kinematik des Relativitätsprinzips*, Annalen der Physik IV, **11**, 1–56 (1909).

[32] Max Born: *Die Relativitätstheorie Einsteins und ihre physikalischen Grundlagen*, Springer, Berlin, 1921, p. 157.

[33] Max Born: *Einstein's Theory of Relativity*, revised edition prepared with the collaboration of Günther Leibfried and Walter Biem, Dover, 1962, pp. 207–214 and 278–289.

[34] W. Pauli: *Pauli Lectures on Physics, Vol. 1. Electrodynamics*, MIT Press, 1972 (reprinted by Dover, 2000), p. 151

[35] *The Feynman Lectures on Physics. The Electromagnetic Field*, Addison-Wesley, 1964, Sect. 28–3 and ff.

[36] Max Born: *My Life. Recollection of a Nobel Laureate*, Taylor & Francis Ltd, 1978, Part 2, IV, pp. 254–255

A.4 B. Kwal: Les expressions de l'énergie et de l'impulsion du champ électromagnétique propre de l'électron en mouvement

B. Kwal: "Les expressions de l'énergie et de l'impulsion du champ électromagnétique propre de l'électron en mouvement," (Expressions for the energy and momentum of the electromagnetic self-field of a moving electron) *J. Phys. Radium* **10**, 103 (1949).

LE JOURNAL DE PHYSIQUE ET LE RADIUM. SÉRIE VIII, TOME X, MARS 1949.

LES EXPRESSIONS DE L'ÉNERGIE ET DE L'IMPULSION DU CHAMP ÉLECTROMAGNÉTIQUE PROPRE DE L'ÉLECTRON EN MOUVEMENT

Par BERNARD KWAL.
Institut Henri Poincaré, Paris.

Sommaire. — L'apparition du facteur 1/3 dans l'expression de l'énergie totale de l'électron en mouvement, résulte de l'emploi simultané dans les calculs d'une grandeur tensorielle (tenseur d'énergie et d'impulsion) et d'une grandeur qui ne l'est pas (élément de volume). La difficulté s'évanouit avec une définition tensorielle de l'élément de volume.

En 1904 Max Abraham (1), qui prônait comme on le sait, l'hypothèse de l'origine purement électromagnétique de la masse de l'électron et a bâti, à cet effet, la théorie de l'électron rigide, remarque que dans la théorie de l'électron de Lorentz l'énergie, totale U du champ de l'électron en mouvement contient un facteur supplémentaire qui prouverait que la masse de l'électron de Lorentz ne peut être considérée comme étant toute d'origine purement électromagnétique.

La démonstration relativiste de ce théorème est souvent reproduite dans les manuels, mais elle ne nous semble pas être tout à fait correcte, car elle se base sur l'évaluation d'une intégrale dans laquelle à côté d'une grandeur qu'on traite tensoriellement, à savoir l'énergie du champ électromagnétique, figure l'élément de volume qui est traité d'une manière différente. On définit en effet l'élément de volume en mouvement à l'aide de l'expression

$$dV = dV^0\sqrt{1-\beta^2}, \qquad (1)$$

d V° étant l'élément de volume au repos (par rapport à l'électron). Cette définition n'est pas tensorielle elle résulte de la manière classique de mesurer les longueurs en mouvement qui subissent la contraction de Lorentz. [Comme corollaire de cette définition de l'élément de volume, nous avons comme on le sait la définition non tensorielle de la force, mesurée dans le système en mouvement

$$F_x = F_x^0, \qquad F_y = F_y^0\sqrt{1-\beta^2}, \qquad F_z = F_z^0\sqrt{1-\beta^2} \quad (2)$$

et d'une manière générale de toutes les grandeurs physiques qui entrent en jeu. Car leur définition doit être adaptée à la définition non tensorielle du volume (1), qui résulte d'une certaine manière d'effectuer les mesures dans le système en mouvement, manière qui ne cadre pas avec la définition des tenseurs]. Examinons maintenant la démonstration ici en cause (2). On prend pour l'expression de l'énergie et de l'impulsion du champ électromagnétique propre de l'électron en mouvement (dans la direction de l'axe OX) les formules suivantes :

$$U = \int T_{44}\,dV, \qquad G_x = \frac{1}{c}\int T_{4x}\,dV, \qquad (2^{bis})$$

T_{ik} étant le tenseur de l'énergie et de l'impulsion

[1] M. ABRAHAM, *Physik. Z.*, 1904, **5**, p. 576.

[2] Cf. R. BECKER, *Théorie des électrons*, § 66. — W. PAULI, *Relativitätstheorie*, 1921, pp. 681 et 751. — M. LAUE, *La théorie de la Relativité*, vol. I, pp. 143-145.

Article published online by EDP Sciences and available at http://dx.doi.org/10.1051/jphysrad:01949001003010301

du champ électromagnétique. On passe alors au référentiel par rapport auquel l'électron est au repos, en faisant subir à l'élément de volume la transformation (1) et aux composantes T_{44} et T_{4x} les transformations des composantes du tenseur T_{ik}

$$\left.\begin{aligned}
T_{11} &= \frac{T^0_{11} + 2\beta T^0_{14} + \beta^2 T^0_{44}}{1-\beta^2},\\
T_{14} &= \frac{T^0_{14} + \beta(T^0_{44} + T^0_{11}) + \beta^2 T^0_{14}}{1-\beta^2},\\
T_{44} &= \frac{T^0_{44} + 2\beta T^0_{14} + \beta^2 T^0_{11}}{1-\beta^2}.
\end{aligned}\right\} \quad (3)$$

Comme dans le référentiel au repos $T^0_{14} = 0$, on obtient, en posant

$$U_0 = \int T^0_{44}\,\mathrm{d}V^0, \quad (4)$$

$$U = \frac{1}{\sqrt{1-\beta^2}}\left(U_0 + \beta^2 \int T^0_{11}\,\mathrm{d}V^0\right),$$

$$G_x = \frac{v_x}{c^2\sqrt{1-\beta^2}}\left(U_0 + \int T^0_{11}\,\mathrm{d}V^0\right).$$

Or

$$T^0_{11} = E^{0\,2}_x, \qquad \text{d'où} \quad \int T^0_{11}\,\mathrm{d}V^0 = \frac{1}{3}U_0.$$

On aboutit ainsi aux formules

$$U = \frac{U_0}{\sqrt{1-\beta^2}}\left(1 + \frac{\beta^2}{3}\right), \qquad \vec{G} = \frac{4}{3}\,\frac{\vec{v}}{c^2}\,\frac{U_0}{\sqrt{1-\beta^2}}. \quad (5)$$

C'est précisément l'existence du facteur 1/3 qui est interprétée comme preuve que la masse de l'électron ne peut pas être considérée comme étant d'origine purement électromagnétique.

Comme nous venons de le dire, nous reprochons à la démonstration ci-dessus le tort de mélanger sans scrupules une grandeur tensorielle avec une grandeur à qui l'on n'attribue pas ce caractère. Pour que des formules (2) on puisse tirer des conclusions correctes, il faut que les grandeurs T_{44} et T_{4x} ne soient pas considérées comme se transformant comme des composantes d'un tenseur, mais qu'elles soient pourvues d'une *variance adaptée* à celle de l'élément de volume dV, comme on le fait, par exemple, lorsqu'on définit d'une manière non covariante, la force ou les différentes grandeurs physiques (température, chaleur) qui interviennent en thermodynamique relativiste.

Nous pouvons néanmoins raisonner sur le tenseur d'énergie-impulsion T_{ik}, à condition bien entendu, d'utiliser une définition covariante de l'élément de volume. Pour le faire, nous allons partir de l'élément de volume au repos dV^0 et nous allons considérer dans le référentiel ou mouvement un pseudo-quadrivecteur dV_i, défini comme suit :

$$\left.\begin{aligned}
\mathrm{d}V_{1,2,3} &= -\,\mathrm{d}V^{1,2,3} = \frac{\vec{v}\,\mathrm{d}V_0}{c\sqrt{1-\beta^2}},\\
\mathrm{d}V_4 &= \mathrm{d}V^4 = \frac{\mathrm{d}V_0}{\sqrt{1-\beta^2}}.
\end{aligned}\right\} \quad (6)$$

A l'aide de cette définition tensorielle, l'énergie et l'impulsion totales du champ électromagnétique se présenteront sous une forme quadri-vectorielle correcte :

$$U_\nu = \int T_{\nu\mu}\,\mathrm{d}V^\mu. \quad (7)$$

Et, nous aurons, dans notre cas :

$$\begin{aligned}
U = U_4 &= \int T_{44}\,\mathrm{d}V^4 + T_{4x}\,\mathrm{d}V^x\\
&= \int (T_{44} - \beta T_{4x})\frac{\mathrm{d}V_0}{\sqrt{1-\beta^2}},\\
G_x = \frac{1}{c}U_x &= \frac{1}{c}\int T_{x4}\,\mathrm{d}V^4 + T_{xx}\,\mathrm{d}V^x\\
&= \frac{1}{c}\int (T_{x4} - \beta T_{xx})\frac{\mathrm{d}V_0}{\sqrt{1-\beta^2}}.
\end{aligned}$$

On vérifiera sans peine que les transformations (3) conduisent maintenant aux relations suivantes

$$U = \frac{U_0}{\sqrt{1-\beta^2}} \qquad \text{et} \qquad \vec{G} = \frac{\vec{v}}{c^2}\,\frac{U_0}{\sqrt{1-\beta^2}}, \quad (8)$$

et non aux relations (5), qui nous paraissent avoir été obtenues par une voie incorrecte.

Manuscrit reçu le 12 novembre 1948.

Le Gérant : Maurice Blondin. 131476 Imp. Gauthier-Villars, 55, quai des Grands-Augustins, Paris. — 1949.

A.5 R. Ruffini: Charges in gravitational fields: From Fermi, via Hanni-Ruffini-Wheeler, to the "electric Meissner effect"

R. Ruffini: "Charges in gravitational fields: From Fermi, via
Hanni-Ruffini-Wheeler, to the 'electric Meissner effect' ," Nuovo Cimento B 119,
785 (2004)

IL NUOVO CIMENTO Vol. 119 B, N. 7-9 Luglio-Settembre 2004
DOI 10.1393/ncb/i2004-10213-8

Charges in gravitational fields: From Fermi, via Hanni-Ruffini-Wheeler, to the "electric Meissner effect"(*)

R. Ruffini(**)

Dipartimento di Fisica, Università "La Sapienza" - I-00185 Rome, Italy
ICRA, International Center for Relativistic Astrophysics, Università "La Sapienza"
I-00185 Rome, Italy

(ricevuto il 17 Novembre 2004)

Summary. — Recent developments in obtaining a detailed model for gamma-ray bursts have shown the need for a deeper understanding of phenomena described by solutions of the Einstein-Maxwell equations, reviving interest in the behavior of charges close to a black hole. In particular a drastic difference has been found between the lines of force of a charged test particle in the fields of Schwarzschild and Reissner-Nordström black holes. This difference characterizes a general relativistic effect for the electric field of a charged test particle around a (charged) Reissner-Nordström black hole similar to the "Meissner effect" for a magnetic field around a superconductor. These new results are related to earlier work by Fermi and Hanni-Ruffini-Wheeler.

PACS 04.20.-q – Classical general relativity.
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

It is a great pleasure to celebrate this eightieth birthday of M.me Choquet-Bruhat. M.me Choquet-Bruhat has collaborated with us in organizing the Marcel Grossman Meeting series for many years, as did Yakov Borisovich Zel'dovich, who had also been a long-standing member of the International Organization Committee of the Grossman Meetings as well as a good friend of both of us. I would like to begin with an anecdote about him. The achievements of Zel'dovich are well known worldwide: he had invented the Katiuscia Rocket which had played an essential role in the history of the Second World War. He then developed both the atomic and thermonuclear bombs of the Soviet

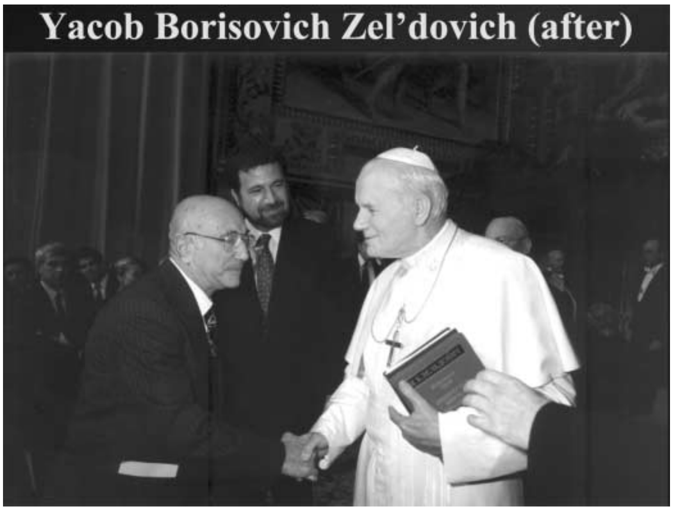
(*) Paper presented at the Elba Conference "Analysis, Manifolds and Geometric Structures in Physics", Elba, June 24-26, 2004.
(**) E-mail: ruffini@icra.it

Fig. 1. – Y. B. Zel'dovich being introduced to the Pope John Paul II by Remo Ruffini (before).

Union with Zakharov and was also instrumental in the development of space research in the Soviet Union. It was not until 1960 that Zel'dovich became interested in relativistic astrophysics and developed his internationally recognized school of research on this topic in the Soviet Union.

A great scientist may contribute to the progress of science not only by his rational scientific works, but also by his mental extravagance. In this sense Zel'dovich triggered one of the greatest discoveries ever in relativistic astrophysics: the gamma-ray bursts (GRBs). For the understanding of these sources it is essential to understand the process of energy extraction from a black hole, an energy we have called blackholic energy. In turn this theoretical research on the vacuum polarization process in the field of a black hole has demanded a deeper understanding of the interrelationships between Maxwell's equations and the Einstein equations. Exactly this problematic convinced us of the need to go back to some of the classic works on the interaction between a charged test particle and a black hole. It has been very fortunate that from this analysis new aspects of physics have surfaced which will be briefly summarized in this paper. I would also like to emphasize that this work finds its origin in a paper by Enrico Fermi which is largely unknown and published only in Italian, only being translated into English this year [1].

I had my first meeting with Zel'dovich in 1968 in Tibilisi, Georgia in the former Soviet Union. I was very impressed by his extensive scientific knowledge and at the same time intrigued by some peculiarities of his character, which immediately surfaced from the first scientific exchange and some anecdotes of his life that he recalled to me. Over the years we became very well acquainted and a great friendship developed between us. Nevertheless, this strangeness and somewhat unexpected manifestation of his character accompanied us all the way to our last meeting. It was in Rome that I had the occasion



Yacob Borisovich Zel'dovich (after)

Fig. 2. – Y. B. Zel'dovich again shaking the hand of the Pope after presentation of his collected papers. Everybody smiles with relief!

to introduce him to Pope John Paul II. Even in that solemn occasion Zel'dovich did not fail to manifest this duality between scientific knowledge and unexpected action. While he was approaching in line, I was introducing other distinguished guests to the Pope, including Bruno Pontecorvo, Roald Sagdeev and Rashid Sunyaev. I noticed that Zel'dovich had hidden under his jacket a voluminous object. This became more and more evident as he was approaching the Pope. You can see the concern in my eyes in fig. 1. When he arrived in front of the Pope, he suddenly opened the jacket and extracted two big red volumes and then handed them to the Pope. The Pope kindly thanked him. But again unexpectedly Zel'dovich took the volumes back from the hand of the Pope, and shouted loudly: "Not just thank you, these are fifty years of my work!" We all realized that these were his collected papers. We then all felt much more relaxed and warmly laughed with great relief. The Pope kept the Zel'dovich books under his arm against his white robe during the entire rest of the audience (see fig. 2).

The topic I will speak about here is again related to the dual activity of Zel'dovich, and it is definitely one of the most astonishing proposals ever made by a *homo sapiens*, which led to the discovery of gamma-ray bursts. It was during the 1950s that Zel'dovich, in order to show the greatness of his scientific achievements and the very large progress in space technology made by the Soviet Union, made a proposal to explode an atomic bomb on the Moon. In his opinion this would have shown the superiority of the Soviet rocketry to reach the Moon before the Americans and would have allowed a large fraction of inhabitants of the Earth to directly see this achievement by the observation of the fireball of the bomb explosion at a very precisely predicted time. Many technicalities hampered the realization of this idea: fortunately, this unacceptable proposal was never implemented. But the possibility of conceiving of such an action had become a reality,

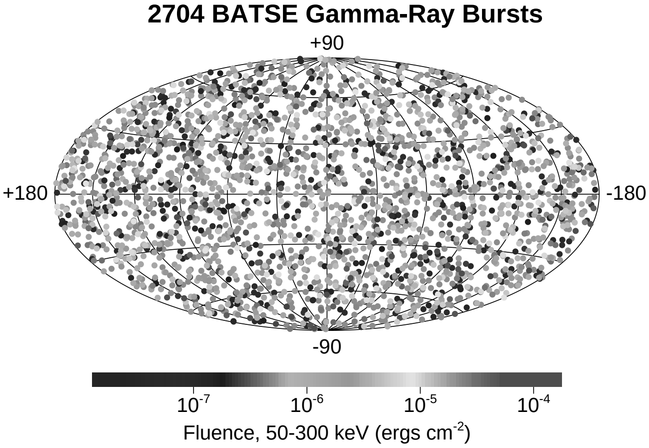

2704 BATSE Gamma-Ray Bursts
+90
+180
-180
-90
10^{-7}
10^{-6}
10^{-5}
10^{-4}
Fluence, 50-300 keV (ergs cm^{-2})

Fig. 3. – Position in the sky, in galactic coordinates, of 2000 GRB events seen by the CGRO satellite. Their isotropy is evident. Reproduced courtesy of the BATSE web site.

and the United States put the *Vela* satellites into very high Earth orbits in order to monitor the nonproliferation agreement. They discovered the gamma-ray bursts. In this case, therefore, even this extravagant and nonscientific proposal of Zel'dovich finally did materialize, fortunately, in a great scientific discovery.

2. – The energetics of gamma-ray bursts

GRBs were detected and studied for the first time using those *Vela* satellites, developed for military research to monitor the violations of the Limited Test Ban Treaty signed in 1963 (see, *e.g.*, Strong [2]). It was clear from the early data of these satellites, which were put at 150000 miles from the surface of the Earth, that the GRBs did not originate either on the Earth or in the Solar System.

The mystery of these sources became more profound as the observations of the BATSE instrument on board the Compton Gamma Ray Observatory (CGRO) satellite([1]) over 9 years proved the isotropic distribution of these sources in the sky (see fig. 3). In addition to these data, the CGRO satellite gave an unprecedented number of details about the structure of GRBs, and on their spectral properties and time variabilities which were recorded in the fourth BATSE catalog [3] (see, *e.g.*, fig. 4). Out of the analysis of these BATSE sources the existence of two distinct families of sources soon became clear (see, *e.g.*, Koveliotou *et al.* [4], Tavani *et al.* [5]): the long bursts, lasting more than one second and softer in spectra, and the short bursts, harder in spectra (see fig. 5).

The situation drastically changed with the discovery of the afterglow by the Italian-Dutch satellite BeppoSAX (Costa *et al.* [6]) and the possibility which led to the optical identification of the GRBs by the largest telescopes in the world, including the Hubble

[1] See http://cossc.gsfc.nasa.gov/batse/

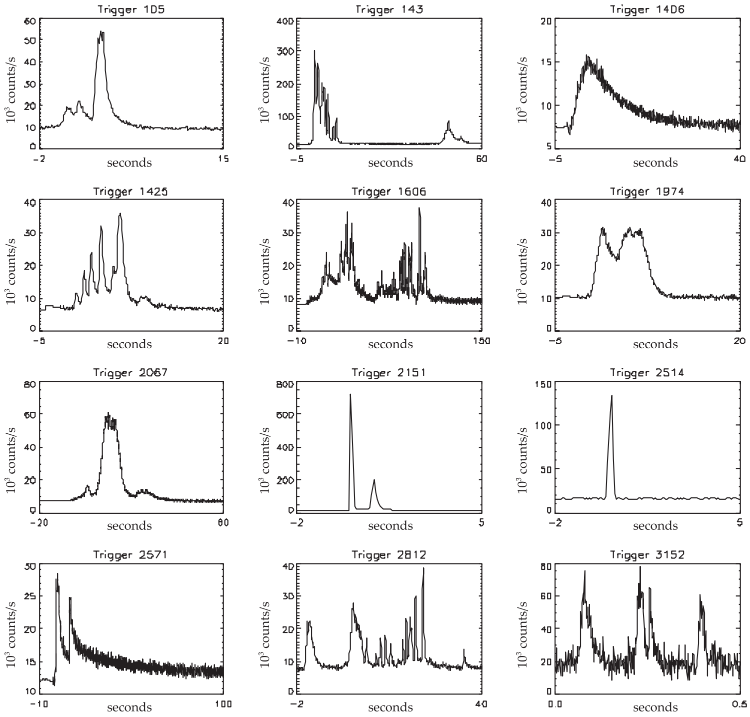
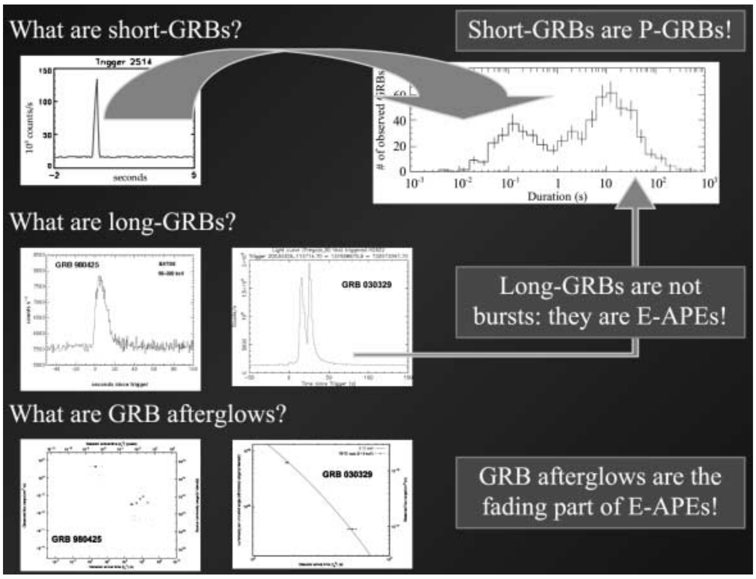


Fig. 4. – Some GRB light curves observed by the BATSE instrument on board the CGRO satellite.

Space Telescope, the Keck Telescope in Hawaii and the VLT in Chile, and allowed as well the identification in the radio band of these sources. The outcome of this collaboration between complementary observational techniques has led in 1997 to the possibility of identifying the distance of these sources from the Earth and their tremendous energy of the order up to 10^{54} erg/s during the burst. It is interesting, as we will show in the following, that energetics of this magnitude for the GRBs had already been predicted out of first principles by Damour and Ruffini in 1974 [7] (see fig. 6).

The resonance between the X- and gamma-ray astronomy from the satellites and the optical and radio astronomy from the ground, had already marked the great success and development of the astrophysics of binary X-ray sources in the seventies (see, *e.g.*, Giacconi and Ruffini [8]). This resonance has been repeated for GRBs on a much larger scale. The use of much larger satellites, like Chandra and XMM-Newton, and dedicated space missions, like HETE-2 and, in the near future, Swift, and the very fortunate circumstance of the coming of age of the development of unprecedented optical technologies for the telescopes offer opportunities without precedent in the history of mankind. In parallel, the enormous scientific interest in the nature of GRB sources and the explo-



What are short-GRBs?

Short-GRBs are P-GRBs!

What are long-GRBs?

Long-GRBs are not bursts: they are E-APEs!

What are GRB afterglows?

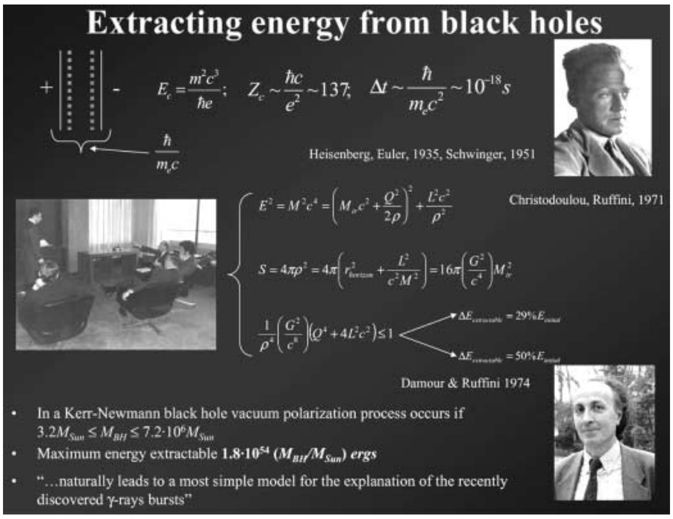
GRB afterglows are the fading part of E-APEs!

Fig. 5. – On the upper right part of the figure are plotted the number of the observed GRBs as a function of their duration. The bimodal distribution corresponding respectively to the short bursts, upper left figure, and the long bursts, middle figure, is quite evident.

ration, not only of new regimes, but also of the totally novel conceptual physical process of blackholic energy extraction, makes the knowledge of GRBs an authentic new frontier in scientific knowledge.

2˙1. *GRBs and general relativity.* – Three of the most important works in the field of general relativity have certainly been the discovery of the Kerr solution [9], its generalization to the charged case (Newman *et al.* [10]) and the formulation by B. Carter [11] of the Hamilton-Jacobi equations for a charged test particle in the metric and electromagnetic field of a Kerr-Newman solution (see, *e.g.*, Landau and Lifshitz [12]). The equations of motion, which are generally second-order differential equations, were reduced by Carter to a set of first-order differential equations which were then integrated using an effective potential technique by Ruffini and Wheeler for the Kerr metric (see, *e.g.*, Landau and Lifshitz [12]) and by Ruffini for the Reissner-Nordström geometry (Ruffini [13], see fig. 7).

All the above mathematical results were essential for understanding the new physics of gravitationally collapsed objects and allowed the publication of a very popular article: "Introducing the black hole" (Ruffini and Wheeler [15]). In that paper, we advanced the ansatz that the most general black hole is a solution of the Einstein-Maxwell equations, asymptotically flat and with a regular horizon: the Kerr-Newman solution, characterized only by three parameters: the mass M, the charge Q and the angular momentum L. This ansatz of the "black hole uniqueness theorem" still today after thirty years presents challenges to the mathematical aspects of its complete proof (see, *e.g.*, Carter [16] and



Extracting energy from black holes

$E_c = \frac{m^2c^3}{\hbar e}$; $Z_c \sim \frac{\hbar c}{e^2} \sim 137$; $\Delta t \sim \frac{\hbar}{m_e c^2} \sim 10^{-18} s$

$\frac{\hbar}{m_e c}$

Heisenberg, Euler, 1935, Schwinger, 1951

$$\begin{cases} E^2 = M^2c^4 = \left(M_{ir}c^2 + \frac{Q^2}{2\rho}\right)^2 + \frac{L^2c^2}{\rho^2} \\ S = 4\pi\rho^2 = 4\pi\left(r^2_{horizon} + \frac{L^2}{c^2M^2}\right) = 16\pi\left(\frac{G^2}{c^4}\right)M^2_{ir} \\ \frac{1}{\rho^4}\left(\frac{G^2}{c^8}\right)\left(Q^4 + 4L^2c^2\right) \leq 1 \end{cases}$$

Christodoulou, Ruffini, 1971

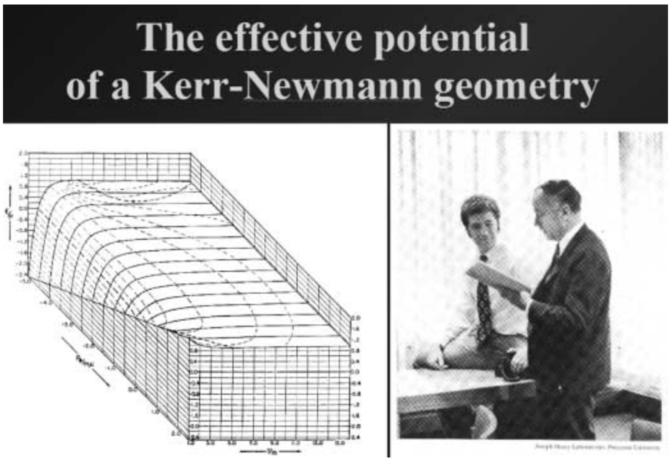
$\Delta E_{extractable} = 29\% E_{initial}$

$\Delta E_{extractable} = 50\% E_{initial}$

Damour & Ruffini 1974

- In a Kerr-Newmann black hole vacuum polarization process occurs if $3.2 M_{Sun} \leq M_{BH} \leq 7.2 \cdot 10^6 M_{Sun}$
- Maximum energy extractable **1.8·10^54 (M_{BH}/M_{Sun}) ergs**
- "…naturally leads to a most simple model for the explanation of the recently discovered γ-rays bursts"

Fig. 6. – Damour.



The effective potential of a Kerr-Newmann geometry

Fig. 7. – The effective potential corresponding to the circular orbits in the equatorial plane of a black hole is given as a function of the angular momentum of the test particle. This diagram was originally derived by Ruffini and Wheeler (right picture). For details see Landau and Lifshitz [12] and Rees, Ruffini and Wheeler [14].

Bini *et al.* [17]). In addition to these mathematical difficulties, in the field of physics this ansatz contains the most profound consequences. The fact that, among all the possible highly nonlinear terms characterizing the gravitationally collapsed objects, only the ones corresponding solely to the Einstein-Maxwell equations survive the formation of the horizon has, indeed, extremely profound physical implications. Any departure from such a minimal configuration either collapses to the horizon or is radiated away during the collapse process. This ansatz is crucial in identifying precisely the process of gravitational collapse leading to the formation of the black hole and the emission of GRBs. Indeed, in this specific case, the Born-like nonlinear term [18] of the Heisenberg-Euler-Schwinger Lagrangian [19,20] are radiated away prior to the formation of the horizon of the black hole (see, *e.g.*, Ruffini *et al.* [21]). Only the nonlinearity corresponding solely to the classical Einstein-Maxwell theory is left as the outcome of the gravitational collapse process.

The same effective potential technique (see Landau and Lifshitz [12]) which allowed the analysis of circular orbits around the black hole was crucial in reaching the equally interesting discovery of the reversible and irreversible transformations of black holes by Christodoulou and Ruffini [22], which in turn led to the mass-energy formula for the black hole

$$E_{\rm BH}^2 = M^2c^4 = \left(M_{\rm ir}c^2 + \frac{Q^2}{2\rho_+}\right)^2 + \frac{L^2c^2}{\rho_+^2}\,, \tag{1}$$

with

$$\frac{1}{\rho_+^4}\left(\frac{G^2}{c^8}\right)\left(Q^4 + 4L^2c^2\right) \leq 1\,, \tag{2}$$

and where

$$S = 4\pi\rho_+^2 = 4\pi\left(r_+^2 + \frac{L^2}{c^2M^2}\right) = 16\pi\left(\frac{G^2}{c^4}\right)M_{\rm ir}^2 \tag{3}$$

is the horizon surface area, $M_{\rm ir}$ is the irreducible mass, r_+ is the horizon radius and ρ_+ is the quasi-spheroidal cylindrical coordinate of the horizon evaluated at the equatorial plane. Extreme EMBHs satisfy the equality in eq. (2).

From eq. (1) there follows that the total energy of the black hole $E_{\rm BH}$ can be split into three different parts: rest mass, Coulomb energy and rotational energy. In principle both Coulomb energy and rotational energy can be extracted from the black hole (Christodoulou and Ruffini [22]). The maximum extractable rotational energy is 29% of the total energy and the maximum extractable Coulomb energy is 50%, as clearly follows from the upper limit for the existence of a black hole, given by eq. (2). We refer to both these extractable energies in the following as the blackholic energy.

The existence of the black hole and the basic correctness of the circular orbits has been proven by the observations of Cygnus-X1 (see, *e.g.*, Giacconi and Ruffini [8]). However, in binary X-ray sources, the black hole only acts passively by generating the deep potential well in which the accretion process occurs. It has become tantalizing to look for astrophysical objects in order to verify the other fundamental prediction of general relativity that the blackholic energy is the largest energy extractable from any physical object.

As we shall see in the next section, the feasibility of extracting the blackholic energy has been made possible by the quantum process of creating, out of classical fields, a plasma of electron-positron pairs in the field of a black hole. This process of energy extraction from the black hole is manifested astrophysically by the occurrence of GRBs.

2˙2. *GRBs and quantum electrodynamics*. – That a static electromagnetic field stronger than a critical value,

$$E_{\rm c} = \frac{m_{\rm e}^2 c^3}{\hbar e}, \tag{4}$$

can polarize the vacuum and create electron-positron pairs, was clearly shown by Heisenberg and Euler [19]. The major effort in verifying the correctness of this theoretical prediction has been directed towards the analysis of heavy-ion collisions (see Ruffini *et al.* [21] and references therein). From an order-of-magnitude estimate, it appears that around a nucleus with a charge

$$Z_{\rm c} \simeq \frac{\hbar c}{e^2} \simeq 137\,, \tag{5}$$

the electric field can be stronger than the critical electric field needed to polarize the vacuum. A more accurate detailed analysis taking into account the bound-state levels around a nucleus increases the value to

$$Z_{\rm c} \simeq 173 \tag{6}$$

for the nuclear charge leading to the existence of a critical field. From the Heisenberg uncertainty principle it follows that, in order to create a pair, the existence of the critical field should last a time

$$\Delta t \sim \frac{\hbar}{m_{\rm e} c^2} \simeq 10^{-18}\,{\rm s}\,, \tag{7}$$

which is much longer than the typical confinement time in heavy-ion collisions which is

$$\Delta t \sim \frac{\hbar}{m_{\rm p} c^2} \simeq 10^{-21}\,{\rm s}\,. \tag{8}$$

This is certainly a reason why no evidence for pair creation in heavy-ion collisions has been found although remarkable efforts have been made in various accelerators worldwide. Similar experiments involving laser beams encounter analogous difficulties (see, *e.g.*, Ruffini *et al.* [21] and references therein).

The alternative idea was advanced in 1975 [7] that the critical-field condition given in eq. (4) could be reached easily, and for a time much larger than the one given by eq. (7), in the field of a Kerr-Newman black hole in a range of masses $3.2 M_\odot \leq M_{\rm BH} \leq 7.2 \times 10^6 M_\odot$. In that paper we generalized the fundamental theoretical framework developed in Minkowski space by Heisenberg-Euler [19] and Schwinger [20] to the curved Kerr-Newman geometry. This result was made possible by the work on the structure of the Kerr-Newman space-time previously done by Carter [11] and by the remarkable

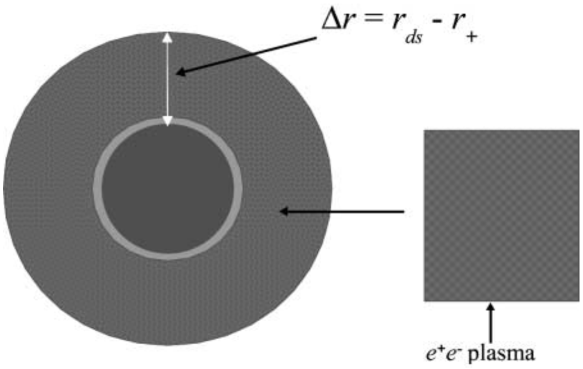


Fig. 8. – The dyadosphere is comprised between the horizon radius and the radius of the dyadosphere. This region is entirely filled with electron-positron pairs and photons in thermal equilibrium. Details in Ruffini [26], Preparata *et al.* [27], Ruffini *et al.* [28].

mathematical craftsmanship of Thibault Damour then working with me as a *post-doc* in Princeton.

The maximum energy extractable in such a process of creating a vast amount of electron-positron pairs around a black hole is given by

$$E_{\max} = 1.8 \times 10^{54}\,(M_{\rm BH}/M_\odot)\ \text{erg}\,. \tag{9}$$

We concluded in that paper that such a process "naturally leads to a most simple model for the explanation of the recently discovered gamma-ray bursts".

At that time, GRBs had not yet been optically identified and nothing was known about their distance and consequently about their energetics. Literally thousands of theories existed in order to explain them and it was impossible to establish a rational dialogue with such an enormous number of alternative theories. We did not pursue further our model until the results of the BeppoSAX mission, which clearly pointed to the cosmological origin of GRBs, implying for the typical magnitude of their energy precisely the one predicted by our model.

It is interesting that the idea of using an electron-positron plasma as the basis of a GRB model was independently introduced years later in a set of papers by Cavallo and Rees [23], Cavallo and Horstman [24] and Horstman and Cavallo [25]. These authors did not address the issue of the physical origin of their energy source. They reach their conclusions considering the pair creation and annihilation process occurring in the confinement of a large amount of energy in a region of dimension ~ 10 km typical of a neutron star. No relation to the physics of black holes nor to the energy extraction process from a black hole was envisaged in their interesting considerations, mainly directed to the study of the opacity and the consequent dynamics of such an electron-positron plasma.

After the discovery of the afterglows and the optical identification of GRBs at cosmological distances, implying exactly the energetics predicted in eq. (9), we returned to the analysis of the vacuum polarization process around a black hole and precisely identified the region around the black hole in which the vacuum polarization process and the subsequent creation of electron-positron pairs occur. We defined this region, using the Greek name dyad for pairs ($\delta\upsilon\alpha\varsigma$, $\delta\upsilon\alpha\delta o\varsigma$), to be the "dyadosphere" of the black hole, bounded by the black-hole horizon and the dyadosphere radius $r_{\rm ds}$ given by (see Ruffini [26], Preparata *et al.* [27] and fig. 8)

$$(10) \qquad r_{\rm ds} = \left(\frac{\hbar}{mc}\right)^{1/2}\left(\frac{GM}{c^2}\right)^{1/2}\left(\frac{m_{\rm p}}{m}\right)^{1/2}\left(\frac{e}{q_{\rm p}}\right)^{1/2}\left(\frac{Q}{\sqrt{G}M}\right)^{1/2} = \\ = 1.12\times 10^8\sqrt{\mu\xi}\,{\rm cm}\,,$$

where we have introduced the dimensionless mass and charge parameters $\mu = M_{\rm BH}/M_\odot$, $\xi = Q/(M_{\rm BH}\sqrt{G}) \le 1$.

At that time the analysis of the dyadosphere was developed around an already formed black hole. In recent months we have been developing the dynamical formation of the black hole and correspondingly of the dyadosphere during the process of gravitational collapse, reaching some specific signatures which may be detectable in the structure of the short and long GRBs (Cherubini *et al.* [29], Ruffini and Vitagliano [30, 31], Ruffini *et al.* [28, 32, 33]).

3. – Reconsideration of a classic Fermi paper

At the very foundation of the GRB phenomena is the vacuum polarization process due to overcritical electric fields of black holes. For these reasons we decided to go back to some of our earlier work and some other classic work in the literature on test particles in gravitational fields, and we have discovered a wealth of new results and opened as well additional new problems for enquiry. We have reconsidered a pioneering paper by Enrico Fermi [34], which has been generally neglected since it was written in Italian. It has only just now been translated into English [1]. In this paper Fermi investigated the electric field generated by a charged particle at rest in a given static and homogeneous gravitational field in the space-time region close to the particle location and then used his result to study the influence of the gravitational field on the charge distribution on an infinitely conducting sphere. He showed that in this case the sphere acquires a dipole electric field and is polarized. In fact, the solution for the electrostatic potential (and the field) can be expressed as the superposition of the solutions corresponding to a point charge and a dipole of suitable moment, in order to satisfy the condition of constancy of the potential on the surface of the sphere.

Fermi uses in his paper the following form of the metric due to Levi-Civita [35] for a uniform gravitational field:

$$({\rm F1}) \qquad {\rm d}s^2 = -(1-2AZ){\rm d}T^2 + {\rm d}X^2 + {\rm d}Y^2 + {\rm d}Z^2 + O(AZ)\,,$$

with the condition $AZ \ll 1$; A denotes the acceleration of gravity. In this metric Fermi considers Maxwell's equations

$$({\rm F2}) \qquad F^{\alpha\beta}{}_{;\beta} = 4\pi J^\alpha, \qquad {}^*F^{\alpha\beta}{}_{;\beta} = 0\,, \qquad F_{\alpha\beta} = 2A_{[\beta;\alpha]}\,.$$

One can introduce (pseudo-) electric and magnetic field quantities

$$({\rm F3}) \qquad E_i = F_{i0}\,, \qquad B^i = \frac{1}{2}\epsilon^{ijk}F_{jk}$$

differing from the physical fields which are instead the orthonormal frame components (and not the coordinate components) of the Faraday 2-form F. In the electrostatic case $B_i = 0$ and $E_{i,0} = 0$, and the vector potential A_μ is determined by the electrostatic potential Φ alone ($A_0 = \Phi$, $A_i = 0$), in terms of which the (pseudo-) electric field components can be written in the form $E_X = -\Phi_{,X}, E_Y = -\Phi_{,Y}, E_Z = -\Phi_{,Z}$.

He then considers a charge q located at the origin of the coordinates as the source term J^α, *i.e.* with current density

$$J^\alpha = \rho u^\alpha \ , \qquad \rho = q\delta(X)\delta(Y)\delta(Z)\,, \tag{F4}$$

where $u = (1 - AZ)^{-1}\partial_T$ is the particle 4-velocity.

In the limit of validity of the metric (F1), Maxwell's equations reduce to the following equation for the electrostatic potential Φ:

$$\nabla^2\Phi + A\partial_Z\Phi = -4\pi(1 - AZ)\rho\,. \tag{F5}$$

The solution corresponding to the source term (F4) is given by

$$\Phi_{\rm part} = \frac{q}{\sqrt{X^2 + Y^2 + Z^2}}\left[1 - \frac{A}{2}Z\right]\,. \tag{F6}$$

Finally, let us assume that the charge is distributed on a conducting sphere of radius R, centered at the origin of the coordinate system. From the condition that the electrostatic potential must be constant on the surface, Fermi finds that a polarization charge density appears on the sphere corresponding to a superposition of a monopolar and a dipolar distribution, explicitly given by

$$\sigma_{\rm F}^R(\theta) = \frac{q}{4\pi R^2} + \frac{qA}{2\pi R}\cos\theta\,. \tag{F7}$$

Fermi also shows that the electric part of the electromagnetic field generated by the electric charge at rest in a homogeneous field of strength A is equal to the electric part of the electromagnetic field which the same charge would produce in the absence of a gravitational field if it moved in accelerated motion with acceleration $A/2$ in the opposite direction with respect to the gravitational field. However, a nonzero magnetic field is present in this latter case, and the Fermi solution corresponds to suitably choosing a gauge in such a way that the vector potential A_μ has the component $A_0 = \Phi$ only. It is interesting that even in this problem there are still some open questions: the factor 1/2 appearing in the acceleration $A/2$ still remains to be completely understood and is very likely connected by some analogy to the same factor 2 occurring in the Thomas precession. This topic is still a matter of active discussion with V. Belinski, D. Bini, J. Elhers and A. Geralico.

4. – Discussions of Wheeler-Hanni-Unruh on the charge near a Schwarzschild black hole

One of the most exciting problems proposed by Johnny Wheeler to students and collaborators at Princeton (see fig. 9) was the problem of a charged test particle at rest near a black hole. The characteristic style of Wheeler has always been to have a

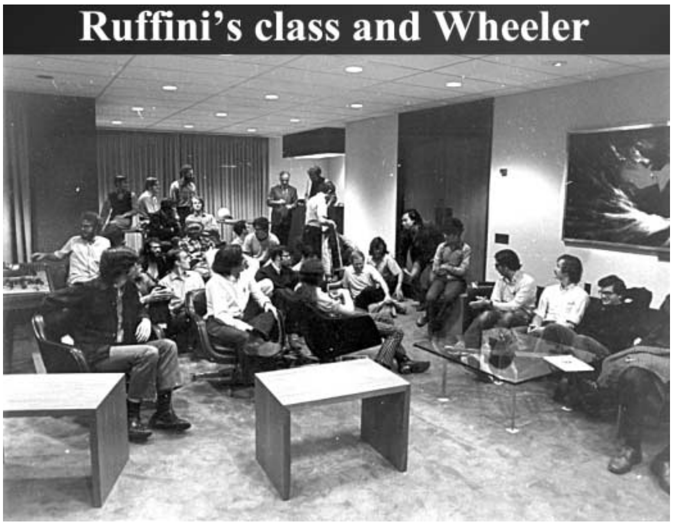


Fig. 9. – My students at Princeton in an evening discussion with John Wheeler. Recognizable on the right are Jim Eisenberg and Rick Hanni. Johnny and I are standing in the back of the room.

strong intuition about the solution of physical problems. His motto, known as "Wheeler theorem number 1", is "Never do a computation without knowing the solution," and he was usually extremely good at guessing the solution of a problem. We had just introduced with Johnny the astrophysical concept of a black hole [15]. It is interesting that the specific case of the charge near a black hole really caught the attention of the students at Princeton, and all of them participated in trying to find solutions of this problem. In particular, Jacob Bekenstein, William Unruh and many others contributed to a lively discussion on the possible outcome of the solution. There were two different possibilities for the field line configurations, as outlined in fig. 10: the one on the left was the first proposal of Johnny, and the one on the right was the one proposed by Unruh, adopting for the Schwarzschild black hole the analogy with an infinitely conducting metal sphere.

While the discussions were polarizing our small scientific community in Princeton, I decided to enter in this issue by bypassing the philosophical and intuitive approach and just solve the corresponding set of equations. It was at that time that Wheeler introduced me to a young very bright undergraduate, Rickard Hanni (see fig. 11).

Recently, reading Fermi's paper, I noticed that the equations we used (see figs. 12 and 13) have the same structure of the ones he used there, except instead of the Levi-Civita metric (F1), describing a uniform gravitational field, we used the Schwarzschild metric

$$\text{(HR1)} \qquad \mathrm{d}s^2 = -f_S(r)\mathrm{d}t^2 + f_S(r)^{-1}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2)\,,$$

where $f_S(r) = 1 - 2\mathcal{M}/r$. The current density corresponding to a charge q placed at the

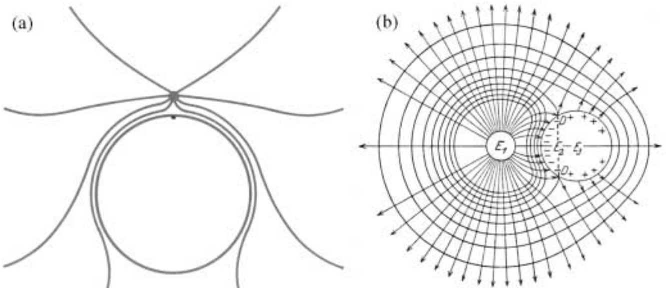
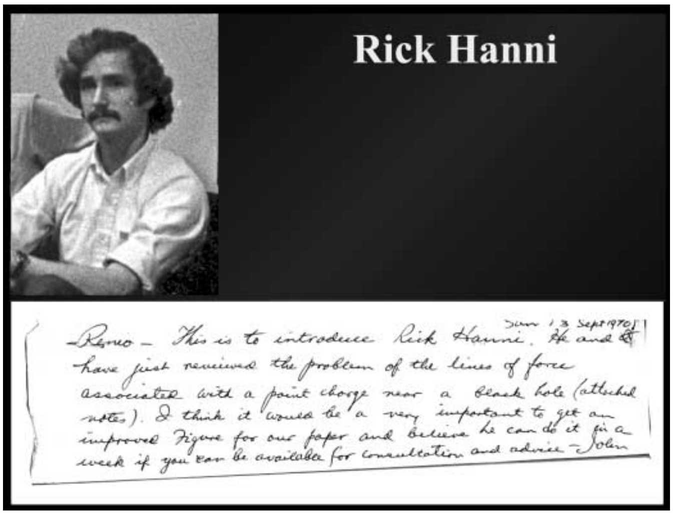


Abb. 26. Influenzierung der isolierten Kugel um E_2 durch eine Punktladung E_1 (darstellbar durch die Punktladung E_3), wobei die Kugel auf konstantem Potential sich befindet (darstellbar durch die Punktladung in E_2). Vgl. Ziff. 69. Der Kreis *DD* ist neutrale Linie.

Fig. 10. – The behavior of the lines of force of the electric field of the test particle as suggested by Wheeler and Unruh ((a) and (b), respectively). The figure on the right is taken from the classical book of Weber on electricity and magnetism.

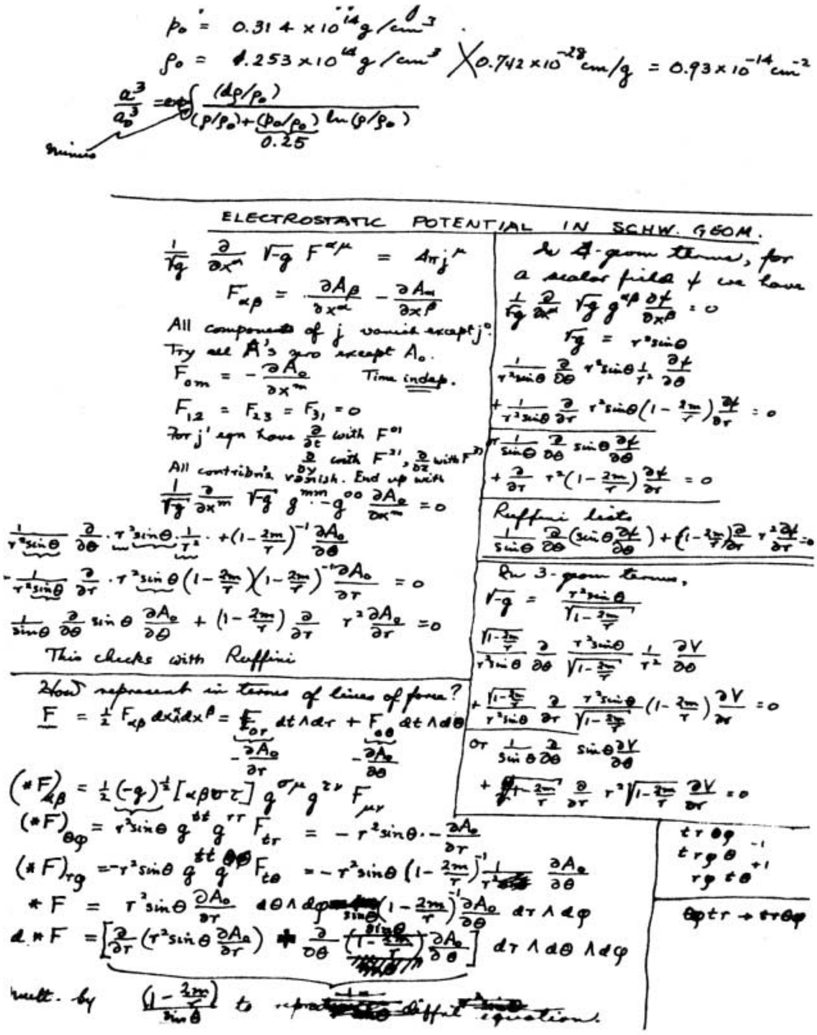


Rick Hanni

Sun 13 Sept 1970

Remo – This is to introduce Rick Hanni. He and I
have just reviewed the problem of the lines of force
associated with a point charge near a black hole (attached
notes). I think it would be a very important to get an
improved Figure for our paper and believe he can do it in a
week if you can be available for consultation and advice – John

Fig. 11. – Rick Hanni at Princeton and the note of Wheeler introducing him. Johnny optimistically, as usual, was expecting the problem of the lines of force to be solved in a week. It took almost one year of very hard work [36].



$p_0 = 0.314 \times 10^{14}\, g/cm^3$.

$\rho_0 = 1.253 \times 10^{14}\, g/cm^3 \times 0.742 \times 10^{-28}\, cm/g = 0.93 \times 10^{-14}\, cm^{-2}$

$$\frac{a^3}{a_0^3} = \exp \int \frac{(d\rho/\rho_0)}{(\rho/\rho_0) + \underbrace{(p_0/\rho_0)}_{0.25} \ln(\rho/\rho_0)}$$

minus

ELECTROSTATIC POTENTIAL IN SCHW. GEOM.

$$\frac{1}{\sqrt{g}} \frac{\partial}{\partial x^\alpha} \sqrt{-g}\, F^{\alpha\mu} = 4\pi j^\mu$$

$$F_{\alpha\beta} = \frac{\partial A_\beta}{\partial x^\alpha} - \frac{\partial A_\alpha}{\partial x^\beta}$$

All components of j vanish except j^0.

Try all A's zero except A_0.

$F_{0m} = -\frac{\partial A_0}{\partial x^m}$ Time indep.

$F_{12} = F_{23} = F_{31} = 0$

For j' eqn have $\frac{\partial}{\partial t}$ with F^{01}

$\frac{\partial}{\partial y}$ with F^{21}, $\frac{\partial}{\partial z}$ with F^{31}

All contrib'ns vanish. End up with

$$\frac{1}{\sqrt{g}} \frac{\partial}{\partial x^m} \sqrt{g}\, g^{mm} \cdot -g^{00} \frac{\partial A_0}{\partial x^m} = 0$$

$$\frac{1}{r^2 \sin\theta} \frac{\partial}{\partial \theta} \cdot r^2 \sin\theta \cdot \frac{1}{r^2} \cdot + \left(1 - \frac{2m}{r}\right)^{-1} \frac{\partial A_0}{\partial \theta}$$

$$+ \frac{1}{r^2 \sin\theta} \frac{\partial}{\partial r} \cdot r^2 \sin\theta \left(1 - \frac{2m}{r}\right)\left(1 - \frac{2m}{r}\right)^{-1} \frac{\partial A_0}{\partial r} = 0$$

$$\frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \sin\theta \frac{\partial A_0}{\partial \theta} + \left(1 - \frac{2m}{r}\right) \frac{\partial}{\partial r} r^2 \frac{\partial A_0}{\partial r} = 0$$

This checks with Ruffini

How represent in terms of lines of force?

$$\underline{F} = \frac{1}{2} F_{\alpha\beta}\, dx^\alpha dx^\beta = \underbrace{F_{0r}}_{-\frac{\partial A_0}{\partial r}} dt \wedge dr + \underbrace{F_{0\theta}}_{-\frac{\partial A_0}{\partial \theta}} dt \wedge d\theta$$

$$({}^*F)_{\alpha\beta} = \frac{1}{2} \underbrace{(-g)^{\frac{1}{2}}} [\alpha\beta\sigma\tau]\, g^{\sigma\mu} g^{\tau\nu} F_{\mu\nu}$$

$$({}^*F)_{\theta\varphi} = r^2 \sin\theta\, g^{tt} g^{rr} F_{tr} = -r^2 \sin\theta \cdot -\frac{\partial A_0}{\partial r}$$

$$({}^*F)_{r\varphi} = -r^2 \sin\theta\, g^{tt} g^{\theta\theta} F_{t\theta} = -r^2 \sin\theta \left(1 - \frac{2m}{r}\right)^{-1} \frac{1}{r^2 \cancel{\sin\theta}} \frac{\partial A_0}{\partial \theta}$$

$${}^*F = r^2 \sin\theta \frac{\partial A_0}{\partial r}\, d\theta \wedge d\varphi \; \sin\theta \left(1 - \frac{2m}{r}\right)^{-1} \frac{\partial A_0}{\partial \theta}\, dr \wedge d\varphi$$

$$d{}^*F = \underbrace{\left[\frac{\partial}{\partial r}\left(r^2 \sin\theta \frac{\partial A_0}{\partial r}\right) + \frac{\partial}{\partial \theta} \frac{\sin\theta}{\left(1 - \frac{2m}{r}\right)} \frac{\partial A_0}{\partial \theta}\right]}\, dr \wedge d\theta \wedge d\varphi$$

mult. by $\frac{\left(1 - \frac{2m}{r}\right)}{\sin\theta}$ to reproduce ~~[illegible]~~ diff'l equation.

In 4-geom terms, for a scalar field ψ we have

$$\frac{1}{\sqrt{g}} \frac{\partial}{\partial x^\alpha} \sqrt{g}\, g^{\alpha\beta} \frac{\partial \psi}{\partial x^\beta} = 0$$

$$\sqrt{g} = r^2 \sin\theta$$

$$\frac{1}{r^2 \sin\theta} \frac{\partial}{\partial \theta} r^2 \sin\theta \frac{1}{r^2} \frac{\partial \psi}{\partial \theta}$$

$$+ \frac{1}{r^2 \sin\theta} \frac{\partial}{\partial r} r^2 \sin\theta \left(1 - \frac{2m}{r}\right) \frac{\partial \psi}{\partial r} = 0$$

$$\text{or } \frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \sin\theta \frac{\partial \psi}{\partial \theta}$$

$$+ \frac{\partial}{\partial r} r^2 \left(1 - \frac{2m}{r}\right) \frac{\partial \psi}{\partial r} = 0$$

Ruffini lists

$$\frac{1}{\sin\theta} \frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial \psi}{\partial \theta}\right) + \left(1 - \frac{2m}{r}\right) \frac{\partial}{\partial r} r^2 \frac{\partial \psi}{\partial r} = 0$$

In 3-geom terms,

$$\sqrt{-g} = \frac{r^2 \sin\theta}{\sqrt{1 - \frac{2m}{r}}}$$

$$\frac{\sqrt{1 - \frac{2m}{r}}}{r^2 \sin\theta} \frac{\partial}{\partial \theta} \frac{r^2 \sin\theta}{\sqrt{1 - \frac{2m}{r}}} \frac{1}{r^2} \frac{\partial V}{\partial \theta}$$

$$+ \frac{\sqrt{1 - \frac{2m}{r}}}{r^2 \sin\theta} \frac{\partial}{\partial r} \frac{r^2 \sin\theta}{\sqrt{1 - \frac{2m}{r}}} \left(1 - \frac{2m}{r}\right) \frac{\partial V}{\partial r} = 0$$

$$\text{or } \frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \sin\theta \frac{\partial V}{\partial \theta}$$

$$+ \sqrt{1 - \frac{2m}{r}} \frac{\partial}{\partial r} r^2 \sqrt{1 - \frac{2m}{r}} \frac{\partial V}{\partial r} = 0$$

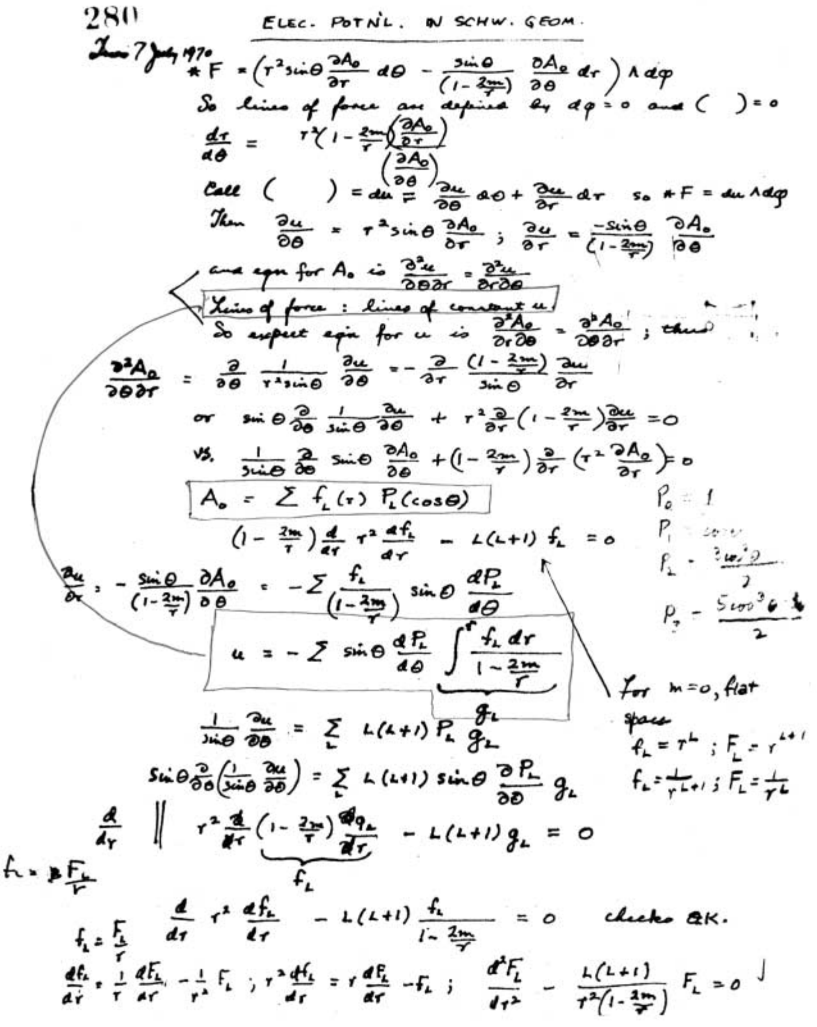
$t\,r\,\theta\,\varphi$

$t\,r\,\varphi\,\theta$ -1

$r\,\varphi\,t\,\theta$ $+1$

$\theta\varphi t r \to t r \theta\varphi$

Fig. 12. – From Wheeler's notebook (1).



280

ELEC. POTN'L. IN SCHW. GEOM.

Tues 7 July 1970

$$*F = \left(r^2 \sin\theta \frac{\partial A_0}{\partial r} d\theta - \frac{\sin\theta}{(1-\frac{2m}{r})} \frac{\partial A_0}{\partial \theta} dr\right) \wedge d\varphi$$

So lines of force are defined by $d\varphi = 0$ and $(\quad) = 0$

$$\frac{dr}{d\theta} = \frac{r^2(1-\frac{2m}{r})\left(\frac{\partial A_0}{\partial r}\right)}{\left(\frac{\partial A_0}{\partial \theta}\right)}$$

Call $(\quad) = du = \frac{\partial u}{\partial \theta} d\theta + \frac{\partial u}{\partial r} dr$ so $*F = du \wedge d\varphi$

Then $\frac{\partial u}{\partial \theta} = r^2 \sin\theta \frac{\partial A_0}{\partial r}$; $\frac{\partial u}{\partial r} = \frac{-\sin\theta}{(1-\frac{2m}{r})} \frac{\partial A_0}{\partial \theta}$

and eqn for A_0 is $\frac{\partial^2 u}{\partial \theta \partial r} = \frac{\partial^2 u}{\partial r \partial \theta}$

$\boxed{\text{Lines of force : lines of constant } u}$

So expect eqn for u is $\frac{\partial^2 A_0}{\partial r \partial \theta} = \frac{\partial^2 A_0}{\partial \theta \partial r}$; thus

$$\frac{\partial^2 A_0}{\partial \theta \partial r} = \frac{\partial}{\partial \theta} \frac{1}{r^2 \sin\theta} \frac{\partial u}{\partial \theta} = -\frac{\partial}{\partial r} \frac{(1-\frac{2m}{r})}{\sin\theta} \frac{\partial u}{\partial r}$$

or $\sin\theta \frac{\partial}{\partial \theta} \frac{1}{\sin\theta} \frac{\partial u}{\partial \theta} + r^2 \frac{\partial}{\partial r}\left(1-\frac{2m}{r}\right)\frac{\partial u}{\partial r} = 0$

vs. $\frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \sin\theta \frac{\partial A_0}{\partial \theta} + \left(1-\frac{2m}{r}\right) \frac{\partial}{\partial r}\left(r^2 \frac{\partial A_0}{\partial r}\right) = 0$

$$\boxed{A_0 = \sum f_L(r) P_L(\cos\theta)}$$

$$\left(1-\frac{2m}{r}\right)\frac{d}{dr} r^2 \frac{df_L}{dr} - L(L+1) f_L = 0$$

$P_0 = 1$

$P_1 = \cos\theta$

$P_2 = \frac{3\cos^2\theta}{2}$

$P_3 = \frac{5\cos^3\theta - 3}{2}$

$$\frac{\partial u}{\partial r} = -\frac{\sin\theta}{(1-\frac{2m}{r})} \frac{\partial A_0}{\partial \theta} = -\sum \frac{f_L}{(1-\frac{2m}{r})} \sin\theta \frac{dP_L}{d\theta}$$

$$\boxed{u = -\sum \sin\theta \frac{dP_L}{d\theta} \underbrace{\int^r \frac{f_L\, dr}{1-\frac{2m}{r}}}_{g_L}}$$

For $m = 0$, flat space

$f_L = r^L$; $F_L = r^{L+1}$

$f_L = \frac{1}{r^{L+1}}$; $F_L = \frac{1}{r^L}$

$$\frac{1}{\sin\theta} \frac{\partial u}{\partial \theta} = \sum_L L(L+1) P_L g_L$$

$$\sin\theta \frac{\partial}{\partial \theta}\left(\frac{1}{\sin\theta} \frac{\partial u}{\partial \theta}\right) = \sum_L L(L+1) \sin\theta \frac{\partial P_L}{\partial \theta} g_L$$

$$\frac{d}{dr} \quad \Big\| \quad r^2 \frac{d}{dr} \underbrace{\left(1-\frac{2m}{r}\right)\frac{dg_L}{dr}}_{f_L} - L(L+1) g_L = 0$$

$f_L = \frac{F_L}{r}$

$$\frac{d}{dr} r^2 \frac{df_L}{dr} - L(L+1) \frac{f_L}{1-\frac{2m}{r}} = 0 \quad \text{checks OK.}$$

$f_L = \frac{F_L}{r}$

$$\frac{df_L}{dr} = \frac{1}{r}\frac{dF_L}{dr} - \frac{1}{r^2} F_L \; ; \; r^2 \frac{df_L}{dr} = r \frac{dF_L}{dr} - F_L \; ; \quad \frac{d^2 F_L}{dr^2} - \frac{L(L+1)}{r^2(1-\frac{2m}{r})} F_L = 0$$

Fig. 13. – From Wheeler's notebook (2).

point $r = b$ on the polar axis $\theta = 0$, with $b > 2\mathcal{M}$, is given by

$$\text{(HR4)} \qquad J^0 = \frac{q}{2\pi r^2}\delta(r-b)\delta(\cos\theta - 1)\,.$$

Maxwell's equations (F2) then reduce to the following equation for the electrostatic potential V for the (pseudo-) electric field $E_r = -V_{,r}$, $E_\theta = -V_{,\theta}$, $E_\phi = -V_{,\phi}$, namely

$$\text{(HR5)} \qquad (r^2 V_{,r})_{,r} + \frac{f_S(r)^{-1}}{\sin\theta}\left[(\sin\theta V_{,\theta})_{,\theta} + \frac{1}{\sin\theta}(V_{,\phi})_{,\phi}\right] = -4\pi r^2 J^0\,.$$

I solved this equation with Hanni (see fig. 14) using a multipole expansion [36]

$$\text{(HR6)} \qquad V = q\textstyle\sum_l [f_l(b)g_l(r)\vartheta(b-r) + g_l(b)f_l(r)\vartheta(r-b)]P_l(\cos\theta)\,,$$

where

$$\text{(11)} \qquad \begin{aligned} f_l(r) &= -\frac{(2l+1)!}{2^l(l+1)!l!\mathcal{M}^{(l+1)}}\frac{(r-2\mathcal{M})^2}{r}\frac{\mathrm{d}Q_l}{\mathrm{d}r}\left[\frac{r-\mathcal{M}}{\mathcal{M}}\right] && l = 0,1,2,\ldots \\ g_l(r) &= \begin{cases} 1 & l = 0 \\ \dfrac{2^l l!(l-1)!\mathcal{M}^l}{(2l)!}\dfrac{(r-2\mathcal{M})^2}{r}\dfrac{\mathrm{d}P_l}{\mathrm{d}r}\left[\dfrac{r-\mathcal{M}}{\mathcal{M}}\right] & l = 1,2,\ldots \end{cases} \end{aligned}$$

and P_l, Q_l are the Legendre functions. We then derived the lines of force by defining the lines of constant flux, obtaining the behavior shown in fig. 15 (details in [36]).

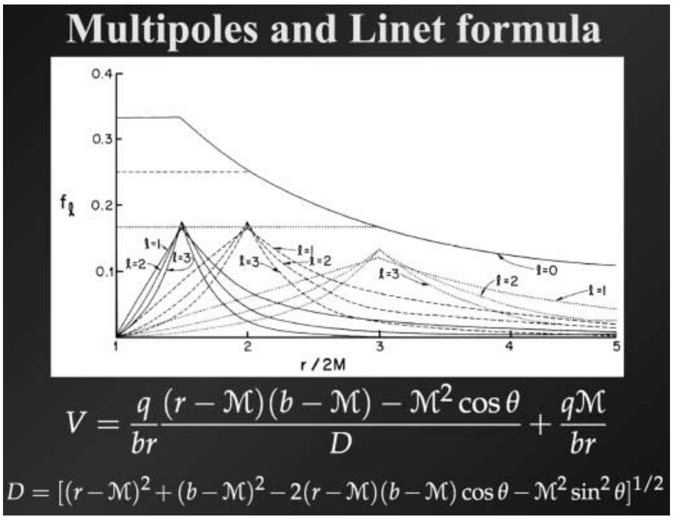
We also defined the concept of the induced charge on the surface of the black-hole horizon, which indeed appears to have some of the properties of a perfectly conducting sphere. If we assume that the test charge and black-hole charge are both positive, at angles smaller than a certain critical angle the induced charge is negative and the lines of force go towards the horizon, while at angles greater than the critical angle the induced charge is positive and the lines of force go away from it. At the critical angle the induced charge density vanishes and the lines of force of the electric field are tangent to the horizon.

This confirmed the Unruh ansatz for the behaviour of the lines of force, but there always remained in my mind the question: "How could it be that the very fertile imagination of Johnny did not enable him to guess *a priori* the correct solution?" As we will show below recent developments may explain that at a deeper level Wheeler was indeed correct to be undecided on this issue. In the meantime my work with Hanni was improved by an important mathematical solution obtained by Linet [37]. He derived a closed form for the electrostatic potential (HR6) by summing over all the multipoles

$$\text{(12)} \qquad V = \frac{q}{br}\frac{(r-\mathcal{M})(b-\mathcal{M}) - \mathcal{M}^2\cos\theta}{D_S} + \frac{q\mathcal{M}}{br}\,,$$

with

$$\text{(13)} \qquad D_S = [(r-\mathcal{M})^2 + (b-\mathcal{M})^2 - 2(r-\mathcal{M})(b-\mathcal{M})\cos\theta - \mathcal{M}^2\sin^2\theta]^{1/2}\,.$$



Multipoles and Linet formula

$$V = \frac{q}{br}\frac{(r-\mathcal{M})(b-\mathcal{M})-\mathcal{M}^2\cos\theta}{D} + \frac{q\mathcal{M}}{br}$$

$$D = [(r-\mathcal{M})^2+(b-\mathcal{M})^2-2(r-\mathcal{M})(b-\mathcal{M})\cos\theta-\mathcal{M}^2\sin^2\theta]^{1/2}$$

Fig. 14. – The radial functions $f_l(r)$ in the multipole moment expression for the potential V given by eq. (11) are shown for a test particle at a selected distance from a Schwarzschild black hole. Below is the closed form of the electrostatic potential V derived by Linet [37].

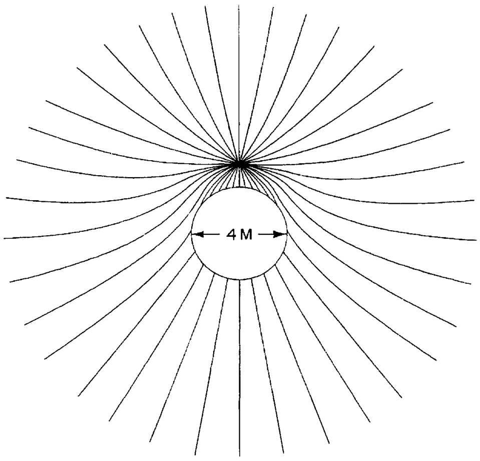


FIG. 4. Lines of force with the test charge at rest at $r = 3M$.

Fig. 15. – Lines of force of the test field with the charged particle at rest on the vertical axis $\theta = 0$ at $r = b$ with $b/\mathcal{M} = 3$ (from [36]).

The induced charge density on the horizon is then easily evaluated

$$\sigma_S^{\mathrm{H}}(\theta) = \frac{q}{8\pi b}\frac{\mathcal{M}(1+\cos^2\theta) - 2(b-\mathcal{M})\cos\theta}{[b-\mathcal{M}(1+\cos\theta)]^2}\,. \tag{HR7}$$

Later on, during the preparation of my volume with Rees and Wheeler, Johnny drew the electric lines of force as they should appear in an embedding diagram of the Schwarzschild solution (see fig. 16 (a)). It is remarkable that he did this free hand, based on his great intuition. It is interesting to compare these same lines of force with those which have been recently recomputed [38] by introducing the explicit computation of the embedding diagram (see fig. 16 (b)). The exterior Schwarzschild solution can be visualized as a 2-dimensional hyperboloid embedded in the usual Euclidean 3-space, by suppressing the temporal and azimuthal dimensions associated with the symmetry. The constant time equatorial slice has the reduced metric

$$\mathrm{d}s^2 = f_S(r)^{-1}\mathrm{d}r^2 + r^2\mathrm{d}\phi^2\,. \tag{14}$$

For $r > 2\mathcal{M}$ the coordinate r is spacelike, so this metric can be embedded in Euclidean space. By employing regular cylindrical coordinates, the Euclidean metric is given by

$$\mathrm{d}s^2 = \mathrm{d}\rho^2 + \rho^2\mathrm{d}\phi^2 + \mathrm{d}z^2\,, \tag{15}$$

with the same azimuthal angle ϕ for both metrics. If we require that the metrics (14) and (15) agree at constant ϕ, we get the condition

$$\mathrm{d}\rho^2 + \mathrm{d}z^2 = f_S(r)^{-1}\mathrm{d}r^2\ . \tag{16}$$

Setting $\rho = r$, this equation can be easily solved for z as a function of r

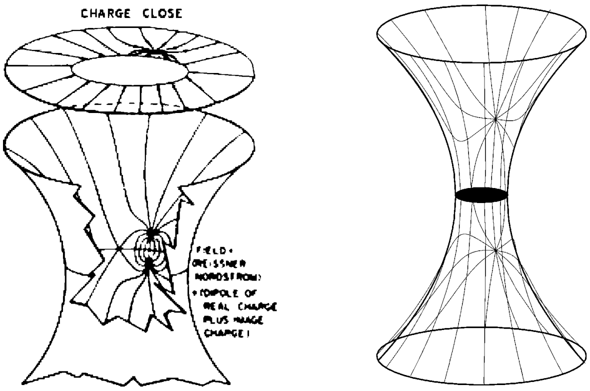
$$z = \int_{2\mathcal{M}}^{r}\left[\frac{2\mathcal{M}}{rf_S(r)}\right]^{1/2}\mathrm{d}r = 2[2\mathcal{M}(r-2\mathcal{M})]^{1/2}\,, \tag{17}$$

with $z(2\mathcal{M}) = 0$. Figure 16 shows the embedding diagram with the electric field lines of the particle; it is in the curved space that the lines of force intersect the event horizon orthogonally (at which the field is strictly radial).

We note that also in this topic there is an open problem still to be resolved: we have also reconsidered [38] the possibility of examining the problem not just of a point particle, but of a conducting sphere in the field of a Schwarzschild black hole, as done by Fermi in the case of a uniform gravitational field. This problem is not yet solved, although using important results obtained by Leaute and Linet [39], we have been able to confirm the Fermi solution at least in the neighborhood of the test particle in Schwarzschild, where in a first approximation the gravitational field can be considered uniform.

5. – On the "electric Meissner effect"

My curiosity of how to justify the fact that Johnny did not succeed in guessing *a priori* the lines of force near a Schwarzschild black hole still intrigued me a few years ago. I decided to look into the matter of a test particle near a Reissner-Nordström spacetime, motivated by results obtained in the mean time by Bicak and coworkers [40] for


CHARGE CLOSE
FIELD = (REISSNER NORDSTROM) + DIPOLE OF REAL CHARGE PLUS IMAGE CHARGE)

(a) (b)

Fig. 16. – Embedding diagram with the electric field lines of the particle shown in fig. 15. (a) is taken from Wheeler's notebook.

a magnetic dipole in the field of an extreme Reissner-Nordström and an extreme Kerr solution. My approach following the "Wheeler theorem number 1" was that in an extreme Reissner-Nordström solution with $Q = \mathcal{M}$ no induced charge could exist, the reason being that any induced charge would make a part of the horizon surface overcritical and would generate a naked singularity. Instead of such catastrophic behaviour I was convinced that nature would have found the way to solve this paradox by not having lines of force crossing the horizon in the $Q = \mathcal{M}$ case. I was also motivated in this thinking by an instant disagreement I felt reading a very publicized article by Parikh and Wilczek [41], where they simply extrapolated the results of a test particle near a Schwarzschild black hole to the case of a Reissner-Nordström metric without understanding the existence of this very profound underlying difference between the two cases.

I then proceeded with D. Bini and A. Geralico [38] to study the set of equations for a test particle in a Reissner-Nordström space-time

$$\text{(BGR1)} \qquad \mathrm{d}s^2 = -f(r)\mathrm{d}t^2 + f(r)^{-1}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2)\,,$$

where $f(r) = 1 - 2\mathcal{M}/r + Q^2/r^2$, with associated electromagnetic field

$$\text{(18)} \qquad F_{\mathrm{RN}} = -\frac{Q}{r^2}\mathrm{d}t \wedge \mathrm{d}r\ .$$

The horizon radii are $r_\pm = \mathcal{M} \pm \sqrt{\mathcal{M}^2 - Q^2} = \mathcal{M} \pm \Gamma$. Maxwell's equations (F2) reduce to the following equation for the electrostatic potential V:

$$\text{(BGR5)} \qquad (r^2V_{,r})_{,r} + \frac{f(r)^{-1}}{\sin\theta}\left[(\sin\theta V_{,\theta})_{,\theta} + \frac{1}{\sin\theta}(V_{,\phi})_{,\phi}\right] = -4\pi r^2 J^0\,.$$

The solution corresponding to a charge q placed at the point $r = b$ on the polar axis $\theta = 0$ (with the same current density as (HR4)) has been derived by Leaute and Linet [42] both as a multipole expansion analogous to (HR6) with

$$(19)\quad f_l(r) = -\frac{(2l+1)!}{2^l(l+1)!l!\Gamma^{(l+1)}}\frac{(r-r_+)(r-r_-)}{r}\frac{\mathrm{d}Q_l}{\mathrm{d}r}\left[\frac{r-\mathcal{M}}{\Gamma}\right] \qquad l=0,1,2,...$$

$$g_l(r) = \begin{cases} 1 & l=0 \\ \dfrac{2^l l!(l-1)!\Gamma^l}{(2l)!}\dfrac{(r-r_+)(r-r_-)}{r}\dfrac{\mathrm{d}P_l}{\mathrm{d}r}\left[\dfrac{r-\mathcal{M}}{\Gamma}\right] & l=1,2,... \end{cases}$$

and in the closed form

$$(20)\qquad V = \frac{q}{br}\frac{(r-\mathcal{M})(b-\mathcal{M})-\Gamma^2\cos\theta}{D_{\mathrm{RN}}} + \frac{q\mathcal{M}}{br}\,,$$

with

$$(21)\qquad D_{\mathrm{RN}} = [(r-\mathcal{M})^2+(b-\mathcal{M})^2-2(r-\mathcal{M})(b-\mathcal{M})\cos\theta-\Gamma^2\sin^2\theta]^{1/2}\,.$$

We also generalized to the Reissner-Nordström case the discussion of the lines of force and associated properties of the horizon presented above for the Schwarzschild case. The induced charge density on the horizon is easily evaluated

$$(\mathrm{BGR7})\qquad \sigma^{\mathrm{H}}_{\mathrm{RN}}(\theta) = \frac{q}{4\pi b}\frac{[\Gamma(1+\cos^2\theta)-2(b-\mathcal{M})\cos\theta]\Gamma}{[b-\mathcal{M}-\Gamma\cos\theta]^2(\Gamma+\mathcal{M})}\,.$$

Note that $\sigma^{\mathrm{H}}_{\mathrm{RN}}(\theta)$ becomes identically zero in the extremely charged case where $\Gamma = 0$. As the hole becomes extreme, an effect analogous to the “magnetic Meissner effect” in the presence of superconductors arises for the electric field, with the electric field lines of the test charge being forced outside the outer horizon (see fig. 17). But this time the effect is not on a magnetic field, but is on the electric field, and it is not due to a superconducting sphere, but to the space-time around an extreme Reissner-Nordström black hole.

6. – On Zerilli’s solution

It must be emphasized that this is only a preliminary result. This behavior must be confirmed by integrating the more general set of equations describing the full Einstein-Maxwell perturbation equations introduced by Zerilli [43]

$$(22)\qquad \tilde{G}_{\mu\nu} = 8\pi\left(T^{\mathrm{part}}_{\mu\nu}+\tilde{T}_{\mu\nu}\right),$$

$$\tilde{F}^{\mu\nu}{}_{;\,\nu} = 4\pi j^{\mu}_{\mathrm{part}},\quad {}^*\tilde{F}^{\alpha\beta}{}_{;\beta} = 0\,,$$

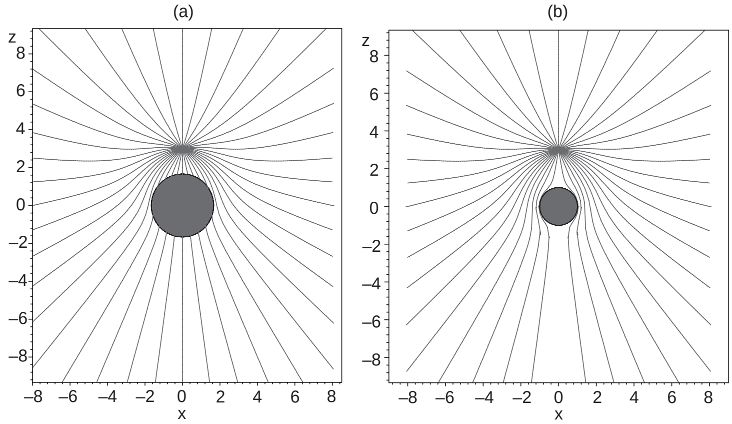


Fig. 17. – (a) shows the behavior of the lines of force of the test field alone with the charged particle at rest on the vertical axis $\theta = 0$ at $r = b$ with $b/\mathcal{M} = 3$, for $Q/\mathcal{M} = 3/4$. (b) corresponds to the extreme case, with the lines forced outside the (outer) horizon, the particle position being the same as in (a). The black circle represents the black-hole horizon.

where the quantities denoted by the tilde refer to the total electromagnetic and gravitational fields, at the first order of the perturbations, *i.e.*

$$\begin{aligned} \tilde{g}_{\mu\nu} &= g_{\mu\nu} + h_{\mu\nu}\,,\\ \tilde{F}_{\mu\nu} &= F_{\mu\nu} + f_{\mu\nu}\,,\\ \tilde{T}_{\mu\nu} &= \frac{1}{4\pi}\left[\tilde{g}^{\rho\sigma}\tilde{F}_{\rho\mu}\tilde{F}_{\sigma\nu} - \frac{1}{4}\tilde{g}_{\mu\nu}\tilde{F}_{\rho\sigma}\tilde{F}^{\rho\sigma}\right]\,,\\ \tilde{G}_{\mu\nu} &= \tilde{R}_{\mu\nu} - \frac{1}{2}\tilde{g}_{\mu\nu}\tilde{R}\,, \end{aligned} \tag{23}$$

where $T^{\text{part}}_{\mu\nu}$ and j^{μ}_{part} are, respectively, the stress-energy tensor and the 4-current associated with a particle of mass m and charge q. The corresponding quantities without the tilde refer to the background Reissner-Nordström metric (BGR1) and its associated electromagnetic field (18).

For a point charge of mass m and charge q at rest at the point $r = b$ on the polar axis $\theta = 0$, the only nonvanishing components of the stress-energy tensor and of the current density are given by

$$\begin{aligned} j^0_{\text{part}} &= \frac{q}{2\pi b^2}\delta\left(r-b\right)\delta\left(\cos\theta - 1\right)\,,\\ T^{\text{part}}_{00} &= \frac{m}{2\pi b^2}f(b)^{3/2}\delta\left(r-b\right)\delta\left(\cos\theta-1\right)\,. \end{aligned} \tag{24}$$

The perturbation equations are obtained from the system (22), keeping terms to first order. That has been done [38] and the existence of the “electric Meissner effect” has been confirmed. There are also some very exciting new results on the two-body solution in a Reissner-Nordström geometry, which we will discuss in the near future.

7. – Conclusions

The examples we have given well illustrate the caution we should apply in stating that black holes behave as perfect conductors and a special care should be used in establishing correspondence and analogies between classical physics and general relativistic regimes. See, *e.g.*, the statement I did in the book in honor of the festschrift of Hagen Kleinert [44]: "The analogies between classical regimes and general relativistic regimes have been at times helpful in giving the opportunity to glance on the enormous richness of the new physical processes contained in Einstein's theory of space-time structure. In some cases they have allowed to reach new knowledge and formalize new physical laws [...]. Such analogies have also dramatically evidenced the enormous differences in depth and physical complexity between the classical physics and general relativistic effects. The case of extraction of rotational energy from a neutron star and a rotating black hole are a good example. In no way an analogy based on classical physics can be enforced on general relativistic regimes. Such an analogy is too constraining and the relativistic theory shows systematically a wealth of novel physical circumstances and conceptual subtleties, unreachable within a classical theory. The analogies in the classical electrodynamics we just outlined are good examples."

It is very interesting that the combined Einstein-Maxwell equations still offer new challenges leading to unexplored physical phenomena. These results offer the possibility of reaching a better understanding of the solutions of both the Einstein and Einstein-Maxwell equations.

In both these topics M.me Choquet-Bruhat has made profound contributions and it is with great pleasure that I present these results to her in honor of her eightieth birthday. I am also very happy to share this celebration by recalling two very good friends of ours, Zel'dovich and Wheeler, both companions with us in the search for a deeper meaning of Einstein's great theory. Last, but not least, after all Johnny was right!

REFERENCES

[1] Gurzadyan V. and Ruffini R. (Editors), *Fermi and Astrophysics* (World Scientific, Singapore) 2005.

[2] Strong I. B., in *Neutron Stars, Black Holes and Binary X-Ray Sources*, edited by Gursky H. and Ruffini R. (Reidel D. Publishing Company, Utrecht) 1975.

[3] Paciesas W. S. *et al.*, *Astrophys. J. Suppl.*, **122** (1999) 465.

[4] Kouveliotou C. *et al.*, *Astrophys. J.*, **413** (1993) L101.

[5] Tavani M., *Astrophys. J.*, **497** (1998) L21.

[6] Costa E. *et al.*, *Nature*, **387** (1997) 783.

[7] Damour T. and Ruffini R., *Phys. Rev. Lett.*, **35** (1975 463).

[8] Giacconi R. and Ruffini R. *Physics and Astrophysics of Neutron Stars and Black Holes* (North-Holland, Amsterdam) 1978.

[9] Kerr R. P., *Phys. Rev. Lett.*, **11** (1963) 237.

[10] Newman E. T., Couch E., Chinnapared K., Exton A., Prakash A. and Torrence R., *J. Math. Phys.*, **6** (1965) 918.

[11] Carter B., *Phys. Rev.*, **174** (1968) 1559.

[12] Landau L. D. and Lifshitz E. M., *The Classical Theory of Fields*, fourth revised edition (Butterworth-Heinemann, Oxford) 2003, p. 352.

[13] Ruffini R., *On the energetics of Black Holes*, in *Black Holes - Les astres occlus*, edited by C. and B. S. De Witt (Gordon and Breach, New York)1973.

[14] Rees M., Ruffini R. and Wheeler J. A., *Black Holes, Gravitational Waves and Cosmology* (Gordon and Breach, New York) 1974.

[15] Ruffini R. and Wheeler J. A., *Phys. Today*, **24** (1971) 30.

[16] Carter B., in *Kerr Fest* (Cambridge University Press) in press.

[17] Bini D., Cherubini C., Jantzen R. T. and Ruffini R., *Prog. Theor. Phys.*, **107** (2002) 967.

[18] Born M., *Proc. R. Soc. London, Ser. A*, **143** (1933) 410.

[19] Heisenberg W. and Euler H., *Z. Phys.*, **98** (1935) 714.

[20] Schwinger J., *Phys. Rev.*, **82** (1951) 664.

[21] Ruffini R., Vitagliano L. and Xue S.-S., *Phys. Rep.*, in preparation.

[22] Christodoulou D. and Ruffini R., *Phys. Rev. D*, **4** (1971) 3552.

[23] Cavallo G. and Rees M. J., *Mon. Not. R. Astron. Soc.*, **183** (1978) 359.

[24] Cavallo G. and Horstman H. M., *Astrophys. Space Sci.*, **75** (1981) 117.

[25] Horstman H. M. and Cavallo G., *Astron. Astrophys.*, **122** (1983) 119.

[26] Ruffini R., *Beyond the Critical Mass: The Dyadosphere of Black Holes*, in *Black Holes and High Energy Astrophysics, Proceedings of the 49th Yamada Conference*, edited by Sato H. and Sugiyama N. (Universal Press Ac., Tokyo) 1998.

[27] Preparata G., Ruffini R. and Xue S.-S., *Astrom. Astrophys.*, **338** (1998) L87.

[28] Ruffini R., Vitagliano L. and Xue S.-S., *Phys. Lett. B*, **559** (2003) 12.

[29] Cherubini C., Ruffini R. and Vitagliano L., *Phys. Lett. B*, **545**, 226 (2002).

[30] Ruffini R. and Vitagliano L., *Phys. Lett. B*, **545** (2002) 233.

[31] Ruffini R. and Vitagliano L., *Int. J. Mod. Phys. D*, **12** (2003) 121.

[32] Ruffini R., Vitagliano L. and Xue S.-S., *Phys. Lett. B*, **573** (2003) 33.

[33] Ruffini R., Fraschetti F., Vitagliano L. and Xue S.-S., to be published in *Int. J. Mod. Phys. D.*

[34] Fermi E., *Nuovo Cimento*, **22** (1921) 176.

[35] Levi Civita T., *Rend. Acc. Lincei* **27** (1918) 3.

[36] Hanni R. and Ruffini R., *Phys. Rev. D*, **8** (1973) 3259.

[37] Linet B., *J. Phys. A: Math. Gen.*, **9** (1976) 7.

[38] Bini D., Geralico A. and Ruffini R., in preparation.

[39] Leaute B. and Linet B., *Int. J. Theor. Phys.*, **22** (1983) 67.

[40] Bicak J. and Ledvinka T., *Electromagnetic fields around black holes and Meissner effect*, in *Proceedings of the III ICRA Network Workshop and IV Italo-Korean Meeting on Electrodynamics and Magentohydrodynamics around Black Holes, Rome-Pescara, July 12-24, 1999*, edited by Ruffini R., *Nuovo Cimento B*, **115** (2000) 739.

[41] Parikh M. K. and Wilczek F., *Phys. Rev. Lett.*, **85** (2000) 5042.

[42] Leaute B. and Linet B., *Phys. Lett. A*, **58** (1976) 5.

[43] Zerilli F. J., *Phys. Rev. D*, **9** (1974) 860.

[44] Ruffini R., *Analogies, new paradigms and observational data as growing factors of Relativistic Astrophysics*, in *Fluctuating Paths and Fields - Festschrift Dedicated to Hagen Kleinert on the Occasion of His 60th Birthday*, edited by Janke W., Pelster A., Schmidt H.-J. and Bachmann M. (World Scientific, Singapore) 2001, p. 771.

Appendix B

Selected papers reprinted from *Il Nuovo Cimento*, Vol. 117B, Nos. 9–11, 1992

This Appendix contains a selection of the articles from the proceedings the meeting "Fermi and Astrophysics" organized at the University of Rome "La Sapienza" and at the ICRANet Center in Pescara October 3–6, 2001 and published in *Il Nuovo Cimento B* **117**, Nos. 9–11. The meeting was focused on the influence of Fermi on astrophysics and general relativity: his activities related to these topics were clustered at the beginning and end of his scientific career. These articles, selected because of their direct commentary on articles by Fermi or related applications of his ideas expressed in those articles, are presented in alphabetical order of their first authors.

Susan Ames discusses the historical background of Fermi's work on cosmic rays, along with current problems and further prospects for the physics of cosmic rays. In particular she points out how the frequently discussed ultra-high cosmic rays cannot be accelerated by the Fermi mechanism. Equipartition between the energy of matter and that of cosmic rays was among the initial points made by Fermi, and in that context Ames mentions also the role of the cosmic microwave background radiation.

Donato Bini and Robert Jantzen give a summary of Fermi's discussion of what we now call Fermi coordinates and Fermi transport with a historical update including Walker's contribution which led to the terminology of "Fermi-Walker transport." This article explicitly estimates the various relativistic contributions to the Fermi-Walker transport for vectors around circular orbits in black hole spacetimes and in their Minkowski limit.

Dino Boccaletti comments on the two papers which resulted from the collaboration of Fermi with Chandrasehkar (see papers 261, 262 of Chapter 4). The first paper is devoted to the study of light dispersion in the polarization plane and using the effect to derive the galactic magnetic field. The second paper contains the generalization of the virial theorem in the presence of a magnetic field. The commentary notes that Fermi was the first scientist to draw attention to the possible existence of a galactic magnetic field.

The review of Andrea Carati, Luigi Galgani, Antonio Ponno and Antonio Giorgilli is devoted to the equipartition problem in the Fermi-Pasta-Ulam paradox both in classical and quantum mechanics. Equipartition is discussed starting from Planck's work and Poincaré's theorem. Numerical results on the dependence of the existence of equipartition and the corresponding time scales on a certain critical energy are mentioned.

Piero Cipriani reviews the work of Fermi in the field of classical analytical mechanics. After a short historical introduction, he emphasizes some aspects of geometrical methods of the description of dynamics and the theory of stochastic differential equations. Interesting recollections on Fermi are quoted.

John G. Kirk reviews the Fermi acceleration mechanism in the context of galactic nuclei and gamma ray bursts, i.e., in processes involving relativistic motion. Diffusive and non-diffusive versions of Fermi's stochastic acceleration are considered, including those predicting a softer spectrum of accelerated particles. The appearance of anisotropy in the accelerated particles with increasing gamma factor is discussed for various astrophysical situations.

Stefano Ruffo reviews evidence for long relaxation time scales in Hamiltonian systems, and shows how complex and diverse is the dynamics of long-range systems. The 'quasi-states' of Fermi-Pasta-Ulam are discussed particularly in the context of two theoretical approaches developed by the author and collaborators, one based on the Vlasov-Poisson equation, and the other based on the averaging of fast oscillations.

Costantino Sigismondi and Francesca Maiolino review an early work by Fermi completed June 20, 1922, the year of his habilitation thesis on statistics at the Scuola Normale Superiore of Pisa, with an application to the case of comets. Fermi studied this case with a coplanar orbit to the one of Jupiter, neglecting the influence of other planets. The probability of ejection of the comet from the solar system (a parabolic or hyperbolic orbit) after interaction with Jupiter is calculated, as well as the probability of an impact with Jupiter. They apply Fermi's results to the case of the Earth in order to recover the time rate of collision of comets with our planet, which reliably produced the extinction of the dinosaurs. In this context the properties of the Oort cloud are discussed as well.

Costantino Sigismondi and Angelo Mastroianni recall that approximately in the same period Fermi studied the formation of X-ray images and presented his first experimental work as a dissertation at the University of Pisa in the spring of 1922. The need for Fermi to make an experimental essay was made mandatory since at that time theoretical physics was not yet considered sufficient to have independent validity. Although his seminal ideas are not among the bibliographical sources investigated by Riccardo Giacconi and Bruno Rossi (1960) when they proposed a telescope using X-rays, Fermi's thesis was the most complete study of X-ray physics in his time. Fermi used the technique of 'mandrels' to form optical surfaces. He anticipated the technique used for the mirrors of the Exosat, Beppo-SAX, Jet-X and XMM-Newton telescopes, a technique which is now a mainstay of optical manufacturing.

Alexei Yu. Smirnov reviews the neutrino flavor transformations in matter, as one of the authors of the original theoretical predictions and related observable effects. In particular, the Sudbury Neutrino Observatory results provide strong evidence of the neutrino flavor conversion. Neutrino conversion is discussed also in the context of supernova neutrinos and the corresponding predictions for the fluxes and energies at the Earth, including the role of the Earth matter effect. The author shows that the data of SN1987 can also be explained by the neutrino oscillations in the matter of Earth as conversions of muon and tau antineutrinos.

George M. Zaslavsky reviews the Fermi-Pasta Ulam problem with an attempt to find the transition from regular to chaotic dynamics. The Fermi acceleration mechanism is considered as a precursor of the Fermi-Pasta-Ulam problem. The Kepler map introduced by Roald Sagdeev and George Zaslavsky and several other problems are considered, demonstrating the role of the Fermi-Pasta-Ulam work in the discretization methods of differential equations and in the study of chaotic systems when the Lyapunov exponent method is not efficient.

B.1 S. Ames: Comments on Fermi's original papers on cosmic ray acceleration

S. Ames: "Comments on Fermi's original papers on cosmic ray acceleration," *Nuovo Cimento B* **117**, 965 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

Comments on Fermi's original papers on cosmic-ray acceleration(*)

S. Ames(**)

Radioastronomy Institute - Auf dem Hügel 71, D-53121 Bonn, Germany

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — In his two 1949 papers on cosmic rays, Fermi introduced the concepts which form the basis of most theories of cosmic-ray acceleration today: magnetic trapping; repeated, small stochastic gains; energy derived fron the large-scale bulk kinetic energy of interstellar plasmas. I consider the historical context in which these concepts were proposed, and compare the questions which Fermi regarded as unresolved with those which we now regard as unresolved.

PACS 95.85.Ry – Neutrino, muon, pion and other elementary particles; cosmic rays.
PACS 98.70.Sa – Cosmic rays (including sources, origin, acceleration, and interactions).
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

In 1949 Fermi published two papers [1,2] which remain today the basis of most theories of the origin of cosmic rays. I should like to consider the influences which may have led him to take the approach he did to tackle the problem, and to reflect on some of his concerns which are still with us today.

Until recently theories of the origin of cosmic rays have created them by invoking a mechanism to accelerate particles to higher energy. Most which remain viable today are based on Fermi's original ideas. As the observational data has improved on ultrahigh-energy cosmic rays with energy above 10^{19} eV, it is, however, questionable whether any acceleration theory can explain these particles. Alternative approaches invoke as yet unknown, decaying GUT particles, created with the GUT energy of about 10^{25} eV, which then cascade down in energy.

(*) Paper presented at the IX ICRA Network Workshop "Fermi and Astrophysics" (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.

(**) E-mail: sames@astro.uni-bonn.de

2. – Historical perspectives

2˙1. *Cosmic rays in 1949*. – Although a mysterious "radiation" with an extraterrestrial origin had been known for about 50 years, very little progress had been made determining either its nature or origin. Their cosmic origin was demonstrated in 1912 by Viktor Hess, using balloon flights to show that the penetrating, ubiquitous, ionizing radiation increases in intensity with altitude. It was not until the 1930's, with increased understanding of nuclear physics, that the "radiation" was recognized to be charged particles. Very little else was known about it, neither what the particles were, nor their energy, nor where they came from, nor how they acquired that energy. The particles were supposed to be subatomic, probably protons. The energy was supposed to be high, considerably suprathermal, since it was penetrating.

Although extensive air showers had been discovered in 1938 by Pierre Auger [3, 4], progress in interpreting them had been hindered by both incomplete theory and inadequate technology. In order to derive the composition and energy of the particle which had caused the shower, a detailed calculation of the nuclear reactions must be made from the original interaction of the particle impinging on the top of the atmosphere, through all the cascades, down to the muons measured on the ground. One single particle at the top of the atmosphere produces millions of particles in the ground. The highest energy particles seen, in fact, produce over 10^9 particles. Clearly not only was more detailed nuclear theory required, but also high-speed computers. Although cosmic rays are a piece of the cosmos which has arrived on our doorstep, and hence local measurements were possible, nevertheless deductions about their properties were inconclusive.

Considering the cosmic-ray's life in the cosmos rather than its local death, one would like to know their spatial extent. Were they confined to the Earth's vicinity, or to the solar system, or did they extend throughout the Galaxy? There was no evidence to favor either of these alternatives. An argument which was invoked to favor more local confinement was that the energy requirements would be enormous otherwise. If the cosmic rays filled interstellar space, a very large amount of energy would have to be channeled from some other source into the cosmic rays. One would not only have to explain how the energy was channeled efficiently, but also find such a large energy reservoir.

3. – Fermi's approach

3˙1. *Influences from other disciplines*. – At Chicago Fermi would have heard from the Yerkes Observatory astronomers of Adams' work on interstellar absorption lines. Although Adams had recently retired as director of Mt. Wilson Observatory, a post he had held for two decades, he had studied and worked at Yerkes in his youth, and would have maintained his ties there. Adams' high dispersion stellar spectra [5] allowed him to identify weak, narrow, absoption lines of molecules at different radial velocites than that of the star. These are due to cold interstellar clouds in the line of sight between us and the star. The concept of cold clouds between the stars, with molecules in them, moving with random velocites of the order of 15 km/s, was new. Although Adams' observations were of only nearby stars, since the stars had to be bright in order to obtain high dispersion spectra, if one were to suppose most of the Galaxy were similar, their kinetic energy represents a very significant energy reservoir.

In 1948 Alfvén visited Chicago. Many of his ideas of magnetic fields in plasmas, and throughout the Universe, were not received well by the physics community at large at that time, but Fermi found his arguments convincing. In particular Alfvén had proposed a decade earlier that there was a large-scale magnetic field throughout the Galaxy, but since the general picture of the Galaxy at that time was of a vacuum between the stars, it was not seen how the electric currents could be supported to maintain such a field. But did not Adams' interstellar absorption lines indicate all was not a vacuum between the stars?

Fermi was familiar with what is sometimes called the Stoermer [6] problem, the motion of a charged particle in a magnetic dipole-field. Scandinavian physicists had been particularly interested in this problem as it applies to the aurora. In the appendix of his book which came out a year later, Fermi [7] considered the magnetic cut-off of charged particles in the Earth's magnetic field.

The ingredients were there: the theory of trapped charged particles moving between magnetic mirrors; Alfvén's conviction that not only was there a large-scale magnetic field in the Galaxy, but also inside interstellar clouds; and observational evidence for clouds moving with random velocities, occupying a significant volume of the Galaxy. One further ingredient for which he had no evidence but which he required, he had to postulate, that of a population of suprathermal seed partcles.

3˙2. *Order-of-magnitude estimate.* – Fermi's approach to the problem clearly reflects the nuclear physicist who thinks in terms of particle decay and lifetimes, rather than the plasma physicist.

Consider a particle of mass m trapped between two moving mirrors. The energy gain per collision is

$$\delta w = B^2 w,$$

where $B = V/c$, and V is the velocity of the mirrors.

After N collisions,

$$w = mc^2 \exp[B^2 N].$$

For losses due to nuclear collisions with a mean free path Λ, a collision loss time T can be defined by

$$\Lambda = Tc.$$

Let the time between collisions with the walls be τ. Then, if the only losses are nuclear collisions, the energy distribution is

$$\pi(w)\mathrm{d}w = (\tau/B^2T)(mc^2)^{\tau/B^2T}\mathrm{d}w/w^{1+\tau/B^2T}.$$

The success of this model is that it produces an inverse power law spectrum. Even then it appeared that was the spectrum which one should try to explain. Although at that time measurements were available only up to particle energies of about 10^{12} eV, the power law in fact extends up to above 10^{19} eV.

Figure 1 [8] shows a recent compilation of data over the range of 13 orders of magnitude in energy. The deviation at low energy is fairly well understood, is a local effect, and is due to the effect of solar modulation. The incoming galactic cosmic rays scatter off magnetic irregularities in the solar wind. The other two features marked have been dubbed anthropomorphically the knee and the ankle. The knee probably results from particles leaking out of the Galaxy. It occurs around the energy for which protons

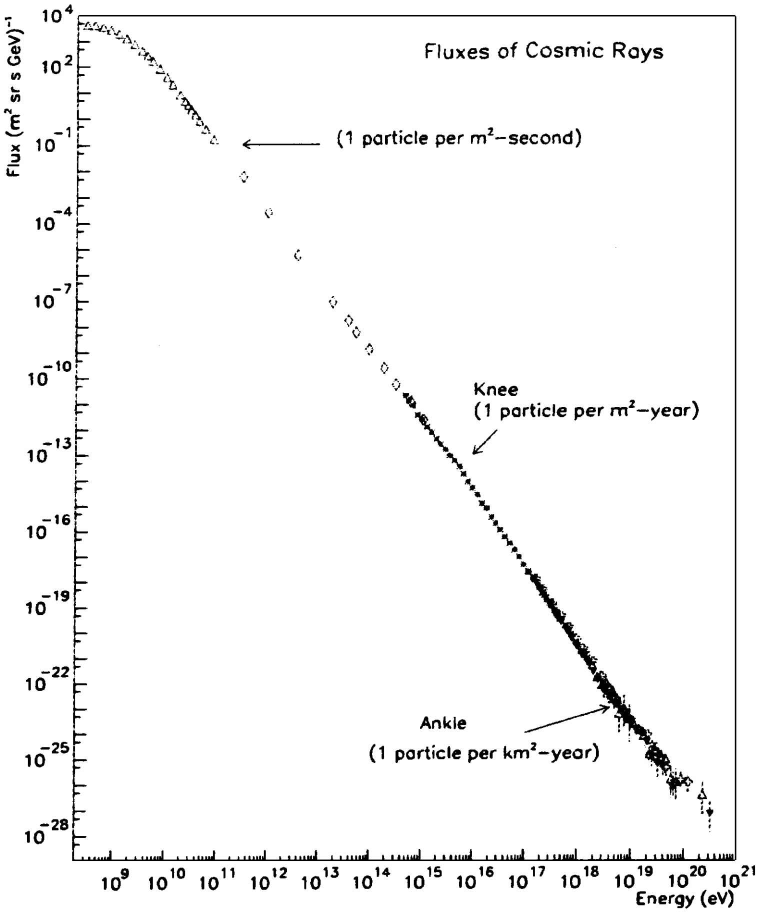


Fig. 1. – Differential energy spectrum of cosmic rays.

can no longer be confined by the magnetic field of the Galaxy. The ankle is less well understood, and may be where a new componant, the ultrahigh-energy cosmic rays, is starting to appear, produced by an entirely different, as yet unknown, mechanism.

4. – Fermi's unfounded concerns

4˙1. *Seed particles.* – Fermi's mechanism requires suprathermal seed particles. They must have energies above about 200 MeV in order to avoid that ionization losses, which are proportional to w^{-2}, are greater than the energy gains. Fermi considered this a major problem for his theory. He proposed a "chain reaction" by spallation, but clearly this can produce only light element cosmic rays, not heavies. If heavy elements were found in the cosmic rays, he felt that would pose a major problem for his model.

At the time solar flares and coronal mass ejections were not known. Figure 2 [9] shows energetic ions measured *in situ* in the solar wind. The sun clearly knows how to inject suprathermal particles into interstellar space, including heavies.

The mass spectrum of the injected particles extends all the way up to iron. Fur-

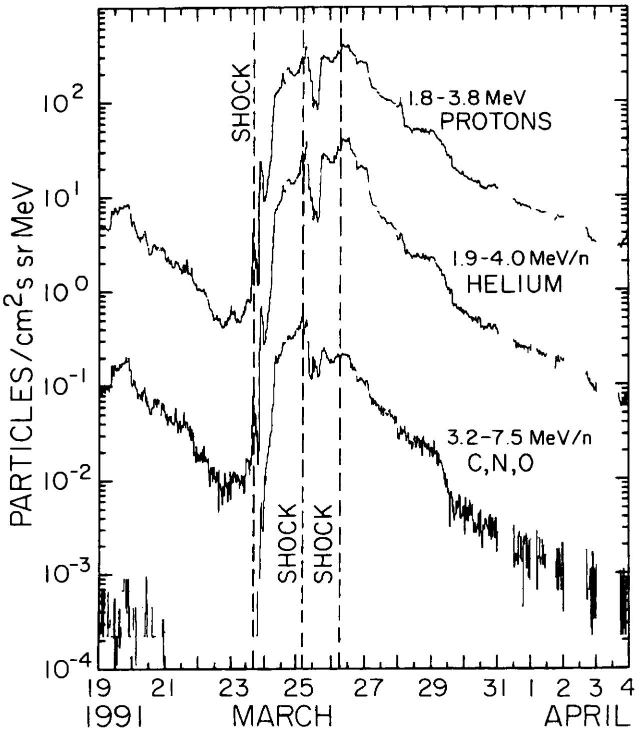


Fig. 2. – Energetic ions measured onboard the Ulysses spacecraft during the period of intense flare activity in March 1991.

thermore the characteristics of these suprathermal particles do not seem to change substantially between about 1 MeV and 100 MeV. Figure 3 [10] shows, for example, the ionic charge of ≈ 1 MeV and ≈ 100 MeV particles ejected from the Sun for the heavy elements up to iron.

Therefore, we now know that in principle the seed particles Fermi needed, and had to postulate in the absence of observational data, are indeed being injected into interstellar space. Their energy is probably derived from magnetic field energy in stellar-type objects. Detailed calculations by Decker and Vlahos [11] are able to produce ions in the range 10–100 MeV in solar flares from 100 keV particles, which can be obtained either from the tail of the distribution of particles directly heated in the flare, or from various direct "prompt" electromagnetic acceleration means in the flare. Providing Fermi with the 200 MeV seed ions he needs does not seem to be a problem in principle, either observationally, nor theoretically.

4˙2. *Heavy nuclei.* – At that time very little was known about the composition of the cosmic rays. Fermi was aware that there was some indication that they were not all protons, and he viewed this as a potential problem. Not only would the injection of heavy seed particles pose a problem, but indeed their entire spectrum. Since loss rates for heavies would be different than for protons, the spectrum would be different. He expected the spectrum to be much steeper for heavies than for protons, so there should be very few at the higher energies.

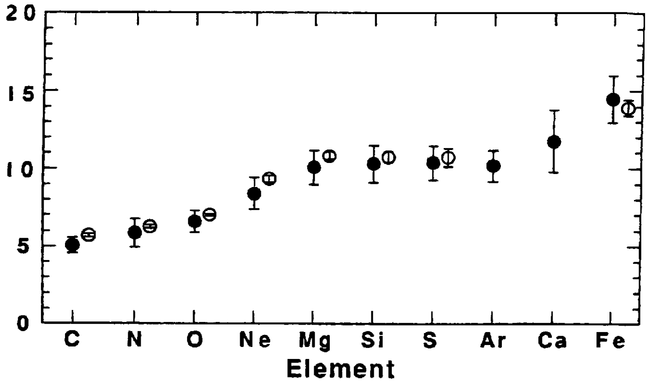


Fig. 3. – Mean ionic charge of solar energetic particles in the energy range 10–100 MeV/nuc (solid symbols) for the events Oct. 31-Nov. 7, 1992. The open symbols show the energy range 0.3–3 MeV/nuc.

Modern measurements show that the heavy elements are well represented. Figure 4 [12] shows that for a wide range of heavy elements the abundance in galactic cosmic rays, extrapolated back to the source, is within an order of magnitude of solar. The details of the deviations from solar abundance are interpreted as clues to the environment in which the particles were accelerated.

Deriving the mass of the primary particle from the air shower data still involves

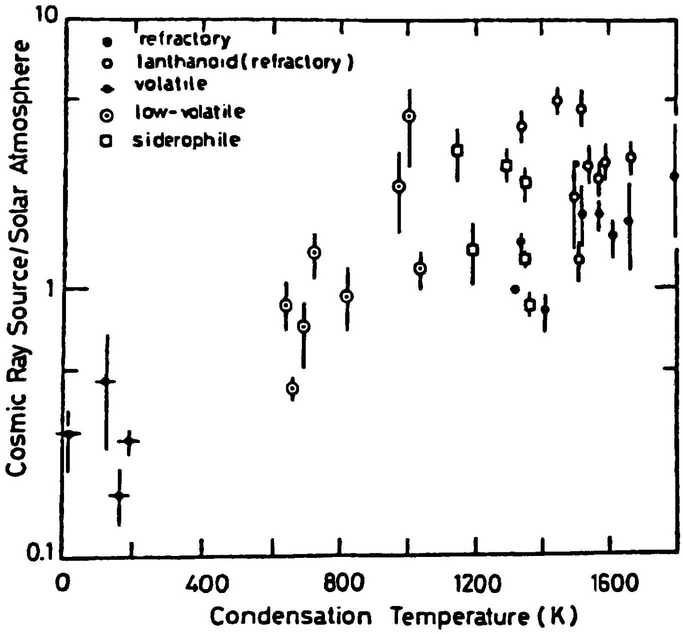


Fig. 4. – The chemical composition of galactic cosmic rays at their sources, as compared with that of the solar atmosphere, as a function of the condensation temperature of the elements.

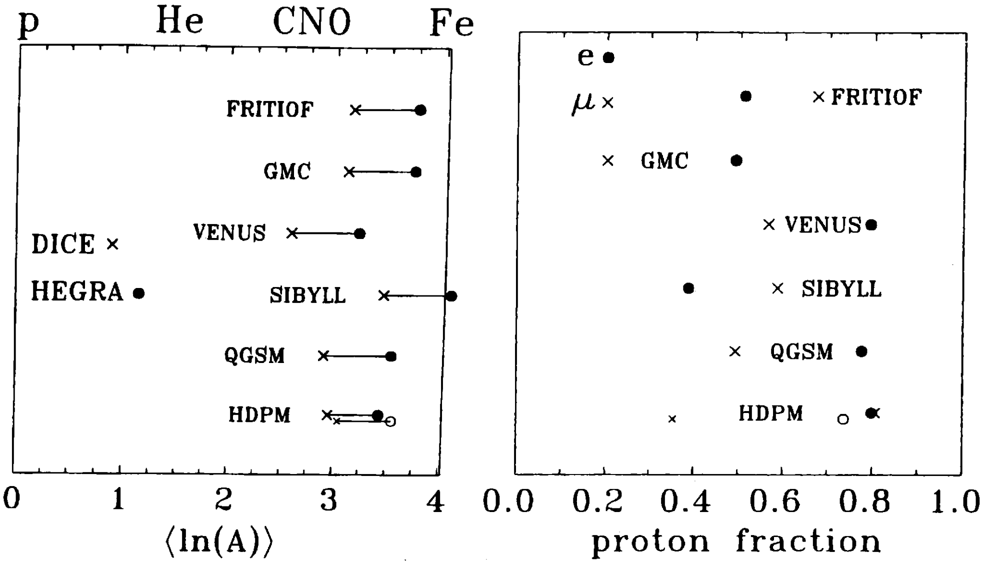


Fig. 5. – (a) Average logarithm of the primary particle mass at 10^{15} eV for different models of high-energy interactions. (b) Proportion of proton- and iron-induced showers in the event sample, obtained from different interaction models.

some uncertainty, both in the model used for the nuclear interactions, and in the details of the interactions themselves. Figure 5 [13] (a) shows the difference for two different experiments, DICE and HEGRA, using 6 different models of the nuclear interactions. The conclusion is that for primaries with energy of 10^{15} eV, a substantial fraction are heavier than CNO nuclei, whichever model or experimental data is used. Figure 5 [13] (b) demonstrates that all of the models agree that the primaries are neither all iron nor all protons.

4˙3. *Energy budget.* – The extent of the confinement region of cosmic rays was not known, and there were no observational constraints. They could have been confined to the region around the Earth, or to the solar system, or to the Galaxy. Fermi's model requires that they fill a large portion, if not all, of the Galaxy. He was concerned that the objection could be raised that this would require the total energy channeled into cosmic rays to be very large. He does not directly counter this objection, except to point out that, therefore, the acceleration mechanism must be efficient.

The argument has often been invoked when new phenomena are discovered in astronomy, that the total energy required would be too large unless the phenomenon is restricted spatially, temporaly, directionally, etc. It has usually proved to be wrong, demonstrating that the Universe is more extreme than most astronomers would like to think it is. The argument that the energy required would be unreasonably high was used when the "nebulae" were proposed to be extragalactic, when quasars were proposed to be at cosmological distances, and, more recently, when gamma-ray bursts were proposed to be at cosmological distances.

In Fermi's model, the cosmic rays draw their energy from the random large-scale kinetic motion of interstellar clouds. The astronomical observations of interstellar absorption lines suggested that this might be a very large energy source. The details of the magnetic mirror, however, depend on the existence of small-scale magnetic irregularities in the cloud, which presumably are indicative of turbulence. Energy dissipation in turbulence, as energy cascades from large eddies down to the smallest scale where it is dissipated in viscosity, is very efficient, probably more efficient than any mechanism to channel energy into cosmic rays could be, since the former is moving the state towards equilibrium and the latter is moving it away. So he postulated that magnetic fields could suppress small-scale turbulence. The energy in turbulence is stored in the largest eddies, but dissipated in the smallest. The problem of the energy balance is determined by the small-scale structure and physics, not the overall extent of the region. The total energy requirement was determined not by the extent of the confinement region *per se*, but by the physics of the dissipation on small scales.

5. – A nuclear physicist before a hydrodynamicist

Fermi's experience as a nuclear physicist influenced how he formulated the problem. Physicists and mathematicians try to reduce a new problem to one which they have solved before. The calculation of the cosmic-ray spectrum is treated as a problem in radioactive decay, and the creation of seed suprathermal particles is formulated as a nuclear chain reaction for which the question becomes whether or not it is self-sustaining. It was only later in collaboration with Chandrasekhar that he became a hydrodynamicist.

In 1949 the existence of interstellar galactic magnetic fields was still controversal. An interstellar magnetic field was necessary for Fermi's cosmic-ray acceleration model to work. The large-scale Galactic magnetic field proposed by Alfvén could also be used to confine the cosmic rays within the Galaxy, if the geometry proved suitable. But was there one? It was just then that some actual observational data became available which confirmed that there indeed was a large-scale galactic field, as Fermi was quick to recognize. In that year Hiltner [14] published measurements of the optical polarization of stars across the sky, and found large-scale patterns. This is attributed to scattering by non-spherical dust grains alligned in a magnetic field in the line of sight to the star. With the first concrete evidence that Alfvén (and he himself as well) was correct, he pursued other consequences of the large-scale interstellar field. Together with Chandrasekhar in 1953 [15] he devised two methods to estimate the field strength using hydrodynamics. One method attributes observed small-scale transverse motions to Alfvén's new waves, and the other estimates the magnetic pressure needed to satisfy hydrostatic equilibrium. The Chandrasekhar-Fermi method is still used today as one of the few ways to estimate interstellar large-scale magnetic-field strength. For a recent review of the Chandrasekhar-Fermi method applied to current-day observations, see Zweibel [16].

6. – Persistent concerns

6˙1. *Composition.* – As for Fermi and his model, the composition is also a crucial test for the current theories of ultrahigh energy (UHE) cosmic rays. Much of the current excitement in cosmic-ray studies focuses on the UHE cosmic rays above 10^{19} eV. Fermi's mechanism cannot explain them, and it is doubtful any modification of the theory would be able to explain them. In fact, they should not exist. Aside from the difficulty of finding a means of accelerating them to that energy, another problem enters for energies above about 5×10^{19} called the Greisen-Zatsepin-Kuz'min cut-off [17, 18]. At those energies they interact with the cosmic microwave background photons, and are destroyed by pion photodissociation. A particle with a relativistic γ of 10^{11} sees a very hard gamma-ray,

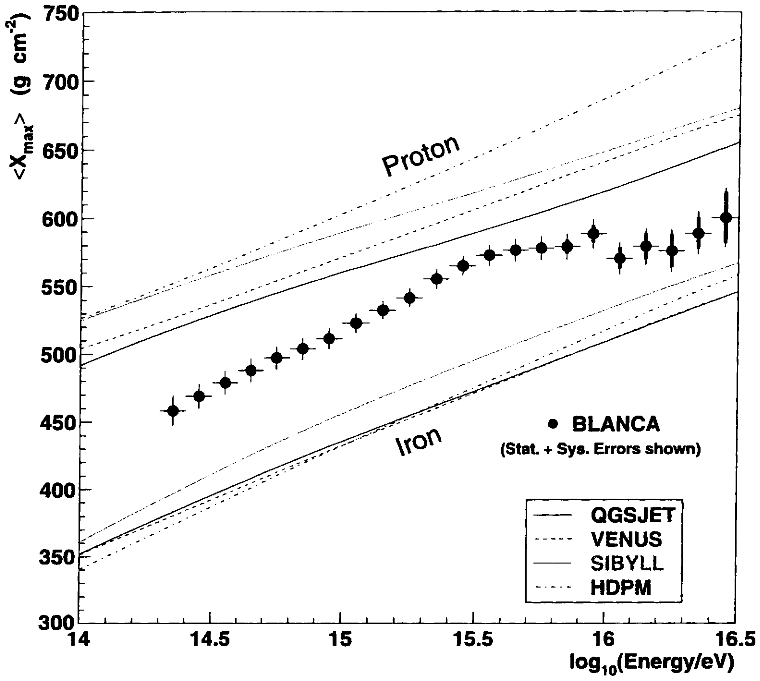


Fig. 6. – The mean depth of shower maximum $X_{\rm max}$ as a function of energy. The curves represent the values for pure proton and pure iron as predicted by different hadronic interaction models.

not a microwave photon. The source must be closer than about 100 Mpc, or it would not have survived. Since particles of such high energy are not deflected by the galactic magnetic field, they must point back to their source. There is, however, nothing in their direction. If it were only 100 Mpc from us and capable of producing 10^{20} eV particles, we should presumably see the photons it also produces.

The so-called "top-down" class of models, which produce particles with energy of the GUT energy of about 10^{25} eV from the decay of exotic supersymmetric particles or topological defects, could not conceivably produce anything heavier than a proton. If heavies are found, top-down exotic models would appear to be ruled out, and the UHE cosmic rays must have been accelerated from thermal matter.

Figure 6 [19] shows the comparison of the data from one air shower array with several different model calculations for the middle energies around the knee. As the energy increases to the highest energies considered by Fermi, the proton fraction increases slightly, as he predicted, although perhaps by considerably less than he would have expected. This is due to the higher losses for heavy nuclei caused by nuclear interactions during propagation through the interstellar medium. By 10^{16} eV, however, the fraction of iron has increased significantly. This is due to light particles escaping from the Galaxy, since the gyroradius in the Galactic magnetic field of a proton is much larger than that of an Fe nuclei. At the highest energies for which there are composition measurements, which correspond to the energies above which even Fe nuclei are not confined in a 3 μG field, there is a tantalizing indication that the proton fraction begins to increase again. Figure 7 [20] shows a similar plot for a different set of model calculations compared with

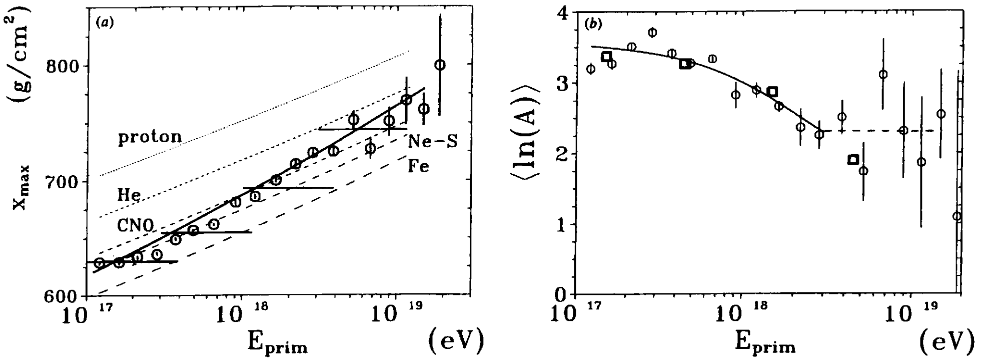


Fig. 7. – (a) Depth of $X_{\max}$ *vs.* the primary energy E_{prim}. (b) $\ln A$ *vs.* E_{prim} derived by two different methods. See original reference for details.

data from the Fly's Eye experiment at the highest energies. It may be that the proton fraction is, indeed, increasing with energy.

6˙2. *Trapping.* – The essential elements Fermi's model supplies are an energy source and particle trapping. The energy in cosmic rays is channeled from the kinetic energy of the random motion of magnetic irregularities, identified with cold interstellar clouds. Particle trapping is the new ingredient, prompted by his work on the Stoermer problem. One step, direct "prompt" acceleration processes, such as in the initial acceleration in a solar flare, are faster and the energy gain per step is much greater, but without a means of trapping the particle within the "accelerator" the total energy gains are modest, since the particle escapes the accelerator. The maximum energy attainable would be considerably below even the highest energies known at the time. Constructing a model which as a byproduct also trapped the particles, allowed him to reach the highest energies he wished. Since he considered only ionization losses, which become negligible for energies above about 200 MeV, he was not worried about the high-energy end of the distribution. He considered the only limitation to the maximum energy attainable was the energy equipartition value, although the process was slow, and it would take a very long time to reach. For the energies known at the time, this was probably so. Now that we know the energy spectrum continues many orders of magnitude above 10^{15} eV (fig. 1), trapping again becomes a problem, and particle leakage becomes the dominant loss process. Theoreticians who try to extend Fermi acceleration to high-energy cosmic rays are forced to resort to postulating intricate, and usually implausible, magnetic field configurations in order to trap the particles. A detailed analysis by Lagage and Cesarsky [21] of the non-linear effects on the magnetic field within the shock, and the resulting changes in the diffusion coefficient, lead to the conclusion that many estimates of the maximum energy attainable are one to two orders of magnitude too large. Even using the most favorable conditions and configurations, 10^{16} eV seems to be a robust upper bound to the maximum energy attainable.

6˙3. *The injection problem.* – The injection of seed particles is a problem which is still with us in the details. The relative abundances measured in cosmic rays do not correspond exactly, either in element nor isotope, with the thermal plasma of any known environment. Presumably this is due to selective acceleration during the injection process. Attempts to find simple correlations, such as with the first ionization potential, always result in several exceptions. Neither have we addressed the question of whether there are enough of them. That requires detailed study of the different galactic stellar populations, and how much each type would contribute.

7. – Self-regulating mechanisms and equipartition

The cosmic-ray energy density is of the same order of magnitude as several other energy densities in the Galaxy. The interstellar medium and the Galaxy as a whole are, however, many orders of magitude removed from equilibrium. The photon energy density is too weak by many orders of magnitude for its spectral distribution or "black-body temperature". Therefore, if two energy densities are found to be equal, one must find a mechanism which makes, and keeps, them equal. Fermi's model provides a natural relation between the random motions in the Galaxy, the dynamical temperature of the disk, and the cosmic-ray energy density. Equipartition is common in astrophysical systems, even if equilibrium is not.

7˙1. *Astrophysical coincidences.* – Fermi was intrigued by puzzles and conundrums, and astrophysics offers many. Fermi's model of cosmic-ray acceleration seeks to make the link

$$\epsilon_{\mathrm{cr}} \approx \epsilon_T,$$

where ϵ_{cr} is the energy density present in cosmic rays, and ϵ_T is the energy density present in large-scale random kinetic motions of gas in the disk. What about

$$\epsilon_{\mathrm{cr}} \approx \epsilon_T \approx \epsilon_*,$$

where ϵ_* is the energy density present in starlight?

To understand

$$\epsilon_{\mathrm{cr}} \approx \epsilon_*,$$

one needs to understand magnetohydrodynamics. The energy density in cosmic rays can influence the formation of interstellar clouds via the Parker instability [22], and the formation of clouds is the first step towards contraction to a protostar. The energy density in cosmic rays enhances the effect of the anisotropic pressure of the magnetic field until the field ballons up out of the disk, and cosmic rays can escape, as, at the same time, the thermal gas falls back and collects at the bottom of the curved field lines to form clouds.

In order for stars to form, the gas must cool. The cosmic rays influence the cooling of the clouds through their non-adiabatic effects within the cloud. The electrons released when a cosmic ray ionizes an ambient atom, then thermalizes and heats the gas. When the atom then recombines it releases a photon which cools the gas. This is a highly non-linear process and depends on the detailed conditions within each cloud, as well as its environment, but there are enough interdependences that the rough equipartition could plausibly be accounted for. The energy density in cosmic rays can influence not only the rate of star formation, but probably at least as important, the Initial Mass Function of the stars formed. The large-scale kinetic energy needed in Fermi's model, the random motion of the clouds, is probably contributed mainly by the higher mass stars, not only through supernovae explosions, and mass ejection at the late stages of stellar evolution, but also by the expansion of an ionization front into the surrounding gas clouds during its main sequence life. The rate of injection of the seed particles needed by Fermi's acceleration mechanism from stellar flares and magnetic activity also depends on the star formation rate and the initial mass function.

There is another "coincidence" which is much more puzzling, however.

$$\epsilon_{\rm cr} \approx \epsilon_{\rm CMB},$$

where $\epsilon_{\rm CMB}$ is the energy density in the cosmic microwave background. In most models of cosmic-ray propagation, there is a halo of cosmic rays around the Galaxy. Radio continuum observations of the synchrotron emission from energetic electrons show that some external galaxies do have a halo of energetic particles, at least of electrons. Diffuse gamma-ray emission in our Galaxy produced by cosmic rays colliding with ambient cold interstellar matter may also indicate a cosmic-ray halo. Such a halo is invoked by theoreticians as a storage ring, where the cosmic rays can propagate with low losses to explain certain isotopic ratios measured in secondary particles produced by spallation in the interstellar medium. Could the extent of the cosmic ray halo be dependent on cosmological epoch?

REFERENCES

[1] Fermi E., *Phys. Rev.*, **75** (1949)1169.

[2] Fermi E., *Nuovo Cimento Suppl.*, Vol. **VI**, no. 3, *International Congress on Cosmic Ray Physics, Como, 11-16 Settembre 1949*, p. 317.

[3] Auger P., Maze R. and Grivet-Meyer T., *C. R. Acad. Sci.*, **206** (1938) 1721.

[4] Auger P. and Maze R., *C. R. Acad. Sci.*, **207** (1938) 228.

[5] Adams W. S., *Astrophys. J.*, **97** (1943) 105.

[6] Stoermer C., *Arch. Sci. Phys. Nat. Ser.* **4** (1911) 32, 117.

[7] Fermi E., *Nuclear Physics* (The University of Chicago Press, Chicago) 1950.

[8] Zavrtanik D., *J. Phys. G*, **27** (2001) 1597.

[9] Sanderson T. R., Marsden R. G., Heras A. M., Wenzel K.-P., *Proc. 22nd ICRC Dublin*, **3** (1991) 173.

[10] Oetliker M. *et al.*, *Proc. 24th ICRC Roma*, **4** (1995) 470.

[11] Decker R. B. and Vlahos l., *Astrophys. J.*, **306** (1986) 710.

[12] Sakurai K., *Proc. 22nd ICRC Dublin*, **2** (1991) 416.

[13] Wibig T., *J. Phys. G*, **27** (2001) 1633.

[14] Hiltner W. A., *Astrophys. J.*, **109** (1949) 471.

[15] Chandrasekhar S. and E. Fermi, *Astrophys. J.*, **118** (1953) 113.

[16] Zweibel E. G., *Polarimetry of the Interstellar Medium*, edited by W. G. Roberge and D. C. B. Whittet, *ASP Conf. Ser.*, **97** (1996) 486.

[17] Greisen K., *Phys. Rev. Lett.*, **16** (1966) 748.

[18] Zatsepin G. T. and Kuz'min V. A., *JETP Lett.*, **4** (1966) 78.

[19] Kampert K-H., *J. Phys. G*, **27** (2001) 1663.

[20] Wolfendale A. W. and Wibig T., *J. Phys. G*, **27** 1625 (2001).

[21] Lagage P. O. and Cesarsky C., *J. Astron. Astrophys.*, **125** (1983) 249.

[22] Parker E. N., *Astrophys. J.*, **145** (1966) 811.

B.2 D. Bini, R.T. Jantzen: Circular holonomy, clock effects and gravitoelectromagnetism: Still going around in circles after all these years. . .

D. Bini, R.T. Jantzen: "Circular holonomy, clock effects and gravitoelectromagnetism: Still going around in circles after all these years. . .," *Nuovo Cimento B* **117**, 983 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

Circular holonomy, clock effects and gravitoelectromagnetism: Still going around in circles after all these years ...(*)

D. Bini(1)(2) and R. T. Jantzen(2)(3)

(1) *Istituto per le Applicazioni del Calcolo "M. Picone", CNR - I-00161 Rome, Italy*
(2) *ICRA, Università di Roma - I-00185 Roma, Italy*
(3) *Department of Mathematical Sciences, Villanova University - Villanova, PA 19085, USA*

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — The historical origins of Fermi-Walker transport and Fermi coordinates and the construction of Fermi-Walker transported frames in black-hole spacetimes are reviewed. For geodesics this transport reduces to parallel transport and these frames can be explicitly constructed using Killing-Yano tensors as shown by Marck. For accelerated or geodesic circular orbits in such spacetimes, both parallel and Fermi-Walker transported frames can be given, and allow one to study circular holonomy and related clock and spin transport effects. In particular the total angle of rotation that a spin vector undergoes around a closed loop can be expressed in a factored form, where each factor is due to a different relativistic effect, in contrast with the usual sum of terms decomposition. Finally the Thomas precession frequency is shown to be a special case of the simple relationship between the parallel transport and Fermi-Walker transport frequencies for stationary circular orbits.

PACS 04.20.-q – Classical general relativity.
PACS 95.30.Cq – Elementary particle processes (astrophysics).
PACS 04.20.Cv – Fundamental problems and general formalism.
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

The geometry of circular orbits in general relativity is so rich that after all these years in which various aspects of it have been studied in many approaches, remarkably there is still something interesting left to say on the matter. Here we describe parallel and Fermi-Walker transport along these curves in black-hole spacetimes after reviewing

(*) Paper presented at the IX ICRA Network Workshop "Fermi and Astrophysics" (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.

the historical origins of Fermi-Walker transport [1, 2] and the other interesting class of curves for which the transport equations can be explicitly solved: general geodesics in stationary axisymmetric spacetimes admitting a Killing-Yano tensor [3-5], a case which includes black-hole spacetimes. For circular orbits in black-hole spacetimes, the total parallel transport or Fermi-Walker angle per orbital revolution is represented in factored form [6], revealing each of the spacetime geometric contributions to the final result, in contrast with the usual sum of terms decomposition associated with the gravitoelectric (GE), gravitomagnetic (GM) and space curvature effects [7, 8].

Born in 1901, Fermi was a boy genius who grew up in Rome, arriving at the University of Pisa already having learned physics and advanced mathematics on his own, and having done many physics experiments with his friend Enrico Persico during high school (including measuring the precise acceleration of gravity in Rome) who together became the first two professors of theoretical physics in Italy in 1927 [9]. Fermi's first three papers (on electromagnetism and relativity) were published in 1921 [10, 11] and 1922 [1] while he still a university student at the University of Pisa caught up in the excitement of the newly born theory of general relativity (1916) and early evolution of quantum mechanics (he was the authority on the latter in Pisa as a student!), after which he returned to Rome briefly. Tullio Levi-Civita had just introduced the notion of parallel transport in 1917 [13, 14] (when Fermi finished high school) and had come to Rome from Padua in 1918 as a full professor of mathematics after having made fundamental contributions to the "absolute differential calculus" [15] largely developed by his mentor Gregorio Ricci-Curbastro (hence the title "Ricci-Calculus" of Schouten's encyclopedic book in 1954 [12]) but improved and applied to physics by Levi-Civita and passed on to Einstein by Marcel Grossmann. More details on these historical figures can be found in the many volumes of Gillispie's standard reference [9].

Starting from Levi-Civita's form of the metric for a uniform gravitational field published in a series of articles of 1917–1918 [16]

$$ds^2 = -a^2 dt^2 + \delta_{ij} dx^i dx^j \ ,$$

Fermi studied the effects of acceleration on Maxwell's equations in his second paper [11] (essentially equivalent to a "variable speed of light" a). He was then led to understand the way in which one could apply these results to a small enough spacetime region around the world line of a test observer in a first-order approximation in the spatial distance from the world line, resulting in his third paper on what later came to be known as "Fermi transport". Strangely making no mention of Fermi's earlier work and citing only Eisenhart [17] (who does not discuss Fermi's work but does list his article in the bibliography) as a reference for Riemann normal coordinates (Eisenhart, appendix 3, see also section 11.7 of [19]), Walker [2] repeated Fermi's calculations for transporting vectors taking a different approach, making explicit the implicit spacetime Fermi transport equation contained in Fermi's paper, and extended the analysis of the effect of curvature on the arclength of nearby curves to second order in the approximation, calculated only to first order by Fermi. Both examined the question on an n-dimensional manifold as an exercise in differential geometry. Eisenhart himself had extended Fermi's discussion a few years after it had appeared to a general symmetric (not necessarily metric) connection [18].

2. – Geometric overview

In a spacetime with coordinates x^α and metric $g_{\alpha\beta}$ (signature $-{+}{+}{+}$), consider an arbitrary timelike world line or spacelike curve $x^\alpha(s)$ parametrized by an arclength paramter s (respectively proper time and proper distance). The unit tangent is $u^\alpha = dx^\alpha/ds$ (4-velocity in the timelike case $\epsilon = -1$), where $u \cdot u = u_\alpha u^\alpha = \epsilon$. The second derivative $D^2x^\alpha/ds = DU^\alpha/ds = a^\alpha$ (acceleration when $\epsilon = -1$) then satisfies $a \cdot u = 0$ (as follows from covariant differentiating $u \cdot u = \epsilon$ along the curve). If v is any vector defined along the curve which is orthogonal to u, *i.e.* $u \cdot v = 0$, its equations of Fermi transport along the curve together with the transport equation for u itself are

$$\frac{Du^\alpha}{ds} = a^\alpha \ , \qquad \frac{Dv^\alpha}{ds} = -\epsilon u^\alpha(a_\beta v^\beta) \ , \tag{1}$$

which were given by Fermi, so the second equation is often referred to as describing Fermi transport, generalizable to any tensors which are orthogonal to u. Vectors transported in this way along the curve remain orthogonal to u.

However, implicit in Fermi's work is the way to transport any vector X along the curve, namely when decomposed orthogonally with respect to u, the part along u is held constant while the orthogonal piece is evolved with this law. Letting

$$X = X^{(||)}u + X^{(\perp)} \ , \quad X^{(||)} = \epsilon X \cdot u \ , \quad X^{(\perp)} \cdot u = 0 \ , \tag{2}$$

then if $X^{(||)}$ is held constant along the curve and $X^{(\perp)}$ is transported by the Fermi transport law, then since $X \cdot a = X^{(\perp)} \cdot a$

$$\begin{aligned}\frac{DX^\alpha}{ds} &= X^{(||)}\frac{Du^\alpha}{ds} + u^\alpha\frac{DX^{(||)}}{ds} + \frac{DX^{(\perp)\alpha}}{ds} \ , \\ &= \epsilon(a^\alpha u_\beta - u^\alpha a_\beta)X^\beta \ ,\end{aligned} \tag{3}$$

which is the general transport law given by Walker, subsequently referred to as Fermi-Walker transport, corresponding to vanishing Fermi-Walker derivative along the curve

$$\frac{D_{(\mathrm{fw})}X^\alpha}{ds} = \frac{DX^\alpha}{ds} - \epsilon(a^\alpha u_\beta - u^\alpha a_\beta)X^\beta = 0 \ , \tag{4}$$

easily extended to any tensor defined along the curve. Relative to parallel transport along the curve, the additional terms $-[\epsilon a \wedge u]^\alpha{}_\beta X^\beta$ generate a rotation/pseudorotation in the plane of u and a by just the amount necessary to keep u tangent to the curve, and for geodesics where $a = 0$, this reduces to parallel transport. The metric itself has vanishing Fermi-Walker derivative along any curve and so in addition to preserving the orthogonal decomposition of the tangent space parallel and perpendicular to u along the curve, all inner products are invariant under this transport.

Fermi derived his transport law by considering first an n-dimensional Riemannian manifold and then specialized his result to $n = 4$ and scaled three coordinates by i to change the signature to $+{-}{-}{-}$. His calculation can be retraced in modern language for the two cases $\epsilon = \pm 1$ simultaneously. Although Fermi described the situation in words without any figures, a picture is worth a thousand words. Figure 1 illustrates the argument for the simpler case of a lying in the plane of u and an orthogonal unit vector v,

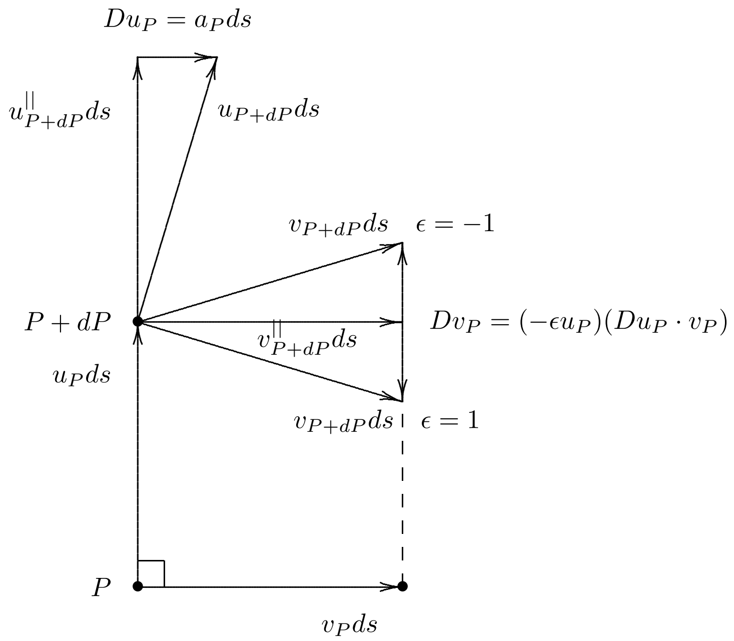


Fig. 1. – The Fermi transport argument. The unit vector v orthogonal to u must undergo a rotation ($\epsilon = 1$) or boost ($\epsilon = -1$) in order to remain orthogonal to u as u itself undergoes this same motion relative to the parallel transported direction $u^{||}$. Only the component of v along the direction in which the tip of u is moving must change, along the direction u by an amount $Du \cdot v$.

showing the "infinitesimal" vectors associated with two infinitesimally separated points on the curve.

Here $u^{(||)}$ and $v^{(||)}$ denote the parallel transported vectors from the point P to the nearby point $P + dP$ on the curve. The unit vector u has (pseudo-)rotated from $u^{(||)}$ by the amount $Du = (u - u^{(||)})ds = ads$, which has the (pseudo-)angle as its magnitude. In order for v to remain orthogonal to u, it must undergo the same (pseudo-)rotation (boost/rotation) in the plane of u and a. The difference vector $Dv = (v - v^{(||)})ds$ must then have the direction $-\epsilon u$ shown in the figure, but since only the component of v in the plane of u and a need change, the scalar amount must be the projection of v along $Du = ads$, namely $v \cdot ads$, so $Dv = (-\epsilon u)(v \cdot ads)$. Dividing through gives the Fermi relation $Dv/ds = -\epsilon u\, a \cdot v$, representing the minimal (pesudo-)rotation necessary to keep v orthogonal to u.

Next Fermi introduces coordinates in the following way. Complete u to an orthonormal frame along the curve by adding the vectors e_i, $i = 1, \ldots, n-1$, with $n = 4$ in the case of a spacetime, where the orthogonal vectors are Fermi transported. For each unit vector $n = n^i e_i$ in the tangent space of a point on the curve at $x^\alpha(s)$, send out a geodesic in its direction and assign coordinates (s, y^i) to points along it, where $y^i = \tilde{s} n^i$ and $\tilde{s}$ is the arclength along the geodesic, nicely illustrated by fig. 13.4 in Misner, Thorne and Wheeler [19] for the more general case of any completion of u to an orthonormal frame. Figure 2 gives the simpler picture analogous to fig. 1. In the spacetime context $\{e_\alpha\} = \{u, e_i\}$ is a locally nonrotating (the spatial axes are fixed with respect to gyro-

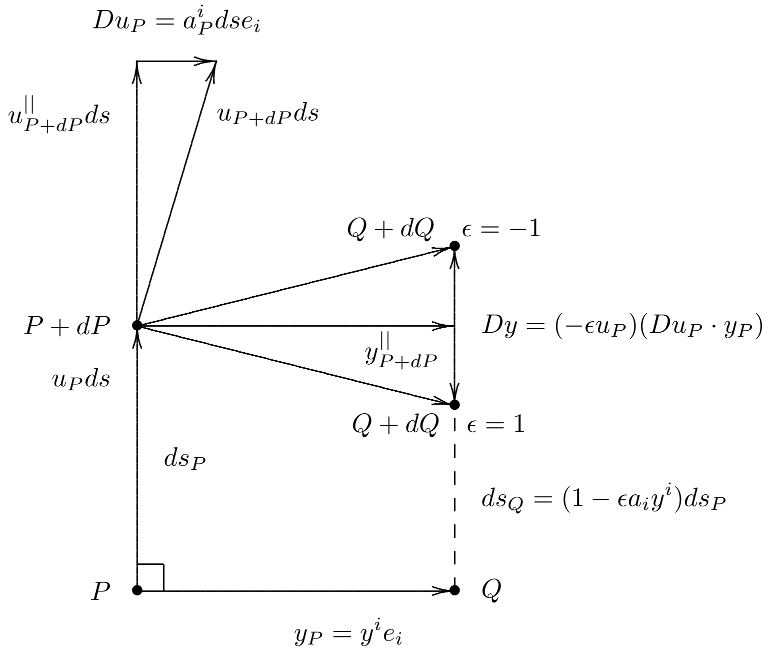


Fig. 2. – The Fermi normal coordinate first-order metric derivation. To first order the coordinates y^i remain orthonormal and orthogonal to s, but the (pseudo-)rotation of u relative to the parallel transported direction $u^{||}$ causes the arclength along a nearby curve of constant (small) y^i (the dashed line) to stretch/shrink compared to the arclength s (ds_Q *vs.* ds_P).

scopes) proper frame for an observer with this world line, naturally called a Fermi frame, adapted to the observer's local time direction and local rest space.

Along the curve itself, these are orthonormal coordinates. As one moves slightly off the curve, the effect of the (pseudo-)rotation of u on a nearby curve of points all sharing the same coordinates y^i causes the arclength ds_Q separating the points Q and $Q + dQ$ to shrink/stretch relative to the arclength ds_P separating corresponding points P and $P + dP$ on the original curve. The change $Du_P = a^i ds\, e_i$ is a (pseudo-)angle times a unit vector, but only the component of $y_P = y^i e_i$ along this unit vector changes by this (pseudo-)angle times its length to get the arclength of the change. Thus the dot product $Du \cdot y_P = a^i ds_P e_i \cdot y^j e_j = a_i y^i ds_P$ gives the amount by which ds_Q differs from ds_Q, with sign ϵ as in the figure: $ds_Q = (1 - \epsilon a_i y^i) ds_P$. To first order the square is just ${ds_Q}^2 = (1 - 2\epsilon a_i y^i){ds_P}^2$. To first order only the metric component along the curve changes

$$\text{(5)} \qquad \text{“}ds^2\text{”} = {ds_Q}^2 + \eta_{ij} dy^i dy^j + O(y^2) = \epsilon(1 - 2\epsilon a_i y^i){ds_P}^2 + \eta_{ij} dy^i dy^j + O(y^2)$$

(quotes to distinguish the line element symbol from the arclength differential along the curve, η_{ij} for the flat orthonormal coordinate 3-metric of the appropriate signature) or dropping the subscript:

$$\text{(6)} \qquad \text{“}ds^2\text{”} = \epsilon(1 - 2\epsilon a_i y^i) ds^2 + \eta_{ij} dy^i dy^j + O(y^2) \ .$$

This is most useful for a timelike curve in spacetime, where $s = \tau$ is the proper time along the world line and this becomes

$$ds^2 = -(1 + 2a_i y^i)d\tau^2 + \delta_{ij} dy^i dy^j + O(y^2) \ , \tag{7}$$

which is eq. (13.71) of Misner, Thorne and Wheeler [19] with zero Fermi rotation of the spatial axes. These are called "Fermi normal coordinates", and represent the locally nonrotating (with respect to gyroscopes) "proper reference frame" of a test observer following this world line. Along a geodesic, the metric is then the usual flat one up to first order in the flat spatial coordinates, characterized by the fact that along the curve itself in these coordinates, the connection components vanish and the metric components are those of flat spacetime in inertial coordinates.

These were first used by Levi-Civita to study geodesic deviation in 1926 [20], as discussed by Manasse and Misner [21], who construct the Fermi frame and compute the Fermi normal coordinate metric up to the second-order terms where the curvature tensor along the world line appears as in Riemann normal coordinates (hence the name "Fermi normal coordinates") and evaluate it for a radial geodesic in the Schwarzschild metric. O'Raifeartaigh [22] investigated how one could duplicate the conditions of locally flat metric and vanishing connection components along an arbitrary accelerated curve, following up earlier work by Schouten [23, 12]. Synge discusses Fermi-Walker transport and Fermi coordinates at length in his 1960 book on general relativity [24].

Ironically Levi-Civita is the original source claiming that Fermi established that one could find local coordinates along an arbitrary curve for which the connection components vanish and even gives a general argument why this should be so (footnote on p. 167 of [15]). This claim is repeated by Schouten, Misner, Pirani [25] and others [26], but Fermi only shows this for a geodesic, not an arbitrary curve, a puzzling fact. In Levi-Civita's discussion of geodesic deviation [20], he makes this general claim about arbitrary curves, but then only explicitly constructs Fermi coordinates for a geodesic. However, the resolution of the puzzle is that for nongeodesics, one must give up Fermi-Walker transport in the construction, using only parallel transport, as explicitly described by the 1-dimensional submanifold specialization of the more general discussion of O'Raifeartaigh [22]. This breaks the link between the adapted coordinates along the world line and the nice orthogonal decomposition of the tangent space associated with the observer's local splitting of space and time, making the construction uninteresting from the point of view of gravitational theory.

Ni and Zimmerman [27] followed up the Misner, Thorne and Wheeler discussion [19] to evaluate the metric up to second order (curvature terms) for the more general rotating Fermi coordinate system along an arbitrary world line, the logical conclusion of the calculations started by Fermi and Walker. Their result shows how the gravitoelectric and gravitomagnetic contributions at the connection level due to the acceleration of the observer and the rotation of the observer spatial frame (the observer's GE and GM fields) and the gravitoelectric and gravitomagnetic parts of the curvature tensor measured by the observer all contribute to the description of the local proper frame of the observer nearby its world line. The metric takes the form

$$\begin{aligned} ds^2 = &-[1 + 2a \cdot x + (a \cdot x)^2 - (\omega \times x)^2 + R_{0ioj} x^i x^j] dx^{02} + \\ &+2[(\omega \times x)_i + \frac{2}{3} R_{0lmi} x^l x^m] dx^0 dx^i + \end{aligned} \tag{8}$$

$$+[\delta_{ij} - \frac{1}{3}R_{iljm}x^l x^m]dx^i dx^j + O(x^3) \ ,$$

where the components of the acceleration a_i of the world line and angular velocity ω_i of the orthonormal spatial frame e_i (relative to a Fermi-transported spatial frame) are expressed in that frame, while the components of the Riemann tensor are evaluated in the orthonormal frame e_0, e_i along the world line at $x^i = 0$, where e_0 is the 4-velocity of the world line.

Even if one follows the Fermi normal coordinate construction exactly, one has problems with coordinate singularities and failure of the coordinates to cover all of spacetime even in the case of flat spacetime, as in the Rindler metric built on an observer family each member of which undergoes constant acceleration. Marzlin [26] examines this difficulty with extending the coordinates away from the world line and suggests a modification of the construction to widen their extendibility. Märzke-Wheeler coordinates [28] are another option for correcting these difficulties, discussed more recently by Pauri and Vallisneri [29] and Dolby and Gull [30], involving radar time.

Walker's approach to Fermi's discussion is based on Riemann normal coordinates at two nearby points along the timelike or spacelike curve in an n-dimensional space, which he uses to compute to second order in the remaining Fermi coordinates y^i the rate ds_Q/ds_P at which arclength changes on a nearby curve traced out by a point Q as in the above discussion, namely $y^i = y^i_0$, s varying. Along a timelike curve in a spacetime, this shows how the tidal curvature affects the proper time of clocks carried in the observer proper frame as their spatial distance from the observer world line increases enough to begin to detect the effects of the curvature of spacetime. This then enables him to find the energy of each point in a small test body with respect to a test observer at any fixed reference point in the body chosen as the given world line, which for a small rigid body reproduces an earlier result, the apparent goal of his investigation.

Walker starts by introducing an arbitrary orthonormal frame $\{e_\alpha\}$ defined along the curve and the corresponding components of the connection induced along the curve

$$\frac{De_\alpha}{ds} = e_\beta W^\beta{}_\alpha \ , \tag{9}$$

which defines a mixed tensor whose totally contravariant form $W^\sharp$ or covariant form $W^\flat$ is antisymmetric, as follows from differentiating $e_\alpha \cdot e_\beta = \eta_{\alpha\beta}$:

$$W_{(\alpha\beta)} = \frac{De_{(\alpha} \cdot e_{\beta)}}{ds} = 0 \ . \tag{10}$$

With hindsight we know that an antisymmetric second-rank tensor can be expressed in terms of its electric and magnetic parts with respect to a unit vector direction

$$W^{\alpha\beta} = [u \wedge A]^{\alpha\beta} + B^{\alpha\beta} \ , \tag{11}$$

where A and B are a vector and second rank antisymmetric tensor both orthogonal to u.

If one desires $u = u^\alpha e_\alpha$ to have constant components in such a frame along the curve, *i.e.* the frame vectors maintain constant angles with respect to the tangent u, then

$$a^\alpha = \frac{Du^\alpha}{ds} = W^\alpha{}_\beta u^\beta = A^\alpha \ , \tag{12}$$

but this leaves B arbitrary, describing the rotation of the frame about the direction u. For a timelike curve in spacetime, if one chooses a frame containing u, then B is the angular velocity of the remaining spatial frame vectors in the local rest space of u, with respect to gyro-fixed axes, also called the Fermi rotation of the frame.

Without saying this, Walker notes that the simplest choice of W amounts to setting $B = 0$, which corresponds to Fermi-Walker transport of the frame along the curve. For a geodesic, he notes that it is natural to pick the orthonormal frame $\{e_\alpha\}$ to contain its tangent u, and for an accelerated curve, the Frenet-Serret frame (containing u) is suggested, in order to construct the family of Riemann normal coordinates used in his calculations (one coordinate system for each point on the given curve, which in general can agree with a Fermi coordinate system only at that given point). In studying nearby curves, he chooses an orthonormal frame containing u whose remaining frame vectors are Fermi-Walker transported along the given curve. It is this frame that is used to evaluate the energies associated with the world lines of nearby curves in the case of a timelike curve in spacetime, as seen by a test observer moving along that curve. (Ni and Zimmerman [27] also give the coordinate accelerations of the world lines of these points, generalized to include rotation.)

3. – Why useful?

Fermi-Walker transport is useful because it describes the behavior of the spin vector S of a torque-free test gyroscope carried by a test observer along its world line with 4-velocity u (see [19])

$$u \cdot S = 0\ , \qquad \frac{D_{(\mathrm{fw})}S}{d\tau} = 0\ . \tag{13}$$

The spatial vectors of a Fermi frame along the world line are therefore locally nonrotating with respect to a set of three independently oriented test gyros carried by the test observer and span the associated local rest space. Along a test particle in free motion along a geodesic, the Fermi frame is parallel transported along the world line and gives a tool for operationally measuring tidal effects of spacetime curvature along the world line. However, in black-hole spacetimes and the larger family of stationary axisymmetric spacetimes containing them, many families of accelerated test observers in uniform circular motion are defined by various aspects of the spacetime geometry and symmetry, so accelerated curves are important in interpreting this geometry and here Fermi-Walker transport is essential. Thus one is interested in the Fermi frame along both geodesics and accelerated curves.

4. – Geodesics in black-hole spacetimes

For a proper time parametrized timelike geodesic world line of a test particle with 4-velocity $u^\alpha = dx^\alpha/d\tau$, 4-momentum $P^\alpha = \mu u^\alpha$, and vanishing acceleration $a^\alpha = Du^\alpha/d\tau = 0$, the rest mass μ provides one constant of the motion $P \cdot P = -\mu^2$. For black-hole spacetimes (and in fact the entire Carter family of type D solutions), the two Killing vectors $\xi[t]$, $\xi[\phi]$ associated with stationary axisymmetry together with the existence of a symmetric Killing tensor $K_{\alpha\beta}$ lead to three additional constants of the motion: $E = -\xi[t] \cdot P$ (energy at infinity) and $L_z = \xi[\phi] \cdot P$ (axial component of angular momentum) and either $\mathcal{Q}$ or $\mathcal{K}$, quadratic constants of the motion associated with the

Killing tensor [31]. These allow the second-order geodesic equations to be reduced to a first-order system of four differential equations which can be interpreted in terms of motion in a potential for each of the Boyer-Lindquist coordinate variables. Marck [3-5], with some initial guidance from Carter, later showed that the closely associated Killing-Yano tensor also allowed one to reduce the construction of a Fermi frame along a geodesic to a single first-order differential equation for a rotation angle.

The tangent to an affinely parametrized timelike geodesic is parallel transported along the geodesic. If one can come up with a second independent parallel transported vector, its spatial projection will also be parallel transported and by normalization, one has the first two vectors of a Fermi frame, leaving the second two in the orthogonal 2-plane defined up to the angle of rotation in that plane.

For a Killing vector ξ, the quantity $\xi \cdot u$ is a conserved quantity (specific energy or angular momentum in our case) along a geodesic

$$\xi_{(\alpha;\beta)} = 0 = a^\alpha \ , \qquad \frac{D}{d\tau}(\xi_\gamma u^\gamma) = \xi_\alpha a^\alpha + \xi_{(\alpha;\beta)} u^\beta u^\alpha = 0 \ , \tag{14}$$

using the chain rule $DX/\tau = u \cdot \nabla X$ for differentiating a field X along the curve.

For a symmetric second-rank tensor which is a Killing tensor

$$K_{[\alpha\beta]} = 0 \ , \quad K_{(\alpha\beta;\gamma)} = 0 \ , \tag{15}$$

one can define a vector $W^\alpha = K^\alpha{}_\beta u^\beta$, a scalar $Q = K_{\alpha\beta} u^\alpha u^\beta = W_\alpha u^\alpha$ and another vector $[P(u)W]^\alpha = (\delta^\alpha{}_\beta + u^\alpha u_\beta) W^\beta = W^\alpha + Q u^\alpha$ which is orthogonal to u. Thus $W = P(u)W - Qu$ is the orthogonal decomposition of W with respect to u. Similar calculations show that Q is a quadratic constant of the motion, and the intrinsic derivatives of W and $P(u)W$ along u are both orthogonal to u, but in general nonzero. This fails to lead to a second parallel-transported direction.

However, black-hole spacetimes also have a second-rank Killing-Yano tensor, which is just an antisymmetric tensor satisfying

$$F_{(\alpha\beta)} = 0 \ , \qquad F_{\alpha(\beta;\gamma)} = 0 \ , \tag{16}$$

found first by Penrose and Floyd [32]. Its square $F^\alpha{}_\gamma F^\gamma{}_\beta$ is automatically a (symmetric) Killing tensor, which is just K itself in this case. This has the advantage that the vector $v^\alpha = F^\alpha{}_\beta u^\beta$ is not only parallel transported along this geodesic but is automatically orthogonal to u (since $v^\alpha u_\alpha = F_{\alpha\beta} u^\alpha u^\beta = 0$), and so only has to be normalized (provided that it is not zero) to get a second Fermi frame vector. These may be completed to an orthonormal frame by adding two spacelike unit vectors defined up to an angle of rotation in the plane orthogonal to the first two vectors. The equations of parallel transport leads to a first-order differential equation for this angle to fix those vectors to also be parallel transported, leading to the Fermi frame. Because of the constants of the motion, this equation may be solved by an integral formula involving those constants.

Marck extends the Fermi frame calculation to null geodesics in his second article and then to spacetimes with two Killing vectors and a Killing-Yano tensor in his third article. He then uses the results to study tidal curvature effects along geodesics in nonrotating black-hole spacetimes [33].

5. – Circular orbits in stationary axisymmetric spacetimes

Accelerated orbits are much more difficult to treat, so one must assume more symmetry in the orbits themselves to make progress, and one must distinguish between parallel transport and Fermi-Walker transport. Circular orbits in stationary axisymmetric spacetimes (especially black-hole spacetimes) are very interesting for many reasons, and turn out to have nice geometry associated with these transports which can be explicitly described without having to integrate any remaining differential equations. Although one can state the final result which solves the equations of parallel transport or Fermi-Walker transport for any vector along such orbits [6], one can build up the same formula by including a number of relativistic effects one by one, leading to the final factored form of that formula. This gives a nice geometric interpretation to the result.

Consider general accelerated but constant speed circular orbits in the equatorial plane $\theta = \pi/2$ of a black hole in Boyer-Lindquist coordinates t, r, θ, ϕ, with mass and specific angular momentum parameters m and $0 \leq a < 1$. This "plane" admits a pair of oppositely rotating periodically intersecting circular geodesics which are timelike outside a certain radius and allow the study of various so-called "clock effects" by comparing either observer or geodesic proper time periods of orbital circuits defined by the observer or the geodesic crossing points. This can be extended to a comparison of parallel transported vectors, corresponding to special holonomy transformations, and with some modifications to Fermi-Walker transport corresponding to a sort of "spin holonomy".

The stationary circular orbits containing every second meeting point of the oppositely rotating circular geodesics departing from an initial crossing point are called geodesic meeting point observers (GMPOs) [6]. Since they are intimately connected to the properties of parallel transport by their definition as having a parallel transported tangent vector, it should be no surprise that the circular geodesics (and these GMPOs) play a key role in the interpretation of the parallel transport transformation along a general circular orbit. If $\zeta_{(\mathrm{geo})-} < 0 < \zeta_{(\mathrm{geo})+}$ are the angular velocities ($d\phi/dt$) of oppositely rotating geodesics, the GMPOs have their average angular velocity $\zeta_{(\mathrm{gmp})} = \frac{1}{2}(\zeta_{(\mathrm{geo})+} + \zeta_{(\mathrm{geo})-})$, which is nonzero only for a rotating black hole where asymmetry exists between the corotating (+) and counterrotating (−) geodesics.

6. – Closed ϕ loops

First we study the various relativistic effects which describe parallel transport around accelerated circular orbits and then for Fermi-Walker transport around the same orbits. The primary effect to consider arises from the geometry of a single ϕ coordinate loop at fixed time t, so that the tangent vector to the orbit lies in the local rest space of the ZAMOs (zero angular momentum observers), also called the locally nonrotating observers, whose 4-velocity is the unit normal n to the t hypersurfaces. Let $R = g_{\phi\phi}^{1/2}$ be the circumferential radius of the orbit at coordinate radius r, and let $\rho = |-(\ln R)_{,r}/g_{rr}^{1/2}|^{-1}$ be its Lie relative curvature [34], or more descriptively, the ZAMO intrinsic radius of turning.

For a tangent vector in the t-ϕ subspace of the tangent space like the initial radial direction in fig. 3, parallel transport around an orbital interval of angle ϕ leads to a compensating (oppositely directed) angle Φ of the parallel transported direction relative to the actual radial direction which cancel each other out in a flat geometry: $d\Phi/d\phi = 1$. When the intrinsic geometry of this plane is not flat, after completing one circuit of

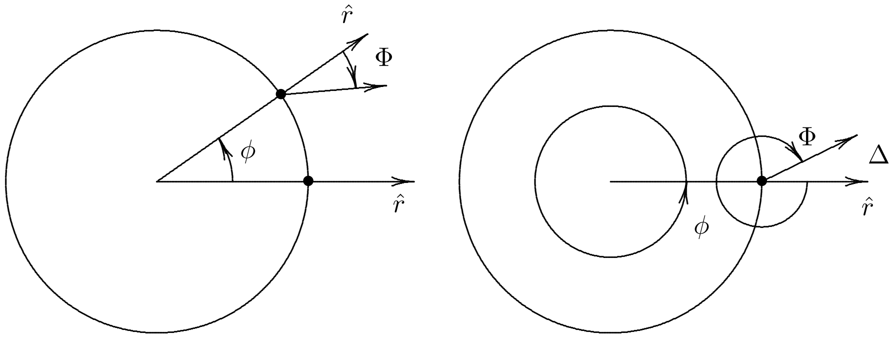


Fig. 3. – Parallel transport around a circular orbit in its own "plane". $d\Phi/d\phi = 1$ in a flat geometry, but curvature causes this relative frequency to differ from 1. When $0 < d\Phi/d\phi < 1$ as in this diagram, the transported direction advances in the orbital direction because the parallel transport angle Φ lags behind the orbital angle ϕ.

the orbit, the radial direction will be rotated forward by the angle $\phi - \Phi$, an advance described by the relative frequency rate $1 - d\Phi/d\phi$. The amount is nicely interpreted by an explanation given by Thorne [35,8] using the tangent cone to the surface of revolution embedding diagram for the intrinsic geometry of the equatorial plane.

When the ratio $R/\rho < 1$, one can imbed this geometry in Euclidean 3-space as a surface of revolution about the z-axis, where R is the distance of a point on this surface from the z-axis and ρ is the distance to the vertex of the tangent cone to the surface which lies on the z-axis as shown in fig. 4. This shows the relationship between the net parallel transport angle per completed orbital circuit

$$\Delta = 2\pi(1 - R/\rho) > 0 \ , \tag{17}$$

advancing in the same direction as the orbital direction. Since the r, θ, ϕ coordinates are orthogonal and ϕ is a Killing coordinate, parallel transport does not affect the component of a spatial vector (in the local rest space of n) perpendicular to the equatorial plane, so this rotation angle in the r-ϕ tangent subspace completely describes the parallel transport around the closed ϕ loops in the intrinsic curved geometry of the plane. This corresponds to a relative frequency rate $0 < d\Phi/d\phi = R/\rho < 1$.

In fact one can express this in 3-dimensional matrix notation in these coordinates as follows. If $X(0)$ is the initial (component) tangent vector and $X(\phi)$ the final parallel transported vector after a change of orbital angle ϕ, one has

$$X(\phi) = e^{\phi A_{(\mathrm{int})}} X(0) \ , \qquad [A_{(\mathrm{int})}]^i{}_j = \frac{R}{\rho}[e^\flat_{\hat{r}} \wedge e^\flat_{\hat{\phi}}]^i{}_j \ , \tag{18}$$

where $e^\flat_{\hat{r}} = g_{rr}^{1/2} dr$ and $e^\flat_{\hat{\phi}} = R d\phi$ are the orthonormal 1-forms, and the matrix $([A_{(\mathrm{int})}]^i{}_j)$ consists of the coordinate components of a mixed second-rank tensor, which is antisymmetric upon lowering its indices ("int" for intrinsic geometry). This exponential representation of the parallel transport transformation arises as the solution of the constant

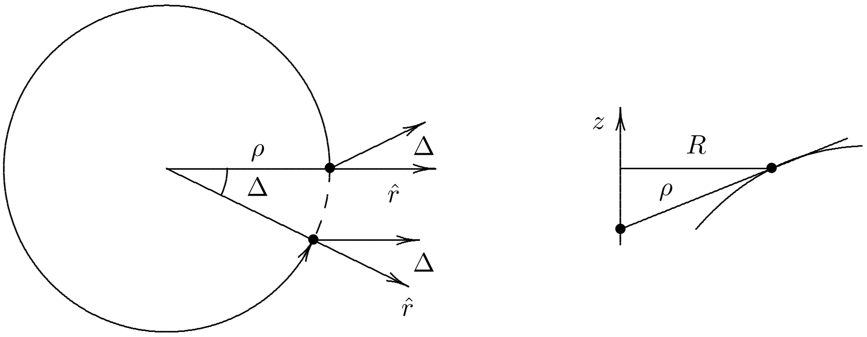


Fig. 4. – The equatorial plane embedding argument. A piece of the cross-section of the embedding surface of revolution is shown together with the tangent cone on the right. The orbit has proper circumference $2\pi R$. Opening the flat tangent cone leads to the circle of radius ρ on the left, where the two bullet points are identified, and having a deficit angle Δ satisfying (solid arc plus dashed arc equals total circumference) $2\pi R + \Delta\rho = 2\pi\rho$ or $\Delta = 2\pi(1 - R/\rho)$. Parallel transport keeps the initial radial direction horizontal, so when it finishes its circuit, it has advanced by the deficit angle with respect to the final radial direction. Thus the deficit angle Δ is exactly the total parallel transport angle for a complete orbital circuit, in the same direction as the orbit.

coefficient linear system of intrinsic parallel transport equations for the coordinate components along the ϕ-parametrized curves

$$\frac{dX(\phi)^i}{d\phi} = A_{(\mathrm{int})}{}^i{}_j X(\phi)^j \ , \tag{19}$$

which corresponds to vanishing intrinsic covariant derivative of X along ϕ

$$\frac{D_{(\mathrm{int})}X(\phi)^i}{d\phi} = \frac{dX(\phi)^i}{d\phi} - A_{(\mathrm{int})}{}^i{}_j X(\phi)^j = 0 \ . \tag{20}$$

By applying this to the vector $e_{\hat{r}}^{\alpha}$, which has the covariant derivative $-A_{(\mathrm{int})}{}^i{}_j e_{\hat{r}}^j$, one sees that

$$\frac{D_{(\mathrm{int})}e_{\hat{r}}}{d\phi} = \frac{R}{\rho} e_{\hat{\phi}} \ , \tag{21}$$

which in turn allows the contravariant form of the tensor $A_{(\mathrm{int})}$ to be re-expressed as

$$A^{\sharp}_{(\mathrm{int})} = e_{\hat{r}} \wedge \frac{D_{(\mathrm{int})}e_{\hat{r}}}{d\phi} \ . \tag{22}$$

However, the spacetime parallel transport issue is distinct from the intrinsic geometry parallel transport, since additional extrinsic curvature terms enter the calculation. It turns out (hindsight) that the radial variation of the tilting of the Killing t-coordinate lines (static observer world lines) away from the normal direction causes the 2-plane of this intrinsic parallel transport rotation to tilt as well in the context of spacetime parallel transport (from the extra Christoffel symbol terms due to the extrinsic curvature). This variation can be described by the angular velocity $\zeta_{(\rm gmp)} = -g_{t\phi,r}/g_{\phi\phi,r}$ of the GMPOs [6], and in the spacetime context, evaluating the parallel transport equations explicitly shows that in the above formula, one adds an additional term to the coefficient matrix which converts $e^\flat_{\hat\phi} = R d\phi$ into

$$Y(\zeta_{(\rm gmp)})^\flat = R(d\phi - \zeta_{(\rm gmp)} dt) = \gamma(\zeta_{(\rm gmp)})^{-1} e_{\hat\phi}(\zeta_{(\rm gmp)})^\flat \ , \tag{23}$$

which is in the ϕ angular direction within the local rest space of the GMPOs, but is the Lorentz contraction of the original unit vector $e_{\hat\phi}$ in the local rest space of the ZAMOs in our intrinsic discussion by the relative gamma factor of the geodesic meeting point observers with respect to the ZAMOs. Since the rotation plane belongs to the local rest space of the GMPOs, their 4-velocity $u(\zeta_{(\rm gmp)})$ is invariant under parallel transport along these orbits.

The spacetime result is then

$$X(\phi) = e^{\phi A} X(0) \ , \qquad A^\alpha{}_\beta = -\Gamma^\alpha{}_{\phi\beta} = \gamma(\zeta_{(\rm gmp)})^{-1} \frac{R}{\rho} [e^\flat_{\hat r} \wedge e_{\hat\phi}(\zeta_{(\rm gmp)})^\flat]^\alpha{}_\beta \ , \tag{24}$$

which is the exponential solution of the linear system

$$\frac{dX(\phi)^\alpha}{d\phi} = A^\alpha{}_\beta X(\phi)^\beta \ , \tag{25}$$

so that the parallel transport angle Φ in the $e_{\hat r}$-$e_{\hat\phi}(\zeta_{(\rm gmp)})$ plane satisfies

$$\frac{d\Phi}{d\phi} = \frac{\gamma(\zeta_{(\rm gmp)})^{-1} R}{\rho} \ . \tag{26}$$

This is the ratio of the Lorentz contraction of the ZAMO intrinsic circumferential radius of the orbit to the local rest space of the GMPOs in which the rotation takes place with the ZAMO intrinsic radius of turning, which sort of seems reasonable. Notice also that the gamma factor increases the slowing down of the parallel transport rotation compared to the orbital rotation of the intrinsic parallel transport alone and when $R/\rho < 1$ (when the orbital “plane” can be embedded in Euclidean 3-space), leads to a prograde rotation with angular velocity $1 - d\Phi/d\phi > 0$ relative to the orbital angle.

As above one has the analogous relation

$$\frac{De_{\hat r}}{d\phi} = \frac{d\Phi}{d\phi} e_{\hat\phi}(\zeta_{(\rm gmp)}) \ , \tag{27}$$

which can be used to re-express (24) as

$$A^\sharp = e_{\hat r} \wedge \frac{De_{\hat r}}{d\phi} \ . \tag{28}$$

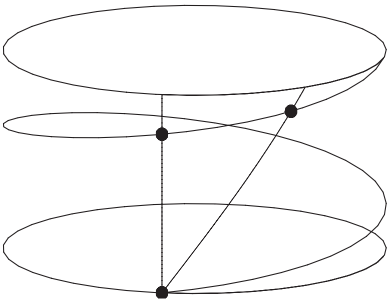


Fig. 5. – The cylinder of circular orbits at a given radius. Once a circular orbit does not close, what constitutes one circuit of the hole depends on which stationary circularly rotating observer is the reference world line.

7. – Helical ϕ loops: observer-dependent circuits

However, the closed ϕ loop orbits are spacelike curves with infinite angular velocity, and we are interested in timelike circular orbits, so we have to now tilt the orbits themselves in time. Moreover, once we "open the loop" by creating a helical curve in spacetime, another problem arises: how to define a complete circuit of the orbit. In fact each stationary circularly rotating observer defines a complete circuit differently, so it must be kept in mind that this concept is clearly observer-dependent, as illustrated by fig. 5.

In a nonrotating (static) black hole, there is an obvious preferred choice of observer: the nonrotating static observers following the t-coordinate lines in the Boyer-Lindquist coordinates. The circuits are then simply characterized by $\Delta\phi = \pm 2\pi$, *i.e.* starting at a given t line and then orbiting one loop around either direction to return to the same line. However, in a rotating black hole, many distinct choices exist all of which reduce to the preferred static observers in the limit of zero rotation:

- static observers or distantly nonrotating observers, following the time coordinate lines along the Killing vector field generating the stationary symmetry which reduces to time translation at spatial infinity,
- ZAMOs or locally nonrotating observers, whose world lines are orthogonal to the time coordinate hypersurfaces,
- geodesic meeting point observers, whose world lines contain every second meeting point of a pair of oppositely rotating geodesic orbits at a given radius,
- extremely accelerated observers (EAOs), for which the magnitude of their acceleration is maximal (except near the hole, where it is minimal),
- Carter observers, whose 4-velocity is the along the intersection of the 2-plane of the 2 repeated null directions of the Riemann curvature tensor with the tangent 2-plane to the t-ϕ cylinder orbits of the stationary axisymmetric symmetry.

Each of these observer families has the same limit at spatial infinity far from the hole, but approaching the hole, each encounters an observer horizon at which their 4-velocity goes null. Some continue to be defined by their geometrical properties as spacelike

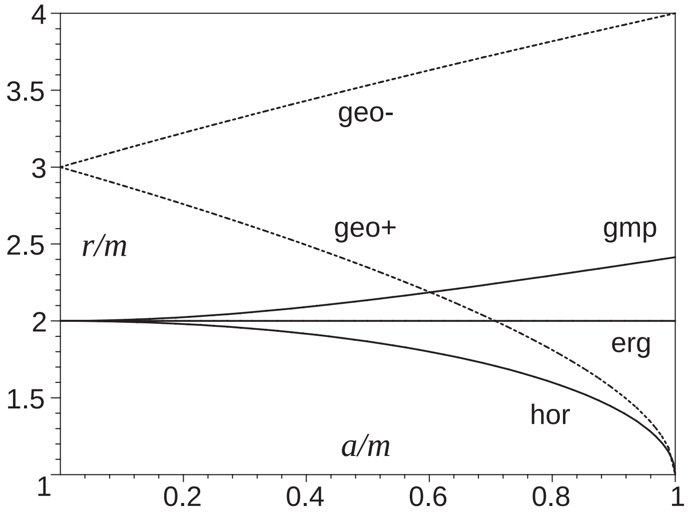


Fig. 6. – The observer horizon radius for each of the geometrically defined stationary circularly rotating observers in the equatorial plane of a Kerr black hole for the physically interesting dimensionless angular momentum parameter interval $0 \le a/m \le 1$. The ZAMOs and Carter observers share the black-hole horizon (hor) as their observer horizon, the static observers have the ergosphere boundary (erg) as their horizon, while the EAOs have their observer horizon at the radius where the counterrotating geodesics go null (geo−).

curves within this horizon, while others do not. For the equatorial plane of our present discussion, fig. 6 compares the various observer horizon radii, all of which lie in the interval $1 \le r/m \le 4$.

Passing to the case of general stationary circularly rotating curves requires the further change in the tangent vector: $\partial_\phi \to \partial_\phi + \zeta^{-1}\partial_t$ and considering values of ζ^{-1} away from zero, tilting the curves in spacetime with respect to the closed ϕ loops. This tangent vector corresponds to using the value of the coordinate ϕ along the curve as its parameter. Switching to t as a parameter along these curves leads to the rescaled tangent vector $\partial_t + \zeta\partial_\phi$, where $\zeta = d\phi/dt$ is the angular velocity of the curve. When nonnull ($\Gamma(\zeta)^{-2} \neq 0$), the tangent can be normalized to a unit vector with corresponding 1-form

$$(29) \qquad \begin{aligned} &u(\zeta) = \Gamma(\zeta)[\partial_t + \zeta\partial_\phi]\ , \qquad u(\zeta)^\flat = \gamma(\zeta)R[d\phi - \bar\zeta dt]\ , \\ &\Gamma(\zeta)^{-2} = -[\partial_t + \zeta\partial_\phi]\cdot[\partial_t + \zeta\partial_\phi]\ , \end{aligned}$$

where $\bar\zeta$ is the angular velocity of the circular orbit with unit tangent $u(\bar\zeta)$ orthogonal to $u(\zeta)$ and $\gamma(\zeta)$ is the corresponding gamma factor with respect to the ZAMOs

$$(30) \qquad u(\zeta) = \gamma(\zeta)[n + \nu^{\hat\phi}(\zeta)e_{\hat\phi}]\ ,\ \gamma(\zeta) = [1 - \nu^{\hat\phi}(\zeta)^2]^{-1/2} = N\Gamma(\zeta)$$

and $N = (-g^{tt})^{-1/2}$ is the lapse function. The relationship between the two relative velocities is a reciprocal one

$$(31) \qquad \nu^{\hat\phi}(\bar\zeta) = 1/\nu^{\hat\phi}(\zeta) = \bar\nu^{\hat\phi}(\zeta)\ ,$$

where

$$\nu^{\hat\phi}(\zeta) = \frac{R}{N}(\zeta + N^\phi) \ , \qquad N^\phi = g_{\phi t}/g_{\phi\phi} \tag{32}$$

relates the relative velocity to the angular velocity.

The additional term in the tangent vector leads to an additional tilt in the angular direction of the plane of the parallel transport rotation as one moves away from $\zeta^{-1} = 0$, corresponding to the angular direction of a new stationary circularly rotating observer. The previous factor picks up an additional term involving the angular velocity $\bar\zeta_{(\rm car)} = -g_{tt,r}/g_{t\phi,r} = a$ of the curves orthogonal to the Carter observers

$$\begin{aligned} Y(\zeta_{(\rm gmp)})^\flat &= R(d\phi - \zeta_{(\rm gmp)}dt) = \gamma(\zeta_{(\rm gmp)})^{-1}e_{\hat\phi}(\zeta_{(\rm gmp)})^\flat \\ &\to R[d\phi - \zeta_{(\rm gmp)}dt - (\zeta_{(\rm gmp)}/\zeta)(d\phi - \bar\zeta_{(\rm car)}dt)] \\ &= \begin{cases} (1-\zeta_{(\rm gmp)}/\zeta)Y(\mathcal{Z}(\zeta))^\flat \ , & \zeta \neq \zeta_{(\rm gmp)} \ , \\ (\bar\zeta_{(\rm car)} - \zeta_{(\rm gmp)})Rdt \ , & \zeta = \zeta_{(\rm gmp)} \ , \\ (1-\zeta_{(\rm gmp)}/\bar\zeta_{(\rm car)})Rd\phi \ , & \zeta = \bar\zeta_{(\rm car)} \ , \end{cases} \end{aligned} \tag{33}$$

where a map $\zeta \to \mathcal{Z}(\zeta)$ picking out the parallel transported direction is defined explicitly by

$$\mathcal{Z}(\zeta) = \zeta_{(\rm gmp)}\frac{\zeta - \bar\zeta_{(\rm car)}}{\zeta - \zeta_{(\rm gmp)}} \overset{g_{\phi t}\to 0}{\longrightarrow} -\frac{g_{tt,r}}{g_{\phi\phi,r}}\frac{1}{\zeta} \ , \tag{34}$$

and satisfies $\mathcal{Z}(\mathcal{Z}(\zeta)) = \zeta$. $\mathcal{Z}(\zeta)$ is the angular velocity of the stationary axisymmetric vector field tangent to the symmetry orbit which is covariant constant along the Killing trajectory with angular velocity ζ, while $Y(\mathcal{Z}(\zeta)) = \gamma(\mathcal{Z}(\zeta))^{-1}e_{\hat\phi}(\mathcal{Z}(\zeta))$ is the angular direction in the orthogonal subspace of the tangent space. The limit $\lim_{\zeta^{-1}\to 0}\mathcal{Z}(\zeta) = \zeta_{(\rm gmp)}$ returns us to the closed ϕ loop case.

Thus for $\zeta \neq \zeta_{(\rm gmp)}$ the parallel transport rotation takes place in the 2-plane spanned by $e_{\hat r} \wedge e_{\hat\phi}(\mathcal{Z}(\zeta))$ and satisfies

$$\frac{d\Phi}{d\phi} = (1 - \zeta_{(\rm gmp)}/\zeta)\gamma(\mathcal{Z}(\zeta))^{-1}\frac{R}{\rho} \ . \tag{35}$$

This corresponds to the exponential solution

$$X(\phi) = e^{\phi A(\zeta)}X(0) \ , \qquad A(\zeta)^\alpha{}_\beta = -(\Gamma^\alpha{}_{\phi\beta} + \zeta^{-1}\Gamma^\alpha{}_{t\beta}) = \frac{d\Phi}{d\phi}[e^\flat_{\hat r} \wedge e_{\hat\phi}(\mathcal{Z}(\zeta))^\flat]^\alpha{}_\beta \tag{36}$$

of the parallel transport equations

$$\frac{dX(\phi)^\alpha}{d\phi} = A(\zeta)^\alpha{}_\beta X(\phi)^\beta \ . \tag{37}$$

Note that the vector field $e_{\hat r}$, which is spatial with respect to all circularly rotating observers, satisfies

$$\frac{De_{\hat r}}{d\phi} = \frac{d\Phi}{d\phi}\, e_{\hat\phi}(\mathcal{Z}(\zeta)) \ , \tag{38}$$

which again implies the relation

$$A(\zeta)^{\sharp} = e_{\hat r} \wedge \frac{De_{\hat r}}{d\phi} \ . \tag{39}$$

The ϕ parametrization of the circular orbit corresponds to defining circuits with respect to the t-coordinate lines. However, the additional factor $(1 - \zeta_{(\mathrm{gmp})}/\zeta)$ which now appears is exactly the factor needed to change the parametrization from the values of the angular coordinate ϕ along the orbit to the values of the new angular coordinate $\tilde\phi = \phi - \zeta_{(\mathrm{gmp})}t$ dragged along by the GMPOs. Since $dt/d\phi = 1/\zeta$ along these orbits, one immediately gets $d\tilde\phi/d\phi = (1 - \zeta_{(\mathrm{gmp})}/\zeta)$. $\tilde\phi : 0 \to \pm 2\pi$ describes one complete circuit corotating or counterrotating with respect to the GMPOs. The frequency rate ratio between parallel transport angle and the GMPO angle is then directly analogous to the closed ϕ loop case

$$\frac{d\Phi}{d\tilde\phi} = \gamma(\mathcal{Z}(\zeta))^{-1}\frac{R}{\rho} \ . \tag{40}$$

Of course as one increases ζ^{-1} from 0 at a given radius, the 4-velocity along the orthogonal direction to the 2-plane of the rotation goes null and then spacelike as one encounters the direction at which the rotation becomes a null rotation and then a boost. Since for geodesics their own 4-velocity is parallel transported, one has $\mathcal{Z}(\zeta_{(\mathrm{geo})\pm}) = \zeta_{(\mathrm{geo})\pm}$ and the 2-plane of the rotation lies in the local rest space of the geodesics themselves. Thus the orbit angular velocity interval where this 2-plane is not spacelike lies somewhere between the two geodesic angular velocities $\zeta_{(\mathrm{geo})\pm}$ when they are both timelike. In fact since $\mathcal{Z}^2 = \mathcal{Z}$, the endpoints are $\mathcal{Z}(\zeta_\pm)$, where $\zeta_\pm$ are the angular velocities of the pair of oppositely rotating null circular orbits at a given radius. As shown in fig. 7, moving towards the black hole, this interval $[\mathcal{Z}(\zeta_-), \mathcal{Z}(\zeta_+)]$ expands until it first encounters on its left side the counterrotating geodesic at the radius where it goes null, while even closer to the hole it then encounters on its right side the corotating geodesic at the radius where it goes null. This interval continues expanding until the radius $r_{(\mathrm{gmp})}$ of the GMPO horizon, where the left endpoint becomes infinite, corresponding to the null rotation which occurs for the closed ϕ loops at that radius, while the right endpoint finally goes to infinity at the black-hole horizon. The left endpoint then re-enters the top right corner of the plot to terminate at the horizon again at infinite velocity.

To explicitly construct a parallel transported orthonormal frame along one of these circular orbits, say for an angular velocity ζ outside the boost zone at a radius outside the GMPO horizon, one can take $u(\mathcal{Z}(\zeta))$ and $e_{\hat\theta}$, which are orthogonal and both parallel transported along the orbit, and the pair of orthonormal unit vectors in the orthogonal 2-plane which result from the parallel transport of the vectors $e_{\hat r}$ and $e_{\hat\phi}(\mathcal{Z}(\zeta))$ from their initial values through an angle Φ (in the counterclockwise direction) related to the change in ϕ (in the clockwise direction) from its initial value by the constant ratio (35). Inside the boost zone $u(\mathcal{Z}(\zeta)$ and $e_{\hat\phi}(\mathcal{Z}(\zeta))$ simply swap causality properties and a boost replaces the rotation. On the boost zone boundary where $u(\mathcal{Z}(\zeta)$ is null, a little more effort is required to get such an orthonormal frame.

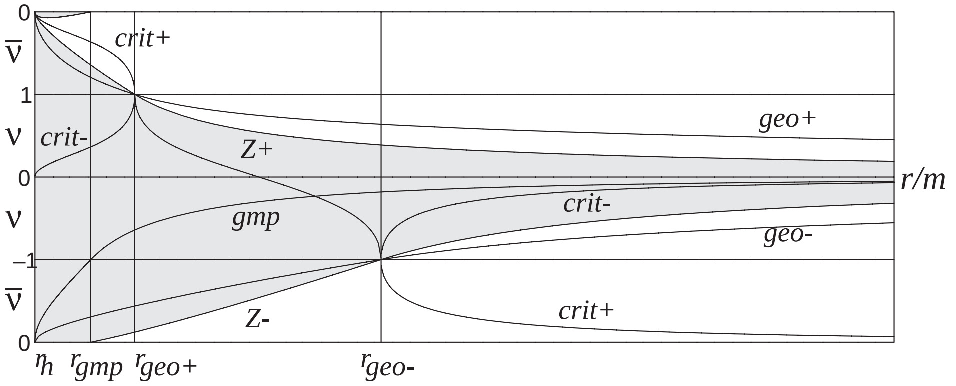


Fig. 7. – The parallel transport boost and rotation velocity profile zones for all timelike, null or spacelike stationary circular orbits in the equatorial plane outside the black-hole horizon at $r = r_h \approx 1.866m$ out to $r/m = 6$, illustrated for $a/m = 0.5$. The horizontal axis is the ZAMO velocity profile ($\nu^{\hat\phi} = 0$); other observer horizons occur where $|\nu^{\hat\phi}| = 1$. The upper and lower borders $\bar\nu^{\hat\phi} = 0$ of the rectangular plot are identified and correspond to the closed ϕ orbits. For spacelike orbits $|\nu^{\hat\phi}| > 1$, the reciprocal velocity $\bar\nu^{\hat\phi} = 1/\nu^{\hat\phi}$ is plotted with decreasing absolute value as one moves away from the horizontal axis. The boost region is shaded, leaving the complementary unshaded region as the rotation zone, separated by the velocity profiles of $\nu^{\hat\phi}(\mathcal{Z}(\zeta_\pm))$, denoted by $Z\pm$, corresponding to null rotations. The circular geodesic profiles cross these velocity curves at their corresponding mutual horizons. The velocity profile of the GMPOs cuts through the middle of the shaded boost region. The three vertical lines inside the plot indicate the GMPO horizon and the corotating (*geo*+) and counterrotating (*geo*−) circular geodesic horizons. Finally the critical velocities (*crit*±) of the magnitude of the acceleration are shown, symmetric about $\nu^{\hat\phi} = 1$ since they satisfy $\nu^{\hat\phi}_{(\rm crit)+}\nu^{\hat\phi}_{(\rm crit)-} = 1$, with the *crit*− curve corresponding to the extremely accelerated observer velocity. For completeness, the spin critical observer [8] velocity profile which connects the two vertices where the *crit*+ and *crit*− curves meet at the two radii $r_{(\rm geo)\pm}$ is also shown.

8. – Fermi-Walker transport

Finally one can construct the corresponding Fermi frames along the timelike circular orbits by considering the relationship between parallel transport and Fermi-Walker transport. If one considers the spacetime Fermi-Walker transport equation (4) applied to a vector X which is orthogonal to u, *i.e.* "spatial", then its Fermi-Walker derivative is also spatial so spatially projecting the equation leads to

$$\frac{D_{(\rm fw)}X}{ds} = P(u)\frac{D_{(\rm fw)}X}{ds} = P(u)\frac{DX}{ds}\ . \tag{41}$$

Thus for spatial vectors, Fermi-Walker transport is just spatially projected parallel transport at the derivative level.

Projecting the covariant derivative of the vector field $e_{\hat r}$ (given by Eq. (38)) into the local rest space of $u(\zeta)$ yields its Fermi-Walker derivative along $u(\zeta)$

$$\frac{D_{(\rm fw)}e_{\hat r}}{d\phi} = P(U(\zeta))\frac{De_{\hat r}}{d\phi} = \frac{d\Phi}{d\phi}P(u(\zeta))\, e_{\hat\phi}(\mathcal{Z}(\zeta)) \tag{42}$$

$$= \frac{d\Phi}{d\phi}\gamma(u(\mathcal{Z}(\zeta)), u(\zeta))\, e_{\hat{\phi}}(\zeta)\ .$$

Thus the ratio between the Fermi-Walker transport angle in the plane of $e_{\hat{r}}$ and $e_{\hat{\phi}}(\zeta)$ of the local rest space of $u(\zeta)$ and the orbital angle is instead

$$\frac{d\Phi_{(\mathrm{fw})}(\zeta)}{d\phi} = \gamma(u(\mathcal{Z}(\zeta)), u(\zeta))\, \frac{d\Phi(\zeta)}{d\phi}\ , \tag{43}$$

with an extra relative gamma factor describing the inverse Lorentz contraction which occurs in the projection, which has unit value for geodesics where this projection reduces to the identity. The relative gamma factor may be expressed in terms of the ZAMO gamma factors [36] as

$$\gamma(u(\mathcal{Z}(\zeta)), u(\zeta)) = \gamma(\mathcal{Z}(\zeta))\gamma(\zeta)[1 - \nu^{\hat{\phi}}(\mathcal{Z}(\zeta))\nu^{\hat{\phi}}(\zeta)]\ . \tag{44}$$

This extra gamma factor for Fermi-Walker transport relative to parallel transport allows $D\Phi_{(\mathrm{fw})}(\zeta)/d\phi > 1$, which can then lead to a retrograde rotation rather than a prograde rotation as in the parallel transport case. Indeed in the flat spacetime limit where $\mathcal{Z}(\zeta) = 0$ (no tilt of the parallel transport plane relative to the Minkowski space time coordinate hypersurfaces), $\zeta_{(\mathrm{gmp})} = 0$ (the limiting GMPOs follow the time coordinate lines) and $R = \rho = r$ (no spatial curvature), only this factor $\gamma(u(0), u(\zeta)) = \gamma(\zeta)$ remains and leads to the retrograde Thomas precession in the local rest space of the circular orbit with angular frequency $d\Phi_{(\mathrm{fw})}(\zeta)/d\phi - 1 = \gamma(\zeta) - 1$ described explicitly in exercise 6.9 of Misner, Thorne and Wheeler [19].

To explicitly construct a Fermi-Walker transported frame along one of these circular orbits, say for an angular velocity ζ outside the boost zone at a radius outside the GMPO horizon, one can take $e_0 = u(\zeta)$ and $e_3 = -e_{\hat{\theta}}$, which are orthogonal and Fermi-Walker transported along the orbit, and the pair of orthonormal unit vectors e_1, e_2 in the orthogonal 2-plane which result from the Fermi-Walker transport of the vectors $e_{\hat{r}}$ and $e_{\hat{\phi}}(\zeta)$ respectively from their initial values through an angle $\Phi_{(\mathrm{fw})}$ (in the counterclockwise direction) related to the change in ϕ (in the clockwise direction) from its initial value by the constant ratio (43).

With the abbreviations $\gamma = \gamma(\zeta)$ and $\nu^{\hat{\phi}} = \nu^{\hat{\phi}}(\zeta)$, one has

$$e_0 = u(\zeta) = \gamma[n + \nu^{\hat{\phi}} e_{\hat{\phi}}]\ , \qquad e_{\hat{\phi}}(\zeta) = \gamma[e_{\hat{\phi}} + \nu^{\hat{\phi}} n]\ , \tag{45}$$

or

$$e_0 \pm e_{\hat{\phi}}(\zeta) = e^{\pm\alpha}[n \pm e_{\hat{\phi}}]\ , \tag{46}$$

where $\nu^{\hat{\phi}} = \tanh\alpha$. Next, letting $e_\pm = e_1 \pm i e_2$, one has to apply the Fermi rotation to the $e_{\hat{r}}$-$e_{\hat{\phi}}(\zeta)$ plane

$$e_+ = e_1 + i e_2 = e^{i\Phi_{(\mathrm{fw})}}[e_{\hat{r}} + i e_{\hat{\phi}}(\zeta)]\ , \tag{47}$$

where one can take $\Phi_{(\mathrm{fw})} = (D\Phi_{(\mathrm{fw})}/d\phi)\,\phi$ along the orbit $\phi = \zeta t + \phi_0$. Letting $e_3 = -e_{\hat{\theta}}$, then e_0, e_1, e_2, e_3 is a spatially righthanded Fermi frame, illustrated in fig. 8 by suppressing the θ direction.

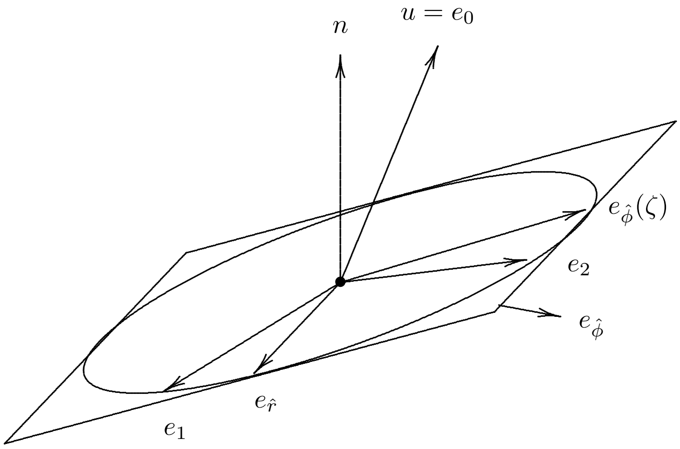


Fig. 8. – The Minkowski spacetime Fermi frame with $e_3 = -e_{\hat\theta}$ suppressed. The local rest space of u is shown relative to the nonrotating (parallel transported) time direction n and its local rest space containing the vectors $e_{\hat r}$ and $e_{\hat\phi}$. The Fermi frame vectors e_1 and e_2 undergo a counterclockwise rotation relative to the boosted axes $e_{\hat r}$ and $e_{\hat\phi}(\zeta)$.

In the flat spacetime case, one has $\Phi_{(\mathrm{fw})} = \gamma\phi$ and $\nu^{\hat\phi} = r\zeta$. This frame was apparently first given by Pirani [39], as well as for the case of circular geodesics in the Schwarzschild spacetime, and later was used by Irvine [38] and Corum [37] to study Maxwell's equations in a uniformly rotating coordinate system in Minkowski spacetime.

If $n, e_x, e_y, e_z = -e_{\hat\theta}$ is the nonrotating orthonormal inertial coordinate frame in Minkowski spacetime (at $\theta = \pi/2$), and therefore parallel transported along any curve, then it can be expressed in terms of the Fermi frame along a circular orbit with angular frequency ζ by a rotation by angle $\gamma\phi$ in the e_1-e_2 plane, followed by a boost with velocity $-\nu^{\hat\phi}$ in the ϕ direction, followed by a rotation by the angle $-\phi$ in the $e_{\hat r}$-$e_{\hat\phi}$ plane

$$\begin{bmatrix} n \\ e_x \\ e_y \\ e_z \end{bmatrix} = R(-\phi)B(-\nu^{\hat\phi})R(\gamma\phi) \begin{bmatrix} u \\ e_1 \\ e_2 \\ e_3 \end{bmatrix} . \tag{48}$$

This is illustrated in fig. 9.

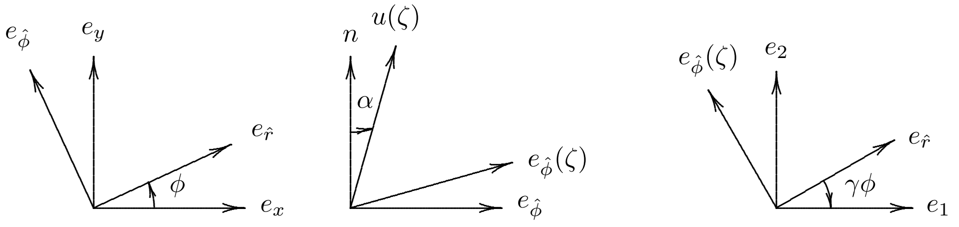
The result of these three transformations, is

$$\begin{gathered} e_2 = e_3 \ , \quad n = \gamma[u - \nu^{\hat\phi}\mathrm{Im}(e^{-i\gamma\phi}e_+)] \ , \\ e_x + ie_y = e^{-i(\gamma-1)\phi}e_+ + i(\gamma-1)e^{i\gamma\phi}\mathrm{Im}(e^{-i\phi}e_+) - ie^{i\gamma\phi}\gamma\nu^{\hat\phi}u \ . \end{gathered} \tag{49}$$

If

$$S = S^n n + P(n)S \ , \ S\cdot u = 0 \ , \ S^n = -S\cdot n \tag{50}$$

is the spin vector of a test gyroscope carried along the circular orbit, belonging to the local rest space of the orbit, then its "laboratory" components, letting $S_\pm = S\cdot e_\pm$,



$R(-\phi)$:
$e_x = ce_{\hat r} - se_{\hat\phi}$
$e_y = se_{\hat r} + ce_{\hat\phi}$
$e_x + ie_y = e^{i\phi}(e_{\hat r} + ie_{\hat\phi})$
$c = \cos\phi,\ s = \sin\phi$

$B(\nu^{\hat\phi})$:
$u(\zeta) = \gamma[n + \nu^{\hat\phi} e_{\hat\phi}]$
$e_{\hat\phi}(\zeta) = \gamma[\nu^{\hat\phi} n + e_{\hat\phi}]$
$u \pm e_{\hat\phi}(\zeta) = e^{\pm\alpha}[n \pm e_{\hat\phi}]$
$\nu^{\hat\phi} = \tanh\alpha$

$R(\gamma\phi)$:
$e_{\hat r} = Ce_1 + Se_2$
$e_{\hat\phi}(\zeta) = -Se_1 + Ce_2$
$e_{\hat r} + ie_{\hat\phi}(\zeta) = e^{-i\gamma\phi}(e_1 + ie_2)$
$C = \cos\gamma\phi,\ S = \sin\gamma\phi$

Fig. 9. – The successive transformations from the parallel transported frame to the Fermi frame in Minkowski spacetime.

$S_x = S \cdot e_x$, etc., are

$$S_z = S_3\ ,\quad S^n = -\gamma\nu^{\hat\phi}\mathrm{Im}(e^{-i\gamma\phi}S_+)\ ,$$
$$S_x + iS_y = e^{-i(\gamma-1)\phi}S_+ + i(\gamma-1)e^{i\gamma\phi}\mathrm{Im}(e^{-i\phi}S_+)\ . \tag{51}$$

Note that the Fermi frame components are constants since the spin vector is Fermi-Walker transported. The case $S_2 = 0$, $S_+ = S_1 = S_3$ reproduces eqn. (6.28) of Misner, Thorne Wheeler [19].

The first term on the right-hand side of the last equation, together with S_z, is the boosted spin vector $B(n,u)S$ actively boosted back from the local rest space of the circular orbit to the local rest space of the nonrotating laboratory frame, while the second term is the distortion in the spin vector due to its relative motion. The active boost identifies $e_{\hat\phi}(\zeta)$ and $e_{\hat\phi}$, removing the passive boost sandwiched between the two passive rotations which simply re-expresses the frames in terms of each other, leading to the pure rotation $R((\gamma-1)\phi)$ of the laboratory boosted spin vector compared to the laboratory axes with the Thomas precession angular velocity

$$(\gamma - 1)\zeta = \frac{\gamma^2\nu^2}{\gamma+1}\zeta\ , \tag{52}$$

a secular precession which grows with time. The distortion term is a mere periodic oscillation which averages out in time.

This decomposition of the measured spatial vector is valid in general

$$P(n)S = B(n,u)S + [\gamma(n,u) - 1][-\hat\nu(n,u)\cdot S]\hat\nu(u,n) \tag{53}$$

and is illustrated in fig. 10 and discussed in refs. [7,36]. Long-term secular effects can be described by the boosted spin vector, which precesses with a constant angular velocity, while the second term is a periodic oscillation analogous to the stellar aberration effect for null directions. In fact in a gyroscope experiment, with the fixed stars as the reference, it is the spin vector boosted into the local rest space of the static observers which gives the secular spin precession formula (subtracting out the stellar aberration effect), since

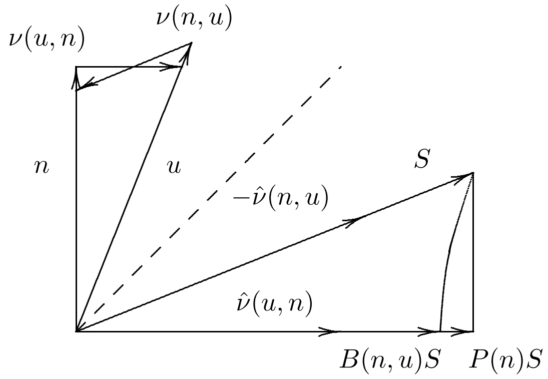


Fig. 10. – The relationship between the projection and boost between local rest spaces along the direction of relative motion. The relative velocity $\nu(u,n)$ of n with respect to u and $\nu(u,n)$ of u with respect to n are related to each other by the relative boost and a sign change. The hatted vectors are the corresponding unit vectors. If S is along the direction of relative motion in the local rest space of u, then since $B(n,u)S = \gamma^{-1}P(n)S$, it follows that $P(n)S = \gamma B(n,u)S = B(n,u)S + (\gamma-1)B(n,u)S$. For an arbitrary vector S in the local rest space of u, this figure applies to its component $[S\cdot\hat\nu(n,u)]\hat\nu(n,u)$ along $-\hat\nu(n,u)$, which satisfies $B(n,u)[-\hat\nu(n,u)] = \gamma\hat\nu(u,n)$, while the components of S orthogonal to the plane of n and u are unchanged by the boost or projection, leading to the general relationship (53) of the text.

it is the static observers which are locked onto the distantly nonrotating observers, *i.e.* the fixed stars. This is calculated exactly for black-hole spacetimes in [8]. It is exactly this factor of $\gamma(n,u)-1$ which occurs in the projection compared to the boost which leads to the relative (difference) precession frequency itself from eq. (42).

The relationship between the parallel transported and Fermi-Walker transported axes remains true for circular orbits in black-hole or even more general stationary axisymmetric spacetimes if for the parallel transported frame one takes $u(\mathcal{Z}(\zeta))$ in place of n and $e_{\hat\phi}(\mathcal{Z}(\zeta))$ in place of $e_{\hat\phi}$ in the above discussion, with γ replaced by the relative gamma factor $\gamma(u(\zeta),u(\mathcal{Z}(\zeta)))$. The secular precession of Fermi-Walker transport compared to parallel transport is governed by the same frequency relationship

$$\frac{d\Phi_{(\rm fw)}(\zeta)}{dt} - \frac{d\Phi(\zeta)}{dt} = [\gamma(u(\zeta),u(\mathcal{Z}(\zeta))) - 1]\zeta \ , \tag{54}$$

for which the Thomas precession is a special case. This relationship should remain valid for helical motion in stationary cylindrically symmetric spacetimes as well.

By considering the natural Frenet-Seret frame along a circular orbit (see [40] and the appendix of [6]), with curvature κ (magnitude of the acceleration in the timelike case) and first torsion τ_1, while the second torsion τ_2 vanishes in the equatorial plane, one easily sees the direct relationship between $u(\mathcal{Z}(\zeta))$ and $u(\zeta)$ can be expressed through their relative velocity

$$\nu^{\hat\phi}(u(\mathcal{Z}(\zeta)),u(\zeta)) = \frac{\kappa}{\tau_1} = \frac{1}{\nu_{gmp}}\frac{(\nu-\nu_+)(\nu-\nu_-)}{(\nu-\nu_{crit+})(\nu-\nu_{crit-})} \ . \tag{55}$$

Here the abbreviations $\nu_\pm$ denote the pair of geodesic relative velocities (zeros of κ), while $\nu_{crit\pm}$ denote the corresponding critical velocities [8] for κ and ν_{gmp} the GMPO relative velocity, all with respect to the ZAMOs. This formula follows from formulas (A.5) of [6] for τ_1 and formulas (4.7) and (4.8) of [8]. It can be directly expressed in terms of the relative velocities with respect to the circular orbit itself as follows:

$$\nu(u(\mathcal{Z}(\zeta)), u(\zeta)) = 2\left(\frac{1}{\nu(u(\zeta_{\mathrm{geo}+}), u(\zeta))} + \frac{1}{\nu(u(\zeta_{\mathrm{geo}-}), u(\zeta))}\right)^{-1} . \tag{56}$$

Since the reciprocal map is the bar map giving the relative velocity of the orthogonal direction in the circular orbit cylinder, *i.e.* the angular direction for a timelike orbit, this states that the relative velocity of the direction along this cylinder giving the orientation of the Lorentz transformation plane is the average of the relative velocities of the local rest space angular directions of the pair of geodesics (when they are timelike).

The extremely accelerated observers have relative velocity ν_{crit-} and not only see the timelike circular geodesics with the same relative speed but opposite directions, but also see the entire boost zone symmetrically [6]. One can show that the map $\mathcal{Z}$ when expressed in terms of relative velocities with respect to these observers (tilde notation) takes the simple form

$$\tilde{\nu} \to -\tilde{\nu}_+\tilde{\nu}_-/\tilde{\nu} \quad (\text{where } \tilde{\nu}_- = -\tilde{\nu}_+\,) \tag{57}$$

characteristic of nonrotating black-hole spacetimes, which implies that the boost zone endpoint velocities then become $\tilde{\nu}(\mathcal{Z}(\zeta_\pm)) = \mp\tilde{\nu}_+\tilde{\nu}_-$. Thus parallel transport along circular orbits links together in various ways not only the GMPOs, Carter observers and static observers but also the extremely accelerated observers.

9. – Circular holonomy and clock effects?

What does all of this have to do with circular holonomy and clock effects? Rotation in black-hole spacetimes introduces an asymmetry between the corotating and counterrotating circular geodesics. The various clock effects measure the difference between the two periods as seen by a given circularly rotating observer for one circuit compared to that observer (either the proper time periods measured by the geodesics themselves, or the observer proper-time periods). Choosing the observers to be the GMPOs leads to the observer-independent clock effect measuring the difference in the geodesic proper periods between every second geodesic crossing point. While all this clock time comparison is going on, it is reasonable to compare gyro spins as well to see how the rotation effects these differently on the pair of geodesics.

Holonomy is the study of how curvature affects vectors during parallel transport around closed loops from a fixed reference point; since parallel transport preserves length, only the direction can change by an element of the Lorentz group with respect to a fixed orthonormal frame in the tangent space at the reference point. The holonomy group at a given point is the subgroup of the Lorentz group which contains all the Lorentz transformations which result from all possible loops starting and stopping at that point. If one restricts the loops to piecewise smooth stationary circular orbits at a fixed radius in the equatorial plane of a black hole, one explores a subset of the holonomy group. The simplest such closed loops are the closed ϕ loops characterized completely by the integer number q of corotating $q > 0$ or counterrotating $q < 0$ circuits.

One can ask when this discrete set of holonomy Lorentz transformations contains the identity. Since parallel transport induces a one-parameter family of such transformations along the orbit relative to the orthonormal spherical frame, the radius must be in the rotation zone in order for these transformations to return to the identity. The rotation zone corresponds to where the GMPOs are timelike, a zone which terminates at their horizon approaching the black hole where $\gamma(\zeta_{(\mathrm{mgp})}) \to \infty$. Then

$$\frac{d\Phi}{d\phi}(2\pi q) = 2\pi p \to \frac{d\Phi}{d\phi} = p/q \tag{58}$$

shows that in this zone where $0 < d\Phi/d\phi < 1$, for each proper fractional rational number value of this relative frequency function (which occurs at a dense set of radii in this zone), parallel transport will return every vector to its original state after a certain number of loops, leading to what Rothman, Ellis and Murugan [41] have called a band of holonomy invariance extending from the GMPO horizon out to infinity. For black-hole spacetimes only, this relative frequency function turns out to be the ratio of the proper (self) period of the GMPOs for one ϕ coordinate loop to the average coordinate time period of the pair of oppositely rotating circular geodesics for the same loop, *i.e.* the time as seen by the distantly nonrotating observers far from the hole.

One can consider a stationary circular orbit which does not close but these are helices in spacetime so one needs two such orbits joined together at an initial point to obtain closed circuits. If one takes one to be one of the various geometrically preferred observers which generalize different aspects of the static spacetime nonrotating observers and the other a timelike geodesic, one can compare how a vector transported around the hole on a circular geodesic compares to the one carried by the observer when they meet. Or one can take a pair of oppositely rotating timelike circular geodesics which start at a common initial point. In the latter case for black holes only, the relative frequency function for each geodesic is also a simple ratio of the proper period (for a circuit defined by the geodesic crossing points themselves) to the average coordinate period for one ϕ coordinate loop.

However, parallel transported vectors along timelike curves are not directly connected with an interesting physical property, while Fermi-Walker transported vectors describe how test gyros behave along these world lines, so it makes sense to extend the notion of holonomy to Fermi-Walker transport in order to compare how gyro spin vectors differ along different circular orbits from the same initial point when they meet again. This requires allowing a relative boost to identify corresponding spin vectors at meeting points since the spin vectors lie in distinct local rest spaces in general. This leads to "spin holonomy". For clock effect oppositely-rotating timelike circular geodesic pairs, this discussion also involves the clock effect periods [6, 42].

Circular orbits have grabbed the imaginations of so many of us over the past century. The present discussion has shown that their geometric richness has still not yet been depleted and has led to further insight about Fermi-Walker transport itself in this context.

* * *

Professor L. Stazi is thanked for his help in obtaining Levi-Civita's articles from the library of the Mathematics Department of the University of Rome "La Sapienza" where Levi-Civita taught for many years.

REFERENCES

[1] Fermi E., *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. & Nat.*, **31** (1922) 21, 51, 101.

[2] Walker A. G., *Proc. R. Soc. Edinburgh*, **52** (1932) 345.

[3] Marck J. A., *Proc. R. Soc. London, Ser. A*, **385** (1983) 431.

[4] Marck J. A., *Phys. Lett.A*, **97** (1983) 140.

[5] Kamran N. and Marck J. A., *J. Math. Phys.*, **27** (1986) 1589.

[6] Bini D., Jantzen R. T. and Mashhoon B., *Class. Quantum Grav.*, **19** (2002) 195 [gr-qc/0111028].

[7] Jantzen R. T., Carini P. and Bini D., *Ann. Phys.*, **215** (1992) 1 [gr-qc/0106043].

[8] Bini D., Carini P. and Jantzen R. T., *Int. J. Mod. Phys. D*, **6** (1997) 143 [gr-qc/0106014].

[9] Gillispie C. C. (Editor), *Dictionary of Scientific Biography* (Scribner's, New York) 1970.

[10] Fermi E., *Nuovo Cimento*, **22** (1921) 199.

[11] Fermi E., *Nuovo Cimento*, **22** (1921) 176.

[12] Schouten J. A., *Ricci-Calculus* (Springer-Verlag, New York) 1954 (see Part III, §8 on Fermi coordinates: p. 166).

[13] Levi-Civita T., *Rend. Circ. Mat. Palermo*, **42** (1917) 173.

[14] Schouten J. A., *Verh. Kon. Akad. AMsterdam*, **12** (1918) 95.

[15] Levi-Civita T., *Lezioni di Calcolo Differenziale Assoluto* (Stock, Rome) 1925; *The Absolute Differential Calculus* (Blackie & Son, London) 1926 (Dover Edition, New York) 1977.

[16] Levi-Civita T., *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. & Nat.*, **26** (1917) 307, 381, 458, 519; **27** (1918) 3, 183, 220, 240, 283, 343; **28** (1919) 3, 101.

[17] Eisenhart L. P., *Riemannian Geometry*, (Princeton Univ. Press, Princeton) 1926.

[18] *Am. Math. Soc. Colloquium Publications* (New York) 1927.

[19] Misner C. W., Thorne K. S. and Wheeler J. A., *Gravitation* (Freeman, San Francisco) 1973 (see §13.8 for Fermi normal coordinates).

[20] Levi-Civita T., *Math. Ann.*, **97** (1926) 291; *Opere matematiche: memorie e note, pubblicate a cura dell'Accademia nazionale dei Lincei*, Vol. **4** (Zanichelli, Bologna) 1954-1960, pp. 433-464.

[21] Manasse F. K. and Misner C. W., *J. Math. Phys.*, **4** (1963) 735.

[22] O'Raifeartaigh L., *Proc. Irish R. Acad. A,* **59** (1958) 15.

[23] Schouten J. A. and Struik D. J., *Einfuhrung in die neueren Methoden der Differentialgeometrie*, I (Noordhoff, Groningen) 1935, p. 106.

[24] Synge J. L., *Relativity, the General Theory* (Amsterdam) 1960.

[25] Pirani F. A. E., *Proc. Irish R. Acad. A*, **61** (1960/1961) 9.

[26] Marzlin K. P., *Gen. Rel. Grav.*, **26** (1994) 619.

[27] Ni W. T. and Zimmerman M., *Phys. Rev. D*, **17** (1977) 1473.

[28] Märzke R. F. and Wheeler J. A., *Gravitation as Geometry I: the Geometry of Space-Time and the Geometrodynamical Standard Meter*, in *Gravitation and Relativity*, edited by H. -Y. Chiu and W. F. Hoffmann (Benjamin W. A., New York) 1964, p. 40.

[29] Pauri M. and Vallisneri M., *Found. Phys. Lett.*, **13** (2000) 401 [gr-qc/0006095].

[30] Dolby C. E. and Gull S. F., *Am. J. Phys.*, **69** (2001) 1257 [gr-qc/0104077].

[31] Carter B., *Phys. Rev.*, **174** (1968) 1559; *Commun. Mat. Phys.*, **10** (1968) 280.

[32] Penrose R., *Ann. N. Y. Acad. Sci.*, **224** (1973) 125.

[33] Luminet J. P. and Marck J. A., *Mon. Not. R. Astron. Soc.*, **212** (1985) 57.

[34] Bini D., Carini P. and Jantzen R. T., *Int. J. Mod. Phys. D*, **6** (1997) 1 [gr-qc/0106013].

[35] Thorne K. S., *in Quantum Optics, Experimental Gravitation, and Measurement Theory* edited by P. Meystre and M. O. Scully (Plenum Press, New York and London) 1981.

[36] Bini D., Carini P. and Jantzen R. T., *Class. Quantum Grav.*, **12** (1995) 2549.

[37] Corum J. F., *J. Math. Phys.*, **18** (1980) 2360.

[38] Irvine W. M., *Physica*, **30** (1964) 1160.

[39] Pirani F. A. E., *Acta Phys. Polon.*, **15** (1957) 389.

[40] Bini D., Merloni A. and Jantzen R. T., *Class. Quantum Grav.*, **16** (1999) 1333.

[41] Rothman T., Ellis G. F. R. and Murugan J., *Class. Quantum Grav.*, **18** (2001) 1217.

[42] Roy Maartens, Bahram Mashhoon and David Matravers, *Class. Quantum Grav.*, **19** (2002) 195.

B.3 D. Boccaletti: Comments on 'Magnetic fields in spiral arms' and 'Problems of gravitational stability in the presence of a magnetic field' by S. Chandrasekhar and E. Fermi

D. Boccaletti: "Comments on 'Magnetic fields in spiral arms' and 'Problems of gravitational stability in the presence of a magnetic field' by S. Chandrasekhar and E. Fermi," *Nuovo Cimento B* **117**, 1009 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

Comments on papers: S. Chandrasekhar and E. Fermi "Magnetic Fields in Spiral Arms" (Ap. J., 118 (1953) 113-115); and "Problems of Gravitational Stability in the Presence of a Magnetic Field" (Ap. J., 118 (1953) 116-141)(*)

D. Boccaletti(**)

Dipartimento di Matematica, Università di Roma "La Sapienza" - Italy

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — As is known, Fermi was one of the first scientists to draw the attention of the astrophysicists to the possible existence of a galactic magnetic field. The first occasion (1949) was that of the famous theory on the origins of cosmic rays. In the succeeding years ('52-'53) Fermi had regular discussions with Chandrasekhar on a variety of astrophysical problems bearing on hydromagnetics. The two papers here commented were the outcome of those discussions. The element of novelty in these papers lies in the introduction of a magnetic field as a fundamental ingredient in the study of of the behaviour of cosmic masses. In the first paper, from the study of the plane of polarization of light one arrives at estimating the magnitude of the ordered component of the galactic magnetic field in the spiral arms. In the second paper the gravitational stability is studied in the presence of a magnetic field, by extending the virial theorem in order to take the forces due to a magnetic field into consideration. The prediction that a gravitating fluid sphere under the influence of a strong magnetic field tends to become unstable and highly oblate by contracting along the axis of symmetry of the field assumes great interest with relation to present problems inherent to neutron stars with strong magnetic fields and emission of gravitational waves. Today it may be also interesting to extend the scheme of Fermi and Chandrasekhar to consider an electric field instead of a magnetic field and then to state for this case the necessary condition for stability. Possible connections with a Reissner-Nordstrom black hole are suggested.

PACS 98.58.Ay – Physical properties (abundances, electron density, magnetic fields, scintillation, scattering, kinematics, dynamics, turbulence, etc.).
PACS 01.30.Cc – Conference proceedings.

[*] Paper presented at the IX ICRA Network Workshop "Fermi and Astrophysics" (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.

[**] E-mail: boccaletti@uniroma1.it

Introduction

Before undertaking to talk about the two Fermi's papers I quote the *incipit* of the Henry Norris Russel Lecture of the American Astronomical Society delivered by Fermi on August 28, 1953: "I became interested in the possible existence of magnetic fields extending through the volume of the galaxy in connection with a discussion on the origin of the cosmic radiation a few years ago"(1). Here Fermi alludes to his famous paper, dating back to 1949, on the origin of cosmic rays [2]. But this sentence, we can say, contains the "philosophy" which always informed Fermi's work. He always starts from concrete and well definite problems to arrive at a theory which explains them. He was not inclined, as those who were familiar with him have reported, to great generalizations.

Fermi was the first to propose that the disc of the galaxy contains a general large-scale magnetic field. The proposal was originated from the existence of cosmic rays having individual energies of typically 1–100 GeV per nucleon. Fermi suggested field strengths of the general order of 10^{-5}–10^{-6} gauss. In the same year (1949), a systematic polarization of the light of the distant stars was discovered by Hiltner and Hall [3, 4]. They found that the polarization was strong and aligned more or less along the plane of the Milky Way for stars lying in the general direction of the galactic centre or anticentre. We must point out that at that time the idea that stars might possess a magnetic field was not so current. In fact, the idea that the Sun had a magnetic field was proved only in 1908 when Hale used the Zeeman effect to prove the existence of magnetic fields in sunspots. Much more later (1951) the difficult task of detecting magnetic fields in stars was undertaken by Babcock who developed equipment of sufficient sensitivity [5]. And just the results of Babcock are quoted in the second of Fermi-Chandrasekhar's papers we are going to comment.

As we said, Fermi, starting from the problem of cosmic rays, went so far as to be interested in the processes occurring in the galaxy and to transform what could be a simple cultural interest in a field of research. As reported by Chandrasekhar in the comments to these two papers in the edition of Fermi's Collected Papers, he regularly met Chandrasekhar in the fall and winter of 1952 and in the spring of 1953 for discussing a variety of astrophysical problems bearing on hydromagnetics and the origin of cosmic radiation. The two papers we will speak about are the outcome of those discussions.

1. – Magnetic fields in spiral arms

The first paper "Magnetic Fields in Spiral Arms" (*Astrophys. J.*, **118** (1953) 113-115) gives an estimate of the magnetic field in the spiral arm in which we are. Two methods are used. The first one is based on an interpretation of the dispersion in the observed plane of polarization of the light of the distant stars. The mean angular deviation of the plane of polarization from the direction of the spiral arm had been found by Hiltner (1951) to be about 0.2 radians. Fermi and Chandrasekhar connected this angle with the magnetic field H, the root-mean-square velocity v of the turbulent motion of the gas masses in the spiral arm and the density ρ of the gas. They found

$$H = \left(\frac{4}{3}\pi\rho\right)^{1/2}\frac{v}{\alpha}. \tag{1}$$

[1] The text of the conference was published in ref. [1].

With $\rho = 2 \times 10^{-24}$ g/cm ^{3}, $v = 5 \times 10^5$ cm/s, $\alpha = 0.2$ radians, one has

$$H = 7.2 \times 10^{-6}\text{gauss}.$$

The second method, more direct, considers the spiral arm as a cylinder of gas of uniform density and requires that, for the stability, at any point

$$P_{\text{grav}} = P_{\text{kin}} + P_{\text{mag}}. \tag{2}$$

That is, on each element of the gas, the gravitational pressure must be counterbalanced by the pressure of the turbulent motion of the gas and by the pressure exerted by the magnetic field assumed as parallel to the axis of the cylinder and of uniform strength. Assuming a radius of the spiral arm of 250 parsecs (7.7×10^{20} cm), they obtained

$$H = 6 \times 10^{-6}\text{gauss}.$$

Therefore, the two independent methods of estimating H agreed in giving essentially the same value for the field strength. Today's estimates for H continue to be of few μgauss. Thus the pioneering work of Fermi and Chandrasekhar provided a practically correct estimate.

2. – Gravitational stability in the presence of a magnetic field

The second paper "Problems of gravitational stability in the presence of a magnetic field" (*Astrophys. J.*, **118** (1953) 116-141), approaches a more general problem since it seeks the necessary condition for the gravitational stability of cosmic masses (assumed to have infinite electrical conductivity) in which there is a prevalent magnetic field. The condition is based on what is known in celestial mechanics as Jacobi stability criterion:

> A necessary condition for the stability is $E_{\text{tot}} < 0$.

Obviously the problem is that of writing such a condition in a significant form. Fermi and Chandrasekhar resort to a simplified form of the virial theorem. Actually, as already Chandrasekhar did in his book of 1939 "An introduction to the study of stellar structure", following Eddington (1916), what they call "virial theorem" is the Lagrange-Jacobi identity in which the total moment of inertia of the system is assumed constant([2]). Essentially, one obtains a relation which links the kinetic energy of the macroscopic motion, the internal energy of the gas, the gravitational potential energy and the magnetic energy:

$$2\mathcal{T} + 3(\gamma - 1)\mathcal{U} + \mathcal{M} + \Omega = 0, \tag{3}$$

where $\mathcal{T}$ is the kinetic energy of mass motion, $\mathcal{U}$ the heat energy of molecular motion, $\mathcal{M}$ the magnetic energy of the prevailing field, Ω the gravitational potential energy and γ denotes the ratio of specific heats. This relation concerns the instantaneous values of the various types of energy, whereas the Clausius virial theorem concerns their time averages on a very long time ($\longrightarrow \infty$). To have stability, the kinetic energy of the macroscopic

([2]) For the Lagrange-Jacobi identity and the virial theorem see, for instance, ref. [6].

motion must vanish and the total energy must be negative. This bring us to obtain the condition

$$(3\gamma - 4)(|\Omega| - \mathcal{M}) > 0, \tag{4}$$

which, for $\gamma > 4/3$, becomes $\mathcal{M}/|\Omega| < 1$.

Fermi and Chandrasekhar then study the condition in the case of two important configurations: an infinite cylinder and a sphere. It is the latter the case we consider more interesting for the developments induced and the generalizations one can obtain. Thus we will dwell on it. The authors assume an incompressible fluid sphere with a uniform magnetic field inside and a dipole field outside and investigate the behaviour of the sphere under a perturbation, so they write the equation of the deformed bounding surface in the form

$$r(\cos\vartheta) = R + \varepsilon P_\ell(\cos\vartheta), \tag{5}$$

where R is the radius of the unperturbated sphere and $P_\ell(\cos\vartheta)$ denotes the Legendre polynomial of order ℓ. The perturbation parameter obviously satisfies $\varepsilon \ll R$. They find that the perturbation of the first order in ε is due to the only P_2-deformation and has a negative sign. Therefore the deformation is in the sense of making the sphere into an oblate spheroid and the final result is given by

$$\frac{\varepsilon}{R} = -\frac{35}{24}\frac{H^2 R^4}{GM^2}, \tag{6}$$

M being the mass and G the gravitational constant. If we roughly approximate the spheroid through an ellipsoid, averaging on the angle ϑ, we obtain

$$r \sim R + \frac{1}{4}\varepsilon = R\Big(1 + \frac{1}{4}\frac{\varepsilon}{R}\Big). \tag{7}$$

If we call a, b, c, the three semi-axes of the ellipsoid, with $a = c$ and b the semi-minor axis in the direction of the magnetic field, we can also write

$$b \sim a\Big(1 - \frac{1}{2}e^2\Big) \tag{8}$$

$e \ll 1$ being the eccentricity. Rewriting the relation (7) as

$$r \sim R\Big(1 - \frac{1}{4}\frac{|\varepsilon|}{R}\Big) \tag{9}$$

and identifying $a = c$ with R, we get $|\varepsilon| = 2Re^2$ and finally

$$e^2 = \frac{35}{48}\frac{R^4}{GM^2}H^2, \tag{10}$$

thus the eccentricity goes linearly with the magnetic field. To consider an interesting example, for $R = 10^6$ cm and $M = 1.4M_\odot$ (the case of a pulsar), the last relation gives

$$e^2 = 0.7 \times 10^{-36} H^2$$

from which $e \sim 10^{-6}$, for $H \sim 10^{12}$ gauss,
$e \sim 10^{-4}$, for $H \sim 10^{14}$ gauss,
$e \sim 10^{-3}$, for $H \sim 10^{15}$ gauss.

With regard to the last relation some considerations can be made. At the time when the pulsars had just been discovered, A. Ferrari and R. Ruffini [7] calculated (through a method different from Fermi-Chandrasekhar's one) which strength of the magnetic field was required to get oblate the spherical neutron star. They found that a field of 10^{15} gauss was necessary to obtain an $\varepsilon \sim 10^{-4}$. The parameter of Ferrari and Ruffini was defined as $\varepsilon = (a-b)/\sqrt{ab}$ and in our case $\varepsilon \sim 10^{-4}$ corresponds to $e \sim 10^{-2}$. Obviously this value was considered unrealistic (we remind that for a pulsar the value currently accepted was $H \sim 10^{12}$ gauss) and the matter was let drop. Today it seems possible that stars with a very strong magnetic field ($H \sim 10^{15}$ gauss) exist, and in this case there may be the prospect of generation of gravitational waves and peculiarities in the dynamics of the stars themselves (precession, etc.)([3]).

3. – Gravitational stability in the presence of an electric field

The seminal character of the paper we are considering can be witnessed also applying the scheme of Fermi and Chandrasekhar by replacing the magnetic field by an electrostatic field. This means to replace the magnetic Maxwell stress tensor

$$\sigma_{ik} = \frac{1}{4\pi}\left(H_i H_k - \frac{1}{2}\delta_{ik} H^2\right)$$

by the electric one

$$\sigma_{ik} = \frac{1}{4\pi}\left(E_i E_k - \frac{1}{2}\delta_{ik} E^2\right)$$

and to put the relevant force components $f_i = \partial\sigma_{ik}/\partial x_\ell$ in the motion equations. Thus they can be written as

$$\rho\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\nabla\left(P + \frac{\mathrm{E}^2}{8\pi}\right) + \rho\nabla V + \frac{1}{4\pi}\nabla\cdot(\mathbf{EE}), \tag{11}$$

where V denotes the gravitational potential and E the intensity of the electric field. Proceeding in the standard way, we denote by $\mathbf{r}$ the position vector of a fluid particle so that $\mathbf{v} = \mathrm{d}\mathbf{r}/\mathrm{d}t$ and multiply eq. (11) scalarly by $\mathbf{r}$ and integrate over the volume τ of the fluid. On integrating by parts, the left-hand side of the resulting equation becomes

$$\int_\tau \rho\mathbf{r}\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}\mathrm{d}\tau' = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\int_\tau \rho\,|\mathbf{r}|^2\,\mathrm{d}\tau' - \int_\tau \rho\,|\mathbf{v}|^2\,\mathrm{d}\tau' \tag{12}$$

and can be immediately rewritten as

$$\frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} - 2\mathcal{T}, \tag{13}$$

[3] See the paper of L. Stella in the IX ICRA Proceedings (World Scientific, 2003) and the references therein.

I being the total moment of inertia and $\mathcal{T}$ the total kinetic energy. Therefore we have

$$\frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} - 2\mathcal{T} = -\int_\tau \mathbf{r}\cdot\nabla\Big(P + \frac{\mathrm{E}^2}{8\pi}\Big)\mathrm{d}\tau' + \frac{1}{4\pi}\int_\tau \mathbf{r}\cdot\nabla(\mathbf{EE})\mathrm{d}\tau' + \int_\tau \rho\mathbf{r}\cdot\nabla V\mathrm{d}\tau'. \tag{14}$$

For the terms on the right-hand side, we obtain

$$\begin{aligned} & -\int_\tau \mathbf{r}\cdot\nabla\Big(\frac{\mathrm{E}^2}{8\pi}\Big)\mathrm{d}\tau' + \frac{1}{4\pi}\int_\tau \mathbf{r}\cdot\nabla(\mathbf{EE})\mathrm{d}\tau' = \\ & = -\int_{\mathbf{S}} \frac{\mathrm{E}^2}{8\pi}\mathbf{r}\cdot\mathrm{d}\mathbf{S} + 3\int_\tau \frac{\mathrm{E}^2}{8\pi}\mathrm{d}\tau' + \frac{1}{4\pi}\int_{\mathbf{S}}(\mathbf{r}\cdot\mathbf{E})\mathbf{E}\cdot\mathrm{d}\mathbf{S} - \frac{1}{4\pi}\int_\tau E^2\mathrm{d}\tau' = \\ & = \frac{1}{8\pi}\int_\tau e^2\mathrm{d}\tau' = \mathcal{E}, \end{aligned} \tag{15}$$

S being the external boundary of the distribution and $\mathcal{E}$ the energy of the electric field (surface integrals vanish); in addition

$$-\int_\tau \rho\mathbf{r}\cdot\nabla V\mathrm{d}\tau' = \Omega, \tag{16}$$

the gravitational potential energy. Finally

$$\begin{aligned} & -\int_\tau \mathbf{r}\cdot\nabla P\mathrm{d}\tau' = -\int_{\mathbf{S}} P\mathbf{r}\cdot\mathbf{S} + 3\int_\tau P\mathrm{d}\tau' = \\ & = 3\int_\tau P\mathrm{d}\tau' = 3\int_\tau \mathrm{R}\rho T\mathrm{d}\tau' = 3(\gamma-1)\int_\tau c_v\rho T\mathrm{d}\tau' = \\ & = 3(\gamma-1)\mathcal{U}, \end{aligned} \tag{17}$$

the pressure being considered vanishing on the boundary surface. Therefore eq. (14) results in

$$\frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} = 2\mathcal{T} + 3(\gamma-1)\mathcal{U} + \Omega + \mathcal{E}. \tag{18}$$

If we consider our system being in a steady state, $\mathrm{d}^2I/\mathrm{d}t^2 = 0$ and, if we exclude bulk motions of matter, also $\mathcal{T} = 0$ and we have that the equation

$$3(\gamma-1)\mathcal{U} + \Omega + \mathcal{E} = 0 \tag{19}$$

must hold. Our system being isolated,

$$E_{\mathrm{tot}} = \mathcal{U} + \Omega + \mathcal{E} = \mathrm{const} \tag{20}$$

and we can write condition (19) as

$$E_{\mathrm{tot}} = -\frac{(3\gamma-4)(|\Omega| - \mathcal{E})}{3(\gamma-1)}. \tag{21}$$

The necessary condition for stability (Jacobi criterion) requires $E_{\text{tot}} < 0$, therefore, for $\gamma > 4/3$,

$$|\Omega| - \mathcal{E} > 0, \tag{22}$$

or

$$\frac{\mathcal{E}}{|\Omega|} < 1. \tag{23}$$

If the system consists of a sphere of uniform mass density and uniform charge density, and M, Q and R are its total mass, total charge and radius, respectively, we know that

$$\Omega = -\frac{3}{5}G\frac{M^2}{R}\,, \qquad \mathcal{E} = \frac{3}{5}G\frac{Q^2}{R}.$$

In units $G = 1$ the inequality (23) results in

$$\frac{Q^2}{M^2} < 1. \tag{24}$$

This is the necessary condition for stability of a sphere of gas of total mass M and total charge Q.

Obviously, as far as we know, electric stars do not exist but there are other systems that can be compared with a charged sphere. If we consider a Reissner-Nordstrom black hole at a great distance (where the space-time is nearly flat), we know that a condition exists which rules that the black hole cannot accept any further charge (extreme black hole). This condition (in $G = 1$ units) is given by $Q^2/M^2 = 1$[4]. Therefore the necessary condition for the stability of a classical charged sphere is the same as the condition for a Reissner-Nordstrom black hole not to become extreme.

REFERENCES

[1] Fermi E., *Astrophys. J.*, **119** (1954) 1.
[2] Fermi E., *Phys. Rev.*, **75** (1949) 1169.
[3] Hiltner W. A., *Astrophys. J.*, **109** (1949) 471.
[4] Hall J. S., *Science*, **109** (1949) 166.
[5] Babcock H. W., *Astrophys. J.*, **114** (1951) 1.
[6] Boccaletti D. and Pucacco G., *Theory of Orbits*, Vol. **1** (Springer Verlag) 2000, Chapt. 3, 2nd edition.
[7] Ferrari A. and Ruffini R., *Astrophys. J.*, **158** (1969) 71.
[8] Misner C. W., Thorne K. S. and Wheeler J. A., *Gravitation*, Part VII (Freeman and Company) 1973.

[4] See, for instance, ref. [8].

B.4 A. Carati, L. Galgani, A. Ponno, A. Giorgilli: The Fermi-Pasta-Ulam problem and the question of the rates of thermalization

A. Carati, L. Galgani, A. Ponno, A. Giorgilli: "The Fermi-Pasta-Ulam problem and the question of the rates of thermalization," *Nuovo Cimento B* **117**, 1017 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

The Fermi-Pasta-Ulam problem(*)

A. Carati(1)(**), L. Galgani(1)(***), A. Ponno(1)(⁂) and A. Giorgilli(2)(⁂)

(1) *Dipartimento di Matematica, Università di Milano*
Via Saldini 50, 20133-Milano, Italy
(2) *Dipartimento di Matematica e Applicazioni, Università di Milano Bicocca*
Via degli Arcimboldi 8, 20126-Milano, Italy

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — A review is given of the Fermi-Pasta-Ulam problem. Its foundational relevance in connection with the relations between classical and quantum mechanics is pointed out, and the status of the numerical and analytical results is discussed.

PACS 05.45.-a – Nonlinear dynamics and nonlinear dynamical systems.
PACS 01.30.Cc – Conference proceedings.

1. – The FPU model and the FPU problem

The Fermi-Pasta-Ulam model is a system of $N+2$ equal particles on a line with mutual interactions between adjacent particles, provided by a potential of the form $V(r) = r^2/2 + \alpha r^3/3 + \beta r^4/4$; certain boundary conditions are also assigned, typically with the two extreme particles fixed. For $\alpha = \beta = 0$ the system is a linear one, and by a familiar linear transformation it can be reduced to a system of N independent harmonic oscillators (called normal modes) with certain frequencies $\omega_j = 2\sin[j\pi/2(N+1)]$, $j = 1, \cdots, N$. The total energy E then reduces to the sum $E = \sum_j E_j$ of the N normal mode energies E_j, which are independent integrals of motion: $E_j(t) = E_j(0)$. When the nonlinear interaction is active, the normal mode energies are no more integrals of motion, and a standard arguments of classical statistical mechanics suggests that their

(*) Paper presented at the IX ICRA Network Workshop "Fermi and Astrophysics" (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.
(**) E-mail: carati@mat.unimi.it
(***) E-mail: galgani@mat.unimi.it
(⁂) E-mail: ponno@mat.unimi.it
(⁂) E-mail: antonio@matapp.unimib.it

time averages $E_j^*(t) = (1/t) \int_0^t E_j(\tau) \mathrm{d}\tau$ should tend to a common value, thus realizing what is usually called the equipartition of energy. More precisely this is expected to occur for almost all initial data with respect to the Gibbs measure, and in the thermodynamic limit, *i.e.* the limit of an infinite system with a finite nonvanishing specific energy ϵ: $N \to \infty, E \to \infty, E/N \to \epsilon > 0$. In such a case, the common value of the time averages of the normal mode energies is identified with the temperature T by $E_j^*(t) \to k_\mathrm{B}T$, where k_B is the Boltzmann constant. Correspondingly, the specific heat (defined as the derivative of energy with respect to temperature) turns out to be a constant, independent of temperature.

The FPU problem consists in establishing whether the dynamics actually leads to equipartition. Typically one considers initial data with the energy given just to some low frequency modes, and one looks for the rate of thermalization *i.e.* for the rate at which energy flows to the high frequency modes. Such a problem was first investigated by Fermi, Pasta and Ulam in the year 1954, through numerical solutions of the equations of motion for $N = 64$, using the facilities of the Los Alamos laboratory [1]. The result they found is that, at least up to the actually observed times, the energy, initially given to the lowest frequency mode, did not appear to flow at all to the highest frequency modes, but was just shared among a small group (or packet) of low frequency modes. The "final" distribution of energy also appeared to decrease more or less exponentially fast with the frequency. The results did not change qualitatively if the initial energy was given not just to the lowest frequency mode, but to a small packet about it.

2. – The significance of the FPU problem: the FPU paradox

The question of the equipartition of energy has a deep foundational meaning for physics, because it is the one that gave rise to quantum mechanics. Indeed, as everyone knows, equipartition of energy (*i.e.* mean energy independent of frequency, and specific heat independent of temperature) is experimentally found to obtain only in the limit of high temperatures and/or low frequencies, and to completely fail in the complementary region. It is actually at this point Planck's constant $\hbar$ entered the game, because it was found by Planck, on October 19, 1900, by fitting the experimental data of black body radiation, that the relevant dimensionless parameter is the quantity $\hbar\omega/k_\mathrm{B}T$, and that the distribution of energy (per oscillator) U *vs.* frequency ω at temperature T is

$$U(\omega, T) = \frac{\hbar\omega}{e^{\hbar\omega/k_\mathrm{B}T} - 1} = kT\,\frac{x}{e^x - 1} \qquad (x = \hbar\omega/k_\mathrm{B}T)\,.$$

Now, as shown by Planck in his second memoir and described in all textbooks, Planck's law is obtained by the usual arguments of statistical mechanics if energy is assumed to be quantized. In particular, for a harmonic oscillator the admitted values of the energy should be $E_n = n\hbar\omega$, $n = 1, 2, \cdots$ (or rather $E_n = (n + 1/2)\,\hbar\omega$, which leads to the addition of the "zero-point energy" $1/2\hbar\omega$). If energy is not quantized, one instead recovers the "classical" equipartition value $U(\omega, T) = k_\mathrm{B}T$. Thus the FPU result appeared as a paradox.

By the way, in our opinion it is not by chance that Fermi happened to study this problem. Indeed his interest for the problem of equipartition of energy goes back to his youth, as is witnessed by the work of the year 1923 in which he had given a subtle mathematical improvement to a theorem of Poincaré [2] (see also [3]). Poincaré was concerned with the number of integrals of motion for a Hamiltonian system, and had proven that “in general” there is just one integral, namely the total energy. Notice that this is a crucial point in connection with the problem of equipartition of energy, because for example in the FPU problem there are N integrals of motion for the linearized system, and one should understand in which mathematical sense can one pass, with the introduction of a nonlinearity, to a situation in which one remains instead with just one integral of motion. The subtle mathematical point addressed by Fermi consisted in paying attention not to the integrals of motion themselves, but rather to single invariant surfaces in the phase space. In this connection let us recall that if there are N independent integrals “in involution”, then the phase space is “foliated” by a continuous set of single invariant N-dimensional surfaces, which in the compact case turn out to be tori. This is actually the frame in which Kolmogorov (just in the same year 1954 of the work of Fermi, Pasta and Ulam) formulated his celebrated theorem, now usually known as the KAM theorem, and it is impressive how a physicist, as Fermi was, might have thought of the problem in such terms. In any case, the theorem of Poincaré was universally interpreted as a dynamical support to the idea that in a generic Hamiltonian system all integrals of motion which are possibly present in an unperturbed system (such as the linearized FPU model) should disappear with the introduction of a generic perturbation; in our case, this would lead to equipartition of energy. The interest for this mathematical problem was clearly the reason for Fermi coming back to the equipartition problem when he happened to have a large computer available. To this historically documented fact, we can add a personal impression, that was formed by a conversation that one of us had several years ago with the late E. Segré (one of Fermi’s pupils) in Berkeley. What Segré said is that in private conversations with friends Fermi would admit that he did not really like the Copenhagen interpretation of quantum mechanics. So, perhaps, the opportunity of going back to the problem in which quantum mechanics itself had originated might have been particularly attractive for him.

3. – The way out of Izrailev and Chirikov; the problem of the energy thresholds

All of us have learned at school that classical mechanics predicts a wrong result (equipartition) in connection with the problem of the distribution of energy for a system of harmonic oscillators, this being exactly the point where the new quantum mechanics originated. So the result of Fermi, Pasta and Ulam appears as a paradox, and it is expected that one should be able to explain and eliminate it by a deeper scrutiny of the problem.

A fundamental contribution in this direction was given by a very deep paper of Izrailev and Chirikov [4] of the year 1966. The main idea was that one should take into account the existence of some energy threshold. Actually, this is a familiar fact in perturbation theory, and possibly it came to the minds of the authors because they were the first physicists that made a connection at all between the FPU problem and the modern results (*i.e.* the KAM theorem) of perturbation theory. The point is that perturbation theory applies at all, *i.e.* the perturbed system is proven to be qualitatively similar to the unperturbed one (there are however some delicate points here, to which we will come back in another section), only if the perturbation is low enough, namely if it is smaller than a certain threshold, which should be suitably estimated in each particular case. In the FPU model the unperturbed system is the linearized one, with its N independent integrals, and what plays the role of the perturbation is essentially the total energy, because the relative “size” of the nonlinearity tends to zero with the total energy. So the idea was that the perturbed system would be qualitatively similar to the unperturbed one only for energies smaller than a certain critical energy, say E^{c}; the result of Fermi, Pasta and Ulam should then be explained as due to the fact that they had considered small initial energies, below the threshold E^{c}. Finally, the paradox would disappear at all if one might prove that the threshold vanishes in the thermodynamic limit.

Actually, the authors even went farther, because they also provided, with some heuristic arguments, an estimate for the threshold. More precisely, they also conceived the idea that the threshold should depend on the choice of the initial data: for an initial excitation of mode j one should correspondingly have a threshold E^{c}_j. So they gave their estimate for the threshold E^{c}_j, or at least two limit expressions for it for the case of low frequencies (small j) and for the case of high frequencies (large j). The key point is now the dependence on the number N of particles because, according to their estimates, the energy thresholds would tend to zero, at least for the case of large j. This result almost eliminates the paradox, because, at least for initial data with excitations of the high frequency modes, in the thermodynamic limit one would always be above threshold, *i.e.* the system would have almost no relation to the unperturbed one and thus would be expected to lead to equipartition.

The job would be totally accomplished if one were able to produce estimates for the thresholds in the case of small j, presenting the same property of vanishing in the thermodynamic limit. In a recent paper by Shepelyansky [5] it is stated that such an accomplishment has now been performed. In his words, the aim of his paper is the following one: "*A possibility that in the FPU problem the critical energy for chaos goes to zero when the number of particles in the chain increases is discussed*". In the introduction the result of Izrailev and Chirikov is mentioned: "*According to Izrailev and Chirikov, in the case of low-mode excitation (nonlinear sound waves) the critical energy increases with the number of oscillators in the chain (or the energy per oscillator is constant)*". It is then discussed how such authors had neglected to take into account certain resonances in their semianalytical estimates, with the conclusion: "*Such resonances not being considered by Izrailev and Chirikov give a sharp decrease of the chaos border in energy which goes to zero with the increase of the number of particles in the lattice. In this sense the long-wave chaos can exist for arbitrarily small nonlinearity*".

4. – The result of Bocchieri, Scotti and Loinger, and the Planck-like distributions

A completely different perspective had been however imagined in the meantime. This occurred in Milano, through the work of Bocchieri, Scotti and Loinger [6] of the year 1970 and a subsequent one [7]. The theoretical group of Milano was then led by Caldirola, who had been among the young physicists influenced by Fermi just before the war, and happened to be particularly interested in problems concerning the foundations of theoretical physics. So in Milano people were particularly sensitive to the possible foundational relevance of the FPU problem, with its implications for the relations between classical and quantum mechanics. Thus, when the FPU problem made its way to Italy through the work of Izrailev and Chirikov, that had been discovered by Loinger, there naturally arose the idea of checking, by numerical solutions of the equations of motion, whether the specific critical energy $\epsilon^{\mathrm{c}} = E^{\mathrm{c}}/N$ vanishes in the thermodynamic limit or not. This was done by Bocchieri, Scotti and Loinger. Actually, the critical energy was investigated just for initial data as in the original work of Fermi, Pasta and Ulam, namely for initial excitations of the lowest mode. The energy threshold was defined in the simplest possible way, by computing the time averages $E^{*}_j(t)$ until they had apparently settled down to some "final" stationary value. It turned out that there was some critical energy above which there was an apparent equipartition, while equipartition did not occur at lower energies. The main result was that the critical energy $E^{\rm c}$ appeared to be proportional to N, *i.e.* there appeared to exist a finite nonvanishing specific critical energy $\epsilon^{\rm c} > 0$. In other terms, it was suggested that according to classical dynamics there is a relevant set of initial data which lead to "final" states not corresponding to equipartition of energy. An interesting fact concerning this work is that the computations were performed with a realistic interatomic potential, namely a standard Lennard–Jones potential $V(r) = 4V_0\big[(\sigma/r)^{12} - (\sigma/r)^6\big]$; this involves two parameters, V_0 and σ, giving the depth and the width respectively of the potential well. The computations were actually performed by taking for the molecular parameters m (the mass of the particles), V_0 and σ realistic values corresponding to Argon as obtained from standard textbooks. The specific critical energy $\epsilon^{\rm c}$ turned out to have a value which is more or less 4 percent of the depth V_0 of the potential well.

Many discussions followed this striking result. Shortly later, an investigation was made [7] of the distribution of energy for the 'final" states that are found below threshold (still with initial data of FPU type), looking for a function that gives the final values E^*_j in terms of the corresponding frequencies ω_j. It was found that the curves were rather well fitted (apart from a short plateau at the very low frequencies) by Planck-like distributions of the form

$$E^*(\omega, E) = \frac{A\omega}{e^{\beta A\omega} - 1}\,.$$

The parameter β was depending on the total energy E more or less as an inverse temperature should, while the parameter A appeared to be a constant. The most striking fact was that, with the realistic values of the molecular parameters corresponding to Argon which had been chosen in the computations, the quantity A turned out to have a value very near to that of Planck's constant. It took some time to understand this point: in brief, Planck's constant had been introduced, somehow by hands, through the molecular parameters. This goes as follows. One immediately checks that the natural action built up from the parameters is just $\sqrt{mV_0}\,\sigma$, so that one has $A = a\sqrt{mV_0}\,\sigma$ with a pure number a; on the other hand, from the textbooks, it turns out that one has, in an incredibly precise way, $\sqrt{mV_0}\,\sigma \simeq 2Z\hbar$, where Z is the atomic number. Thus the numerical computations had just provided an estimate of the pure number a, which turned out to be of the order of magnitude of $1/50$. Shortly later an interesting contribution was also given by Cercignani, who suggested [8] that there might be an analogy between energy thresholds and quantum zero-point energy.

5. – The problem of the relaxation times; old and modern aspects

One thus remains with the problem of deciding between two possible alternatives. Denote by "freezing" the FPU qualitative phenomenon that the "final" distribution of energy is near the initial one, so that in particular equipartition does not hold for the final distribution if the initial one corresponds to excitation of the very low frequency modes; such a freezing is expected to hold below some energy threshold. Then the alternative is whether such a freezing persists in the thermodynamic limit or not, *i.e.* whether it is relevant for physics or not. At first sight, one might be tempted to say that the school of Chirikov would bet for the second alternative, and the group of Milano for the first one. But perhaps we are misinterpreting our colleagues, and their hopes might be not so dissimilar from ours.

Passing from hopes to facts, or to theorems, it turns out that the answer is not at all simple, and actually has not yet been afforded. The main difficulty resides in providing a clear definition for the freezing, especially in connection with the question of the times involved (*i.e.* the size of the relaxation times in relation to the observation times). Another point concerns the meaning that should be attributed to the notion of an “energy per oscillator”. These are indeed quite delicate problems on which we are presently actively working, and we limit here ourselves to some comments, mainly addressed to the first problem, *i.e.* that of the times involved.

The general physical problem of the dependence of the results on the observation time turns out to have a a strong counterpart in perturbation theory. Indeed, in general, in perturbation theory one aims at proving that a certain system is “similar” to another “ unperturbed” one, but the mathematical implementation of such an idea requires that preliminarily a time t should be fixed up to which the similarity should hold; such a time is the counterpart of the physical observation time. On the other hand, it is a general fact that the similarity can be proven to exist only if the perturbation is below a certain threshold, so that correspondingly the threshold turns out to depend on the given observation time.

Now, the KAM theorem refers to an infinite observation time, and all the available estimates indicate that it should not apply in the thermodynamic limit (we do not discuss here the problem of the existence of invariant low-dimensional tori, which is now so popular in the mathematical literature, and is studied for example by Kuksin and by Bambusi). One can instead make reference to finite times, and Nekhoroshev [9] has thought us [10] how to deal with them in a perspicuous efficient way (see also [11]). Many numerical and analytical studies have been performed from this point of view on the FPU and related models [12], on which we do not have time to enter. What we want to stress here is that finally the scientific community seems to have come to agree that taking into account the observation time is a physically relevant requirement even in the FPU and related problems. Asking whether one has equipartition or not without an accurate discussion of the times involved is extremely naive and unphysical.

By the way, it has also been realized that the relevance of the observation time in this connection was actually well known since the “old times” , because it was Boltzmann himself [13] that for the first time conceived that the phenomenological lack of equipartition in crystals and polyatomic molecules could be explained as corresponding to the fact that equipartition had not been achieved within the actual observation times; the relaxation times to equipartition would be much longer than the experimental observation times. This idea was pursued by Jeans [14], and then discussed at the first Solvay conference [15], particularly by Nernst, who declared that such long relaxation times had never been observed in experiments. The story has been described elsewhere [16]. In short, “long” relaxation times (even of the order of one second) have actually been observed in the phenomenon of the dispersion and anomalous absorption of sound in diatomic molecules [17], and is rather well accounted for by the classical theory, although some delicate problems are still open. Moreover, the circumstance that one should have some “time dependent specific heat” is presently accepted even as a trivial fact [18]. The situation is however rather delicate, especially in connection with the problem of understanding from this point of view the standard static measurements of the specific heats.

The discussion would become here really intricate, and at the moment we do not have a clear answer available. We have however a main qualitative perspective. The general idea is that in classical models of crystals and of polyatomic molecules one might meet with situations qualitatively analogous to those which are met in the phenomenology of glasses, spin glasses and polymers, where an essential role is played by the fact that there exist relaxation times differing from each other by huge orders of magnitude. So, one might have an essentially rapid relaxation to some kind of metaequilibrium state, which should last for an extremely long time; the final relaxation to a standard Maxwell-Boltzmann equilibrium, and thus to equipartition, might then occur only over such huge time scales. Something like this was suggested for the first time in a work [19], where the interaction of a FPU system with a heat reservoir was studied numerically. Such an idea had also been pursued for the case of polyatomic molecules [20]. A special attention had there been given to the so-called Landau-Teller model [21, 22], which takes there the role of the FPU model. In particular, it was recently observed that in the Landau-Teller model of molecular collisions the energy of the internal vibrations performs a kind of random walk in which there occur rare conspicuous jumps, somehow analogous to those occurring in Lévy processes [23]. One should then meet there with the phenomenon of the anomalous diffusion, which might thus be expected to occur also in the FPU model.

In connection with the physical necessity of taking into account the observation times, one also meets with a quite delicate problem of interpretation pointed out by Boltzmann. The problem concerns the identification, which is usually made in statistical mechanics, between thermodynamic energy and mechanical energy. According to Boltzmann, one should declare in advance which is the chosen observation time. Then, considering a system which possesses a certain mechanical energy in virtue of some initial conditions, its thermodynamic energy should be identified with the fraction of the mechanical energy that the system can actually exchange with the measurement instrument up to the given observation time. As pointed out by Nernst [24] (see also [25]), in such a way one might have a situation in which there is equipartition of energy for the mechanical energy, just in virtue of the choice of the initial data (according to the Maxwell-Boltzmann distribution), and instead a Planck distribution for the exchangeable energy (see also [26]). This is a crucial point if one wants to interpret the phenomenon of the freezing of the high frequencies modes when one considers initial conditions of a generic type (*i.e.* according to Maxwell-Boltzmann), and not just of the special type corresponding to an excitation of the low frequency modes.

6. – Some recent results

We now quickly describe some recent results, which are mostly still unpublished. The first one is of numerical type, and is already in print [27]. For initial data of the FPU type, a strong evidence is given of the fact that the results depend on the specific energy ϵ in the following way. There exists a critical specific energy ϵ^{c} such that for $\epsilon > \epsilon^{\mathrm{c}}$ equipartition is obtained within a time that increases as an inverse power of ϵ with decreasing ϵ. Instead, below threshold, *i.e.* for $\epsilon < \epsilon^{\mathrm{c}}$, one meets with two time scales: in a short time there is formed a "natural packet" that extends up to a maximal frequency $\overline{\omega}(\epsilon)$ proportional to $\epsilon^{1/4}$. Only on a much longer time scale would one get equipartition. Just in these days indications are being found that such a large time scale might increase as a stretched exponential of $1/\epsilon$.

The natural packet mentioned above is presumably to be identified with what in the year 1972 was considered to be the "final state" providing a Planck–like distribution. So it is of particular interest to obtain any possible analytical information about it. This has now been afforded, and will in a short time be written down. By arguments related to the description of the FPU model in terms of solitons, along the lines of the celebrated work of Zabusky and Kruskal [28] (see also [29]), it is shown that the results actually depend on the specific energy, and an explicit analytical formula is given for the natural packet, which is confirmed to extend up to a maximal frequency $\overline{\omega}(\epsilon)$ proportional to $\epsilon^{1/4}$. The analytical form of the packet is found to fit in an extremely good way the numerical data.

Some progress was also made in the direction of getting rigorous analytical results in the thermodynamic limit, because for the first time it has been possible to perform a finite number of perturbative steps in that limit. This required the establishment of a suitable measure-theoretic framework for perturbation theory itself, and in particular a clarification of what should be meant by "energy per oscillator", in that limit.

7. – Final comments

So, let us come back to the problem of deciding between the two alternatives, namely whether the original FPU result is relevant for physics or not. In the light of the recent results just mentioned, we are confident that the FPU paradox cannot be eliminated and that it has a deep physical meaning.

The general perspective mentioned above naturally leads to the following interpretation of the paradox. Before Fermi, Pasta and Ulam the alternative was between classical mechanics, which should be wrong, and quantum mechanics, which is correct. But this makes no reference to times. We would instead suggest: up to "short" times classical mechanics might qualitatively agree with quantum mechanics, and only later on might they differentiate. Indeed, according to quantum mechanics Planck's law is the final equilibrium distribution, while, apparently, according to classical mechanics it might just describe a metaequilibrium distribution which only over much longer glassy-like time scales would finally evolve to the "classical" Maxwell-Boltzmann equilibrium.

By the way, such a perspective seems to be in the way of becoming a rather popular one. Indeed it is presently often stated that there should exist some characteristic Ehrenfest time up to which classical and quantum mechanics agree in predicting motions of "ordered" type; later on they would instead differentiate, because nothing would happen according to quantum mechanics, while "chaotic" motions would occur according to classical mechanics. This is actually qualitatively analogous to the perspective proposed here.

So much for what concerns the possible logical relations between classical and quantum mechanics, in connection with the problem of equipartition of energy. It would be very interesting to know what is the actual status of the experiments concerning measurements of the specific heats of crystals and of polyatomic molecules over extremely long times.

REFERENCES

[1] Fermi E., Pasta J. and Ulam S., in *Fermi E.*, *Collected Papers* (University of Chicago Press, Chicago) 1965, and *Lect. Appl. Math.*, **15** (1974) 143.

[2] Fermi E., *Nuovo Cimento*, **25** (1923) 267; *Phys. Z.*, **24** (1923) 261.

[3] Benettin G., Ferrari G., Galgani L. and Giorgilli A., *Nuovo Cimento B*, **72** (1982) 137; Benettin G., Galgani L. and Giorgilli A., *Poincaré's non-existence theorem and classical perturbation theory in nearly-integrable Hamiltonian systems*, in *Advances in Nonlinear Dynamics and Stochastic Processes*, edited by R. Livi and A. Politi (World Scientific, Singapore) 1985.

[4] Izrailev F. M. and Chirikov B. V., *Sov. Phys. Dokl.*, **11** (1966) 30.

[5] Shepelyansky D. L., *Nonlinearity*, **10** (1997) 1331.

[6] Bocchieri P., Scotti A., Bearzi B. and Loinger A., *Phys. Rev. A*, **2** (1970) 2013.

[7] Galgani L. and Scotti A., *Phys. Rev. Lett.*, **28** (1972) 1173.

[8] Cercignani C., Galgani L. and Scotti A., *Phys. Lett. A,* **38** (1972) 403; Galgani L. and Scotti A., *Riv. Nuovo Cimento*, **2** (1972) 189.

[9] Nekhoroshev N. N., *Russ. Math. Surv.*, **32** (1977) 1; in *Topics in Modern Mathematics: Petrovskii Sem.*, no. 5, edited by O. A. Oleinik (Consultant Bureau, New York) 1985.

[10] Benettin G., Galgani L. and Giorgilli A., *Celestial Mech.*, **37** (1985) 1; *Nature*, **311** (1984) 444.

[11] Fucito F., Marchesoni F., Marinari E., Parisi G., Peliti L., Ruffo S. and Vulpiani A., *J. Phys. (Paris)*, **43** (1982) 707; Parisi G., *Europhys. Lett.*, **40** (1997) 357.

[12] Livi R., Pettini M., Ruffo S. and Vulpiani A., *J. Stat. Phys.*, **48** (1987) 539; Benettin G., Galgani L. and Giorgilli A., *Commun. Math. Phys.*, **121** (1989) 557; Galgani L., Giorgilli A., Martinoli A. and Vanzini S., *Physica D*, **59** (1992) 334; Escande D., Kantz H., Livi R. and Ruffo S., *J. Stat. Phys.*, **76** (1994) 605; Poggi D., Ruffo S. and Kantz H., *Phys. Rev. E*, **52** (1995) 307; De Luca J., Lichtenberg A. J. and Ruffo S., *Phys. Rev. E*, **60** (1999) 3781; Casetti L., Cerruti-Sola M., Modugno M., Pettini G., Pettini M. and Gatto R., *Riv. Nuovo Cimento*, **22** (1999) 1; Perronace A. and Tenenbaum A., *Phys. Rev. E*, **57** (1998); Kramers P. R., Biello J. A. and Lvov Y., *Discr. Cont. Dyn. Syst.–B*, in print.

[13] Boltzmann L., *Nature*, **51** (1895) 413; *Lectures on Gas Theory*, translated by S. G. Brush (University of California Press, Berkeley) 1964.

[14] Jeans J. H., *Philos. Mag.*, **35** (1903) 279.

[15] Langevin P., de Broglie M. (Editors) *La théorie du rayonnement et les quanta* (Gauthier-Villar, Paris) 1912.

[16] Carati A., Galgani L. and Pozzi B., *The problem of the rate of thermalization, and the relations between classical and quantum mechanics*, in *Mathematical Models and Methods for Smart Materials*, edited by M. Fabrizio, B. Lazzari and A. Morro (World Scientific, Singapore) 2002.

[17] Herzfeld K. F. and Litovitz T. A., *Absorption and Dispersion of Ultrasonic Waves* (Academic Press, New York and London) 1959; Kneser H. O., in *Rendiconti della Scuola Internazionale di Fisica "Enrico Fermi": XXVII, Dispersion and absorption of sound by molecular processes* (Academic Press, New York and London) 1963; Rapp D. and Kassal T., *Chem. Rev.*, **64** (1969) 61; Bhatia A. B., *Ultrasonic Absorption* (Clarendon Press, Oxford) 1967; Lambert J. D., *Vibrational and Rotational Relaxation in Gases* (Clarendon Press, Oxford) 1977; Krasilnikov V. A., *Sound and Ultrasound Waves* (Moscow 1960, and Israel Program for Scientific Translations, Jerusalem 1963); Kneser H. O., *Schallabsorption und Dispersion in Gases*, in *Handbuch der Physik XI-I* (Springer-Verlag, Berlin) 1961; Rapp D. and Kassal T., *Chem. Rev.*, **64**(1969) 61.

[18] Birge N. O. and Nagel S. R., *Phys. Rev. Lett.*, **54** (1985) 3674; Birge N. O., *Phys. Rev. B*, **34** (1986) 1631.

[19] Carati A. and Galgani L., *J. Stat. Phys.*, **94** (1999) 859.

[20] Benettin G., Galgani L. and Giorgilli A., *Phys. Lett. A*, **120** (1987) 23.

[21] Landau L. D. and Teller E., *Phys. Sowjetunion Z.*, **10** (1936) 34, in *Collected Papers of Landau L. D.*, ed. ter Haar (Pergamon Press, Oxford) 1965, p. 147.

[22] Baldan O. and Benettin G., *J. Stat. Phys.*, **62** (1991) 201; Benettin G., Carati A. and Sempio P., *J. Stat. Phys.*, **73**, (1993) 175; Benettin G., Carati A. and

Gallavotti G., *Nonlinearity*, **10** (1997) 479; Benettin G., Hjorth P. and Sempio P., *J. Stat. Phys.*, **94** (1999) 871.

[23] Carati A., Galgani L. and Pozzi B., *Phys. Rev. Lett.*, **90** (2003).

[24] Nernst W., *Verh. Dtsch. Phys. Ges.*, **18** (1916) 83.

[25] Galgani L. and Benettin G., *Lett. Nuovo Cimento*, **35** (1982) 93; Galgani L., *Nuovo Cimento B*, **62** (1981) 306; *Lett. Nuovo Cimento*, **31** (1981) 65; Galgani L., in *Stochastic Processes in Classical and Quantum Systems*, edited by S. Albeverio, G. Casati and D. Merlini, *Lecture Notes in Physics*, **262** (Springer-Verlag, Berlin) 1986.

[26] Carati A. and Galgani L., *Phys. Rev. E*, **61** (2000) 4791; *Physica A*, **280** (2000) 105; in *Chance in Physics*, edited by J. Bricmont *et al.*, *Lecture Notes in Physics* (Springer-Verlag, Berlin) 2001.

[27] Berchialla L., Galgani L. and Giorgilli A., *Discr. Cont. Dyn. Syst.–B*, in print.

[28] Zabusky N. J. and Kruskal M. D., *Phys. Rev. Lett.*, **15** (1965) 240.

[29] Ponno A., Galgani L. and Guerra F., *Phys. Rev. E*, **61** (2000) 7081.

B.5 P. Cipriani: Enrico Fermi's excursions through the fields of classical physics: Watching the landscapes of phase space and the nature of dynamical paths, looking for ergodicity

P. Cipriani: "Enrico Fermi's excursions through the fields of classical physics: Watching the landscapes of phase space and the nature of dynamical paths, looking for ergodicity," *Nuovo Cimento B* **117**, 1017 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

Enrico Fermi's *excursions* through the fields of classical physics: Watching the landscapes of phase space and the nature of dynamical paths, looking for ergodicity(*)

P. Cipriani(**)

INOA - Largo E. Fermi, 6, 50125 Firenze, Italy
ICRA - P.le Repubblica, 10, 65122 Pescara, Italy

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — The (relatively few) works of Fermi within the fields of classical physics had, and still have, a deep impact on our understanding of the structure of phase space of generic nonlinear systems, along with the relative implications on the justification of a statistical description of macroscopic systems and the approach towards equilibrium. One of those milestones along the path to reconcile microscopic dynamics with macroscopic description is the first *inverse experiment*, performed by Fermi, Pasta, Ulam and Tsingou on an one-dimensional anharmonic chain (FPU model). After a brief historical introduction whose aim is mainly to show how that revolutionary experiment frames perfectly into Fermi's personality, I discuss how this model, and particularly the *philosophy* beyond it, can be considered, still today, a valid *conceptual paradigm.* I show how to obtain analytical estimates of dynamic and geometric quantities through which it is possible to generalize the existing definitions of chaoticity indicators and of the *threshold* marking the onset of strong chaos. Nevertheless, as far as some of the most recent successful approaches to FPU problem are concerned, I outline how these cannot be generalized painlessly. Discussing in some details why they work for FPU-like models, we meet with the difficulties and troubles emerging when trying to applying them to *peculiar* Hamiltonian systems, for which these methodologies can give, at most, just some hints on their macroscopic behaviour. In particular, I review some conceptual and technical aspects of the combined use of the *geometrical transcription of dynamics* and the theory of *stochastic differential equations*, pointing out the issues preventing a direct extension to more general systems.

(*) Paper presented at the IX ICRA Network Workshop "Fermi and Astrophysics" (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.

(**) Permanent *postal* address: C.S.S., Via E.Nardi, 14-16 - 02047 Poggio Mirteto (RI), Italia. E-mail: cipriani@ino.it; cipriani@icra.it

Notwithstanding, this analysis gives noteworthy hints even on the much more controversial issue of a statistical description of gravitationally interacting N-body systems, furthermore allowing to understand some seemingly inconsistent results existing in the literature.

PACS 01.65.+g – History of science.
PACS 05.20.-y – Classical statistical mechanics.
PACS 05.45.-a – Nonlinear dynamics and nonlinear dynamical systems.
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

Enrico Fermi, allegedly, spent most of his research activities on atomic and nuclear physics [1,2]; nevertheless his *digressions* from that main path revealed outmostly fruitful and stimulating, both because of the originality of issues addressed and the novelty of the approaches proposed. To this it must be added the usual depth and *clarity of physical reasoning* (using Feynman's words [3]) of almost all Fermi investigations.

Among his most relevant studies within the fields of classical (*i.e.* non-quantum) physics, it should be certainly mentioned the extension of the *Poincaré theorem* [4] on the non existence of analytic first integrals for (generically) perturbed Hamiltonian systems, who renforced the belief that, from a practical point of view, a generic nonlinear system had to be ergodic, and consequently it could be successfully described by Statistical Mechanics (SM) methods. That theorem was probably one of the inspirations for the first *inverse experiment* performed with Pasta, Ulam and Tsingou devised to verify the approach towards equilibrium of *nonlinear* many degrees of freedom (mdof) systems [5].

There are, of course, many other Fermi's seminal contributions to classical dynamics, many of them with direct astrophysical implications: we just recall how his interest on the origin of high energy *cosmic rays*, led to the formulation of a very simple and elegant model for particle acceleration [6, 7], which became, again, a paradigmatic example for the interpretation of a wide class of phenomena, in the fields of fluid dynamics and chaotic systems (see, *e.g.*, ref. [8]). He, young, gave also important contributions to the interpretation of some aspects of general relativity (when even that theory was almost in its infancy), strictly linked to one of its most physically pregnant concepts, *i.e.* the *Equivalence Principle* [9, 10]. His *statistical* model of (heavy) nuclei [11, 12], found later interesting extensions to the description of gravitationally collapsed objects [13].

Furthermore, the attraction felt by Fermi towards SM and thermodynamics is witnessed not only by his celebrated fundamental studies [14] on the distribution law of half-integer spin particles(1) (owing him their name), but also by his involvement in writing down the *Statistical Mechanics* item for the *Enciclopedia Italiana* [16] and several monographies and textbooks on the same subject, on thermodynamics and even on

(1) And it is worth mentioning that, contrary to what usually reported, Fermi was led to this very important result by his aim at obtaining a satisfactory derivation of the Sackur-Tetrode formula for the entropy of an ideal monoatomic gas, rather than by its (presumed) interests in the conduction in metals [15]. This confirms the persistent attention maintained by Fermi towards the basic concepts of thermodynamics.

fluid dynamics [17].

Nowadays, however, the work of Fermi, referred to as the most thought-provoking for modern studies on the links between classical dynamics and SM foundations, is certainly the *pioneering inverse experiment*([2]) mentioned above [5], whose goal was to find a dynamical (microscopic) justification of the occurrence and effectiveness of thermodynamic (macroscopic) relaxation processes in generic non-integrable dynamical systems([3]).

Very good discussions about the implications of the FPU experiment on SM foundations can be found in this volume (see, *e.g.*, ref. [18]), also in connection with the issue of relaxation in stellar systems [19].

Therefore, in the following, rather than again commenting on the outcomes of that *experiment* and of its *technologically improved* successive repetitions, I will put more emphasis on the interpretations and generalization of those approaches which can be relevant for the long-standing issue of a *rigorous* justification of the statistical approach for mechanical systems [20].

More specifically, I deal with the *signatures* accompanying the onset of chaos in mdof Hamiltonian systems, using *alternative* tools to characterize and quantify the degree of instability. Within the phenomenology of the FPU model, it is possible to introduce a (rather elementary) generalization of Lyapunov exponent (or LCN), which, despite its simplicity, allows to overcome the ambiguities raising in some settings, mainly of astrophysical or cosmological interest. On its grounds, the implications of the onset of chaos in general mdof Hamiltonian systems are discussed, with some emphasis on the gravitationally bound *N*-body problem, where important *caveats* have to be kept in mind. On this light, I review critically the approaches which revealed very fruitful for the study of FPU-like models, and discuss the points which deserve a critical reconsideration, when trying to extend the above frames to more *critical* Hamiltonians. I derive some analytical (or semi-analytical) expressions for relevant quantities related to *dynamical*, *geometrical* and *statistical* scales of time and energy for the FPU model, and outline why similarly reliable estimates cannot be obtained so easily for the much more complex gravitational *N*-body system, discussing analogies and differences.

In a sense, the key ingredient of the analyses that follow has its foundation in the belief that the relevant, macroscopic (*i.e.* thermodynamic) properties of mdof systems can be obtained, in the generic case, by relatively simple and general physical considerations, supported by rather elementary mathematical computations. Obviously, *peculiar* systems and/or further investigations on detailed aspects can instead require articulated physical argumentations, and perhaps also (more) sophisticated mathematical treatment.

The approach just outlined is an (arduous and humble) attempt to follow what was

([2]) I prefer to term the *numerical integrations* of the equations of motion of model systems as *inverse experiments.* Without going into detailed (and here inopportune) discussions, a loose rationale is that there, instead of starting from observations, from where trying to formulate a tentative general law, open to successive refinements and improvements through *laboratory experiments*, one starts from a *certain* hypothetical law to obtain *experimental* data, to be compared (hopefully!) with observations of the real physical system whose modellization is sought.

([3]) The depth of Fermi insights into the issue of relaxation and the foundations of the statistical description of mdof systems earned him a very deep esteem even by one of the *founding fathers* of SM, *i.e.* Ehrenfest, who manifested repeatedly his interest in a collaboration with Fermi, which however has never been finalized [15].

Fermi's initial attack to complicated problems. As universally recognized(^{4}), Fermi researches, and teaching as well, were characterized by the already mentioned *clarity of physical reasoning* [3], which is undoubtedly connected with his ability to keep explanations simple, emphasizing conceptual understanding rather than calculations [22]. He was firmly convinced that, exposing physics in simple terms, forces the clarification of his own comprehension, and taught, even in his informal lectures, that knowledge of physics is achieved gradually, insisting that a deep understanding profits mostly by *intuitive and geometric, rather than analytic* arguments (see ref. [1], p. 673). Fermi repeated often that a first attack to solve a new and difficult problem must proceed through simplifications and analogies with known situations, and that every logical step had to be made with due reflection, even at the cost to proceed slowly.

There are a lot of appealing aspects in Fermi's trait: the preference he showed towards problems whose solutions can be guessed by simple calculations, based mainly on order of magnitude estimates [23]; his deep and active involvement in the experimental setup and his genial intuitions about the possibility to project and realize completely new kind of *experiments*, using tools which were appointed before for different tasks (see [24], p. 19). Also it must be mentioned his enthusiasm which led him to actively participate even in the *practical implementation* of his projects; *e.g.*, he was so much involved in the task of testing the validity of the ergodic hypothesis that, in a few weeks, he was able to effectively contribute in *programming electronic machines.*

The above repeatedly mentioned preference for heuristic and qualitative approaches should not be interpreted absolutely as a lack of rigor; as Ulam wrote (see ref. [24] p. 15), *Strangely enough,* [*Fermi*] *started as a mathematician.* [...]. *When he wanted to, he could do any kind of mathematics.".*

To further exemplify the innate Fermi attitude towards a *deep, conceptual*, and, at the same time, *physical and intuitively understandable* comprehension of *real* phenomena, it suffices to rememeber his opinion (after his period in Göttingen, around 1925) against the *operational and formal* foundations of the rising Quantum Mechanics, especially in the form of a *Mechanics of Matrices*: *"according to my taste, I feel that they are going too far along the tendence to renonce to understand the things"* (cited in ref. [25]). Coherently with this dislike for methods and approaches *that work, though it is not well understood why*; Fermi's projects for his, unfortunately never arrived, *"old age"* included the writing of a Physics book addressing all those difficult points usually concealed behind phrases like *"it is well known that..."* [21].

Fermi's *genius* was always moderated by a serious, systematic preparation, that led Feynman to feel himself affected by *confusion*, facing with *"the clarity of the exposition and the perfection* [...] *to make everything look so obvious and beatifully simple"*, whenever Fermi gave a lecture *"about any subject whatever he had thought before"* [3].

When necessary, after the first heuristic and qualitative attack, Fermi was always able to go beyond this initial framing of the problem, completing all the detailed physical arguments and mathematical steps to arrive at a comprehension of the phenomena at hand as complete as possible. Thus, presenting, six months before the submission of Dirac article on the same subject, his quantum theory of the ideal gas, *"Fermi worked out its consequences in more detail* [*and*], [...]*, showed that at low temperatures the equation of*

[4] See the tribute to Fermi [21], from where I have extracted some of the memoirs that follow.

state of the Fermi-Dirac gas has the form (from ref. [26], pp. 160-1)

$$p = \frac{ah^2n^{5/3}}{m} + \frac{bmn^{1/3}k_B^2T^2}{h^2} + \dots \,. \tag{1}$$

All the above is an attempt to outline Fermi's way to physics, starting from heuristics and simple approaches, arriving at an almost complete solution, through all the necessary steps.

In what follows the reader will not find, instead, *a* definite answer to (perhaps none of) the issues addressed. Nevertheless, I will present a general, though quantitative enough, critical reconsideration of some allegedly established beliefs and, for the issues left open, are indicated the directions along which the *right* answers could be possibily obtained.

2. – The Fermi-Pasta-Ulam experiment

The results of the FPU experiment are usually alleged to disprove the validity of the ergodic hypothesis and, consequently, of the Poincaré-Fermi theorem [25], as numerical integrations of the equations of motion showed an *almost periodic* behaviour of the istantaneous energy distribution among the modes of different wave numbers, *"To our surprise, the string started playing a game of musical chairs, only between several low notes,* [...]*, afterwhat would have been several hundred ordinary up and down vibrations, it came back almost exactly to its original shape"* (from [24], p. 19).

Subsequent numerical works showed instead that this almost astonishing result (*"a little discovery"*, Fermi said([5])) was mainly due to the small energy given to the system and to the too short integration time.

By now it is largely agreed that there are some *thresholds* separating different regimes in the dynamics of FPU chains. From a thermodynamical viewpoint, the lower threshold, related to the KAM behaviour, is of negligible relevance, as it tends to zero (quickly, *i.e.* almost exponentially) as the number of degrees of freedom increases. It is more interesting to investigate the applicability of the *Nekhoroshev theory*, which deals with the finite time conservation of the *actions*. This latter theory, of undeniable utmost relevance from the point of view of analytical mechanics, has been invoked [27] to explain the existence of a threshold in the scaling of the relaxation times for FPU (and similar) chains. However some inconsistencies survive, related mostly to the N-dependence of the predicted threshold. It is thus generally recognized, at least in the Physics community (though, admittedly, no rigorous proof exists), that a further threshold exists, whose location depends, for a given Hamiltonian, only on the energy density and which does not vanish in the Thermodynamic Limit (TL). This so-called *strong stochasticity threshold* (SST) is clearly linked to the stochastic properties of the dynamics and the *failure*, with respect to Fermi's expectations, of the *experiment* performed on the MANIAC computer can be attributed to the energy densities used, all below the *critical energy density*, $\varepsilon_c \doteq (E/N)_c$, characterizing the SST. Indeed it has been found [27-31], that FPU chains show a quasi regular behaviour, associated to very long relaxation times, for $\varepsilon < \varepsilon_c$, and a strongly chaotic dynamics, in turn leading to fast equilibration, above the threshold.

([5]) How important Fermi actually considered those results was witnessed by his intention to focus his planned *Gibbs Lecture* on them. Unfortunately, the cancer prevented him to achieve that aim.

Furthermore, it has been found that ε_c is an *intensive* quantity, *i.e.* does not depend on N. One can correctly speak about a *threshold* because of the *qualitatively different* behaviours below and above ε_c. In fact, the dynamic properties undergo a rather abrupt modification, witnessed, *e.g.*, by a change in the slope of the maximal Lyapunov exponent as a function of the energy density (see below for details), and the statistical behaviour is drastically influenced as well, as pointed out by the very different scaling of the relaxation times, changing from a strong (for $\varepsilon < \varepsilon_c$) to a very weak (if any) dependence on ε, above the threshold.

3. – Strong stochasticity threshold and geometric properties of configuration manifold

The existence of a SST has been proposed initially on the basis of numerical simulations, and has found later some theoretical justifications. In the last decade, moreover, the geometrical transcription of dynamics([6]), revived [34] in the 80ths, helped to show that the above threshold has a clean geometrical counterpart, which allows to obtain a very convincing scenario for the onset of strong chaoticity in the FPU chains and similar mdof Hamiltonian systems [31, 35, 32, 36].

As shown in fig. 1, already the first numerical integrations of the FPU model devoted to investigate the nature of the SST gave some support to its possible relationship with Nekhoroshev theory; though the predicted N-dependence of the critical value of the energy densisty was not confirmed. Subsequent works (see, *e.g.*, refs. [37-40]) seem to support the interpretation that SST and Nekhoroshev theory describe two different kinds of transition, and this is confirmed also by the geometrical approach [31, 41, 32], which predicts, for the SST, a value which does not depend on N, in particular it remains finite (*i.e.* does not vanish) in the TL.

Without going into details (see, *e.g.*, the recent review [36]), we just recall the basic steps of the *Geometrodynamical* approach (GDA), needed to illustrate the results obtained for FPU-like systems and the assumptions which have to be checked before to extend those results to more general dynamical systems (as self-gravitating N-body ones).

Within the GDA, the dynamical evolution of a Hamiltonian system with N degrees of freedom is rephrased in terms of a geodesic flow over a suitable manifold([7]), whose stability properties are determined by the Jacobi-Levi-Civita equation for geodesic spread:

$$\frac{\nabla}{\mathrm{d}s}\left(\frac{\nabla z^a}{\mathrm{d}s}\right) + \mathcal{H}^a_c z^c = 0 \quad (a = 1, \ldots, N), \tag{2}$$

where $\nabla/\mathrm{d}s$ stands for the covariant derivative *along* the flow, $\{z^a\}$ is an arbitrary perturbation to the reference geodesic, and the mixed tensor, $\boldsymbol{\mathcal{H}}$, describe the curvature properties of the dynamical manifold. Under some rather *mild* assumptions [32], the behaviour of a generic perturbation to a *given* geodesic (and then the stability of the

([6]) Developed, among others, by Levi-Civita, Synge, Eisenhart and others (see ref. [32] for references to these historical works) before 1930, and whose relevance for the issue of relaxation in many-body systems was first realized by Krylov [33].

([7]) Which can be either Riemannian [34,41,31,32], pseudo-Riemannian [35,42,36] or Finslerian [42,43].

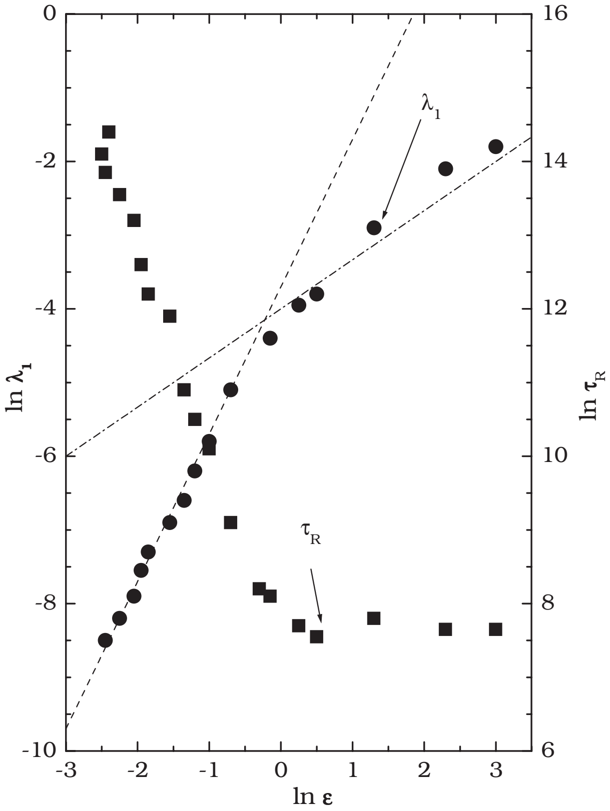


Fig. 1. – Scaling of maximal Lyapunov exponent λ_1 (circles, scale on the left) and of relaxation time τ_R (squares, scale on the right) with energy density. The figure has been generated from data taken from ref. [27].

flow) can be *reasonably*([8]) described using a single scalar *effective* equation, instead of the N equations (2), which reads

$$\frac{\mathrm{d}^2 z}{\mathrm{d}s^2} + k_R[s] z \cong 0 \ , \tag{3}$$

where $z \doteq (g_{ab} z^a z^b)^{1/2}$ is the norm of the perturbation (g_{ab} being the metric over the manifold) and $k_R[s] \doteq Ric[\mathbf{q}(s), \mathbf{u}(s)]/(N-1) \equiv \mathcal{H}^a_a/(N-1)$ is the Ricci curvature per degree of freedom in the $(N-1)$ two-directions determined by the flow.

For non-singular dynamics (*i.e.* Hamiltonian systems whose potential energy has a finite lower bound whose absolute value increases at most linearly with N), like FPU and similar models([9]) the above effective equation (3) can be written [31, 41, 35, 32],

([8]) And the crucial point is just to determine what these two words, *mild* and *reasonably*, actually mean, case by case.

([9]) Many-body Hamiltonians with Lennard-Jones interactions, short-range coupled rotators, $\lambda\Phi^4$-models, and so on.

equivalently([10]), as an evolution equation in terms of the *Newtonian* time t, related to the affine parameter of the geodesic s, by a transformation $\mathrm{d}s = \mathcal{A}\,\mathrm{d}t$, where the explicit form of the conformal factor $\mathcal{A}$ depends on the manifold used [42]:

$$\frac{\mathrm{d}^2Y}{\mathrm{d}t^2} + Q(t)Y \cong 0 \ , \tag{4}$$

where $Y(t)$ and $Q(t)$ are simply related [41,42,36], to $z[s(t)]$ and $k_R[s(t)]$, respectively.

Having made a long story short, we arrived at eq. (4) which has been the starting point of a very elegant and effective method [45, 35] of computation of the maximal Lyapunov exponent for mdof Hamiltonian systems. This approach, which make use of the theory of stochastic differential equations [8], has been enormously successful [36].

Briefly, under suitable assumptions, analyzed below, the average growth rates of a solution $y(t)$ of eq. (4), are determined by the equation for the moments [8,45]:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \langle y^2\rangle \\ \langle \dot{y}^2\rangle \\ \langle y\dot{y}\rangle \end{pmatrix} = \begin{pmatrix} 0 & 0 & 2 \\ \sigma_Q^2\tau_Q & 0 & -2Q_0 \\ -Q_0 & 1 & 0 \end{pmatrix}\begin{pmatrix} \langle y^2\rangle \\ \langle \dot{y}^2\rangle \\ \langle y\dot{y}\rangle \end{pmatrix} \ , \tag{5}$$

where $Q_0 \doteq \langle Q\rangle$, $\sigma_Q^2 \doteq \langle Q^2\rangle - Q_0^2$ and τ_Q represent, respectively, the average, the variance and the autocorrelation time of the (hypothetically assumed Gaussian) stochastic process $Q(t)$. The eigenvalues of the above equations can be easily calculated [8] and, for the present purposes, can be conveniently written in the following form [41]:

$$\mu_1 = 2\sqrt{\frac{Q_0}{3}}\left(g^{1/3} - g^{-1/3}\right) \tag{6}$$

and

$$\mu_{2,3} = -\sqrt{\frac{Q_0}{3}}\left[g^{1/3} - g^{-1/3} \pm i\sqrt{3}\left(g^{1/3} + g^{-1/3}\right)\right] \ , \tag{7}$$

where, for brevity of notation, it has been introduced the quantity g, defined as

$$g \doteq \xi + \sqrt{1+\xi^2} \geq 1 \ , \tag{8}$$

and, in turn,

$$\xi \doteq \frac{9\,\sigma_Q^2\,\tau_Q}{8\sqrt{3}\,Q_0^{3/2}} \ . \tag{9}$$

From the formulae above a series of results follow immediately:

- A particular form of conservation of volumes in the space of moments; that is $\sum_{i=1}^{3}\mu_i \equiv 0$.

[10] This point is somewhat controversial, in that there can be, in principle, the possibility that the reparametrization leads to different answers as far as the stability of the flow is concerned [44]. There are however rather stringent arguments to exclude that this could happen for non singular systems, whereas the issue is open for peculiar (*e.g.*, gravitational) N-body systems.

- As it is always $g \geq 1$, it follows that $\mathsf{Re}(\mu_1) \geq 0 \geq \mathsf{Re}(\mu_{2,3})$, with equality if and only if $g = 1 \Longleftrightarrow \xi = 0$.

- From the previous point it follows that, in the space of moments, there are always two *contracting* and one *expanding* (or three *neutral* in the case $\xi = 0$) *eigendirections.*

- This means that, under the assumptions adopted to derive eq. (5), the average asymptotic growth rate of $|y(t)|$, *i.e.* the maximal Lyapunov exponent, is

$$\lambda_1 \equiv \mu_1/2 \equiv \sqrt{\frac{Q_0}{3}}\left(g^{1/3} - g^{-1/3}\right) . \tag{10}$$

- Thus, the instability exponent increases from zero, for constant, positive *effective frequency*, $Q(t)$, monotonously as the amplitude of fluctuations grows.

Before to discuss the limits of validity of this approach, let us showing how well it works *within* these limits. The reliability of the GDA is witnessed by fig. 2, where, in the upper panel, the maximal Lyapunov exponent is computed according to the Van Kampen-Pettini formula, eq.(10), using the time averages of the geometrical observables [41, 32]. As it follows immediately from a comparison with fig. 1, where the Lyapunov exponents are computed using the standard BGS algorithm [46], in the $\beta\varepsilon$ region of overlap, the agreement is very good([11]) and the GDA give a much faster and perturbation independent method, which allows to extend the energy and N ranges of the simulations. However, the simulations performed up to very high energy densities [41] allow to correct the first claims about a scaling of the $\lambda_1(\beta\varepsilon) \propto (\beta\varepsilon)^{2/3}$. Indeed, figure 2 shows clearly that, while the scaling $\lambda_1 \propto \varepsilon^2$ for $\varepsilon \ll \varepsilon_c$ is confirmed, above the threshold, it is instead $\lambda_1 \propto \varepsilon^{1/4}$. This results has very deep implications on the nature of chaos of FPU model; firstly because it raises some questions against the *explanations* of the nature of stochasticity based on *Random Matrices Approximation.* Moreover, the scaling above the SST suggests that the dynamics is in a regime of fully developed stochasticity, in which the diffusion of orbits in phase space proceeds at the maximal rate allowed by the dynamics. The bottom panel of fig. 2 shows indeed the same data of the plot above, where the Lyapunov exponent $\lambda_1(\beta\varepsilon)$ is multiplied by the *dynamical time*, $t_D(\beta\varepsilon)$, of the system, which has been introduced in ref. [32] and whose precise meaning is described below. Furthermore, the extension of numerical simulations to larger N, confirms the N-independence of the threshold, thus giving further support to the idea that the SST marks a different transition to chaos with respect to the one predicted on the basis of Nekhoroshev theory.

The above interpretation of a strong chaotic regime above the SST is further confirmed by the comparison of the λ_1 values reported in fig. 2, in which the parameters entering eqs. (6), (10) are computed as time average along numerically integrated trajectories, with the analogous results obtained by Pettini and coworkers [45,35], who used the same formula([12]), using instead the phase space averages of the same parameters. The results

([11]) Notice that data in fig. 1 are expressed in *natural* logarithms.

([12]) And the Eisenhart geometrization, rather than the Jacobi one. However, for FPU-like systems and for $N \gg 1$ they are equivalent, as discussed in detail below.

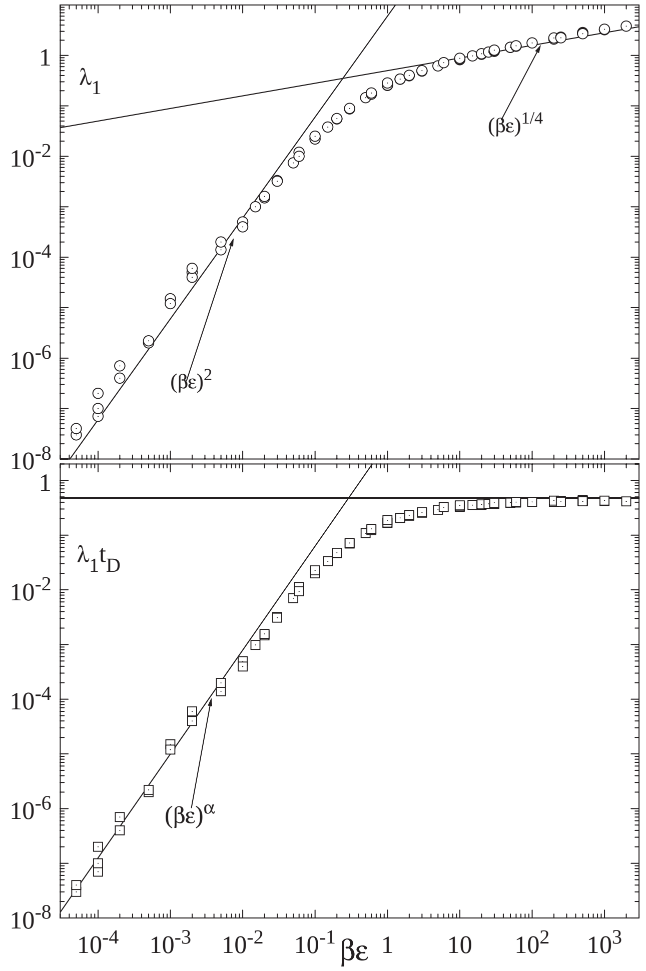


Fig. 2. – The upper plot shows the energy density dependence of the maximal Lyapunov exponent for different FPU-β chains. For a given $\beta\varepsilon$, are reported the values of λ_1 computed according to equation (10), for N ranging from 50 to 450 and anharmonicity parameter β varying in the interval $[0.05, 0.2]$. The significant parameter is clearly $\beta\varepsilon$. The lines indicate the slopes $(\beta\varepsilon)^2$ and $(\beta\varepsilon)^{1/4}$ and cross each other (conventionally) in correspondence of the SST. In the lower plot the λ_1 values are multiplied by the dynamical time, $t_D(\beta\varepsilon)$, (see text), showing that the quantity $\lambda_1 t_D$ is virtually constant above the SST. Below the threshold the slope is $\alpha \simeq 2$.

agree completely([13]), thus proving that, above the threshold, strongly chaotic dynamics is accompanied by *physical ergodicity*, relaxation is fast for every N and the equilibrium values do not show any dependence on N (provided that $N \gg 1$). On the contrary, at low energy densities there is a weak $N-$dependence, both in the values of the Lyapunov exponents and in the relaxation rates towards equilibrium.

([13]) Except at very low energy density, thus confirming a lack of ergodicity in the quasi-integrable limit, or, more probably, very long ergodicity times [47], though, from a physical viewpoints, either interpretation does not make a relevant difference.

3˙1. *Geometric signature of the SST: numerical and analytical evidence.* – Up to now we have mainly reviewed and commented the emergence of the SST from dynamical and statistical viewpoints. It is obvious however that the results based on the computation of the maximal Lyapunov exponent through eq. (10) shows how the geometrical features of the manifold detect, at least indirectly, the transition.

Therefore it is natural to see whether the SST has also a more direct signature within the GDA. We will see that this is indeed the case and, moreover, this will lead also to clarify the meaning of the dynamical time introduced above.

Referring to the literature (*e.g.*, ref. [32]) for the details of the steps leading from eq. (2) to eq. (3), let us start from the explicit expression of the *effective* frequency, as it can be determined, without loss of generality([14]), within the Jacobi GDA, for a N-degrees of freedom *natural* Hamiltonian system, reads

$$Q(t) = \frac{\Delta U}{N} + \frac{(\nabla U)^2}{NW} + \frac{1}{N}\left[\frac{\ddot{W}}{W} - \frac{3}{2}\left(\frac{\dot{W}}{W}\right)^2\right] , \tag{11}$$

where $H = (1/2)a_{ij}\dot{q}^i\dot{q}^j + U(\mathbf{q}) \equiv E$ is the conserved Hamiltonian, $W \doteq (E - U)$ is a shorthand for the total kinetic energy and, as usual, dots stand for (Newtonian-)time derivatives.

For the FPU-β system, that is, for chains of N coupled anharmonic oscillators with Hamiltonian

$$H = \sum_{i=1}^{N}\left[\frac{p_i^2}{2m} + U(q_{i+1} - q_i)\right] \quad \text{with} \quad U(x) = \frac{1}{2}x^2 + \frac{\beta}{4}x^4 , \tag{12}$$

and, *e.g.*, periodic boundary conditions, $q_{N+1} = q_1$ and $q_0 = q_N$, in the large-N limit, the last two terms in the expression of $Q(t)$ give essentially no contributions, neither to the average Q_0, not to the amplitude of fluctuations σ_Q. Also the (squared) gradient term, though with a comparatively weaker N-dependence, tends to disappear in the thermodynamic limit. All this is well evident from fig. 3, where the numerically computed values of the (time) averages of (rescaled) Ricci curvature (squares) $\hat{k}_R \doteq 2W^2 k_R$, effective frequency Q (circles) and Laplacian of the potential energy per degree of freedom $\Delta U/N$ (triangles) are shown, for FPU chains of different lengths (from $N = 50$ to $N = 450$), against the parameter $\beta\varepsilon$, which measures the departure from integrability. From the figure it emerges clearly that the average values of the three quantities are almost the same, unless conditions very near to integrability are considered and/or the number of

[14] It is obviously out of place here, but it can be easily shown [42] that the different GDA's, based on Jacobi (Riemannian), Eisenhart (pseudo-Riemannian) or Finsler geometries, differ only in the ranges of applicability; and that, for *natural and non singular* Lagrangian systems, which can be *geometrized* in all the settings, all them give the same results *as long as the number of degrees of freedom is large!* The differences between the various geometrizations indeed vanish in the large N limit (at least) as E/N^2, where E is the total energy. If the governing interaction is *stable* [48], then it follows that, in the large N limit, all the results are independent from the particular geometrization adopted. This, again, is not necessarily true if the interaction potential is not stable, in which case some differences could survive.

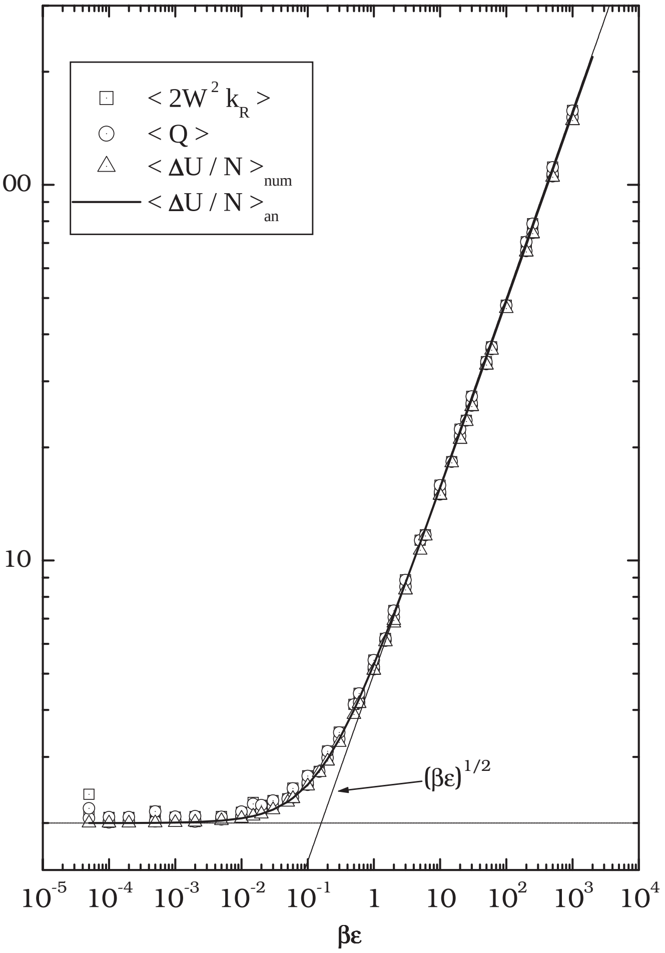


Fig. 3. – Energy density dependence of curvature related quantities. Circles, squares and triangles refer, respectively, to the averages of $\hat{k}_R$, Q and $\Delta U/N$, computed along numerically integrated trajectories of FPU chains with N ranging from 50 to 450 and anharmonicity parameter β varying in the interval $[0.05, 0.2]$. Data are taken from ref. [41].The significant parameter is clearly $\beta\varepsilon$. The thick curve refers to the analytical formula described in the text, while the thin lines are simply a guide for the eyes and help to locate the geometrical (counterpart of the) threshold.

degrees of freedom is small([15]). From this and the inspection of eq. (11), it is clear that, at least for FPU chains, the terms additional to the Laplacian in the expression for $Q(t)$ give a negligible contribution to the average. It can be verified [41] that, as far as the amplitude of fluctuations is concerned, the situation is almost the same, with some greater discrepancies for low N at very small energy densities. Thus we can conclude that, at least enough far from the quasi-integrable regime and for sufficiently large N,

([15]) For a given value of $\beta\varepsilon$, multiple values of dependent variables are plotted, for different values of N, though in most cases they virtually coincide.

the stability of motions depends only on the behaviour of the Laplacian term appearing in $Q(t)$([16]).

In the case of FPU models, for many geometric quantities it is also possible to obtain reliable (semi-)analytical estimates. Indeed, if we define $\langle \delta q \rangle \doteq \langle q_{i+1} - q_i \rangle_i$ and put

$$\eta \doteq \frac{\langle \delta q^4 \rangle - \langle \delta q^2 \rangle^2}{\langle \delta q^4 \rangle} \; ; \tag{13}$$

using the virial theorem in its most general form, we find that the ratio between anharmonic and harmonic potential energies depends only on $\beta\varepsilon$ and η

$$\frac{\langle U_4 \rangle}{\langle U_2 \rangle} = \frac{1}{3}\left(\sqrt{1 + 3\beta\varepsilon(1+\eta)} - 1\right) \; . \tag{14}$$

Using again the virial, an analytic expression for the Laplacian is found:

$$\frac{\Delta U}{N} = 2 + \frac{4}{1+\eta}\left(\sqrt{1 + 3\beta\varepsilon(1+\eta)} - 1\right) \; . \tag{15}$$

From numerical simulations it is easy to obtain a numerical estimate of the correlation term η and, finally, the closed form:

$$\frac{\Delta U}{N} = 2\left[1 + \left(\sqrt{1 + 6\beta\varepsilon} - 1\right)\right] \; , \tag{16}$$

which is plotted as a continuous (thick) line in the same figure 3 and is in very good agreement with *experimental* data.

Moreover, we see that the GDA give a consistent, direct and convincing evidence of the existence and location of the SST. It can be conventionally located by the crossing of the two asymptotic behaviours of the Laplacian, and a direct comparison with the plot of the scaling of the maximal Lyapunov exponent of fig. 2 show the good agreement between the dynamical and geometrical signatures of the threshold. Along with the equally well convincing correspondence between dynamic and statistical mechanical thresholds, this completes the path from SM to GDA.

3˙2. *The dynamical time scale and the adimensional instability indicator.* – The steps which allow an analytical estimate of the scaling of $\Delta U(\beta\varepsilon)/N$, turn out to help the explicit computation of the above introduced dynamical time scale, $t_D(\beta\varepsilon)$; which, at a heuristic level, gives the overall rescaling, when the energy density changes, of any dynamical process occurring in the chain.

([16]) Nevertheless, it should be emphasized that the approach necessarily fails in the quasi-periodic regime and at small N, where, for example, negative values of the curvature k_R occur more frequently than in the high energy regime, and persist for very long transient periods. This is clearly due to the large collective oscillations accompanying the virialization process, during which the terms containing the time derivatives in eq. (11) can assume relatively large values. The combined effects of small N (these terms vanish in the TL at least as $N^{-1/2}$), along with the very slow *phase mixing* due to quasi periodic behaviour [41, 32] invalidate the starting assumptions at the grounds of the approach leading to eq. (10).

It is well known, indeed, that in FPU-like models it is possible to introduce the normal modes frequencies, which describe the hierarchy of time scales associated with the dynamics of phonon-like excitations. Those frequencies do not depend however on the energy density, have always a maximum which is $\mathcal{O}(1)$, whereas the lowest frequencies scale as $\mathcal{O}(1/N)$. To the extent that the chain is near the integrable (*e.g.*, harmonic) limit, normal modes frequencies suffice to describe the dynamics of the system; however, when the departure from integrability is strong, *i.e.* when the anharmonic interaction tends to become comparable with the harmonic one, the normal modes hierarchy does not more describe correctly the internal dynamics (and even the very definition of normal modes loses partially its meaning).

There are several ways to extract an energy dependent overall time scale for anharmonic systems, many of them can be based simply on dimensional arguments and give the correct qualitative behaviour [41]. In order to obtain a more *quantitative* estimate, we proceed in complete analogy with what is done in stellar dynamics, where a proper (dynamical) time scale is defined as the *typical* time required for a star to cross the system. In that case it is assumed a sort of energy equipartition between the stars, and in this case we have to make a similar assumption, which, however, *cannot be extended to the normal modes*, just because for them it is clear that equipartition holds *only after the systems has relaxed* and the conditions for the equipartition among them is one of the main issues. The simplest and most safe assumption, rigorously verified by *inverse experiments* of any kind, is that each oscillator has, on the average, the same energy. As the FPU system is an extensive one, independently of the N and ε values, we can assume that this average energy is of the order of the energy density itself. Simple dimensional arguments indicate that the $t_D(\beta\varepsilon)$ scaling is related to the ratio of the anharmonic to the harmonic potential energies. We write, for the maximal values of these potential contributions to the total energy, respectively, $U_{4_{\rm MAX}}$ and $U_{2_{\rm MAX}}$, and the rather intuitive parametric pair of equations:

$$(17)\qquad U_{2_{\rm MAX}}(\eta) = E\cos^2\varphi \quad \text{and} \quad U_{4_{\rm MAX}}(\eta) = E\sin^2\varphi \ ,$$

where η has been defined in eq. (13) and $\varphi = \varphi(\eta)$ is an obvious measure of the departure from the harmonic regime.

Using the virial theorem, in complete analogy as before for the determination of the analytic formula for ΔU, a quantitative expression for the dynamical time scale is found:

$$(18)\qquad t_D(\beta\varepsilon) = 2\sqrt{2}\frac{\left[\sqrt{1+4\beta\varepsilon(1+\eta)}-1\right]^{1/2}}{[\beta\varepsilon(1+\eta)]^{1/2}}\int_0^1 \frac{\mathrm{d}x}{\sqrt{1-x^2\,\cos^2\varphi - x^4\,\sin^2\varphi}} \ .$$

Taking into account that from numerical simulations we have a precise measure of η, the expression for t_D can be written in closed form as

$$(19)\qquad t_D(\beta\varepsilon) = 4\,[\Theta(\beta\varepsilon)]^{1/2}\ \mathcal{E}_K\left[[1-\Theta(\beta\varepsilon)]^{1/2}\right] \ ,$$

where $\mathcal{E}_K(x)$ is the complete elliptic integral, defined in terms of the incomplete elliptic integral,

$$\mathcal{E}_F(z,x) = \int_0^z \frac{\mathrm{d}t}{\sqrt{1-t^2}\sqrt{1+x^2\,t^2}} \ ,$$

as $\mathcal{E}_K(x) \doteq \mathcal{E}_F(1,x)$; and we defined moreover,

$$\Theta(\beta\varepsilon) = \frac{\sqrt{1+8\beta\varepsilon}-1}{4\beta\varepsilon} \ .$$

Despite its seemingly complicate expression, the essential $\beta\varepsilon$-dependence in the above equation is almost completely represented by the $\Theta^{1/2}$ term. Indeed, the argument of the elliptic integral varies only in $[0,1]$ and the variation of $\mathcal{E}_K(x)$ within this range is very small. In the harmonic limit, $\beta\varepsilon \to 0$, it is evident that $\Theta \to 1$, and then

$$\lim_{\beta\varepsilon\to 0} \mathcal{E}_K[(1-\Theta)^{1/2}] = \mathcal{E}_K(0) = \frac{\pi}{2} \cong 1.5708\ldots \ .$$

On the other hand, when $\beta\varepsilon \to \infty$, $\Theta(\beta\varepsilon) \to (\beta\varepsilon)^{-1/2}$, so that

$$\lim_{\beta\varepsilon\to \infty} \mathcal{E}_K[(1-\Theta)^{1/2}] = \mathcal{E}_K(1) = \frac{\pi^{3/2}\sqrt{2}}{4\left[\Gamma\left(\frac{3}{4}\right)\right]^2} \cong 1.311\ldots \ .$$

Incidentally, as $\Theta(\beta\varepsilon) \sim (\beta\varepsilon)^{-1/2}$ at high energy, eq.(19) shows that in the high energy density regime it is $t_D \propto (\beta\varepsilon)^{-1/4}$. Recalling the scaling of the maximal Lyapunov exponent above the SST, this result *explains* the significance of the lower panel in fig. 2, *i.e.* why $\gamma_1 \doteq \lambda_1 t_D$ is almost exactly constant above the threshold, an observation of crucial relevance for the interpretation of the nature of the chaotic behaviour above the SST.

It must be observed that the result of the above computation differs very much *in the form* from those based on more naive approaches. For example, an estimate based on simple dimensional arguments leads to the much simpler expression

$$t_{D_a}(\beta\varepsilon) = \left[1 + \frac{2}{3}\left(\sqrt{1+3\beta\varepsilon}-1\right)\right]^{-1/2} \ , \tag{20}$$

which, as shown in fig. 4, gives values appreciably different from those obtained from the *exact* equation (19). Nevertheless, if the dimensional arguments are complemented by numerical estimates on the degree of correlations among nearby sites, then the results obtained are in a much better agreement with the *exact* ones, though, formally, the corresponding expressions still differ from eq. (19).

To illustrate this fact, in fig. 4 we report, along with the *exact* and the *naive* curves predicted by eqs. (19) and (20), respectively, also two other semi-analytical estimates, based on the virial theorem and numerical estimates of suitable correlation functions, reading as

$$t_{D_b}(\beta\varepsilon) = \left[1 + \frac{2}{3}\left(\sqrt{1+6\beta\varepsilon(1+\zeta)}-1\right)\right]^{-1/2} \tag{21}$$

and

$$t_{D_d}(\beta\varepsilon) = t_{D_d}(0)\left[\frac{\sqrt{1+4\beta\varepsilon(1+\eta)}-1}{\frac{4}{3}\beta\varepsilon - \frac{2}{9(1+\eta)}\left[\sqrt{1+3\beta\varepsilon(1+\eta)}-1\right]}\right]^{1/2} \ , \tag{22}$$

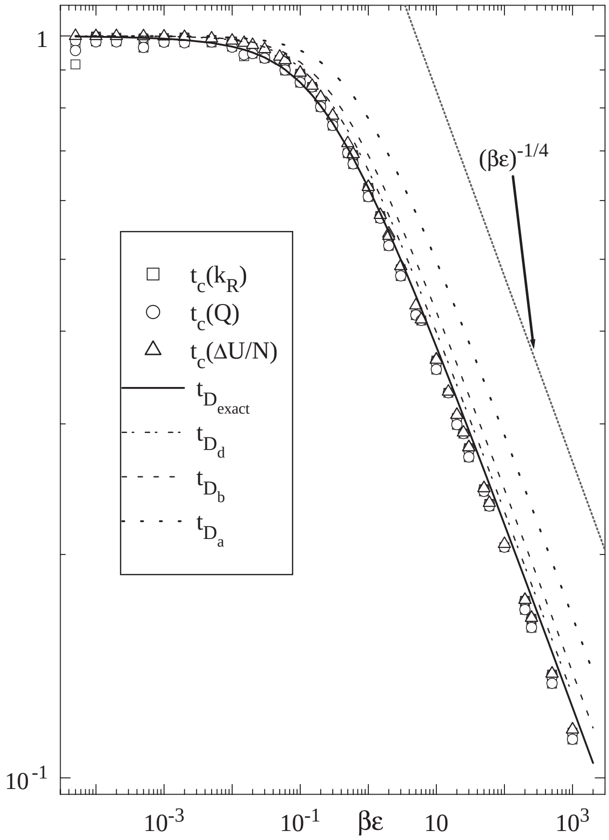


Fig. 4. – The curves show the behaviour of the analytical estimates of the dynamical time scale, $t_D(\beta\varepsilon)$, according to the various formulas discussed in the text. The line on the right indicates the asymptotic behaviour of all the above estimates, $t_D(\beta\varepsilon) \propto (\beta\varepsilon)^{-1/4}$ and the symbols represent the numerical estimates of the dynamical periods associated to the *effective* frequencies governing the evolution of perturbations within the GDA.

where η is the same quantity defined above, eq. (13), and ζ is related to the correlations of the form $\langle q_i q_{i+1} \rangle$ and similar ones.

Despite that rather different look, the last two estimates are in satisfactory agreement with the exact one; and, moreover, they agree rather well even with *experimental* values, shown also in fig. 4. The different symbols represent the *experimental*, numerical determinations of the correlation times for the geometrical and dynamical quantities determining the evolution of perturbations, that is $\hat{k}_R$, Q and $\Delta U/N$. The analytical estimates, except the one given by eq. (20), are in a complete agreement with the numerical data. The *exact* expression is only slightly better than those based on dimensional arguments alone, pointing out how the virial theorem, in its more general form [41,32,49], captures all the essential time-energy scaling of internal dynamics.

The previous results show that, starting from two completely separate approaches, it is found that the scaling of the instability time, as measured by the (reciprocal of the) maximal Lyapunov exponent, and the dynamical time, for the FPU-β model, scale, in the strong stochastic region, exactly at the same rate([17]).

This evidence, suggests a new *invariant* criterion to detect the SST, or, more precisely (and generally), the onset of the regime in which the spreading of orbits proceeds at the maximal rate allowed by the underlying dynamics. It can be defined using the adimensional chaoticity indicator $\gamma_1 \doteq \lambda_1 t_D$: the onset of fully developed stochasticity is detected by the constancy of the indicator γ_1.

This rather trivial generalization has some relevant consequences in those frameworks where the time scales of the systems under study change rapidly varying some parameters (*e.g.*, the energy, the external field, etc.), and also for systems undergoing intermittent evolution, where quiescent and irregular phases are intermingled and characterized by very different time scales. Furthermore, *the introduction of an adimensional instability measure, get rid of most of the ambiguities raising in those settings where the choice of the appropriate evolution parameter (e.g., in general relativistic dynamical systems) is an issue.*

Among the most important problems which can get some important hints from the above discussion, I mention briefly the issue of the statistical mechanical and thermodynamical description of N-body self-gravitating systems. For them the above analysis has some noteworthy consequences [50, 51]: it is found that, with all the remarks and warnings appropriate to their intrinsically peculiar nature([18]), it can be guessed that they are, at any binding energy, in a state of strong stochasticity. On the basis of a *fast mixing hypothesis* [50, 51], this allows to propose a coherent thermodynamic setting, which give a dynamical justification to many phenomenological descriptions adopted previously.

Referring to the cited bibliography for a complete account of the suggestions and results which can be obtained from the GDA in order to understand better the onset of instability in more general Hamiltonian systems than FPU-like models, I will conclude this contribution trying, as said in the Introduction, to deepen the investigation of the *dark side*: in the next section a list of concrete and practical remarks are presented which should be taken into account, *before* to carry out any direct extension of the approach.

4. – Check of the hypotheses at the ground of the analytical computation of the maximal Lyapunov exponent

Already the steps illustrated in the previous section point out that a critical reconsideration of the general setting is in order if an extension of the framework to systems of different nature than FPU-like models is sought. Therefore, we recall the key hypotheses at the grounds of the method and proceed to verify their limits of validity:

- The fundamental assumption for the application of the Van Kampen formulae, is that the (squared) *frequency*, $Q(t)$ represents a faithful realization of a stochastic process. This amounts to say that the distribution of the Q values should be reasonably well described by a gaussian and that the autocorrelation function of

[17] This derivation, incidentally, devoids of any content the assert, sometimes reported, according to which the dynamical time is *defined* as the reciprocal of the maximal LCN: the dynamical time exists and is finite, according to the above prescription, even for completely integrable systems, when the Lyaunov time is infinite!

[18] And already pointed out by Gurzadyan and coworkers in a series of papers [52, 53].

$Q(t)$ can be approximated, on the time scales of the evolution, as a $\delta-$function, *e.g.*,

$$\langle Q(t+t')Q(t)\rangle \cong Q_0^2 + \sigma_Q^2\,\tau_Q\,\delta(t').$$

– From a physical perspective, this amounts to say that the evolution of $Q(t)$ must be fast with respect to the dynamical time scales. Geometrically, this means that the *effective* curvature is a strongly fluctuating quantity.

– To the above arguments, it must be added the following trivial consideration: in all the derivations it is assumed that $Q(t)$ is a non-negative quantity, although it can be argued that, if the probability of occurrence of negative values is sufficiently small, then the growth rate of the solution is accurately represented by λ_1. Thus, it follows that, if the amplitude of fluctuations exceeds the average value, $\sigma_Q > Q_0$, then the frequency cannot longer be assumed as positive definite, and the approach does not work anymore.

– A further obvious breakdown of the method occurs when the dynamic evolution is quasi-periodic. In such a case, indeed, independently of the amplitude of fluctuations, the values of ξ and g can be arbitrarily large, and so λ_1 can formally increase even if the underlying dynamics is not chaotic([19]).

– As, in practice, the $Q(t)$ is self-consistently generated by the dynamics itself, and depends on the evolution of *state space* coordinates of the system, it is clear that a full justification of the validity of the approach must be always checked *à posteriori*. That is, if the underlying dynamics is *not enough chaotic*, then the fluctuations of $Q(t)$ cannot be assumed as stochastic, and the instability rate predicted by the use of Van Kampen formula can be, at best, an upper limit to (twice) the true Lyapunov exponent.

– Stated otherwise, eq. (6) is hardly valid in the quasi-constant curvature limit. It could be guessed that there must exist some threshold ξ_+, below which the predicted instability exponent cannot be exact. However the estimate of this possible limiting value is not so immediate. Indeed, if it is clear that the shorter the correlation time τ_Q, the greater must be the fluctuations, in order to keep constant the value of ξ (and, then, of g), on physical grounds, it is instead plausible that, for very short correlation times, even small fluctuations can lead to stochasticity. Indeed, it follows from eq. (10) that, in the small fluctuation and small correlation times, that is, for $\xi \to 0$, it is $\mu_1 \propto Q_0^{1/2}\xi$.

– It is otherwise clear that in the opposite limit, that is for $\xi \gg 1$, the above formula is not more justified. In fact, neglecting numerical factors, we have

$$\xi \sim \left(\frac{\sigma_Q^2}{Q_0^2}\right)\frac{\tau_Q}{t_G}\ ,$$

([19]) This is indeed what is really observed for nearly integrable conditions and when the number of degrees of freedom is not large enough, so that statistic relative fluctuations, of order $\mathcal{O}(N^{-1/2})$, mimic true curvature fluctuations. In these cases the Van Kampen-Pettini formula, eq. (10), clearly overestimates the instability rate.

where $t_G \doteq Q_0^{-1/2}$ sets a sort of *geometrodynamical time scale.* Thus, values of ξ greater that a few units, immediately imply the failure of at least one of the assumptions: if $\xi \gg 1$, then, necessarily it must be either $\sigma_Q \gg Q_0$ or $\tau_Q \gg t_G$ (or both). In the first case the hypothesis of a positive *frequency* is not more verified, in the second even the very definition of *δ-correlated stochastic process* could be questioned. Furthermore, in the limit $\xi \gg 1$, the expression for the instability exponent can be rewritten as

$$\lambda_1 \sim t_G^{-1} \xi^{1/3} ,$$

which, to my knowledge, has never been observed. What happens, in FPU-like models, is that, increasing the energy, and then the degree of chaos, the correlation time goes to zero (as shown in the previous section and in fig. 4) so that the limit $\xi \gg 1$ is never reached, except when the number of degrees of freedom is so small that the existence of conserved quantities introduces quasi periodicities and spurious long time correlations, even in the presence of chaotic dynamics.

As we have seen, in the case of FPU model it is easy to test the validity of the approach against the possible sources of inconsistency listed above. Both the Gaussian distribution of the Q values can be checked and the autocorrelation time τ_Q can be either analytically estimated and numerically computed [41]. The limits of validity of the Van Kampen-Pettini formula are easily located and the results obtained are in a comfortable agreement with those obtained with other approaches.

Figure 5 shows, for an intermediate energy density, that the gaussian distribution of *curvatures* and *frequencies* is indeed a rather well satisfied hypothesis, both for what concerns the their values and (still better) for their time-derivatives as well. It must be mentioned that, while the distributions of the Laplacian and of Hill's frequency Q share common averages and widths, the k_R-distribution is much broader. This evidence hints at reconsider the applicability of the Van Kampen-Pettini approach within the *purely* geometrical setting related to the Jacobi metric. Though this remark is not of so much relevance for the FPU models, it can have deep impact on the general applicability of the method. In particular, in the case of large fluctuations all the results appealing to the existence of an *average* curvature lose their justification.

Let us then analyze from where the most serious hindrances to the application of the method above can originate, considering that we have in mind the extension of the approach to the gravitational N-body problem.

- For systems with non smooth interactions, the curvature fluctuations are usually much bigger than for FPU-like potentials, unless there is a minimum in the interaction potential (*e.g.*, LJ or Morse systems) and the energy density is slightly above that minimum. This amounts to say that fluctuations are comparable with mean values as soon as the departure from the *integrable limit* (*if it exists!*) is appreciable.

- Nevertheless, non smooth potentials usually imply a very large spectrum of time scales and often also non homogeneous distributions. This partially can compensate the consequences of the previous remark, as it can lead to very fast decay of correlations, that is to very small correlation times.

- Obviously, large fluctuations imply that the assumption of an everywhere positive curvature is hardly fulfilled. In the case in which the frequency of negative values

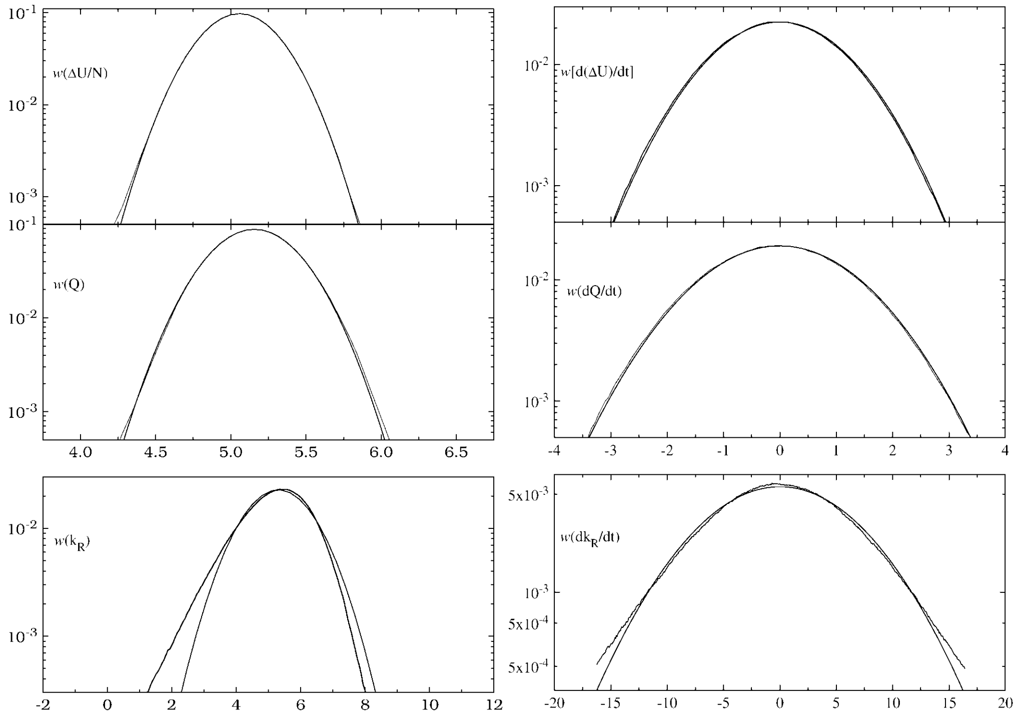


Fig. 5. – Check of the Gaussian distribution of the values of *curvature* and *frequencies*. A) In the left column, from bottom to top, are shown, respectively, the distributions of k_R, Q and $\Delta U/N$. Thick curves represent the actual distributions as obtained from long time numerical integrations of dynamics (data taken from [41]), while thin curves represent the best fits with a Gaussian (N.B.: notice that vertical scales are in log-scale). B) Distributions of the time derivatives of the quantities represented in A). Except for the Ricci curvature k_R, where a clear asimmetry is evident, all other quantities are well distributed according to a Gaussian. Data refer to an intermediate value of $\beta\varepsilon$, near the SST.

is appreciable, instability can originate both from mechanisms like those described above, and also from the local instability related to negative curvature. Any attempt to estimate analitically the instability growth rate is in such cases hopeless. At the other extreme, if the curvature is almost everywhere negative, then the estimate of the instability time is possible, though this is not the case for almost any realistic model of physical many degrees of freedom system([20]).

– In systems with non extensive interactions (*e.g.*, self-gravitating N-body systems), a strongly unstable evolution, with fast *phase mixing* and decay of correlations, can coexist with very long time correlations of collective degrees of freedom, associated, for example, with the Virialization process (*e.g.*, with the *Violent Relaxation*

([20]) Though there is no rigorous proof of this statement. For instance, V. Gurzadyan (private conversation, see also ref. [34]) confirmed to the author that spherical stellar systems constitute probably an example of a system for which the assumption of almost everywhere negative curvature can be verified rigorously.

phase). A signature of this phenomenon, though with much weaker consequences, is present even in the FPU model, where the probability of negative curvature values is relatively high when the system is left to evolve from initial conditions far from (global) equilibrium. After that equilibrium is attained, that probability becomes vanishingly small [32].

- Obviously, systems with singular interactions, in consequence of what stated above, support even more reliably the assumption of uncorrelated fluctuations, just because a singular two body interaction can occur almost indepedently from a previous one [54]. Furthermore, the hypothesis of uncorrelated fluctuations depends crucially on the dimensionality of the systems, being much more easily fulfilled in $3D$-systems than in one-dimensional chains.

- Most of the results on FPU-like models, and in particular the analytical computations of λ_1 on the basis of eq. (10), greatly profited of the possibility of canonical estimates of geometric quantities [35]. This is legitimate in systems like FPU, provided that ergodicity (and mixing) can be safely assumed. Indeed, in this case, time averages coincide with microcanonical phase averages. As the system is *tempered* and *stable*, then the *rigorous theorems* [48] assert that microcanonical and canonical averages coincide in the TL (differences being at most of order $\mathcal{O}(1/N)$). Indeed the results obtained in [41, 32], through dynamical (*i.e.* microcanonical) numerical integrations of equations of motion show a very good agreement with those obtained in [45, 35, 36], using phase space averages. Clearly both the logical steps above can be questioned for systems with non compact phase space and non-extensive interactions: ergodicity and mixing cannot be rigorously defined and the equivalence between microcanonical and canonical ensembles does not hold. In refs. [50, 51] possible solutions to these issues are discussed.

- The above points can be summarized saying that, while for FPU-like models it is clear that fluctuations are responsible for the onset and development of chaos, this simple paradigm cannot be extended to *peculiar* Hamiltonians. This also because, in the most general case, the simple order of magnitude estimates which allow to claim, for FPU, that in the large-N limit the fluctuations come essentially only by the first two terms in the right hand side of eq. (11), are not so easy, and the relative weight, in the fluctuations, of various terms must be checked carefully.

- Although the dynamics of singular systems can be, in a sense, strongly chaotic, it is possible that some relevant collective quantities, included those used in the GDA, can have virtually infinite ergodicity times, so that phase and time averages lead to conflicting indications.

Notwithstanding all the remarks above, it is worth to emphasize that the GDA helps to deepen the understanding of the interplay between the energy and time scales, leading naturally to operate a due distinction between *mathematical and physical ergodicities*, arguing that the latter (alone) can be relevant for a statistical mechanical description. As remarked already before, as long as the system is above the SST, that is, if the dynamics is strongly chaotic, then a complete and precise agreement exists between the time and phase averages of any geometrical and dynamical observable. However, the agreement remains very good even below that threshold if the number of degrees of freedom is large enough. The *ergodicity times*, however, increase rapidly with N. It is puzzling to observe that the same happens for, *e.g.*, self-gravitating systems, notwithstanding they seem to be in a dynamic regime at least as chaotic as FPU chains well above the SST. This points out once more that the *analogic method* is often a good guide to guess the behaviour of more complex problems, but, if left alone, without any further rigorous investigation, can sometimes lead to dangerous conclusions.

5. – Epilogue

I started my Ph.D. as an astrophysicist, (at least, this is what I was believing). For a true trick of the fate, I met the FPU problem, without intention, and this problem became afterward my principal interest for two years, forming then the core of (more than half of) my Ph.D. Thesis.

I cannot say whether the crossing of my life with the FPU problem was lucky or not, what is sure is that it marked a *threshold*, and as such, I feel still today, fascinated and at the same time frightened by its immense richness of faces and traps. And this is amazing, because I was attracted by the FPU problem because of its deceptive simplicity, as, perhaps [55], many others. Yes, indeed what surely I like is the Fermi's first approach to problems, *"based on simple math"*; and I tried, in the pages above, to learn the Fermi's lesson, avoiding to take for granted what is not demonstrated, without however to give up to being guided, in the first investigations, by intuition and analogies. Probably I didn't succeed more than slightly; nevertheless I hope that the few good remarks can be useful to my own future works and, perhaps, also to other people.

People who worked with Fermi learned from him how to face physical problems; they better than others can pass to us the atmosphere around him at work.

It seems appropriate to finish with these few notes, by Stan Ulam [24], which was involved with Fermi in the project and realization of the study which has been the central point of this contribution. They point out once more how Fermi's personality reflected coherently in his work.

"His [Fermi's] eyes, darting at times, would be fixed reflectively when he was considering some questions. He would try to elucidate other persons thoughts by asking questions in a Socratic manner, [...] . I think he had a supreme sense of the important. He did not disdain work in the so-called smaller problems; at the same time, he kept in mind the order of importance of things in physics. This quality is more vital in physics than in mathematics, which is not so uniquely tied to reality". (p. 15)([21]).

"As soon as the machines were finished, Fermi, with his great common sense and intuition, recognized immediately their importance for the study of problems in theoretical physics, astrophysics and classical physics". (p. 19).

And, after a personal reflection [24] (p. 19), *"Now Banach, Fermi and von Neumann were dead—the three great men whose intellects impressed me the most."*; another Ulam's memoir on Fermi, reported by a close friend of him [24] (p. 27): *"[Ulam] admired Fermi's genius for solving physical problems with the minimum amount of math. Since that time Fermi remained for him the ideal of a scientist. I his old age he liked to repeat that Fermi had been the last physicist".*

* * *

To Maria Teresa, for her patience and for having forced me to go a little beyond

[21] That is, a physicist should not *be fashinated* by elegant mathematical approaches, unless they improve the physical understanding of the phenomena.

simple calculations. To R. Ruffini for his impromptu flashing suggestions, whose number has been (unfortunately) decreasing over the years, because of the divergence between our interests. This however did not prevent him to unselfishly support my activities. To M. Pettini and L. Galgani, for continuous moral encouragements and many interesting discussions. To A. Politi, for many seemingly unrelated discussions, not really so; and also for his tolerance. To V. Gurzadyan: I have profited very much, since the beginning of my study in this field, from periodic exchange of ideas with him, although often without each other agreeing with other's opinion. I remember the many repeated efforts, made together with G. Pucacco and D. Boccaletti, trying to understand a little more on geometry, statistical and classical mechanics. The author is partially supported by CSS under the initiative n.2002B:CaPMeP.

REFERENCES

[1] Fermi E., *Note e Memorie* (Accad. Naz. dei Lincei & Univ. of Chicago Press) 1965.

[2] Fermi L., *Atoms in the Family: My Life with Enrico Fermi* (Univ. of Chicago Press) 1961.

[3] Mehra J., *The Beat of a Different Drum: the Life and Science of Richard Feynman* (Oxford Univ. Press) 1994.

[4] Fermi E., *Nuovo Cimento*, **25** (1923) 267; 271; **26** (1923) 105.

[5] Fermi E., Pasta J. and Ulam S., Los Alamos Report LA-1940, reprinted in ref. [1].

[6] Fermi E., *Phys. Rev.*, **75** (1949) 1169.

[7] Zalsavsky G. M., *Chaos in Dynamic Systems* (Harwood Acad. Publ.) 1985.

[8] Van Kampen N. G., *Phys. Rep.*, **24** (1976) 171; *Stochastic Processes in Physics and Chemistry* (North-Holland) 1981.

[9] Fermi E., *Rend. Accad. Lincei*, **31** (1922) 21, 51, 103.

[10] Fermi E., *Nuovo Cimento*, **22** (1921) 176; *Rend. Accad. Lincei*, **14** (1923) 114.

[11] Fermi E., Rend. Accad. Lincei, **6** (1927) 602; **7** (1928) 342.

[12] Thomas L. H., *Proc. Cambridge Philos. Soc.*, **23** (1927) 542.

[13] Ferreirinho J., Ruffini R. and Stella L., *Phys. Lett. B*, **91** (1980) 314; Merloni A., Ruffini R. and Torroni V., *Nuovo Cimento B*, **113** (1998) 123.

[14] Fermi E., *Rend. Accad. Lincei*, **3** (1926) 145; *Z. Phys.*, **36** (1926) 902.

[15] Belloni L., *Eur. J. Phys.*, **15** (1994) 102.

[16] Fermi E., *Meccanica Statistica*, in *Encicl. Italiana di Scienze, Lettere ed Arti* (Ist. Treccani), **32** (1936) 518.

[17] Fermi E., *Thermodynamics* (Prentice-Hall) 1937; *Rend. Accad. Lincei*, **32** (1923) 395; Fermi E. and Von Neumann J., *Taylor Instability of Incompressible Liquids*, Document AECU-2979, parts I & II (1951-3).

[18] Carati A. and Galgani L., this issue, p. 1017.

[19] Allahverdyan A. E. and Gurzadyan V. G., this issue, p. 947.

[20] Ehrenfest P. and T., *The Conceptual Foundations of the Statistical Approach in Mechanics*, 1912 (Cornell Univ. Press) 1959.

[21] Lan B. L., *Eur. Phys J.*, **23** (2002) L29.

[22] Dresselahaus M. S., *Fermi as a Teacher*, Fermilab Notes (2001).

[23] Goldberger M., *Phys. Persp.*, **1** (1999) 328.

[24] Cooper N. G. (Editor) *From Cardinals to Chaos* (Los Alamos Science) 1987.

[25] Gallavotti G., in *Conoscere Fermi*, edited by C. Bernardini and L. Bonolis, 76 (see in particular § 4), Soc. Italiana di Fisica (2001).

[26] Brush S. G., *Statistical Physics and the Atomic Theory of Matter* (Princeton Univ. Press) 1983.

[27] Pettini M. and Landolfi M., *Phys. Rev. A*, **41** (1990) 768.

[28] Livi R. *et al.*, *Phys. Rev. A*, **31** (1985) 1039.

[29] Livi R. *et al.*, *Phys. Rev. A*, **31** (1985) 2740.

[30] Livi R. *et al.*, *J. Stat. Phys.*, **48** (1987) 539.
[31] Pettini M., *Phys. Rev. E*, **47** (1993) 828.
[32] Cipriani P. and Di Bari M., *Planet. Space Sci.*, **46** (1998) 1499.
[33] Krylov N. S., *Works on Foundations of Statistical Physics* (Princeton Univ. Press) 1979.
[34] Gurzadyan V. G. and Savvidy G. K., *Astron. Astroph.*, **160** (1986) 203.
[35] Casetti L., Livi R. and Pettini M., *Phys. Rev. Lett.*, **74** (1995) 375.
[36] Casetti L., Pettini M. and Cohen E. G. D., *Phys. Rep.*, **337** (2000) 237.
[37] Pettini M. and Cerruti-Sola M., *Phys. Rev. A*, **44** (1991) 975.
[38] Casetti L. *et al.*, *Phys. Rev. E*, **55** (1997) 6566.
[39] Dauxois T. *et al.*, *Physica D*, **121** (1998) 109.
[40] Lepri S., *Phys. Rev. E*, **58** (1998) 7165.
[41] Cipriani P., Ph.D Thesis, Univ. of Rome "La Sapienza" (1994).
[42] Di Bari M. and Cipriani P., *Planet. Space Sci.*, **46** (1998) 1534.
[43] Cipriani P. and Di Bari M., *Phys. Rev. Lett.*, **81** (1998) 5532.
[44] Gurzadyan V. G., private conversation (2002).
[45] Casetti L. and Pettini M., *Phys. Rev. E*, **48** (1993) 4320.
[46] Benettin G., Galgani L. and Strelcyn J. M., *Phys. Rev. A*, **14** (1976) 2338.
[47] Parisi G., *EuroPhys. Lett.*, **40** (1997) 357.
[48] *Statistical Mechanics: Rigorous Results* (Addison-Wesley) 1969.
[49] Cerruti-Sola M., Cipriani P. and Pettini M., *Mon. Not. R. Astron. Soc.*, **328** (2001) 339.
[50] Cipriani P. and Pettini M., *Astrophys. Space Sci.*, **283** (2003) 347.
[51] Cipriani P., submitted (2003).
[52] Gurzadyan V. G. and Kocharyan A. A., *J. Phys. A*, **27** (1994) 2879.
[53] Gurzadyan V. G., in *Structure and Dynamics of Globular Clusters*, *ASP Conf. Ser.*, **50** (1993) 127.
[54] Hayes W. B., *Phys. Rev. Lett.*, **90** (2003) 054104 and references therein.
[55] Ciccotti G., private conversation (1999).

B.6 J.G. Kirk: Particle acceleration at astrophysical shock fronts: From supernova remnants to gamma-ray bursts

J.G. Kirk: "Particle acceleration at astrophysical shock fronts: From supernova remnants to gamma-ray bursts," *Nuovo Cimento B* **117**, 1117 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

Particle acceleration at astrophysical shock fronts: From supernova remnants to gamma-ray bursts(*)

J. G. Kirk(**)

Max-Planck-Institut für Kernphysik - Postfach 10 39 80, 69029 Heidelberg, Germany

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — In two seminal papers, Fermi outlined the stochastic theory of particle acceleration in astrophysical environments. Fifty years later, a direct descendant of this theory is still the favoured explanation for the problem which motivated Fermi—the acceleration of cosmic rays in our galaxy. More recently, the same basic ideas have been generalised to apply to situations involving relativistic motion, such as active galaxies and gamma-ray bursts. This paper presents Fermi's characteristically simple and powerful ideas, describes their generalisation and assesses their impact on the current status of our ideas concerning the origins of galactic cosmic rays and gamma-ray bursts.

PACS 98.65.Cw – Galaxy clusters.
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

Charged particles change their energy when allowed to move in an electric field. Naturally enough, therefore, early theories of the acceleration of cosmic rays concentrated on locating the electric fields which could be responsible for such high energy particles. This was a nearly impossible task, since the required potentials are enormous (see, for example Swann [1]). Fermi's seminal contribution to this problem [2] was to note that acceleration also results if a charged particle interacts with a magnetic field that is constrained to move because it is frozen into clouds of ionised gas that wander around in interstellar space. His theory was the first stochastic acceleration mechanism. Stochastic theories have a crucial advantage, because the fundamental process responsible for

(*) Paper presented at the IX ICRA Network Workshop "Fermi and Astrophysics" (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.
(**) E-mail: John.Kirk@mpi-hd.mpg.de

an energy increase—in this case reflection from a moving cloud—can affect any particle with a probability essentially independent of its energy. Although the energy increase per event is very small, it can accumulate over a very large number of events; Fermi's theory implies a distribution of particle energies that is a power law extending to energies far in excess of those which could then be observed.

Because of Fermi's work, stochastic theories of particle acceleration which involve small individual energy changes are usually referred to as "Fermi mechanisms". Despite the fact that huge electric potentials ($\sim 10^{17}\,$V) capable of performing the required acceleration in a single step are now know to be present around rotating, magnetised neutron stars, almost all theories of the acceleration of cosmic rays adopt a Fermi-type stochastic mechanism, In particular, the theory of diffusive acceleration at shock fronts, independently proposed in 1977/8 by four different groups [3-6], falls into this category, as does the generalisation of this mechanism to relativistic shock fronts [7], which is potentially important, for example, in understanding the non-thermal radiation from gamma-ray bursts. In this paper, I will briefly review each of these topics, mentioning recent theoretical developments such as the inclusion of anomalous transport models [8] and ultra-relativistic dynamics [9, 10]. I will also briefly review the observational status of two specific applications of these mechanisms: the acceleration of cosmic rays by supernova remnants and the production of gamma-ray burst afterglows.

2. – Stochastic particle acceleration

Fermi [2] computed the change ΔE in a particle's energy upon scattering off a magnetised cloud of speed u and found it to be of first order in the small parameter u/v, where v is the particle speed. Averaging over all scattering angles, he showed that the first-order term vanishes, leaving only a second order contribution. Thus, Fermi's 1949 paper [2] proposed what is now called a "second-order" Fermi mechanism. Acceleration of cosmic rays was balanced by energy losses due to collisions with particles in the interstellar gas. As a result, the theory predicted different spectra for protons and heavy nuclei, contrary to observation. To alleviate this problem, Fermi subsequently proposed that the main process competing with acceleration was escape from the Galaxy [11]. This necessitated a faster acceleration mechanism, which led Fermi to suggest that acceleration occurred primarily in "traps", where the motion of the scattering centres is convergent. In such a situation the first-order contribution to ΔE survives the averaging process, leading to a "first-order" Fermi mechanism.

2˙1. *Diffusive acceleration at shock fronts.* – The theory of diffusive acceleration at shock fronts gives a convincing physical realisation of Fermi's "traps", because the inherent dissipation at the shock front naturally leads to a convergent fluid flow. Scattering centres anchored in such a flow would, therefore, yield a first-order Fermi mechanism. Fluctuations in the the embedded magnetic field could take on such a role, and these are expected to be generated in the vicinity of a collisionless shock front. Thus, the random cloud motion of Fermi's first paper is replaced by ordered fluid motion through a single shock front and the stochastic component is provided by scattering off slowly moving MHD waves in the up and downstream plasmas, rather than off interstellar clouds. This diffusive acceleration mechanism operates on much smaller spatial scales than those envisaged by Fermi and is very much faster. However, the real advantage over Fermi's original theory is that the competing process of escape is now no longer independent of the acceleration mechanism, but is controlled by the same set of scattering centres.

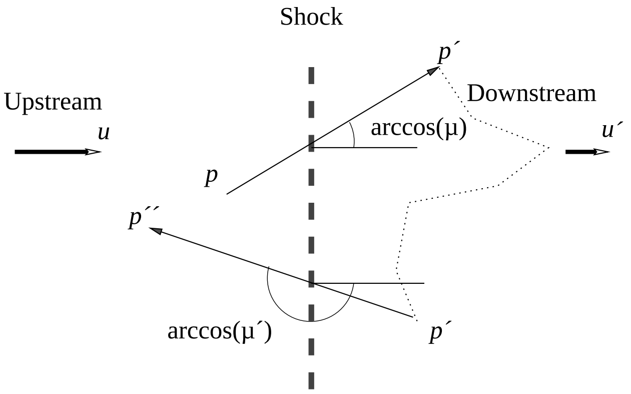


Fig. 1. – Sketch of a particle trajectory at a shock front. Starting upstream with momentum of magnitude p measured in the local rest frame of the plasma, the particle crosses the shock—without interaction—to achieve p' as measured in the local rest frame downstream. Elastic scattering then takes it to a position where it may recross into the upstream region, upon which its momentum in the local fluid frame changes from p' to p'' and the cycle is complete.

The situation is sketched in fig. 1 for a shock front with upstream plasma speed u and downstream plasma speed u', both directed along the shock normal. The simplest approach is to work with the magnitude p of the particle momentum; then, elementary kinematics gives

$$\frac{\Delta p}{p} = \frac{\Delta u}{v}(\mu - \mu') \tag{1}$$

to lowest order in $\Delta u/v$, where $\Delta u = u - u'$ and μ and μ' are the direction cosines shown in fig. 1. The basic assumption of the theory is that the stochastic scattering process keeps the distribution close to being uniform and isotropic, which implies *diffusion* of particles in space. Averaging over an isotropic distribution of particles gives a non-vanishing first-order term in the average gain g:

$$g \equiv \frac{\langle \Delta p \rangle}{p} = \frac{4\Delta u}{3v}. \tag{2}$$

Spatial diffusion implies that the particle density at the shock front equals that far downstream. Then, counting the number of particles entering the downstream region per second over the shock front, and comparing it with the number advected away downstream gives the escape probability per cycle of crossing/recrossing:

$$P_{\rm esc} = \frac{4u'}{v} \tag{3}$$

Following Fermi's reasoning [2] the combination of an average gain given by eq. (2) and an escape probability given by eq. (3) leads to a power-law distribution of accelerated particles:

$$\frac{\mathrm{d}N}{\mathrm{d}p} \propto p^{-1-P_{\mathrm{esc}}/g} \tag{4}$$

that is independent of the details of the scattering process, being determined solely by the compression ratio u/u' of the shock front. This attractive result has motivated a considerable amount of research. One of the main concerns has been the development of a "non-linear" theory, in which the energy imparted to the scattered particles is self-consistently extracted from the plasma motion, producing a modification of the flow and, concomitantly, of the particle density [12].

2˙2. *Non-diffusive acceleration at shock fronts.* – The basic assumption that scattering of charged particles by MHD waves keeps the distribution function almost uniform in space and nearly independent of the direction of the particle velocity is convenient but not convincing, since particle transport in the strongly turbulent plasmas surrounding collisionless shock fronts is known to be "anomalous" rather than diffusive in nature (see, for example, Annibaldi *et al.* [13]).

One particularly important property is that charged particles tend to move along magnetic field lines more easily than across them. This has led to the picture of particle transport as a superposition of diffusion along a field line, which itself has a stochastic component to its direction, causing it to "wander" [14-16]. If this operates at a shock front, the process remains a first-order Fermi mechanism, since each individual shock crossing produces a gain in energy and the first-order contribution to the energy gain cannot average out. Furthermore, the distribution of angles at which a particle crosses the shock is not affected by the fact that the field line upon which it is located wanders in space. As a result, the gain found in eq. (2) is unchanged. However, the statistics of a particle's return to the shock front is modified, especially if the shock normal is perpendicular to the average field direction. This situation has been investigated by Duffy *et al.* [17] and Kirk *et al.* [8], who use propagators appropriate for sub-diffusion, rather than conventional diffusion, and predict a softer spectrum of accelerated particles than found in the diffusive case. These investigations have recently been extended to cover non-isotropic distributions and oblique shocks [18, 19].

2˙3. *Acceleration at relativistic shock fronts.* – Yet another first-order Fermi mechanism at a shock front appears if the plasma speed u is relativistic. In this case, even if we maintain the picture of a stochastic scattering process which drives the distribution towards one that is uniform in space and isotropic in momentum, the simple approximation of spatial diffusion cannot be made. This is because the near-isotropy required in the diffusion approximation can be achieved only if $u/v \ll 1$. In the case of a relativistic flow—and also in the case of mildly suprathermal particles at non-relativistic shocks [20, 21]—there is no alternative but to attempt a solution of the full problem including the angular dependence of the distribution. This was done using an eigenfunction expansion by Kirk and Schneider [7], an approach which has recently been optimised [9] to cover all shock speeds up to (and including) the ultra-relativistic limit $u \to c$. Because the problem is formulated in terms of a scattering operator, a Monte-Carlo approach has also proved illuminating [22, 23, 10].

The angular distribution of those particles accelerated into a power-law distribution $f \propto p^{-s}$, found using the eigenfunction approach, is depicted in fig. 2. As the upstream

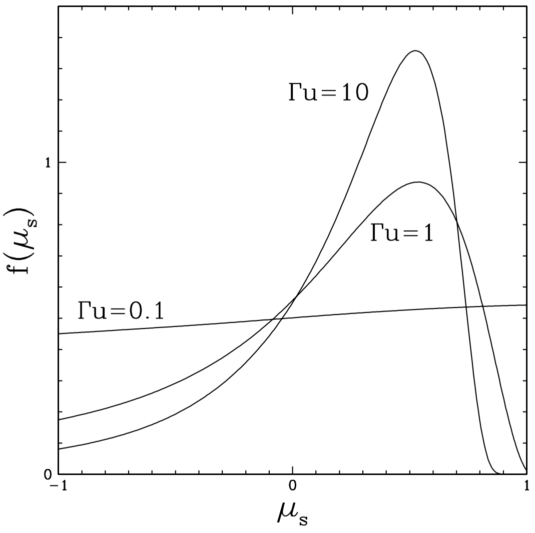


Fig. 2. – The distribution of particles accelerated at a relativistic shock front as a function of the cosine μ_s of the angle between the particle velocity and the shock normal, measured in the frame in which the shock front is stationary and the upstream flow speed u is directed along the normal. Particles moving along the normal from upstream to downstream have $\mu_\mathrm{s} = 1$. The jump conditions assume a gas in thermal equilibrium (Synge/Jüttner equation of state) downstream and a cold upstream gas. Distributions are shown for three different upstream speeds u and the corresponding Lorentz factor Γ. For $\Gamma u = 0.1c$ the distribution is almost isotropic, whereas for $\Gamma u = 10c$ it is strongly peaked and almost indistinguishable from the asymptotic distribution for $u \to c$.

plasma speed increases (Γu in this figure is the spatial part of the upstream 4-speed in units of c) the anisotropy becomes more strongly pronounced. The exact form depends, of course, on the details of the scattering operator. Here we have chosen the case of isotropic diffusion in pitch angle [9]. A good approximation to the angular distribution is provided by the first term in the eigenfunction expansion:

$$f \propto (1 - \mu_\mathrm{s} u/c)^{-s} \exp\left[-\frac{1+\mu_\mathrm{s}}{1-\mu_\mathrm{s} u/c}\right], \tag{5}$$

where μ_s is the cosine of the angle between the particle velocity and the shock normal, measured in the frame in which the shock front is stationary. Fortunately, the predicted energy spectrum seems to be insensitive to the scattering operator, although this statement must be interpreted cautiously, since it relies on testing with a limited set of operators. An interesting aspect of ultra-relativistic shocks in unmagnetised plasma is that as $u \to c$, the compression ratio $r = u/u' \to 3$, independent of the temperature and composition of the plasma. This is mirrored in a unique asymptotic value s_asy of the power law index, as illustrated in fig. 3. For isotropic pitch angle diffusion, this has the value $s_\mathrm{asy} = 4.23 \pm 0.01$.

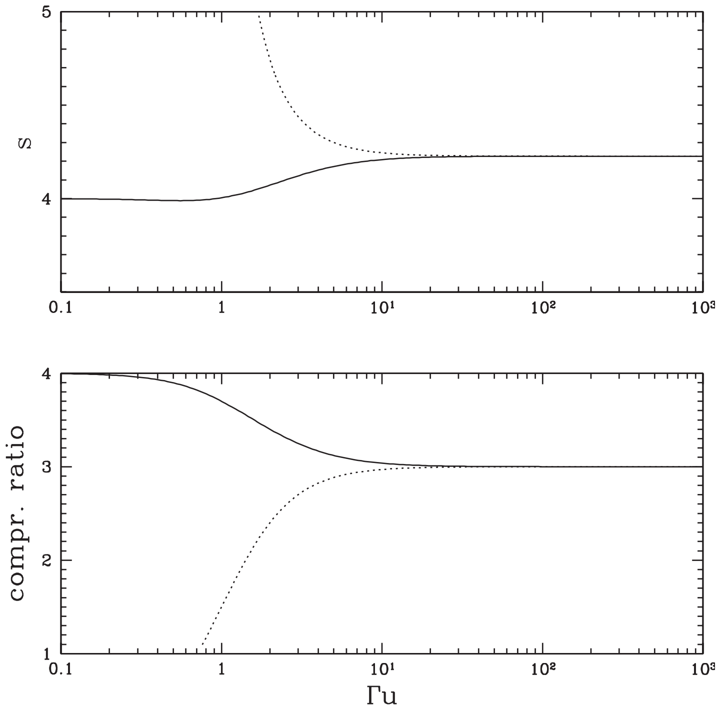


Fig. 3. – The power-law index of the distribution function of particles accelerated at an unmagnetised, relativistic shock front as a function of the spatial component of the upstream 4 speed Γu. Results are shown for two cases: the solid line represents a strong shock (negligible upstream pressure) and the dotted line a relativistic gas (energy density=pressure/3). The lower panel shows the corresponding compression ratios. As the compression ratios tend to the limiting value of 3, the power-law index tends to 4.23.

3. – Acceleration of cosmic rays in supernova remnants

Fermi's motivation for proposing his particle acceleration mechanism was to explain the acceleration of cosmic rays. Supernova are currently the preferred source of the energy required to maintain the Galactic cosmic ray population, and the diffusive acceleration process operating at supernova remnant shocks is the main mechanism under discussion in this connection [24]. This theory predicts a gamma-ray signature [25] which should be visible by current ground-based detectors using the imaging Čerenkov technique. However, because an unambiguous signal has not yet been detected (see, for example Aharonian *et al.* [26]), attention is increasingly being focused on weaknesses in the theoretical approach and the ways in which these might be eliminated [27].

A particularly promising development is the realisation that cosmic rays themselves are likely to generate a magnetic field at a supernova remnant shock front that is much stronger than the field swept up in the interstellar medium [28, 29]. An amplified field has been assumed for several years in the modelling of the radio emission from supernovae (*e.g.*, Chevalier [30]), but has not as yet been incorporated into state-of-the-art calculations of diffusive acceleration at shocks [31]. One immediate implication is that the maximum energy to which a particle can be accelerated within the lifetime of the supernova shock front is increased. This is important, because a major problem with the current theory is that it predicts the maximum energy at a point where the observed spectrum shows no indication of a change in the source behaviour. An amplified field admits the possibility that supernova remnants could accelerate particles to higher energy—perhaps up to the "knee" at 10^{15} eV in the observed cosmic ray spectrum, where a change in the source properties is no longer excluded.

However, field amplification also has a much deeper implication. Because magnetic field is generated predominantly perpendicular to the normal of the shock front, the transport properties of the cosmic rays are more likely to be described by the anomalous process referred to in subsect. **2**˙2, than by the approximation of spatial diffusion underlying the diffusive acceleration theory. The predicted spectral index is substantially changed by anomalous transport, being much softer in the presence of the sub-diffusive effects expected at a perpendicular shock front. This could bring the overall picture of acceleration at supernova remnants more into line with gamma-ray observations [32], which favour a soft spectrum, and also explain why the TeV flux is lower than expected. Support for this idea also comes from recent progress in the theory of cosmic ray propagation, which indicates that the source spectrum should be softer than that predicted by the diffusive acceleration theory [33].

4. – Gamma-ray bursts

Because of their high flux, gamma-ray bursts are thought to contain material moving with very high bulk velocity towards the observer (*e.g.*, Baring and Harding [34]). They therefore join relativistic jets and pulsar winds as favourite objects to which to apply the theory of particle acceleration by the first-order Fermi process at relativistic shocks described in subsect. **2**˙3. However, despite the large number of known bursts, the difficulties of making detailed observations of any single object leave much room for speculation. Three aspects of the explosion which produces a gamma-ray burst are under investigation: the initial energy release, the production of the gamma-rays themselves in the relativistic outflow and the interaction of the outflow with the surrounding medium to produce the "afterglow". Particle acceleration has little to say about the first aspect, which can be modelled without reference to nonthermal particles [35]. Concerning the other two, the theory described in subsect. **2**˙3, if it is to be applied, relies on the existence of a relativistic collisionless shock front. This requires that those particles flowing outwards at high Lorentz factor interpenetrate either the background, or streams of different Lorentz factor, over a spatial scale that is small compared to other length scales of the flow—such as the distance from the site of the explosion, or the lateral size of a collimated component. If this is the case, a fluid treatment of the outflow is appropriate; if not, a multi-fluid or even a kinetic approach is needed (see, for example, Pohl and Schlickeiser [36]). Because of the transformation properties of the magnetic field, a region of interpenetrating relativistic flows is almost certain to correspond to a perpendicular shock front [37]. Particle-in-cell simulations have been performed of configurations similar to this [38,39], suggesting that the region of interpenetration has a thickness of a few ion inertial lengths.

However, the theory as described above concerns only those particles that see the shock front as a discontinuity. Internal shocks—separating regions of different Lorentz factor in the outflow—could conceivably exist in an electron/positron plasma. In this case, those leptons which acquire an energy E somewhat larger than that implied by the relative velocity u_{rel} of the streams (*i.e.* $E > E_{\mathrm{rel}} = m_{\mathrm{e}}c^2/\sqrt{1 - u_{\mathrm{rel}}^2/c^2}$, where m_{e} is the electron mass) could have a gyro length large compared to the shock thickness and be accelerated into a power-law distribution. Observational evidence for this is, however, scant. Gamma-ray burst spectra are most successfully modelled using a spectrum containing multiple power-laws [40] and the index associated with the highest energy particles does not appear to be unique [41], although at least one of them displays precisely the predicted slope [42].

Unless it is capable of sustaining a high rate of pair production in its precursor [43], the blast wave separating the relativistic outflow and the interstellar medium of the host galaxy, will consist of an electron/ion plasma. In this case, an electron must first achieve an energy somewhat larger than $E_{\mathrm{rel}}m_{\mathrm{p}}/m_{\mathrm{e}}$ (m_{p} is the proton mass) before being accelerated into a power-law distribution by the first-order Fermi process. A mechanism which could perform such pre-acceleration for positrons in an electron/ion/positron plasma has been proposed by Hoshino *et al.* [38], and may operate in the Crab Nebula [44], but no such mechanism has yet been found for electrons.

However, accepting that relativistic shocks exist and that electrons are pre-accelerated into the regime of first-order Fermi acceleration, the problem of computing the observational signature of the mechanism remains remains formidable. In part, this is due to the complicated mapping of the emission events in the accelerating flow onto the world line of the observer [45]. Recently, Downes *et al.* [46] have performed relativistic hydrodynamic simulations which take full account of this effect and compute the evolution of a spherically symmetric fireball. A novel feature or their work is the incorporation of a model of particle acceleration by the first-order Fermi mechanism, in which they adopt the “universal” power law index of 4.23. They find that the deceleration of the flow results in a rather harder high energy photon spectrum than would naively be expected, although still close to the observed optical to X-ray afterglow spectra. They also find that results concerning the time evolution of spectral breaks obtained using simplified pictures of the hydrodynamics [47] require significant modification.

5. – Conclusions

Fermi’s ideas on particle acceleration have spawned a substantial amount of activity in astrophysics, and still underlie the dominant paradigms for the acceleration of cosmic rays as well as the acceleration of particles in more exotic, relativistic environments. In the next few years, observations of TeV gamma-rays from supernova remnants by experiments such as HESS, CANGAROO and VERITAS could well demonstrate that these objects are responsible for accelerating cosmic rays, most likely by a first-order Fermi mechanism involving anomalous particle transport at shock fronts. In the case of gamma-ray bursts, rapid progress in our understanding can be expected from multi-wavelength observations of afterglows. However, much theoretical work on the (magneto-)hydrodynamics of the explosion and the nature of the shock precursor is needed before a convincing application of the first-order Fermi process can be contemplated in these objects.

REFERENCES

[1] Swann W. F. G., *Phys. Rev.,* **43** (1933) 217.
[2] Fermi E., *Phys. Rev.,* **75** (1949) 1169.
[3] Krymsky G. F., *Sov. Phys. Dokl.,* **22** (1977) 327.
[4] Bell A. R., *Mon. Not. R. Astron. Soc.,* **182** (1978) 147.

[5] Axford W. I., Leer E. and Skadron G., *Proc. 15th. Int. Cosmic Ray Conf. (Plodiv),* **11** (1977) 132.

[6] Blandford R. D. and Ostriker J. P., *Astrophys. J. Lett.,* **221** (1978) L29.

[7] Kirk J. G. and Schneider P., *Astrophys. J.,* **315** (1987) 425.

[8] Kirk J. G., Duffy P. and Gallant Y. A., *Astron. Astrophys.,* **314** (1996) 1010.

[9] Kirk J. G., Guthmann A. W., Gallant Y. A. and Achterberg A., *Astrophys. J.,* **542** (2000) 235.

[10] Achterberg A., Gallant Y. A., Kirk J. G. and Guthmann A. W., *Mon. Not. R. Astron. Soc.,* **328** (2001) 393.

[11] Fermi E., *Astrophys. J.,* **119** (1954) 1.

[12] Malkov M. A. and Drury L. O'C., *Rep. Prog. Phys.,* **64** (2001) 429.

[13] Annibaldi S. V., Manfredi G. and Dendy R. O., *Phys. Plasmas,* **9** (2002) 791.

[14] Getmantsev G. G., *Sov. Astron. J.,* **6** (1963) 477.

[15] Jokipii J. R., *Astrophys. J.,* **183** (1973) 1029.

[16] Chuvilgin L. G. and Ptuskin V. S., *Astron. Astrophys.,* **279** (1993) 278.

[17] Duffy P., Kirk J. G., Gallant Y. A. and Dendy R. O., *Astron. Astrophys.,* **302** (1995) L21.

[18] Gieseler U. D. J. and Kirk J. G., *Proc. 26th. Int. Cosmic Ray Conf. (Salt Lake City),* **4** (1999) 427.

[19] Ragot B. R., *Astrophys. J.,* **547** (2001) 1010.

[20] Kirk J. G. and Schneider P., *Astron. Astrophys.,* **225** (1989) 559.

[21] Malkov M. A. and Völk H. J., *Astron. Astrophys.,* **300** (1995) 605.

[22] Bednarz J. and Ostrowski M., *Phys. Rev. Lett.,* **80** (1998) 3911.

[23] Bednarz J. and Ostrowski M., *Mon. Not. R. Astron. Soc.,* **310** (1999) L11.

[24] Kirk J. G., Melrose D. B. and Priest E. R., *Plasma Astrophysics*, *Saas-Fee Advanced Course 24*, edited by A. O. Benz and T. J.-L. Courvoisier (Springer-Verlag, Berlin) 1994.

[25] Drury L. O'C., Aharonian F. A. and Völk H. J., *Astron. Astrophys.,* **287** (1994) 959.

[26] Aharonian F. A. *et al.*, *Astron. Astrophys.,* **373** (2001) 292.

[27] Kirk J. G. and Dendy R. O., *J. Phys. G,* **27** (2001) 1589.

[28] Lucek S. and Bell A. R., *Mon. Not. R. Astron. Soc.,* **314** (2000) 65.

[29] Bell A. R. and Lucek S., *Mon. Not. R. Astron. Soc.,* **321** (2001) 433.

[30] Chevalier R., *Astrophys. J.,* **259** (1982) 302.

[31] Völk H. J., Berezhko E. G., Ksenofontov L. T. and Rowell G. P., *Proc. 27th. Int. Cosmic Ray Conf. (Hamburg),* **2** (2001) 470.

[32] Gaisser T. K., Protheroe R. J. and Stanev T., *Astrophys. J.,* **492** (1998) 219.

[33] Ptuskin V. S., *Space Sci. Rev.,* **99** (2001) 281.

[34] Baring M. G. and Harding A. K., *Astrophys. J.,* **491** (1997) 663.

[35] Ruffini R., Bianco C. L., Fraschetti F., Xue S. and Chardonnet P., *Astrophys. J.,* **555** (2001) L113.

[36] Pohl M. and Schlickeiser R., *Astron. Astrophys.,* **354** (2000) 395.

[37] Kirk J. G. and Duffy P., *J. Phys. G,* **25** (1999) R163.

[38] Hoshino M., Arons J., Gallant Y. A. and Langdon A. B., *Astrophys. J.,* **390** (1992) 454.

[39] Smolsky M. V. and Usov V. V., *Astrophys. J.,* **531** (2000) 764.

[40] Preece R. D., Briggs M. S., Mallozzi R. S., Pendleton G. N., Paciesas W. S. and Band D. L., *Astrophys. J. Suppl.,* **126** (2000) 19.

[41] Lloyd-Ronning N. M. and Petrosian V., *Astrophys. J.,* **565** (2002) 182.

[42] Galama T. J., Wijers R. A. M. J., Bremer M., Groot P. J., Strom R. G. Kouveliotou C. and van Paradijs J., *Astrophys. J.,* **500** (1998) L97.

[43] Thompson C. and Madau P., *Astrophys. J.,* **538** (2000) 105.

[44] Gallant Y. A., van der Swaluw E., Kirk J. G. and Achterberg A., to appear in *Neutron Stars and Supernova Remnants, ASP Conf. Series*, edited by P. O. Slane and B. M. Gaensler (2002).

[45] Ruffini R., Bianco C. L., Fraschetti F., Xue S. and Chardonnet P., *Astrophys. J.,* **555** (2001) L107.

[46] Downes T. P., Duffy P. and Komissarov S., to appear in *Mon. Not. R. Astron. Soc.*, astro-ph/0102271.

[47] Sari R. and Piran T., *Astrophys. J.,* **485** (1997) 270.

B.7 S. Ruffo: The Fermi-Pasta-Ulam-Tsingou numerical experiment: Time-scales for the relaxation to thermodynamical equilibrium

S. Ruffo: "The Fermi-Pasta-Ulam-Tsingou numerical experiment: Time-scales for the relaxation to thermodynamical equilibrium," *Nuovo Cimento B* **117**, 1161 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

The Fermi-Pasta-Ulam-Tsingou numerical experiment: Time-scales for the relaxation to thermodynamical equilibrium(*)

S. Ruffo(**)

Dipartimento di Energetica "S. Stecco", Università di Firenze - via S. Marta 3, Firenze, Italy
INFN and INFM - I-50139 Firenze, Italy
ENS-Lyon, Laboratoire de Physique - 46 Allée d'Italie, 69364 Lyon Cedex 07, France

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — The approach to equilibrium of an isolated system is the basic principle of thermodynamics: the so-called zero-principle. Fermi, Pasta and Ulam (FPU) performed the first numerical study of this process for a chain of anharmonically coupled oscillators. The FPU "experiment" has been an amazingly rich source of problems in modern dynamical system theory. Recent results have shown the presence of increasingly long time-scales of the relaxation process as the energy is decreased. States previously classified as "frozen" have been instead discovered to approach very slowly the equipartition state. The dependence of the diffusive time-scale τ_{D} on energy E and number of degrees of freedom N has been found both analytically and numerically for some classes of initial conditions. An interesting extension of the FPU experiment concerns systems with long-range interactions that simulate gravity. Here unconventional thermal behaviors have been found to persist for times which increase with system size.

PACS 05.45.-a – Nonlinear dynamics and nonlinear dynamical systems.
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

The Fermi, Pasta, Ulam (FPU) and Tsingou([1]) numerical "experiment" [1] was the first historical attempt to check the predictions of classical statistical mechanics on the

(*) Paper presented at the IX ICRA Network Workshop "Fermi and Astrophysics" (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.

(**) E-mail: ruffo@avanzi.de.unifi.it

([1]) M. Tsingou contributed to the numerical work and then did not participate in the writing of the report.

long-time behavior of a nonlinear Hamiltonian system with a large number N of degrees of freedom.

It is interesting to recall the motivations of the FPU experiment in Ulam's words [2]:

> *As soon as the machines [the first computers] were finished, Fermi, with his great common sense and intuition, recognized immediately their importance for the study of problems in theoretical physics, astrophyscics and classical physics. We discussed this at length and decided to attempt to formulate a problem simple to state, but such that a solution would require a lengthy computation which could not be done with pencil and paper or with the existing mechanical computers. After deliberating about possible problems, we found a typical one requiring long-range prediction and long-time behavior of a dynamical system. It was the consideration of an elastic string with two fixed ends, subject not only to the usual elastic force of strain proportional to stress, but having, in addition, a physically correct non-linear term. The question was to find out how this non-linearity after very many periods of vibrations would gradually alter the well-known periodic behavior of back and forth oscillations in one mode; how other modes of the string would become more important; and how, we thought, the entire motion would ultimately thermalize, imitating perhaps the behavior of fluids which are initially laminar and become more and more turbulent and convert their macroscopic motion into heat.* pp. [225,226]

As comes out clearly from these words, the experiment plays a relevant role for the basic understanding of the so-called *Zeroth law* of thermodynamics, which states that:

> *An isolated system will, in the course of time, approach a state of "thermal" equilibrium in which all macroscopic variables have reached steady values.*

Fermi *et al.* tried to detect thermal equilibrium by looking at energy equipartition among the quadratic modes (phonons) of an anharmonic oscillator chain (the equipartition "principle" of Boltzmann). Other observables could have been chosen, like temperature, or specific heat, as more recently done in ref. [3]. One should first remark that the number of "physically interesting" thermodynamical observables (the parameters defining a thermal state, or macroscopic variables) is much smaller than the number of degrees of freedom.

The result of this experiment was a big surprise for the authors:

> *The results were entirely different qualitatively from what even Fermi, with his great knowledge of wave motions, had expected. The original objective had been to see at what rate the energy of the string, initially put into a single sine wave (the note was struck as one tone), would gradually develop higher tones with the harmonics, and how the shape would finally become "a mess" both in the form of the string and in the way the energy was distributed among higher and higher modes. Nothing of the sort happened. To our surprise the string started playing a game of musical chairs, only between several low notes, and perhaps even more amazingly, after what would have been several hundred ordinary up and down vibrations, it came back almost exactly to its original sinusoidal shape.*
>
> *I know that Fermi considered this to be, as he said, "a minor discovery". And when he was invited a year later to give the Gibbs lecture (a great honorary*

> *event at the annual Americal Mathematical Society meeting), he intended to talk about this. He became ill before the meeting, and his lecture never took place. But the account of this work, with Fermi, Pasta and myself as authors, was published as a Los Alamos Report.* pp. [226,227]

As emphasized above by Ulam, the expected relaxation to energy equipartition was not revealed, during the time of observation and with low energy initial excitations. Directly from the Los Alamos Report:

> *Certainly, there seems to be very little, if any, tendency towards equipartition of energy among all degrees of freedom at a given time. In other words, the system certainly do not show mixing.*

This also implied that *ergodicity* or the stronger property of *mixing* were not an obvious consequence of the non-existence of analytic first integrals of the motion besides the total energy. What the authors observed, after an initial growth of the energy in the neighbouring modes, was that energy sharing was restricted only to the first few modes, which showed a quite regular dynamics. They did not detect, as expected, a gradual and continuous energy flow from the first excited mode to the higher ones. Even more surprisingly, at later times, almost all the energy was flowing back into the initially excited mode, so that the system displayed a sort of quasiperiodic behavior.

The field remained dormant for several years until, in a pioneering paper, Chirikov and Izrailev [4] showed that, at sufficiently high energy, the FPU model did relax to equipartition. Similar experiments were repeated in Italy [5]. It then became clear that the system had qualitatively different behaviors as the energy E, fixed by the initial condition, was varied.

These results stimulated numerical studies aiming at the determination of the dependence of the different observed behaviors of the FPU system on the number N of degrees of freedom (see ref. [6] for a review).

After several years of investigations, the qualitative picture emerging from all these studies can be summarized as follows. The transition between a *quasi-integrable* behavior to a *mixing* one, is controlled by the energy E. At small energies the motion is weakly chaotic with positive but small Lyapunov exponents, revealing the presence of thin "stochastic" layers in the phase-space, which is mostly filled with Kolmogorov-Arnold-Moser (KAM) [7] tori. This is the energy range where a continuum description of the FPU chain has been found in terms of the modified Korteweg-de Vries equation that displays soliton solutions [8]. We will not comment further on this approach, which has led to results of paramount importance in nonlinear science and solid-state physics. On the contrary, at higher energies the maximum Lyapunov exponent and the Kolmogorov-Sinai entropy (the sum of the positive Lyapunov exponents [9]) rise considerably, revealing the growth of "stochastic" regions. This happens in an energy range which is extensive with N, *i.e.* one can define a "threshold" energy density $\epsilon_{\mathrm{c}} = E/N$ above which chaos is well developed in phase space [10] (this is also known under the name of "strong stochasticity threshold" [11]). Well above the transition region $\epsilon \gg \epsilon_{\mathrm{c}}$ all the signatures of large-scale chaos are present: the number of positive Lyapunov exponent increases with N, the orbit shows a fast diffusion on the constant energy hypersurface, spatio-temporal correlations rapidly decay. Hence, in this region the system reaches thermal equilibrium pretty fast.

In this paper we briefly review the most recent results concerning the scaling, with energy E and number of degrees of freedom N, of the, loosely speaking, "diffusive" time-scale τ_{D} for the relaxation to thermodynamical equilibrium (detected by energy equipartition among linear normal modes) in the FPU model for small energy densities ϵ. We also add a short comment on some recent results found for systems with infinite range interactions, the opposite extreme of the FPU model that considers only nearest neighbour interactions on a lattice. For such long-range forces, we report on the existence of extremely long-lived out-of-equilibrium states.

2. – Transition to equipartition

The FPU system is an approximate model for analyzing the behavior of a classical solid at low temperatures. The reduction of complexity in comparison to the real physical situation is considerable. Only one spatial dimension is considered and the interaction (typically of the Lennard-Jones type) is expanded for small displacements around the equilibrium positions of the molecules, the lattice is weakly anharmonic.

Fermi, Pasta and Ulam considered a one-dimensional chain of oscillators with unit mass and weakly nonlinear nearest-neighbour interaction (the lattice spacing is also taken of unitary length). Calling q_i and p_i the coordinates and, respectively, the momenta of the oscillators, the model is defined by the following Hamiltonian:

$$H = \sum_{i=1}^{N} \frac{p_i^2}{2} + \sum_{i=1}^{N} V(q_{i+1} - q_i) \ , \tag{1}$$

where

$$V(x) = \frac{1}{2}x^2 + \frac{\alpha}{3}x^3 \tag{2}$$

for the so-called α-FPU model, and

$$V(x) = \frac{1}{2}x^2 + \frac{\beta}{4}x^4 \tag{3}$$

for the β-FPU model. A further case was considered by Fermi *et al.*, called "broken linear" for which

$$\begin{aligned} V &= \frac{1}{2}ax^2, \quad |x| < d \\ &= \frac{1}{2}bx^2 + \frac{1}{2}(a-b)d^2, \quad |x| > d \ , \end{aligned}$$

with $a > b$. In this case the nonlinear spring joining two particles has two linear regions with different slope a for small strain and b for stronger strain. This model has never been reconsidered later on, but it would indeed deserve some interest because "exact" event driven numerical codes could be developed for its simulation. Periodic or fixed boundary conditions have been considered. At fixed energy, the coupling constant $\alpha(\beta)$ determines the amount of nonlinearity in the model. Conversely, for a fixed value of $\alpha(\beta)$, the increasing departure from the harmonic behavior is controlled by increasing the energy. It can be easily shown that the dynamics depends only on the parameter $\alpha\sqrt{E}$ (βE).

Hamiltonian (1), written in linear normal coordinates (Q_k, P_k) (phonons) becomes

$$H = \frac{1}{2}\sum_k \left(P_k^2 + \omega_k^2 Q_k^2\right) + V_{\rm I}(\mathbf{Q}) \ , \tag{4}$$

where ω_k are the phonon frequencies, with $\omega_k = 2\sin(\pi k/N)$ for periodic boundary conditions and $\omega_k = 2\sin(\pi k/2(N+1))$ for fixed boundary conditions. The harmonic energy of mode k is defined by $E_k = (P_k^2 + \omega_k^2 Q_k^2)/2$. The potential $V_{\rm I}$ describes the weak interaction among the phonons and typically all phonons interact, although some selection rules are present [13].

The FPU experiment aimed at showing the progressive decorrelation of the system during its temporal evolution. To this end, the authors chose a far from equilibrium initial condition, giving all the energy to the lowest $(k = 1)$ normal mode only, and then calculating the instantaneous energies $E_k(t)$ of all modes. They expected to see a progressively uniform redistribution of energy among all modes, caused by the small anharmonic coupling $V_{\rm I}$ among them. On the contrary they observed the well-known FPU recurrent behavior: energy was flowing back regularly to mode $k = 1$ after an initial share. Return to the initial condition was not exact, but the possibility that relaxation was present on longer times was ruled out by a following numerical experiment, which revealed the "superperiod" phenomenon [12].

In the Los Alamos Report Fermi *et al.* recommended to perform further numerical simulations using directly the equations of motion for the Q_k's. The difficulty consists in that all modes interact, making the calculation of the force extremely lengthy. A method to overcome these difficulties has been recently devised [13] and such simulations have been performed, revealing a pattern of interesting invariant submanifolds, sometimes containing exact solutions.

At higher energies, the equipartition state is reached in a relatively short time. A transition is present from a low-energy region where the system appears not to be approaching equipartition, showing recurrent behavior in time, to a higher-energy region where, on the contrary, equipartition is quickly reached.

The results presented below mostly refer to the β-FPU model, but extensions to the α-FPU sometimes exist.

A useful tool to characterize the approach to equilibrium is the so-called "spectral entropy" [10]. Let us define a weight as the ratio $p_k = E_k(t)/\sum_k E_k(t)$ between the energy of phonon k and the total harmonic energy (as it should be, this is non-negative and $\sum_k p_k = 1$). Define then a Shannon entropy using this weight $S = -\sum_k p_k \ln p_k$. This is a function of time which measures the effective number of excited normal modes by $n_{\rm eff}(t) = \exp[S(t)]$. One usually looks at the fraction of excited modes $n_{\rm eff}/N$, which varies from $O(1/N)$ when a few modes are excited to $O(1)$ at equipartition.

The first clear numerical evidence of the existence of a transition region and of its stability with N was obtained in ref. [10] using "spectral entropy". Above $\epsilon_{\rm c}$, the "spectral entropy" was shown to increase in time, reaching asymptotically its maximal value $\ln(N)$. Below $\epsilon_{\rm c}$ the spectral entropy remained instead close to the value of the initial state. After a convenient normalization, the points showed a tendency to accumulate on a N independent curve. While the behavior above $\epsilon_{\rm c}$ is confirmed also by the most recent numerical simulations, below $\epsilon_{\rm c}$ a much slower relaxation processes to equipartition has been more recently discovered and it is ruled by much longer time-scales. Its origin will be briefly described in the next section.

The transition to equipartition had been indeed suggested by Chirikov and Izrailev [4] using the "resonance overlap" criterion. Let us give a brief sketch of the application of this powerful criterion. The Hamiltonian of the β-FPU model can be rewritten in action-angle variable, considering as an approximation just one Fourier mode (this is justified when most of the energy is still residing in this mode, *e.g.*, at the beginning of the time evolution)

$$H = H_0 + \beta H_1 \approx \omega_k J_k + \frac{\beta}{2N} (\omega_k J_k)^2 \ , \tag{5}$$

where $J_k = \omega_k Q_k^2$ is the action variable (in practice only the nonlinear self-energy of a mode is considered in this approximation) and H_0, H_1 are the unperturbed (integrable) Hamiltonian and the perturbation, respectively. If the energy is placed initially in mode k, then $\omega_k J_k \approx H_0 \approx E$. It is then easy to compute the nonlinear correction to the linear frequency ω_k, obtaining the renormalized frequency ω_k^{r}

$$\omega_k^{\mathrm{r}} = \frac{\partial H}{\partial J_k} = \omega_k + \frac{\beta}{N}\omega_k^2 J_k = \omega_k + \Omega_k. \tag{6}$$

When $N \gg k$

$$\Omega_k \approx \frac{\beta H_0 k}{N^2}. \tag{7}$$

If the frequency shift is of the order of the distance between two resonances

$$\Delta\omega_k = \omega_{k+1} - \omega_k \approx N^{-1} \ , \tag{8}$$

(the last approximation being again valid when $N \gg k$), *i.e.*

$$\Omega_k \approx \Delta\omega_k \tag{9}$$

(the last equation expresses the resonance overlap criterion), one obtains an estimate of the threshold energy density multiplied by β, which is the control parameter for the development of sizeable chaotic regions,

$$\beta\epsilon_{\mathrm{c}} \approx k^{-1} \ , \tag{10}$$

with $k = O(1) \ll N$. In other words, a threshold energy density exists below which primary resonances are weakly coupled inducing an extremely slow relaxation process to equipartition. In the beginning the belief was that no relaxation was present and that the states were "frozen" out of equilibrium. Above ϵ_{c}, on the contrary, a fast relaxation to equipartition is observed.

3. – Time-scales

A rapid increase of the relaxation time to equipartition at ϵ_{c} was first revealed in ref. [14]. Later on, a rapid decrease of the maximal Lyapunov exponent $\lambda_{\max}$ was found at ϵ_{c} [11], in correspondence of the transition region. No strong dependence on N of the $\lambda_{\max}$ *vs.* ϵ curve was detected, at sufficiently large N. At low energies $\lambda_{\max} \sim \epsilon^2$, implying that the "Lyapunov time" $\tau_\lambda = \lambda_{\max}^{-1}$, which measure the rate at which microscopic chaotic instabilities develop in time, increases as $\tau_\lambda \sim \epsilon^{-2}$ as ϵ is decreased. This was the first signature of the presence of a power-law for a "typical" time below the strong stochasticity threshold ϵ_c.

Concerning perturbation theory results, the findings obtained using Nekhoroshev estimates have been summarized in ref. [15]. Nekhoroshev theory allowed to evaluate lower bounds for the time variation of the unperturbed actions on times that, though being finite, increase exponentially as the perturbation parameter is decreased. It is possible, using this approach, to find results valid for initial conditions on open sets in the phase space, as opposed to methods based on the Kolmogorov-Arnold-Moser theorem (on the other hand the latter has the advantage to give statements valid for all times). The stability time τ_S of the single unperturbed actions (or action "freezing" time) is found to scale as

$$\tau_S = \tau_* \exp\left[\frac{\beta_*}{\beta}\right]^d , \tag{11}$$

where, in general, both τ_*, β_* and d depend on N. The most important dependence on N is that of d: the best estimates so far obtained for FPU gives $d \simeq N^{-1}$, a result confirmed also by numerical simulations (the estimate appears to be optimal). This result suggests that in the thermodynamic limit the freezing times might become short, or even vanishing, and the region of violation of energy equipartition could disappear. However, such estimates are valid in an energy region that shrinks to zero as N is increased and might therefore be irrelevant for the problem at hand, since the strong stochasticity threshold is found to persist in the thermodynamic limit.

Normal form theory has been used in ref. [16] to find an effective Hamiltonian that describes the interaction among a reduced number of long wavelength modes. The main result is that above a critical energy E_c the system reaches equipartition on a time proportional to N^2; below this critical energy the time needed increases even faster with N (perhaps exponentially). This holds when the initial excitation is given to a subset of low modes whose center k and packet size Δk do not increase with N: the so-called "mechanical" class [17]. If instead $k \propto N$, so-called "thermodynamical" class, the typical time-scale to equipartition increases like N. These predictions were also supported by numerical simulations. The construction of an effective Hamiltonian begins by performing a large N expansion of the dispersion relation

$$\omega_k = 2\sin\left(\frac{\pi k}{2(N+1)}\right) \approx \frac{k}{N} - \left(\frac{k}{N}\right)^3 , \tag{12}$$

we are treating now the system with fixed ends and neglecting all factors $O(1)$ or $O(\pi)$ in the approximation. A four-wave resonance relation ($k_1 + k_2 + k_3 + k_4 = 0$) is then considered producing in the resonant normal form those angles which are slowly (adiabatically) varying; these latter are found to be $\theta_s = \theta_1 + \theta_3 - 2\theta_2$ and $\theta_{sp} = \theta_2 + \theta_4 - 2\theta_3$. This corresponds to the presence of a modified linear frequency

$$\omega_k^l = \omega_1 + \omega_3 - 2\omega_2 \approx \frac{k}{N^3} . \tag{13}$$

Then the frequency shift (7) can be written as

$$\Omega_k \approx R\omega_k^1 \, , \tag{14}$$

with $R = \beta N H_0$ the new resonance overlap parameter, $R \approx 1$ corresponding to the resonance overlap condition. This parameter controls the deformation of the actions, monotonic in the energy. The angles $\theta_{\rm s}$ and $\theta_{\rm sp}$ are slowly evolving with the frequency $\Omega_k \approx \beta k H_0/N^2$; the latter determines the characteristic evolution time for the actions $\tau \sim N^2/H_0$ for $k \simeq$ const, while $\tau \sim N/H_0$ if $k \propto N$. These results are consistent with numerical simulations [18] obtained by looking at the time evolution of the "spectral entropy". Moreover, since the resonance overlap parameter is proportional to $H_0 N$, chaos is present at very small energy if N is big enough [19]. This last result is consistent with the behavior of the maximal Lyapunov exponent [11].

Energy transfer to higher modes is present, but takes place on much longer times. Actually, it is also known that the energy fraction transferred to the highest modes is exponentially small in mode number [20]. This explains the FPU observation that the dynamics was apparently restricted to only a few long wavelength modes. A nonlinear frequency shift can be also estimated for high modes $\Omega_k^{\rm h} \approx k/N^2$ [16], and when it becomes of the order of Ω_k a Melnikov-Arnold type of argument gives an estimate for the critical energy $E_{\rm c} \approx 1/\beta$ below which no transfer to high modes should be present. This critical energy is however irrelevant in the thermodynamic limit, in this limit the transfer to high modes is always present. Moreover, since the truncated Hamiltonian, studied in ref. [19], does not evolve to equipartition, maintaining an exponential Fourier spectrum for all times, the coupling to high modes is really the crucial effect for the slow evolution towards equipartition below $\epsilon_{\rm c}$. This is also confirmed by a different approach based on the derivation of the breakdown (shock) time $\tau_{\rm shock}$ for the non-dispersive limit of the modified Korteweg-de Vries equation [21], which turns out to be $\tau_{\rm shock} \approx \Omega_k^{-1}$. The development of the shock on the lattice produces the formation of fast spatial oscillations at the shock border, which corresponds to a growth of the short-wavelength Fourier components, a phenomenon which is well known to happen in the integration schemes of the Burgers equation in the zero dissipation limit [22]. It is also reasonable to conjecture that this is the reason why models of the electromagnetic field in a cavity or string models do not show evolution to equipartition, because the linear dispersion relation prevents from coupling high-frequency modes.

This new approach to the study of the time-scale to equipartition below the "strong stochasticity threshold", where primary resonances do not overlap and hence chaos is "weak" can be summarized as follows for the β-FPU model. One can define a "diffusive" time scale to equipartition $\tau_{\rm D}$ looking at the "typical" evolution time of the "spectral entropy", or better $n_{\rm eff}/N$. For "mechanical" initial conditions, k and Δk fixed, *i.e.* not increasing with N, one finds that

$$\tau_{\rm D} \sim \frac{\sqrt{N}}{\epsilon}. \tag{15}$$

This is why these initial conditions are called "mechanical", they are initial conditions of the "mechanical" finite-N system, which do not scale properly to the thermodynamic limit. As $N \to \infty$ they tend to live forever, never reaching the thermal state of equipartition (to this class did belong the initial condition used in the original FPU experiment). These initial conditions do not respect the *Zeroth law* of thermodynamics and they must be excluded by hand in the construction of thermodynamics from mechanics. These are "special" initial states that, in the Boltzmann approach to the foundations of thermodynamics, live in the "less probable" part of the phase-space and never (as $N \to \infty$) flow to the "most probable" larger part. On the contrary for "thermodynamical" initial conditions, $k \propto N$ and Δk fixed or growing with N

$$\tau_{\mathrm{D}} \sim \frac{1}{\epsilon^3}, \tag{16}$$

i.e. the diffusive time scale is *intensive*, it is finite in the thermodynamic limit. It increases quite sharply as the energy density is decreases, this is why previous numerical studies led to the belief that such initial states where "frozen".

This result is based on a *model* [17] where two crucial assumptions are made: i) that a low-mode set forms where the random phase approximation is valid allowing for the calculation of $\Omega_k \sim \beta E k/N^2$ as for the single mode initial condition; ii) that selective transfer to high-modes (denoted by "h") happens only if Ω_k is bigger than $k\Delta h/N^2$ (Melnikov-Arnold argument). After a short but tricky calculation one gets an effective equation for modal energies

$$\frac{\mathrm{d}E_k}{\mathrm{d}t} = -\left(\frac{2\beta}{N}\right)\omega_k \frac{\beta E}{2\pi} E_k E_{\mathrm{h}} , \tag{17}$$

which gives the diffusive time-scale

$$\tau_{\mathrm{D}} = \frac{2\pi}{(\beta\epsilon)^3} \ln\left(\frac{\pi}{2\beta\epsilon}\right) . \tag{18}$$

Numerically, the logarithmic correction is not detected and the result is consistent with a $1/\epsilon^3$ divergence of the time-scale. This time-scale was first suggested in ref. [23] and is in sharp contrast with the Lyapunov time-scale, which diverges at small energy as $1/\epsilon^2$. Hence, we can conclude that the process of relaxation to equipartition in the FPU model is not regulated by the microscopic chaotic instability, but by the typical time in which an orbit diffuses in phase-space, which is determined by the interaction among the phonons.

4. – Non-equilibrium states

The last part of the FPU report, very rarely discussed in the literature, contains some speculations that could justify the persistence of the oscillator chain in a non equilibrium state. In the authors words:

> *What is suggested by these special results is that in certain problems, which are approximately linear, the existence of quasi-states may be conjectured.*

The authors recall the results of the Frobenius-Perron (FP) theorem for products of random matrices with positive elements, which states that an asymptotic fixed vector is obtained from the product, contrary to what happens for ergodic motion. They even conjecture a theorem, that to my knowledge has never been proven, which generalizes the FP theorem to nonlinear transformations. However, what we have discussed in the previous section seems to exclude the presence of non-equilibrium states in the FPU models if the system is large enough (thermodynamic limit) and for a large class of initial conditions.

Recently, there are indications that such states could be present in models with long-range interactions [25]. This could be of evident interest for gravitational systems. We have for instance considered a system of particles of unit mass moving on a circle with Hamiltonian

$$H = \sum_{i=1}^{N} \frac{p_i^2}{2} + \frac{1}{N} \sum_{i<j=1}^{N} \cos(q_i - q_j) \ , \tag{19}$$

where $q_i \in [0, 2\pi)$ is the position of the i-th particle and p_i its conjugate momentum. Observe that particles interact with a repulsive cosine potential and the sum is extended to all the particles. This is an infinite range potential and the interaction is rescaled by $1/N$ following Kac's prescription [24]. When the kinetic energy is small and the particles are initially homogeneously distributed on the circle a striking clustering phenomenon takes place which leads to the formation of a density profile with two peaks at distance π on the circle [25]. Such a state, which is not predicted by both microcanonical and canonical statistical mechanics, show a very weak degradation as time goes on, but its stability increases as the number N of particles is increased. Simulations performed with $N = 10000$ display the fast formation of this "bicluster" state, which is stable in the course of time in very long computer simulations (up to times of order 10^8 in proper units). It appears to be a realization of the FPU "quasi-states". We have developed two theoretical approaches to the understanding of this phenomenon. The first one [26] is based on the solution of the Vlasov-Poisson system of equations, that we obtain by taking the $N \to \infty$ mean-field (*i.e.* at fixed volume) limit of Hamiltonian (19). It is important to remark that this approach is based on an exchange in the order of the $t \to \infty$ and $N \to \infty$ limits. Statistical mechanics performs the two limits in the order above, because it requires first ergodicity and then the thermodynamic limit. The Vlasov-Poisson equation is instead found by first taking the $N \to \infty$ limit and only afterwards the $t \to \infty$ limit if one is interested in asymptotic solutions. This exchange of limits might be at the origin of the observed bicluster "quasi-states", which could then be related to the underlying solutions of the Vlasov-Poisson system: in a sense an analogue situation to the one of the soliton solutions of the modified Korteweg-de Vries equation, which is the continuum limit of the FPU model at small energies. The other theoretical approach [27] begins with the observation that the medium is found to display fast oscillations. By averaging over this fast time scale one obtains an effective Hamiltonian for the slow motion, whose statistical mechanical equilibrium solution predicts the formation of a bicluster. Non-equilibrium "quasi-states" thus result as true equilibria of an effective Hamiltonian. We are currently trying to extend these results to interactions that decay in space, like gravity.

5. – Conclusions

In the spirit of Fermi's pionieristic attitude, I would like to conclude by mentioning the fact that Fermi would have liked to study on the computer another problem that nobody could formulate well and work on to my knowledge. Again in Ulam's authobiography [2] one finds that Fermi said one day:

> *It would be interesting to do something purely kinematical. Imagine a chain consisting of very many links, rigid, but free to rotate around each other. It would be curious to see what shapes the chain would assume when it is thrown on a table, by studying purely the effects of the initial energy and the constraints, no forces.* p. 229

I hope that someone will take over this proposal in the next future and that the study of such a "kinematical system" will lead to similar advancements as those stimulated by the FPU "dynamical system".

* * *

I acknowledge useful discussions with A. Vulpiani and the financial support of CNRS, France. I thank T. Dauxois for reading this manuscript.

REFERENCES

[1] Fermi E., Pasta J. and Ulam S., Los Alamos Report LA-1940 (1955), later published in *Collected Papers of Enrico Fermi*, edited by E. Segré, Vol. II (University of Chicago Press, Chicago) 1965, p. 978; also reprinted in *Nonlinear Wave Motion*, edited by A. C. Newell, *Lect. Appl. Math.*, Vol. **15** (AMS, Providence, Rhode Island) 1974; also in *The Many-Body Problem*, edited by D. C. Mattis (World Scientific, Singapore) 1993.

[2] Ulam S. M., *Adventures of a Mathematician* (Charles Scribner's Sons, NY) 1976.

[3] Escande D., Kantz H., Livi R. and Ruffo S., *J. Stat. Phys.*, **76** (1994) 605.

[4] Izrailev F. M. and Chirikov B. V., *Dokl. Akad. Nauk SSSR*, **166** (1966) 57 (*Sov. Phys. Dokl.*, **11** (1966) 30).

[5] Bocchieri P., Scotti A., Bearzi B. and Loinger A., *Phys. Rev. A*, **2** (1970) 2013.

[6] Benettin G., *Ordered and chaotic motions in dynamical systems with many degrees of freedom*, in *Molecular Dynamics Simulation of Statistical Mechanical Systems*, *Proceedings of International School of Physics "Enrico Fermi", Course XCVII*, edited by G. Ciccotti and W. G. Hoover (North-Holland, Amsterdam) 1986.

[7] Kolmogorov A. N., *Dokl. Akad. Nauk SSR* **98** (1954) 527; Arnold V. I., *Russ. Math. Surv.*, **18** (1963) 9; Moser J. K., *Nachr. Akad. Wiss. Göttingen, Math.-Phys. Kl.*, **2** (1962) 1.

[8] Zabusky N. J. and Kruskal M. D., *Phys. Rev. Lett.*, **15** (1965) 240.

[9] Ruelle D., *Chaotic Evolution and Strange Attractors* (Cambridge University Press) 1989; Eckmann J. P. and Ruelle D., *Rev. Mod. Phys.*, **57** (1985) 617.

[10] Livi R., Pettini M., Ruffo S., Sparpaglione M. and Vulpiani A., *Phys. Rev. A*, **31** (1985) 1039; Livi R., Pettini M., Ruffo S. and Vulpiani A., *Phys. Rev. A*, **31** (1985) 2740.

[11] Pettini M. and Landolfi M., *Phys. Rev. A*, **41** (1990) 768.

[12] Tuck J. L., Los Alamos Report LA-3990 (1968); Tuck J. L. and Menzel M. T., *Adv. Math.*, **9** (1972) 399.

[13] Poggi P. and Ruffo S., *Physica D*, **103** (1997) 251.

[14] Kantz H., *Physica D*, **39** (1989) 322.

[15] Galgani L., Giorgilli A., Martinoli A. and Vanzini S., *Physica D*, **59** (1992) 334.

[16] De Luca J., Lichtenberg A. J. and Lieberman M. A., *Chaos* **5** (1995) 283.

[17] De Luca J., Lichtenberg A. J. and Ruffo S., *Phys. Rev. E*, **60** (1999) 3781.

[18] Kantz H., Livi R. and Ruffo S., *J. Stat. Phys.*, **76** (1994) 627.

[19] Shepelyansky D. L., *Nonlinearity*, **10** (1997) 1331.

[20] Fucito F., Marchesoni F., Marinari E., Parisi G., Peliti L., Ruffo S. and Vulpiani A., *J. Phys. (Paris)*, **43** (1982) 707; Livi R., Pettini M., Ruffo S., Sparpaglione M. and Vulpiani A., *Phys. Rev. A*, **28** (1983) 3544.

[21] Poggi P., Ruffo S. and Kantz H., *Phys. Rev. E*, **52** (1995) 307.

[22] Levermore C. D. and J-G Liu, *Physica D*, **99** (1996) 191.

[23] Casetti L., Cerruti-Sola M., Pettini M. and Cohen E. G. D., *Phys. Rev. E*, **55** (1997) 6566.

[24] Kac M., Uhlenbeck G. and Hemmmer P. C., *J. Math. Phys.*, **4** (1963) 216; **4** (1963) 229; **5** (1964) 60.

[25] Dauxois T., Holdsworth P. and Ruffo S., *Eur. Phys. J. B*, **16** (2000) 659.

[26] Barré J., Dauxois T. and Ruffo S., *Physica A*, **295** (2001) 254.

[27] Barré J., Bouchet F., Dauxois T. and Ruffo S., *Birth and long-time stabilization of out-of-equilibrium coherent structures*, submitted to *Eur. Phys. J. B* (2001) [cond-mat/0203013]; Barré J., Bouchet F., Dauxois T. and Ruffo S., *Out-of-equilibrium states as statistical equilibria of an effective dynamics*, preprint (2002) [cond-mat/0204407]; Leyvraz F., Firpo M.-C. and Ruffo S., *Inhomogeneous quasi-stationary states in a mean-field model with repulsive cosine interaction*, to be published on *J. Phys. A* [cond-mat/0204255].

B.8 C. Sigismondi, F. Maiolino: Enrico Fermi and the statistics of comets

C. Sigismondi, F. Maiolino: "Enrico Fermi and the statistics of comets," *Nuovo Cimento B* **117**, 1207 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

Enrico Fermi and the statistics of comets(*)

C. Sigismondi(1)(**) and F. Maiolino(2)(***)

(1) *Dipartimento di Fisica, Università di Roma "La Sapienza"*
P.le A. Moro 5, 00185 Rome, Italy
ICRA - P.zza della Repubblica 10, 65100 Pescara, Italy
(2) *Dipartimento di Biologia, Università di Roma "La Sapienza"*
P.le A. Moro 5, 00185 Rome, Italy

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — For his abilitation thesis at the "Scuola Normale Superiore" of Pisa, Enrico Fermi presented in 1922 a theorem of statistics with an application to the case of comets. He studied comets with coplanar orbit to that one of Jupiter, and neglected the influence of other planets. The probability of ejection of the comets from the solar system after interacting with Jupiter is calculated, as well as the probability of impact on Jupiter. We discuss those results comparing them with modern issues in solar system cosmogony (Oort Cloud, Kuiper Belt). We apply the calculation of Fermi to the case of the Earth, in order to recover the time rate of comets collision with our planet, which reliably produced the extinction of the dinosaurs.

PACS 95.55.-n – Astronomical and space-research instrumentation.
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

In sect. **2** we present the theorem demostrated by Fermi and his application to the comets interacting with the Sun and Jupiter. In sect. **3** are shown all the data, known to Oort, relative to the dynamics of comets. Jan Oort (1900-1992) was one of the most prominent experts of comets' dynamics in the last century. The main difference with the simple model presented by Fermi is that comets belong to a cloud surrounding the

(*) Paper presented at the IX ICRA Network Workshop "Fermi and Astrophysics" (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.
(**) E-mail: sigismondi@icra.it
(***) E-mail: francesca@videor.it

solar system at a distance of (5–10) $\cdot 10^4$ Astronomical Units (AU), the Oort Cloud, and they enter the solar system with a random inclination. We revisited the comets' impacts applying Fermi's law to the Earth.

2. – A theorem of probability applied to comets

On June 20th, 1922 Enrico Fermi presented a theorem of probability with some astronomical applications as abilitation thesis to the "Scuola Normale Superiore" of Pisa. While such theorem had been published in 1926 in the Journal "Il Nuovo Cimento" [1], the astronomical case remained unpublished until 1959, when the original manuscript has been found in the Domus Galileiana in Pisa [2].

Fermi proved a lemma of a Laplace's theorem, nowadays known as *Central Limit Theorem* and applied this lemma to the dynamics of the comets.

2˙1. *Fermi's lemma.* – Let $y_1, y_2, ..., y_n$ be n random variables uncorrelated with the same given statistical distribution $\rho(y)$, the probability $P_{>a}$ that at least one of the quantities $\{y_1, y_1 + y_2, ..., \sum_1^n y_n\}$ exceeds a, where $a > 0$, is given by

$$P_{>a} = 2/\sqrt{\pi} \int_{\frac{a}{\sqrt{2nk^2}}}^{\infty} e^{-x^2} \mathrm{d}x \tag{1}$$

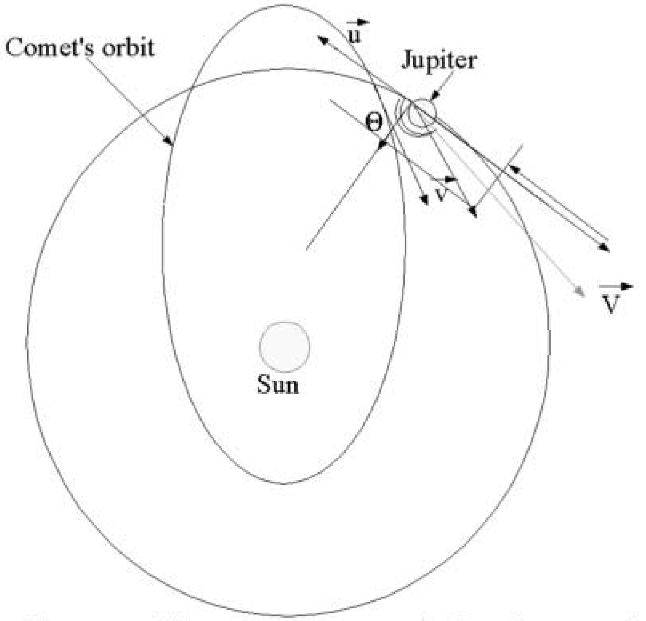
if $a \gg k$ and k^2 is the square average of y, say $k^2 = \int_{-\infty}^{+\infty} y^2 \rho(y) \mathrm{d}y$; in particular, if $n \to \infty$, $P_{>a} \to 1$.

2˙2. *Astronomical application.* – Fermi applied this theorem to the motion of a comet, under the influence of Jupiter. The interaction of the comet with Jupiter is a typical "Circular Restricted Three-Bodies Problem" in the solar system. The efforts of many famous mathematicians have been devoted to this difficult problem, including Euler and Lagrange (1772), Jacobi (1836), Hill (1878), Poincaré (1899), Levi-Civita (1905), and Birkhoff (1915). In 1772, Euler first introduced a synodic (rotating) coordinate system. Jacobi (1836) subsequently discovered an integral of motion in this coordinate system (which he independently discovered) that is now known as the Jacobi integral. Hill (1878) used this integral to show that the Earth-Moon distance remains bounded from above for all time (assuming his model for the Sun-Earth-Moon system is valid), and Brown (1896) gave the most precise lunar theory of his time [3]. Fermi made the following assumptions:

- The mass of Jupiter (m) is negligible respect to that one of Sun (M).
- The orbit of the comet is coplanar to Jupiter's one.
- The orbit of Jupiter is circular.
- The comet has an infinitesimal mass, so that it perturbs neither Jupiter nor the Sun.

If u is the velocity of Jupiter, V the velocity of the comet relative to Jupiter, then the absolute velocity v of the comet and the relative angle θ are related by

$$v^2 = u^2 + V^2 + 2uV \cos\theta. \tag{2}$$



Geometry of the reference systems in Fermi's approach

Fig. 1. – Geometry of the encounters of the comets with Jupiter.

In fig. 1 is represented the geometry described in this paragraph. The modulus of the velocity V is constant during the approach of the comet with Jupiter, the comet is deviated from its original orbit and only the angle θ varies. The changes of this angle θ assume the role of the quantity k of the above lemma in the application of Fermi. The roles of the quantities y of the lemma 2.1 are assumed by the sum of the deviation angles $\Sigma\delta\theta$ of the comet's orbit after every encounter with Jupiter. Once fixed V, θ_0 is the particular value of θ for which the orbit becomes hyperbolical ($v^2 = 2u^2$), and it is given by

$$\cos\theta_0 = \frac{u^2 - V^2}{2uV}. \tag{3}$$

The probability that the comet's orbit becomes hyperbolical crossing n times the orbit of Jupiter is

$$P_{\rm hyp} = 2/\sqrt{\pi}\int_{H/\sqrt{n}}^{\infty} e^{-x^2}{\rm d}x, \tag{4}$$

where

$$H = \frac{\theta^* - \theta_0}{\sqrt{8mh/\pi RV^2 \sin\theta_0}}, \tag{5}$$

h is a parameter of value $h = 2.5$, R is the radius of Jupiter's orbit, and finally the angle $\theta^* \geq \theta_0$ is for an initially elliptical orbit.

This probability goes to unity when n tends to infinite.

Another application of this theorem is the calculation of the probability that the comet hits Jupiter at its first encounter. The probability that the collision happens during the first crossing is given by

$$P_{\text{first}} = \frac{1}{\pi R \sin\theta_0}\sqrt{\rho^2 + \frac{2Gm\rho}{V^2}}, \tag{6}$$

where ρ is the sum of Jupiter and comets radii; while the probability that the collision happens after n times is

$$P_{n^{th}} = \frac{2e^{-n\cdot P_{\text{first}}}}{\sqrt{\pi}} P_{\text{first}} \int_0^{H/\sqrt{n}} e^{-x^2}\mathrm{d}x. \tag{7}$$

Finally the probability that the collision never happens is

$$P_{\text{never}} = e^{-2P_{\text{first}}H}. \tag{8}$$

3. – The Oort Cloud and the dynamics of comets

Jan Oort, who worked at the University of Leiden from 1924 to 1992, studied stellar dynamics with Jacobus C. Kapteyn at Gröningen. In 1927 Oort confirmed Bertil Lindblad's hypothesis on galactic rotation analyzing motions of distant stars. Oort found evidence for differential rotation and founded the mathematical theory of galactic structure. During World War II Oort started with Hendrik C. van de Hulst the successful search for a radio spectral line and after the war Oort led the Dutch group who used the 21 cm line to map hydrogen gas in the Galaxy. They found the spiral structure, the galactic center, and the motion of gas clouds. In 1950 Oort proposed the model for the origin of comets [5], which is nowadays generally accepted. He later showed that light from the Crab Nebula is polarized, confirming Iosif S. Shklovskiis suggestion of synchrotron radiation. He continued researching galaxies and their distribution until shortly before his death at 92 remaining a leader in European astronomy and playing a major role in the rise of many international organizations.

3˙1. *The formation of Oort Cloud and the laws found by Fermi.* – In his second Halley lecture, delivered at Oxford on May 6^{th} 1986 Jan H. Oort [4] reviewed the achievements on cometary dynamics and in particular the origin and dissolution. This lecture followed the first one after 35 years, held by the same author [5], where he showed that there are not comets with negative semiaxes, coming from outside the Solar System along a hyperbolic orbit [5]. Once presented the data on the distribution of the semiaxes of long-period comets, which range between $4\cdot 10^4$ and $2\cdot 10^5$ AU, he discussed where the comets have been formed and how their orbital parameters are distributed. Since the density of pre-solar nebula at the above distances cannot explain the formation of kilometer-sized bodies as the nuclei of comets, their origin has to be found in the inner part of the solar system in a region where water could condensate, at distances like those of Saturn or Uranus.

Being bodies smaller than the planets, the orbits of comets may have not been perfectly circular, therefore susceptible of expulsion by planets. Some comets left the Solar System as a consequence of a single encounter with a planet, but in most cases they have been gradually diffused into larger elongated orbits. The planetary perturbations

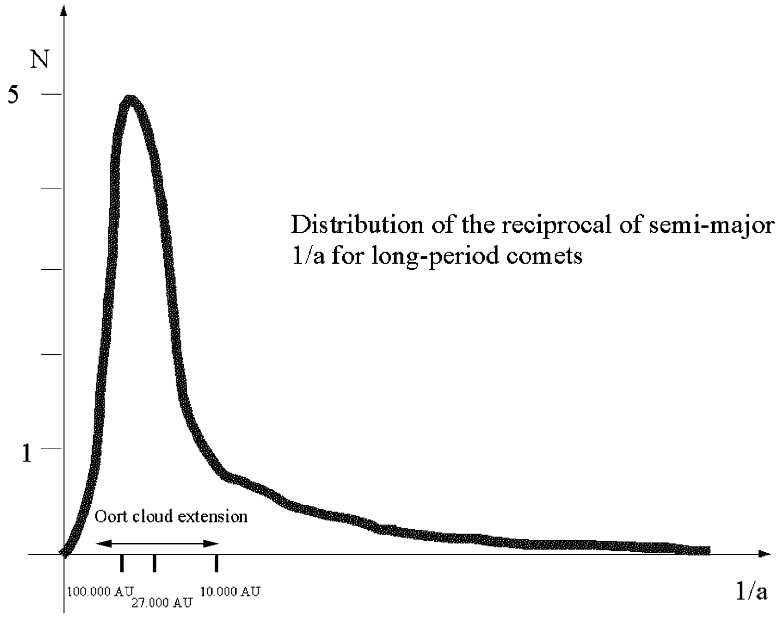


Fig. 2. – Distribution of semiaxes' reciprocal and identification of Oort Cloud.

affected the semi-major axes, while the other orbital elements (perihelion and aphelion, inclination) have been changed by stellar perturbations, when the comets orbits extended approximately to $3 \cdot 10^4$ AU.

The case of coplanar orbits contemplated by Fermi is, therefore, not only a simplification of the problem, but it reproduces the conditions of the early Solar System, when the comets were belonging to the protoplanetary disk.

The Oort Cloud is depleted by new comets going inward; the solar sistem comets have the chance to break-up during a passage near the Sun, moreover they loose continuously matter becoming invisible at subsequent returns. In this way the net flux of comets inside the Solar System is larger than the flux outwards, but the distribution of fig. 2 is maintained constant by the above break-ups and consumptions.

3˙2. *The Kuiper Belt.* – The so-called Edgeworth-Kuiper Belt is a region of the outer Solar System first postulated by Edgeworth [7] and then by Kuiper [8] and is the dynamical reservoir of Jupiter's family of comets. Elongated orbits and enough coplanarity with the planets are characteristics of the object belonging to this structure, known as Trans-Neptunian Objects. Although the density of the protosolar nebula could have been sufficient to allow the formation of such bodies, the dynamical history of the Kuiper Belt can also be described within Fermi's theory, as the sketch in fig. 3 suggests.

4. – Application of Fermi's law to the case of a comet impact to the Earth

Let us consider the probability that a comet knocks against the Earth. Supposing coplanar orbits, we rescale Fermi's equation for P_{first}, with $u = V$ in the case of Earth's

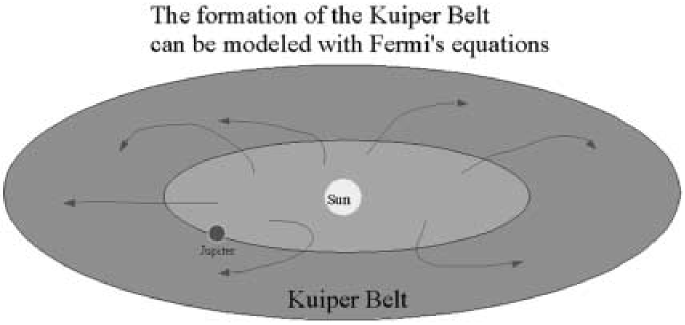

The formation of the Kuiper Belt
can be modeled with Fermi's equations
Sun
Jupiter
Kuiper Belt

Fig. 3. – Mechanism of formation of the Kuiper Belt.

orbit to have

$$P_{\text{first, Earth, coplanar}} = 7 \cdot 10^{-6} \text{ per comet.} \tag{9}$$

The distribution of inclinations for the orbits of new comets being homogeneus, the probability to have a coplanar orbit is that one of having an orbit whose inclination does not exceed the angular extension of the Earth's radius as seen from the Sun, *i.e.* $\alpha \in [-0.024°, +0.024°]$, say $2.7 \cdot 10^{-5}$ of the whole range of possible angles. Therefore, with 20 new comets per year visiting the inner Solar System we can expect a probability of impact per year

$$P_{\text{first, Earth}} = 4 \cdot 10^{-9} \; y^{-1} \tag{10}$$

which corresponds to one impact every 250 million years. This number increases considering further encounters.

D. Steel [6] calculates the same probability considering the cross-sectional area of the Earth, the tangent area for an impact, which is one part over $4 \cdot (1.5 \cdot 10^5/6.4)^2$ of the area of a sphere with radius equal to 1 AU, obtaining the value

$$P_{\text{impact,Earth}} \sim 4.5 \cdot 10^{-10} \text{ per comet.} \tag{11}$$

with a rate of 20 new comets per year entering the inner Solar System. Such formula yields a rate of one cometary impact every 100 millions years, in agreement with our previous estimate with Fermi's formula.

From both approaches it is evident that comets' impacts have characterized the history of life on our Planet, and they were probably the responsible of mass extinctions as that one of dinosaurs. Today such an impact rate every 100 million years is commonly accepted [4].

4˙1. *The case of Comet Shoemaker-Levy 9.* – A recent example of the impact of a comet on Jupiter is the impact of the comet Shoemaker-Levy 9 on July 1994 (see fig. 4). It occurred after a close encounter with the planet during July 1992 at about 95000 km from the center of the planet ($1.33\,R_{\text{J}}$).

Such comet has been discovered about eight months after the 1992 encounter, when it was already trapped around Jupiter with a two years elongated orbit of eccentricity

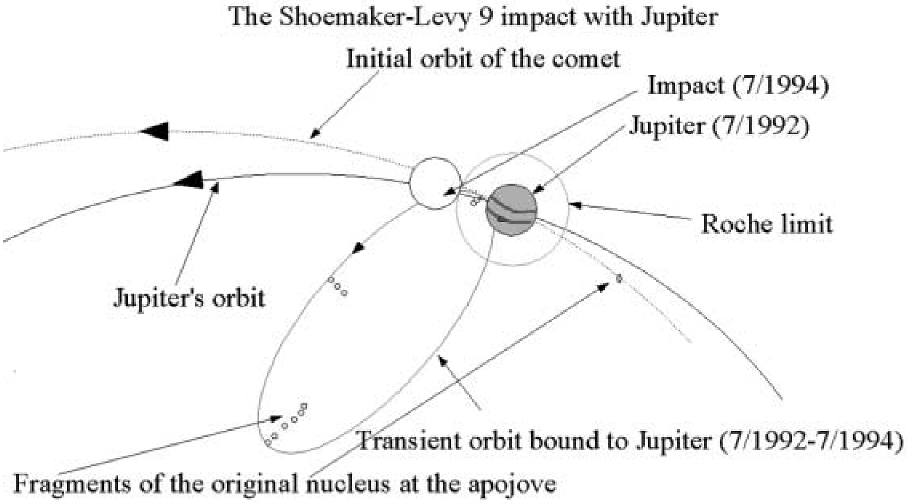

The Shoemaker-Levy 9 impact with Jupiter
Initial orbit of the comet
Impact (7/1994)
Jupiter (7/1992)
Roche limit
Jupiter's orbit
Transient orbit bound to Jupiter (7/1992-7/1994)
Fragments of the original nucleus at the apojove

Fig. 4. – A sketch of the two-years dynamics of the impact on Jupiter of the comet Shoemaker-Levy 9.

0.9965 as a consequence of such close enconter, when the original body was disrupted by tides in 21 large fragments 1–2 km sized [9].

The comet in July 1992 approached Jupiter within the Roche limit [10] L,

$$L \sim R_{\rm J} \cdot (2\rho_{\rm J}/\rho_c)^{1/3} = 1.43\text{–}2.06 R_{\rm J}. \tag{12}$$

At the minimum distance from Jupiter (Perijove) the gravitational influence of Jupiter was about 10^4 times that one of the Sun, while at the Apojove this ratio was near unity. Therefore the initial orbit around the Sun was transformed into a very elongated transient orbit around Jupiter. The combinated action of Jupiter and the Sun at the Apojove was the responsible of the final impact on the planet.

In this case we can not use the criterion of Tisserand [11] to identify an invariant quantity T, as function of the changing orbital parameters,

$$T = \frac{a_p}{a} + 2\sqrt{\frac{a}{a_p}(1-e^2)} \cdot \cos(\hat{\imath}) \tag{13}$$

because of two reasons: 1) the enconter is so close that the mass of Jupiter cannot be posed to zero as to get the formula of Tisserand's invariant T; 2) the initial orbit was bound to the Sun, the final was bound to Jupiter, while in Tisserand's approach the comet is always bound to the same body and perturbed by a body whose semi-major axis is a_p (a, e, $\hat{\imath}$ are comet's orbital parameters).

The treatise on Celestial Mechanics of Tisserand [11] is an update of Pierre-Simon Laplace's work on the same subject, and it is still used as a sourcebook by authors writing on celestial mechanics. It has been quoted also by Fermi, who considered that the study on the influence of Jupiter on comet's dynamics was done only in view of explaining the capture of comets with parabolic orbits when they pass close to Jupiter. Later, as we have seen in sect. **3**, no parabolic comets have been observed. This fact supports the

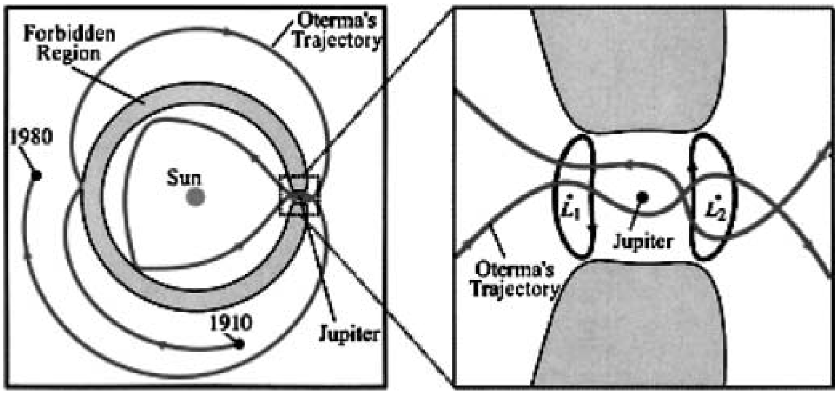


Fig. 5. – Left panel: Plot of a resonance switching orbit of comet Oterma in a coordinate system rotating with Jupiter (as that one used by Fermi). Right panel: expanded view of the L_1 and L_2 region. From Jaffé *et al.*(2002).

approach of Fermi to the problem of the influence of Jupiter to the minor bodies of the solar system: it is a good tutorial for understanding the dynamical history of the inner solar system and its relationship with the outer clouds of comets.

To complete the scenario of dynamical interaction between Jupiter and minor bodies of the solar system it is to mention the case of the Jupiter's family of comets such as Oterma and Gehrels 3. Resonant transitions occur between orbital periods in proportion 2:3 to the period of Jupiter and those 3:2. The comets in those transitions pass through the Lagrangian points L_1 and L_2 (see fig. 5). A new statistical approach [12] has been proposed to study the mass flux of impact's ejecta temporarily orbiting around Mars. This mechanism has been suggested as responsible of spreading the life in the solar system, following the debate on Mars meteorites found in Anctartica [13].

5. – Conclusions

The work of Fermi presented here has been buried in the Domus Galileiana until 1959, about a decade after the first theorization of Oort's Cloud for the origin of comets. Fermi's approach was rather complete to afford the problem of comets' origin once known precise data on their orbits. We applied the formulae derived by Fermi to some data nowadays known with better precision than in 1922, finding agreement with more recent approaches. The comparison with the Oort Cloud and the Kuiper Belt, and the rate of cometary impacts on the Earth are examples of the actuality of Fermi's work, either from a didactic and tutorial point of view.

REFERENCES

[1] Fermi E., *Nuovo Cimento*, **3** (1926) 313.
[2] Fermi E., *Fermi Note e Memorie, Accademia Nazionale dei Lincei*, Vol. **I** (1965) 227.
[3] Szebehely V. G., *Theory of Orbits: The Restricted Problem of Three Bodies* (Academic Press, New York) 1967.
[4] Oort J. H., *The Observatory*, **106** (1986) 186.
[5] Oort J. H., *The Observatory*, **71** (1951) 129.

[6] Steel D., *Rogue Asteroids and Doomsday Comets* (John Wiley & Sons Inc., New York) 1995.

[7] Edgeworth K. E., *J. Br. Astron. Assoc.*, **53** (1943) 181; *Mon. Not. R. Astron. Soc.*, **109** (1949) 600.

[8] Kuiper G. P., in *Astrophysics: A Topical Symposium*, edited by J. A. Hynek (McGraw Hill, New York) 1951, p. 357.

[9] Sekanina Z., Chodas P. W. and Yeomans D. K., *Astron. Astrophys.*, **289** (1994) 607.

[10] Rickman H., *The Rev. Mex. A. A. (Serie de Conferencias)*, **4** (1996) 35.

[11] Tisserand F., *Traité de Mecanique Céleste*, 4 vol. (Gauthier-Villars, Paris) 1889-96, second edition (1960).

[12] Jaffé C., Ross S. D., Lo M. W., Marsden J., Farrelly D. and Uzer T., *Phys. Rev. Lett.*, **89** (2002) 011101.

[13] Thomas-Keprta K. L., Wentworth S. J., McKay D. S., Taunton A. E., Allen C. C., Romanek C. S. and Gibson E. K., *Meteoritics & Planetary Science*, **32** (1997) A128.

B.9 C. Sigismondi, A. Mastroianni: Enrico Fermi and X-ray imaging: An overview from his thesis work (1922) to astronomical applications

C. Sigismondi, A. Mastroianni: “Enrico Fermi and X-ray imaging: An overview from his thesis work (1922) to astronomical applications,” *Nuovo Cimento B* **117**, 1217 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

Enrico Fermi and X-ray imaging: An overview from his thesis work (1922) to astronomical applications(*)

C. Sigismondi(1)(2)(**) and A. Mastroianni(1)(***)

(1) *Dipartimento di Fisica, Università di Roma "La Sapienza" - P.le A. Moro 5, 00185 Rome*
(2) *ICRA, International Center for Relativistic Astrophysics*
P.zza della Repubblica 10, 65100 Pescara, Italy

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — Enrico Fermi studied the formation of images with X-rays and presented his first experimental work as dissertation at University of Pisa in the spring of 1922 (*Nuovo Cimento*, **24** (1922) 133 and **25** (1923) 63). Although those seminal ideas are not present in the sources investigated by Riccardo Giacconi and Bruno Rossi (*J. Geophys. Res.*, **65** (1960) 773) when they firstly proposed a telescope for imaging with X-rays, the thesis of Fermi was the most complete on X-rays physics at his time. Fermi used the technique of "mandrels" to form optical surfaces. He was a forerunner to the technique used for the mirrors of Exosat, Beppo-SAX, Jet-X and XMM-Newton telescopes, and this technique is now a mainstay of many optical manufacturing techniques.

PACS 95.55.-n – Astronomical and space-research instrumentation.
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

In occasion of the celebrations of the centennial birthday of Enrico Fermi, born in 1901, we have studied his thesis work made at the University of Pisa in 1922 at the end of undergraduate studies. We have examined his work with the perspective of X-rays Astronomy in mind.

Enrico Fermi entered the University of Pisa in 1918, at age 17, and prepared his thesis work in X-rays imaging four years later. Roberto Vergara Caffarelli [1] has found

(*) Paper presented at the IX ICRA Network Workshop "Fermi and Astrophysics" (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.

(**) E-mail sigismondi@icra.it

(***) E-mail: mastroianni@icra.it

in 1990 the original manuscript in Pisa at "Domus Galileiana" in 1990. Fermi published two papers describing his experimental methods, and the original results [2,3], on those papers we will focus our attention.

In sect. **2** we analyze the methods of production and imaging of X-rays adopted by Fermi, sketching the main physical underlying principles (X-rays production of iron K_α, Bragg diffraction, and the technical solution of casting curved mirrors of Mica with mandrels. Fermi used Bragg diffraction for generating his images, in an astigmatic geometry which is similar to the Rowland circle, used in modern X-ray spectroscopes.

X-ray Astronomy started in early sixties when Riccardo Giacconi and Bruno Rossi designed a X-rays rocket-borne telescope. The exploited reflection of X-rays under grazing incidence, discovered after the thesis of Fermi by Arthur H. Compton in 1923 [4]. The geometry of the grazing-incidence optics was proposed by Hans Wolter [5,6] in 1952 for applications in microscopy, following the studies of Paul Kirkpatrick [7] in 1950. Those themes are outlined in sect. **3**. Incidentally in 1947 Fermi studied a subject similar to X-ray grazing incidence: the reflection of neutrons on mirrors [8].

In sect. **4** we consider Exosat, Beppo-SAX, Jet-X and XMM-Newton X-ray observatories as exemple where the mandrel's technique has been exploited to cast the reflecting surfaces. In this respect Fermi can be considered a forerunner.

In sect. **5** we present some documents of Fermi and Franco Rasetti related to the thesis work of Fermi at Pisa.

Finally in the conclusions we point out the commitment of Fermi with X-rays in the '20s.

2. – Fermi's thesis work

One of the goals of the experimental part of Fermi's thesis was the realization of images with X-rays. Fermi applied the methods of Maurice De Broglie and M. G. Gouy [9,2], who suggested to exploit the Bragg reflection over a cylinder of Mica. M. De Broglie showed the reflection of X-rays over a convex surface of Mica without obtaining real focuses; M. G. Gouy proposed theoretically the geometry of the cylinder to obtain real focuses. The necessity for using Bragg reflection is clarified directly by Fermi's words:

"I raggi Röntgen non subiscono né rifrazioni né riflessioni [...] Ne segue che nell'ottica dei raggi X il problema di ottenere immagini non può, come nell'ottica ordinaria, risolversi per mezzo di lenti o di specchi sferici" ("Röntgen rays are neither refracted nor reflected [...] It follows that in X-rays optics the problem of obtaining images cannot be solved by means of lenses or spherical mirrors, as in the ordinary optics.)" [3].

In Gouy's paper cited by Fermi, the Bragg relation

$$n\lambda = 2dsin\alpha, \tag{1}$$

where n is an integer, λ is the wavelength, d is the distance among the reticular planes, α is the complement of the incidence angle, reads as

$$\frac{1}{\lambda} = \frac{n}{4Rd}\sqrt{x^2 + 4R^2}, \tag{2}$$

where R is the radius of the cylinder and x is the distance showed in fig. 1.

The X-rays source is posed in A, the n-th order image is formed in AB after the reflection on the "circular belt" (dashed in fig. 1) of the cylinder, where the Bragg relation is verified at the n-th order.

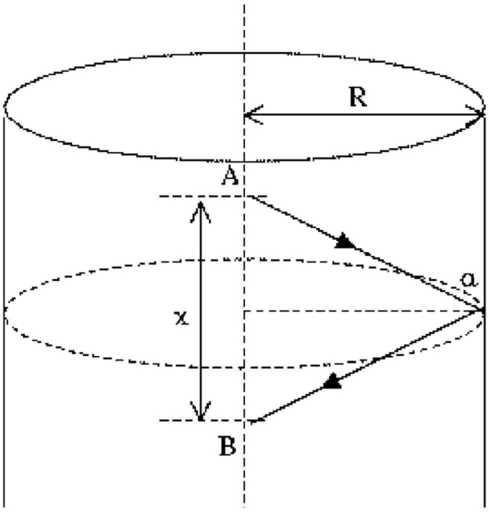


Fig. 1. – Gouy's geometrical construction. This device produces astigmatic images, *i.e.* rays from a monochromatic point source do not pass through a single point in the focal plane, but instead have different foci in the dispersion and in the cross-dispersion directions.

Fermi, instead, considered the more general case in which the source S and the image I are not placed exactly on the axis of the cylinder. He applied the formulas used for spherical mirrors to the projections of the rays on the plane of the "belt", as showed in fig. 2.

The relation between r, r' and the radius of curvature R is given by the equation

$$\frac{1}{r\cos\vartheta}+\frac{1}{r'\cos\vartheta}=\frac{2}{R}, \tag{3}$$

where $\vartheta = \pi/2 - \alpha$ is the angle of incidence. This optical system is astigmatic.

Bragg diffraction is used nowadays for X-rays grating and crystal spectrometers, exploiting the Rowland circle geometry [10,11]. Henry Augustus Rowland (1848-1901) was an American physicist; he determined the value of the ohm and the mechanical equivalent of heat and invented the Rowland diffraction grating for spectroscopy. He published the preliminary results in The Observatory in 1882 [12]. This geometry is very similar to that one used by Fermi and it is shown in fig. 3. The Rowland circle introduces aberrations of order of l^2/R^2, where l is the length of the grating and $L = 2R$ is its radius of curvature. It is an astigmatic optical system, and its principles have been exploited in the design of X-ray grating spectrometers as XMM-Newton [11]. The circle passing per A and A' in fig. 2 represents the grating surface, and it has its center at S.

Fermi posed the source near the axis of a cilynder of radius R, so considering as the grating the cilynder of fig. 2, Rowland's circle has a radius $r = 1/2R$.

Fermi had experienced some X-rays spectroscopy since he was a student. Franco Rasetti, one of the best friends and fellow of Fermi in Pisa, recalled the period in which Fermi learned the basic X-rays tubes techniques on which he based his dissertation:

"In the fall of 1920, three students, Enrico Fermi, Nello Carrara and Franco Rasetti were [...] admitted to the Physics Department at Pisa. Professor Luigi Puccianti, director of the Physics Laboratory, allowed them freedom of initiative to a degree seldom

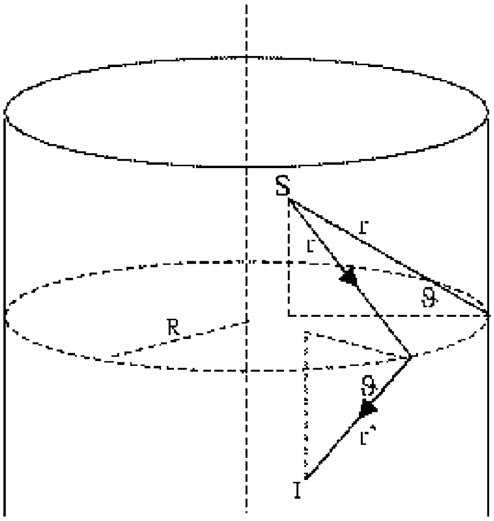


Fig. 2. – Fermi's geometrical construction. The source and the image are both off axis. There are three rays: two of them covering a distance r from the source to the Mica, and another focusing in I at a distance r' from the Mica, determined by eq. (1). The Bragg reflection occurs on the whole dashed circle.

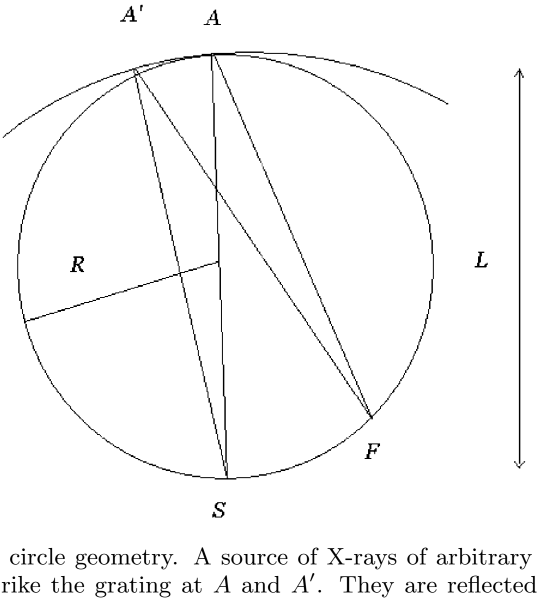


Fig. 3. – The Rowland circle geometry. A source of X-rays of arbitrary wavelength is located at S, and their rays strike the grating at A and A'. They are reflected via Bragg diffraction toward the common focus at F. The grating is curved, and $L = 2R$ is its radius of curvature. The system is astigmatic, and if the length of the grating is $l = A'A$, an aberration of order l^2/R^2 is produced by this geometry. For a given concave diffraction grating in a spectrograph, the circle along which the entrance slit (S in our case), grating (AA'), and focal points (F) of various wavelengths lie is Rowland's circle.

granted to students in Italy or elsewhere. They [...] received keys to the library and instruments cabinets, and were given permission to try any experiment they wished with the apparatus contained therein. Carrara and Rasetti, who [...] had come to recognize Fermi's immense superiority in the knowledge of mathematics and physics, henceforth regarded him as their natural leader, looking to him rather than to the Professors for instruction and guidance" [13, 14].

The experimental tools were used most for didactic purposes, but they were able to modify the devices with their own hands:

"Fermi, after much reading of the pertinent literature, decided that X-rays were the field that offered the best chance for original research, and suggested that all three learned some of the technique. The tubes available were of the gas-filled type [...]. The first task that Fermi set for the group was to produce a Laue photograph [...]. It soon appeared that sealed-in tubes were not fitted for research and the experimenters decided to build their own tubes. The glass part was made to specification by a glass blower, while the physicists had to seal windows and electrodes. [...] Considerable time was spent before these tubes could be satisfactorily operated, but eventually the K radiations of several elements were obtained and observed by Bragg reflection" [13, 14].

The apparatus used by Fermi was made by a cathode sending electrons on a metal surface, the anticathode, which is the X-rays source. The K emission lines series was "made with the lines emitted after the jump of an electron from any upper shell to the K shell" [2].

The anticathode itself played the role of the slit in the ordinary spectroscopy. The system was surrounded by the crystalline mirror and a photographic plate was placed where images should form. The whole system was enclosed in a vacuum tube.

It is to remind that in 1922 spectroscopy was still based on the Bohr-Sommerfeld atomic theory. The complete quantum mechanics came only in 1925-1926, with the works of Werner Heisenberg, Max Born, Pascual Jordan, Erwin Schrödinger, Paul Adrien Maurice Dirac.

The Mica crystal was curved by Fermi himself using a *mandrel*'s technique: the Mica was fixed around a metal cylinder (a brass one, for example) by means of a sealing wax layer (see fig. 4). When the sealing wax cooled, the mirror was ready to be used. Fermi made mirrors of about 4×6 cm^2 excluding the irregular parts, which he checked with ordinary light.

With this experimental apparatus, Fermi was able to produce images with some K lines of iron. In his paper [3] he showed also a bidimensional image of an "X", formed by the X-rays emission of two copper wires posed on the anticathode to form a cross.

3. – Grazing incidence of X-rays on metal

The problem of obtaining imaging properties exploiting grazing incidence has been dealed firstly by Hans Wolter [5, 6]. It is possible to obtain images using a system of two coaxial mirrors, *i.e.* a parabolic and a hyperbolic mirror with the same axis, such as the first focus is coincident with the focus of the second one (see fig. 5).

This configuration obeys to the Abbe sine conditions for aplanaticity [15]. In addition to that condition, fig. 6 shows that with two mirrors in the Wolter 1 configuration a shorter focal length is recovered with respect to only one mirror.

Besides the Bragg diffraction, also the reflection of X-rays on a high Z metal surface is possible under grazing incidence condition. The discovery of this effect is due to Arthur H. Compton, who published his work in 1923 [4]. Using the Drude-Lorentz theory of

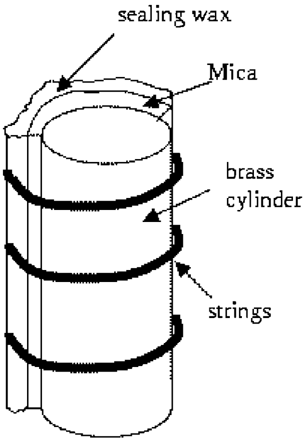


Fig. 4. – Fermi's technique to cast a Mica mirror.

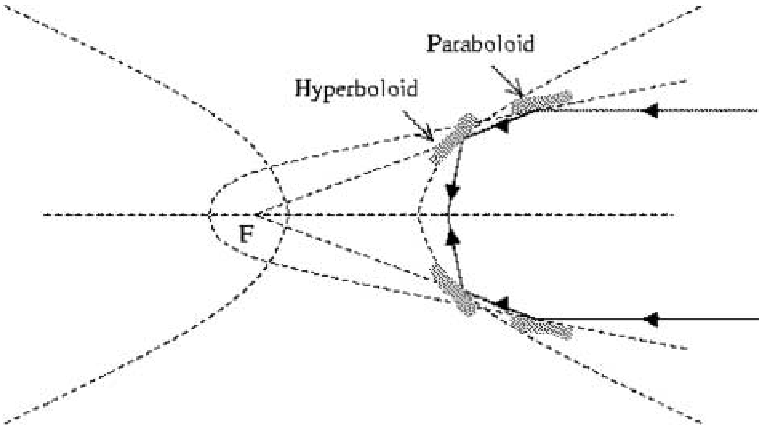


Fig. 5. – Wolter mirror-configuration type 1: Paraboloid and Hyperboloid.

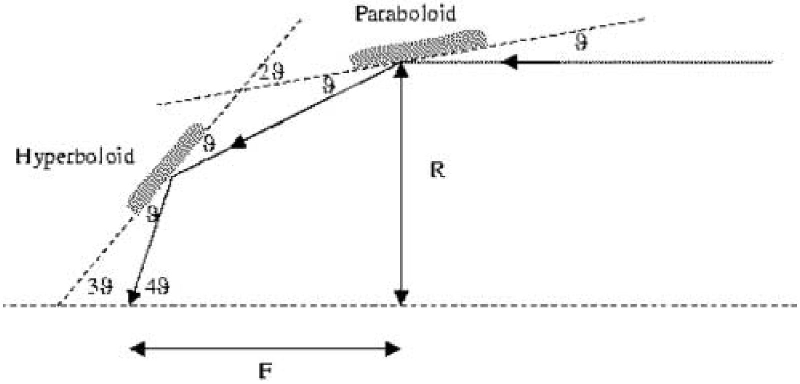


Fig. 6. – The reduction of focal length produced by the second mirror.

Table I. – *Values of ϑ_c for different metals.*

Metal	Z	ϑ_c (degrees) at 1 keV	ϑ_c (degrees) at 8 keV
Al	13	1.7	0.21
Cu	24	2.4	0.30
Ni	28	2.4	0.30
Pt	78	3.6	0.45
Au	79	3.4	0.43

dispersion, Compton explained the deviation from Bragg's law at small grazing angles (the grazing angle is the one between the ray and the surface). It is to remark that Puccianti [16] was ready to feedback Compton's paper in the same year, probably thanks to the work made by Fermi in his laboratory. Puccianti considered that at small grazing angles, the projected reticular distance between atoms is smaller, and a Bragg effect should modulate the new effect discovered by Compton. This effect was expected to be different from metal to metal.

The critical angle can be derived from the straightforward argument of the dependence of the electronic polarizability on the frequency (as reported in almost any textbook on the calssical electromagnetic field). From the polarizability it is easy to obtain an expression for the refraction index n and then for the critical angle.

The model is described by the Newton equation of motion of an electron, bound to the nucleus by an elastic term $-m_e\omega_0^2\vec{x}$ and forced by the electric field $\vec{E}(t) = \vec{E}_0 e^{i\omega t}$.

Putting the solution $x(t)$ that we obtain neglecting the oscillation with ω_0 in the expression of the polarizability

$$\alpha = \frac{Zex(t)}{E(t)} \tag{4}$$

we get for the refraction index in the case of X-rays ($\omega \gg \omega_0$)

$$n \simeq 1 - \frac{2\pi NZe^2}{m_e\omega^2}. \tag{5}$$

from which we get the following expression for the critical angle:

$$\vartheta_c = \sqrt{\frac{4\pi NZe^2}{m_e\omega^2}}. \tag{6}$$

Introducing the atomic weight M and the density ρ of the metal and using the classical radius of the electron $r_0 = e^2/m_e c^2 = 2.82 \cdot 10^{-13}$ cm, we can find an expression whch gives ϑ_c in degrees, provided the energy measured in keV:

$$\vartheta_c = \frac{1.65^\circ}{E(\text{keV})}\sqrt{\frac{Z\rho}{M}}. \tag{7}$$

whose values are shown in table I.

A work by Paul Kirkpatrick [7] and two papers by Hans Wolter [5,6] carried the main highlights considered by Giacconi and Rossi [17] for making the first successful mission with a rocket-borne X-rays telescope in 1962 [18-21].

4. – "Mandrels" techniques in X-rays astronomy

Fermi adopted the technique of mandrels to cast the mirrors for X-rays in his thesis work. Nowadays this technical solution is a mainstray of many optical manufacturing techniques. In X-rays astronomy we have the examples of Exosat LE Mirrors (www.ecc.), Beppo SAX (www.ecc.), Jet-X (www.ecc.) and XMM-Newton (www.ecc.). For example, the telescope of the Italian-Dutch Satellite Beppo SAX is made by 30 confocal and coaxial mirrors nested into each other [22]. The geometry of the mirrors is the tronco-conical approximation of the exact Wolter 1 geometry, therefore the telescope does not focus exactly in one point a source at infinity on axis, but they concentrate in a small spot the photons. Form this property it is called Concentrator. See the paper of one of us [23], for a more complete description of the imaging properties of Beppo SAX Concentrator with extended sources. The procedure for producing each mirror is no longer of bending the material around the mandrels, but the golden mirrors are electroformed around the aluminum mandrels, coated of nickel. Different expansion coefficients with temperature allow the separation of the mirrors from the mandrels [24]. In this way a golden surface with an accuracy of ~ 1 Å has been realized.

5. – Excerpta from E. Fermi and F. Rasetti

In a letter of January 1922 to his friend Enrico Persico, Fermi wrote:
"Caro Enrico, io sto facendo il conferenziere, il relativista, il fisico, [...] In questi giorni ho avuto un po' da fare perché ho dovuto scrivere la mia conferenza sulla relatività". ("Dear Enrico, I am doing the lecturer, the relativist, the physicist, [...]. In these days I was very busy because I had to write my conference on relativity."

In March 1922, closed to his dissertation, in another letter to Persico, we can't notice particular enthusiasm for his thesis work:
"In questi giorni ho avuto e ho parecchio da fare per la mia tesi che, fra parentesi, è venuta una porcheria delle più solenni. Essenzialmente sarà costituita dalle seguenti parti: introduzione con cenno storico e riassunto dello stato attuale della questione; parte teorica consistente in alcuni studi sopra il potere risolutivo nella riflessione sopra cristalli molto sottili in luce curva e nello studio completo dell'effetto dei moti termici sulla riflessione dei raggi X; parte sperimentale consistente nell'ottenere per mezzo di riflessione sopra lamine di mica curve, fotografie dell'anticatodo alla Lockyer. Come vedi il programma è molto modesto. In compenso ha il pregio di essere ormai quasi completamente eseguito". ("In these days I was and I am very busy with my thesis, that, incidentally, has turned out a hearty rubbish. Essentially it will follow this scheme: introduction with a short historical account and summary of the actual knowledge; theoretical part consisting in some studies of the resolving power during reflection on very thin crystals with curved light and in the complete study of the thermal motion effects on X-rays reflection; experimental part consisting in obtaining, by means of reflection on Mica curved surfaces, photographic images of the anticathode à la Lockyer. As you can see the program is very modest. In return, it has the merit of beeing nearly completed".)

We can surely agree with Franco Rasetti, who justified the choice by Fermi of a thesis so far from his interests in that period with this words, in the prologue of Fermi's thesis article that appeared on the collected papers of Enrico Fermi

("At the time in Italy theoretical physics was not recognized as a discipline to be taught in universities, and a dissertation in that field would have been shocking at least to the older members of the faculty. Physicists were essentially experimentalists, and only an experimental dissertation would have passed as physics").

Rasetti stressed also the rare characteristics, typical of Fermi, to excel either in theoretical or in experimental physics, as well as the fact that "Fermi loved most of all to alternate the two activities" [13].

Conclusions

Fermi used X-rays physics again only in his research on molecules and crystals [25] (in the '30s) and in the cited work on the reflection of neutrons [8]. It seems that Fermi was never really involved in X-rays, at least ompared with the deep level reached in his other great contributions. As we have seen, at the time of his thesis, Fermi was extremely active in theoretical physics. In 1922 Fermi was preparing a theoretical thesis for the prestigious Scuola Normale Superiore in Pisa, that he was attending at the same time as the University: the demonstration of some theorems of probability theory to be applied to the comet's motion [26]. Besides, Fermi was really an authority in quantum and relativistic theories, despite of his young age.

Soon after the thesis, he published some papers in relativity theory, among which the one on "Fermi coordinates" in general relativity. Then, after the degree in physics and the diploma at Scuola Normale Superiore (1922), Fermi came back to Rome and worked again in theoretical physics. He will spend some months in Gottingen producing new results in analitical mechanics, then, in Leida, some papers preluding the discovery of the Fermi-Dirac quantum statistics (1926). During the period 1927-1934 about, Fermi became the leader of the celebrated team in via Panisperna, producing the artificial radioactivity with slow neutrons. In 1933 he produced perhaps the most important and elegant result of his career as a theoretist: the theory of beta decay. In 1938 after the Nobel Prize, Fermi went in USA and became an expert in nuclear fission, then in the rising fields of particle physics and computer simulations. He founded also another famous school of physics in Chicago.

Those works of Fermi in X-rays, while containing many seminal ideas, have not directly contributed to the first steps of X-rays astronomy, but it is worth to repeat that the thesis of Fermi was the most complete work on X-rays physics at his time.

REFERENCES

[1] Vergara Caffarelli R., *Nuovo Saggiatore*, **17** (2001) 8.
[2] Fermi E., *Nuovo Cimento*, **24** (1922) 133.
[3] Fermi E., *Nuovo Cimento*, **25** (1923) 63.
[4] Compton A. H., *Philos. Mag.*, **45** (1923) 1121.
[5] Wolter H., *Ann. Phys. (Leipzig)*, **10** (1952) 94.
[6] Wolter H., *Ann. Phys. (Leipzig)*, **10** (1952) 286.
[7] Kirkpatrick P., *Nature*, **166** (1950) 251.
[8] Fermi E. and Zinn W. H., *Reflection of neutrons on mirrors*, Physical Society Cambridge Conference Report, **92**, Chicago (1947).
[9] Gouy M. G., *C. R. Acad. Sci.*, **161** (1915) 765.
[10] Paerels F., *Future X-ray Spectroscopy Missions*, in *X-ray Spectroscopy in Astrophysics*, edited by van Paradijs J. and Bleeker J. A. M., *Lecture Notes in Phyiscs*, Vol. **520** (Springer-Verlag, Berlin Heidelberg) 1999 p. 347.

[11] Willingale R., *New Developments in X-ray Optics*, in *X-ray Spectroscopy in Astrophysics*, edited by J. van Paradijs and J. A. M. Bleeker, *Lecture Notes in Physics*, Vol. **520** (Springer-Verlag, Berlin Heidelberg) 1999, p. 435.

[12] Rowland H. A., *The Observatory*, **5** (1882) 224.

[13] Fermi E., *Note e Memorie*, Vol. I (Accademia Nazionale dei Lincei) 1962.

[14] Segrè E., *Enrico Fermi, fisico* (Zanichelli, Bologna) 1987.

[15] Wilson R. N., *Reflecting Telescope Optics I* (Springer) 1996.

[16] Puccianti L., *Nuovo Cimento*, **25** (1923) 307.

[17] Giacconi R. and Rossi B., *J. Geophys. Res.*, **65** (1960) 773.

[18] Giacconi R., Gursky H., Paolini F. R. and Rossi B., *Phys. Rev. Lett.*, **9** (1962) 439.

[19] Gursky H., *X-rays Astronomy. The first decade*, in *Exploring the Universe*, edited by Gursky H., Ruffini R. and Stella L. (World Scientific Pub., Singapore) 2000.

[20] Rossi B., *Astronomia in raggi X*, *Adunanze Straordinarie per il Conferimento dei Premi Antonio Feltrinelli - Accademia Nazionale dei Lincei*, **1** (1972) 199.

[21] Van Speybroeck L. P., *Historical development of X-rays optics for astronomy*, in *Exploring the Universe*, edited by H. Gursky, R. Ruffini and L. Stella (World Scientific Pub. Singapore) 2000.

[22] Conti G. *et al.*, *Engineering Qualification Model of the SAX X-rays Mirror Unit. Technical data and X-rays imaging characteristics*, *Proc. SPIE*, **2011** (1993) 118.

[23] Sigismondi C., *Nuovo Cimento*, **112** (1997) 501.

[24] Conti G. *et al.*, *Lacquer Coated Mandrels for Production of Replicated X-rays Optics*, *Proc. SPIE*, **1140** (1989) 376.

[25] Fermi E., *Molecole e cristalli* (Zanichelli, Bologna) 1938.

[26] Fermi E., *Un teorema di calcolo delle probabilitá ed alcune sue applicazioni* (Scuola Normale Superiore di Pisa), *Fermi Note e Memorie*, **1** (1922) 227.

B.10 A.Yu. Smirnov: Neutrino mass spectrum and neutrino astrophysics

A.Yu. Smirnov: “Neutrino mass spectrum and neutrino astrophysics,” *Nuovo Cimento B* **117**, 1237 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

Neutrino mass spectrum and neutrino astrophysics(*)

A. Yu. Smirnov(**)

The Abdus Salam ICTP - Strada Costiera 11, 34100 Trieste, Italy
Institute for Nuclear Research - RAS, Moscow, Russia

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — Neutrino astrophysics, and in particular study of solar and supernova neutrinos, plays an important role in reconstruction of the neutrino mass and mixing spectrum. Observable effects of neutrino mass and mixing are based on neutrino transformations in matter. The Fermi coupling constant determines immediately the scale of phenomena and applications of the effects. We consider the status the solar neutrino problem. The MSW LMA solution with $\Delta m^2 = (6–7) \cdot 10^{-5}$ eV^2 and $\tan^2\theta = 0.35–0.4$ gives the best fit of the data. In the case of LMA solution the Earth matter effects allow to explain some features of signals from SN1987A. Future studies of neutrino signals from supernova, and in particular the Earth matter effects on these neutrinos will allow to select or confirm the solution of the solar neutrino problem, identify the type of mass hierarchy (ordering) of the neutrino spectrum and measure or restrict the mixing parameter U_{e3}.

PACS 98.65.Cw – Galaxy clusters.
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

A number of important developments in the neutrino astrophysics is related to the neutrino flavor transformations in matter [1, 2]. The scale of phenomena is given immediately by the Fermi coupling constant, G_F. Namely, G_F determines the minimal width of matter,

$$\mathrm{d} \equiv \int n(r)\mathrm{d}r \ , \tag{1}$$

(*) Paper presented at the IX ICRA Network Workshop “Fermi and Astrophysics” (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.
(**) E-mail: smirnov@ictp.trieste.it

required for strong flavor transformations:

$$d > d_0 \approx G_{\mathrm{F}}^{-1} \ . \tag{2}$$

The width should be larger than inverse Fermi coupling constant. In eq. (1), $n(r)$ is the matter density and the integration is over the neutrino trajectory.

Precise inequality is [3]

$$d > \frac{1}{G_{\mathrm{F}} \tan 2\theta}, \tag{3}$$

where θ is the mixing angle.

Numerically, the minimal width equals

$$d_0 \sim G_{\mathrm{F}}^{-1} \approx 4 \cdot 10^8 A \ \mathrm{cm}^{-2} \ , \tag{4}$$

here A is the Avogadro number. Interestingly, d_0 is of the order of the average width of the matter in the Earth.

$$G_{\mathrm{F}}^{-1} \sim d_{\mathrm{Earth}}. \tag{5}$$

Some insight: at low energies neutrinos undergo, mainly, refraction which is described by the effective potential V. The matter effect on mixing of two neutrinos is given by difference of the potentials: $\Delta V \equiv V_e - V_\mu \sim G_{\mathrm{F}} n_e$. The difference of potentials leads to appearance of the phase difference

$$\Delta\Phi = \int V \mathrm{d}r. \tag{6}$$

The condition for strong matter effect (2) corresponds then to $\Delta\Phi \approx \pi$. The condition (2) is satisfied apart from the Earth for the Sun, the supernovae and in the Early Universe [3].

Reconstruction of the neutrino mass and flavor spectrum is one of the fundamental problems in particle physics. In this paper we consider how conversion effects in matter and astrophysical observations can contribute to determination of neutrino parameters.

1˙1. *Neutrino mass and flavor spectrum.* – We consider neutrino mass spectra with mixing of three flavors: $\nu_\alpha = U_{\alpha i} \nu_i$ $(\alpha = e, \mu, \tau, \quad i = 1, 2, 3)$ which satisfy the following phenomenological conditions.

1) The spectra lead to $\nu_\mu - \nu_\tau$ oscillations with parameters

$$|m_3^2 - m_2^2| \equiv \Delta m_{\mathrm{atm}}^2 = (1.5\text{–}4) \cdot 10^{-3} \mathrm{eV}^2, \qquad \sin^2 2\theta_{\mu\tau} > 0.9, \tag{7}$$

as the dominant mode of the atmospheric neutrino conversion.

2) The spectra reproduce one of the large mixing solutions of the solar neutrino problem, namely, LMA MSW, LOW MSW, Vacuum oscillations (VO) or quasi-vacuum oscillations (QVO), so that

$$|m_2^2 - m_1^2| \equiv \Delta m_\odot^2, \qquad \theta_{12} = \theta_\odot \ . \tag{8}$$

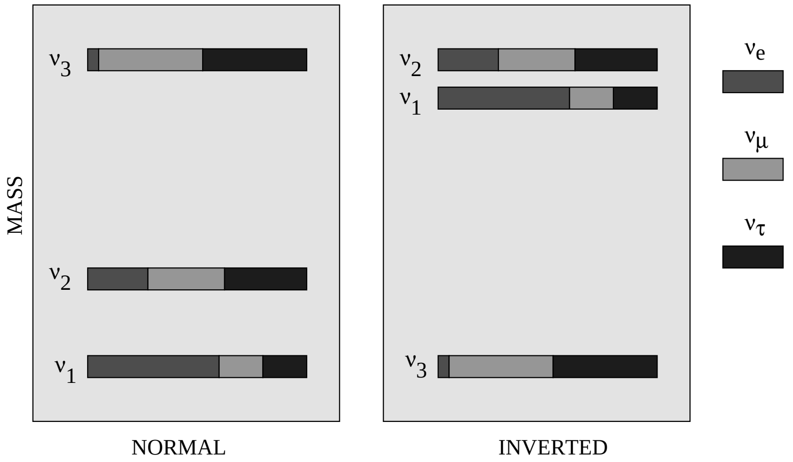


Fig. 1. – Neutrino mass and flavor spectra.

In all the cases

$$\Delta m^2_{\rm atm} \gg \Delta m^2_{\odot}.$$

We will call the pair of the mass states ν_1, ν_2 with split $\Delta m^2_{\odot}$ as the solar pair.

3) The admixture of the electron neutrino in the third mass eigenstate ν_3 satisfies the CHOOZ [4], Palo Verde [5] bounds:

$$|U_{e3}|^2 \lesssim 0.02 \ . \tag{9}$$

These phenomenological constraints lead to two possible spectra (fig. 1) which differ by the type of mass hierarchy or type of ordering of mass eigenstates for the non-hierarchical spectra. In the case of *normal* mass hierarchy (left panel) the solar pair is lighter than ν_3: $m_3 > m_2, m_1$. In the case of inverted mass hierarchy the states of the solar pair are heavier than ν_3: $m_3 < m_2 \approx m_1$ (right panel).

There are still the following unknown in the spectrum:

- Type of the neutrino mass spectrum: hierarchical, non-hierarchical (all masses are of the same order), degenerate. The determination of the type of spectrum is related to the question of the absolute mass scale.
- Type of mass hierarchy or ordering (in the case of non-hierarchical or degenerate spectra): normal or inverted;
- Value of U_{e3};
- Parameters of the solar pair, $\Delta m_{\odot}$, $\theta_{\odot}$, which depend strongly on specific solution of the solar neutrino problem;
- CP-violation phases.
- Existence of additional mass eigenstates.

In what follows we will discuss items 2–4, that is, the identification of the solution the solar neutrino problem, the type of mass hierarchy and value of U_{e3}.

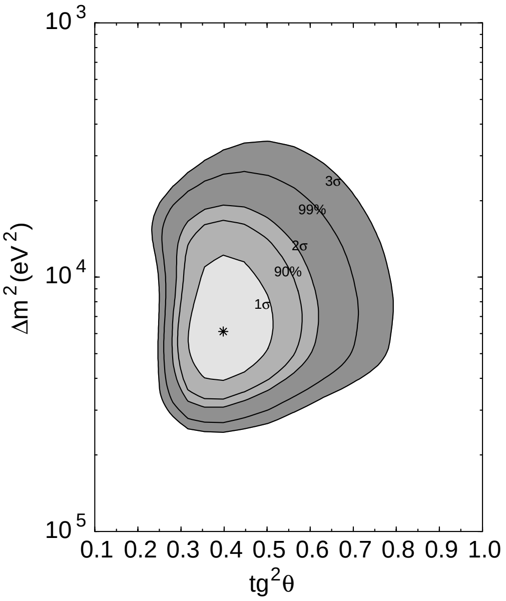


Fig. 2. – The global LMA MSW solution. The boron neutrino flux is considered as free parameter. The best fit point is marked by a star. The allowed regions are shown at 1σ, 90% C.L., 2σ, 99% C.L. and 3σ.

2. – SNO results and solution of the solar neutrino problem

The latest SNO results which include the day and night energy spectra of events [6], give very strong evidence of the neutrino flavor conversion. As far as specific mechanism of conversion is concerned, the data further favor the LMA MSW solution of the solar neutrino problem.

The global analysis of all available data leads to the best fit point [7]

$$\Delta m^2 = 6.2 \cdot 10^{-5}\ \text{eV}^2, \quad \tan^2\theta = 0.40. \tag{10}$$

In fig. 2 we [7] present the region of the LMA solutions in the (Δm^2- $\tan^2\theta$) plot. Shown are contours of confidence level with respect to the best fit point.

The VO-QVO and LOW solutions are accepted at about 3σ level with respect to the best fit LMA solution.

Further progress in this field will be related to KamLAND experiment and thento BOREXINO experiment. They are able to identify the solution and to measure the solar oscillation parameters with rather high accuracy.

Unfortunately not too much can be sad from these studies about other unknown of the spectrum. Both the solar neutrino data (from present experiments) and KamLAND have rather weak sensitivity to U_{e3} in the allowed region. Consequently, the sensitivity to the mass hierarchy is also weak. In what follows we will show that studies of the neutrino bursts from supernova can resolve these problems.

3. – Conversion of supernova neutrinos. General picture.

The neutrino conversion in supernovae has been extensively discussed before [8-28]. In particular, one expects the disappearance of the ν_e flux due to conversion $\nu_e \to \nu_\mu, \nu_\tau$ inside the star. At the cooling stage, when the fluxes of all neutrino species are produced, the conversion inside the star leads to partial or complete permutation of the $\bar{\nu}_e$ and $\bar{\nu}_\mu, \bar{\nu}_\tau$ spectra [19, 20]. This causes the appearance of an high energy tail in the $\bar{\nu}_e$ spectrum which contradicts the SN1987A observations [20].

3˙1. *Inside the star.* – A supernova is source of fluxes of neutrinos and antineutrinos of all the three flavours. These fluxes, F^0_α and $F^0_{\bar{\alpha}}$ $(\alpha = e, \mu, \tau)$, are characterized by the hierarchy of their average energies,

$$\langle E_e \rangle < \langle E_{\bar{e}} \rangle < \langle E_\mu \rangle \ , \tag{11}$$

and by the equality of fluxes of the non-electron neutrinos (which will be denoted as ν_x):

$$F^0_\mu = F^0_{\bar{\mu}} = F^0_\tau = F^0_{\bar{\tau}} \equiv F^0_x \ . \tag{12}$$

The original integral (over the time of burst)) fluxes produced in the star can be described by a Fermi-Dirac spectra:

$$F^0_\alpha(E, T_\alpha, L_\alpha) \propto \frac{E^2}{e^{E/T_\alpha} + 1} \ , \tag{13}$$

where E is the energy of the neutrinos, L_α is the total energy released in ν_α, and T_α is the temperature of the ν_α gas in the neutrinosphere. According to the hierarchy (11) the indicative values $T_e = 3.5$ MeV, $T_{\bar{e}} = 5$ MeV and $T_x = 8$ MeV will be taken as reference in our calculations. Often equipartition of the energy between the various flavours is assumed, so that $L_\alpha \simeq E_{\rm B}/6$, with $E_{\rm B}$ the binding energy emitted in the core collapse of the star: $E_{\rm B} \simeq 3 \cdot 10^{53}$ ergs.

In the presence of neutrino mixing and masses the neutrinos undergo flavour conversion on their way from the production point in the star to the detector at Earth. Matter effects dominate the conversion inside the star, where a wide range of matter densities is met.

The masses and mixings determine the pattern of level crossings in the star [21]. There are two resonances (level crossings) in the schemes under consideration (1):

- The high density (H) resonance, determined by the parameters $\Delta m^2_{\rm atm}$ and U_{e3}. The conversion in the region of this resonance is described by the Landau-Zener–type probability, P_H, of transition between the mass eigenstates ν_2 and ν_3.
- The low density (L) resonance with parameters of the solar pair: $\Delta m^2_\odot$, $\sin^2 2\theta_\odot$. For the LMA solution the propagation in the L resonance is adiabatic, so that probability of $\nu_2 \to \nu_1$ transition associated to this resonance is zero.

Depending on the type of mass hierarchy the resonances appear in different channels [21]:

i) for normal mass hierarchy both the resonances are in the neutrino channel.

ii) for inverted mass hierarchy the H resonance is in the antineutrino channel, whereas the L resonance is in the neutrino channel.

As we will see, these different possibilities correspond to different conversion effects both inside the star and in the matter of the Earth.

3˙2. *Crossing the Earth.* – The possibility of oscillations of supernova neutrinos in the matter of the Earth has been discussed long time ago [9]. It was marked that the effect of oscillations can be significant for values of parameters: $\Delta m^2 \sim 10^{-6}$–$6 \cdot 10^{-5}$ eV^2 and $\sin^2 2\theta > 2 \cdot 10^{-2}$. The effect is different for detectors with different trajectories of the neutrinos inside the Earth, and studying the oscillation effects in these detectors one can restore the direction to the supernova [9].

Further studies of the Earth matter effect on supernova neutrinos have been performed in ref. [21] in connection with the role the supernova neutrinos can play in the reconstruction of the neutrino mass spectrum. In particular, it was shown that the very fact of the detection of the Earth matter effect in the neutrino and/or antineutrino channels will allow to establish the type of the mass hierarchy and to restrict the element U_{e3} of the mixing matrix.

In connection with the fact that the LMA gives the best global fit of the solar neutrino data, the interpretation of the SN1987A data has been revisited [22-25]. The regions of the oscillation parameters have been found [23] in which the Earth matter effects can explain the difference of the K2 and IMB spectra. Such an interpretation also favors the normal mass hierarchy case or very small values of U_{e3} for the inverted mass hierarchy [26, 23, 24].

Recently, the Earth matter effects were considered also in ref. [27] where the expected spectra of events at SuperKamiokande (SK) and SNO have been calculated in three neutrino context with a two-layers approximation for the Earth profile.

Due to the short duration of the burst, and the spherical symmetry of the Earth, for a given detector the trajectories of neutrinos (and therefore the regeneration effect) can be completely described by the nadir angle θ_n of the supernova with respect to the detector: if $\cos\theta_n > 0$ the detector is shielded by the Earth. The angle θ_n depends i) on the location of the supernova in Galaxy, ii) on the time t of the day at which the burst arrives at Earth and iii) on the position of the detector itself.

We first consider a supernova located in the galactic center (declination([1]) $\delta_s = -28.9°$) and three detectors [29]: LVD [30], SNO [31] and SK [32]. Figure 3 a) from [28] shows the dependence of $\cos\theta_n$ on the time t for the three detectors. The horizontal line at $\cos\theta_n = 0.83$ corresponds to the trajectory tangential to the core of the Earth ($\theta_n = 33.2°$), so that trajectories with $\cos\theta_n < 0.83$ are in the mantle of the Earth. For $\cos\theta_n > 0.83$ the trajectories cross both the mantle and the core.

From the figures it appears that [28]

1. For most of the arrival times the supernova is seen with substantially different nadir angles at the different detectors, so that one expects different Earth matter effects observed.

2. At any time t the neutrino signal arrives at Earth, at least one detector is shielded by the Earth ($\cos\theta_n > 0$) and therefore will see the regeneration effect. Earth

([1]) We define δ_s as the the angle of the star with respect to the equatorial plane of the Earth.

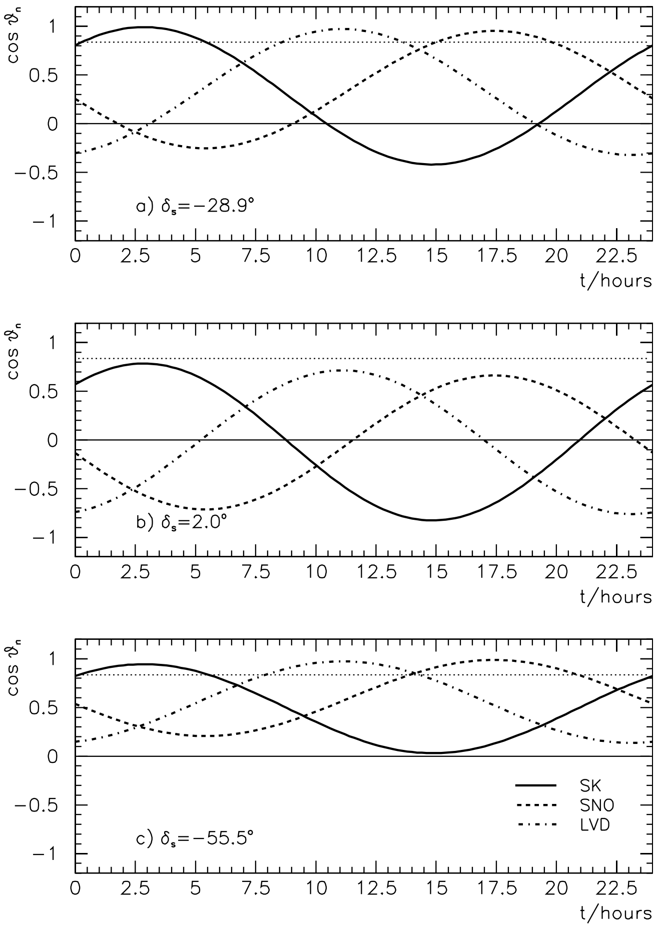


Fig. 3. – The cosines of the nadir angles θ_n of SuperKamiokande, SNO and LVD detectors with respect to the supernova as functions of the arrival time of neutrino burst. The three panels refer to three different locations of the star in the galactic plane (given by the declination angle δ_s). We fixed $t = 0$ as the time at which the star is aligned with the Greenwich meridian.

shielding is verified even for two detectors simultaneously for a large fraction of the times.

3. At any possible arrival time t one of the detectors is not shielded by the Earth. So that, once the direction to the supernova is known, one can identify such a detector and use its data to reconstruct the neutrino energy spectrum without regeneration effect.

4. For a substantial fraction of the times for one of the detectors the trajectory crosses the core of the Earth.

In fig. 3 b), c), we show similar dependences of $\cos\theta_n$ on the time t for other locations of the star in the galactic plane.

4. – Supernova neutrino fluxes at the Earth

4˙1. *The schemes with normal mass hierarchy.* – In the case of normal hierarchy, there is no level crossing in the high resonance region in the antineutrino channel, so that the antineutrino flux at the detector does not depend on the jump probability P_H (see [21] for details):

$$F_{\bar{e}}^D = F_{\bar{e}} + (F_{\bar{e}}^0 - F_x^0)(\bar{P}_{1e} - |U_{e1}|^2) \ , \tag{14}$$

where

$$F_{\bar{e}} \approx F_{\bar{e}}^0 - (F_{\bar{e}}^0 - F_x^0)(1 - |U_{e1}|^2) \tag{15}$$

is the $\bar{\nu}_e$ flux arriving at the surface of the Earth (without Earth matter effect) and the fluxes F_α^0 are defined in eq. (13). Here $\bar{P}_{1e}$ denotes the probability of $\bar{\nu}_1 \to \bar{\nu}_e$ conversion inside the Earth and $\bar{P}_L$ is the jump probability in the L resonance.
The relative Earth effect can be characterized by the ratio

$$\bar{R} \equiv \frac{F_{\bar{e}}^D - F_{\bar{e}}}{F_{\bar{e}}} \ . \tag{16}$$

From eqs. (14), (15) we find

$$\bar{R} = \bar{r}\bar{f}_{\rm reg} \ , \tag{17}$$

where $\bar{r}$ is the ("reduced") flux factor:

$$\bar{r} = \frac{F_{\bar{e}}^0 - F_x^0}{F_{\bar{e}}^0|U_{e1}|^2 + F_x^0(1 - |U_{e1}|^2)} \ . \tag{18}$$

and $\bar{f}_{\rm reg}$ the regeneration factor:

$$\bar{f}_{\rm reg} \equiv (\bar{P}_{1e} - |U_{e1}|^2) \ . \tag{19}$$

The flux factor, eq. (18), determines the sign and the size of the effect. Due to the hierarchy of energies, eq. (11), a critical energy $\bar{E}_c$ exists at which $\bar{r} = 0$. We have $\bar{r} > 0$ below the critical energy, $E < \bar{E}_c$, and $\bar{r} < 0$ for $E > \bar{E}_c$. For realistic temperatures of the neutrino fluxes one gets

$$\bar{E}_c = (25\text{–}28) \ \mathrm{MeV} \ . \tag{20}$$

From eqs. (18), (20) it follows that at very high, as well as at very low energies, the relative regeneration effect becomes independent of the original fluxes.

The regeneration factor, eq. (19), describes the propagation effect inside the Earth and is analogous to the regeneration factor which appears for solar neutrinos. $\bar{f}_{\rm reg}$ corresponds to genuine matter effect: it is zero in vacuum.

The dynamics of the conversion inside the Earth is described by the regeneration factor $\bar{f}_{\rm reg}$, eq. (19). For LMA parameters the Earth matter effect consists in an oscillatory modulation of the neutrino energy spectrum.

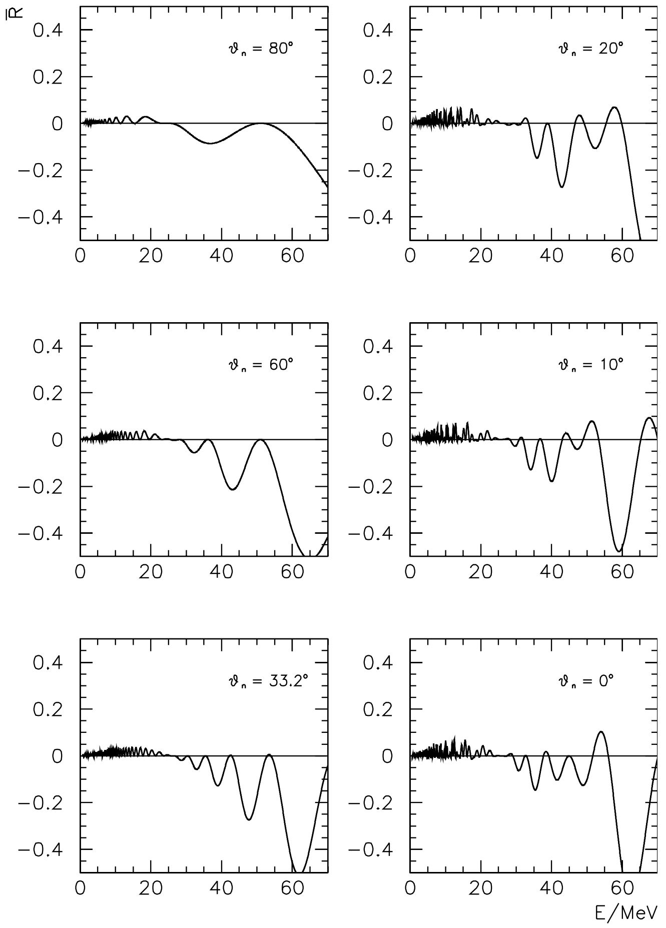


Fig. 4. – The relative Earth matter effect in $\bar{\nu}_e$ channel, $\bar{R}$, as a function of the antineutrino energy for LMA oscillation parameters and various values of the nadir angle θ_n. We have taken $\Delta m^2_\odot = 5 \cdot 10^{-5}$ eV^2, $\sin^2 2\theta_\odot = 0.75$; $T_{\bar{e}} = 5$ MeV, $T_x = 8$ MeV. The figure refers to normal mass hierarchy (or inverted hierarchy with $P_H = 1$).

Figure 4 from [28] shows the ratio $\bar{R}$ as a function of the neutrino energy for various values of θ_n. For mantle crossing trajectories, $\theta_n > 33.2°$, the effect is mainly due to the interplay of oscillations and adiabatic evolution. The factor is positive in the whole energy spectrum, so that the sign of the matter effect is determined by the flux factor (18): we have $\bar{R} > 0$ for $E < \bar{E}_c$ and $\bar{R} < 0$ for $E > \bar{E}_c$.

As a result of adiabatic evolution, the depth of oscillations of the regeneration factor is determined by the electron number density at the surface of the Earth, n_e^0:

$$\bar{D}_f \approx 2\sqrt{2} G_F n_e^0 \frac{E}{\Delta m^2_\odot} \sin^2 2\theta_m^0 \ . \tag{21}$$

Here θ_m^0 is the mixing angle of the solar pair in matter at the surface.

The depth $\bar{D}_f$ has a resonant dependence on the quantity $x \equiv 2E|V|/\Delta m^2_\odot$, with

V being the matter potential. Both $\bar{D}_f$ and l_m increase as the system approaches the resonance; correspondingly, the period $\Delta E/E$ increases. For neutrinos propagating in the mantle and $\Delta m^2_\odot = 5 \cdot 10^{-5}$ eV^2 (which is used in fig. 4) the resonance is realized at $E = E_R \simeq 150$ MeV. Thus the Earth effect is larger in the highest energy part of the spectrum.

For normal mass hierarchy, the H resonance is in the neutrino channel and the ν_e flux at the detector depends on P_H [21]:

$$F_e^D \simeq F_e + (F_e^0 - F_x^0) P_H (P_{2e} - |U_{e2}|^2) \ , \tag{22}$$

where the ν_e flux arriving at the surface of the Earth equals:

$$F_e \simeq F_e^0 - (F_e^0 - F_x^0)(1 - P_H |U_{e2}|^2) \ . \tag{23}$$

Here P_{2e} is the probability of the transition $\nu_2 \rightarrow \nu_e$ inside the Earth.

From eqs. (22), (23) one finds the relative Earth matter effect, $R \equiv (F_e^D - F_e)/F_e$, and the flux factor, r:

$$R = r P_H f_{\text{reg}} \ , \tag{24}$$

$$r = \frac{F_e^0 - F_x^0}{F_e^0 P_H |U_{e2}|^2 + F_x^0 \left[1 - P_H |U_{e2}|^2\right]} \ . \tag{25}$$

The flux factor, r, eq. (25), changes sign at lower critical energy with respect to the case of antineutrinos, since the original ν_e spectrum is softer than the $\bar{\nu}_e$ spectrum. We get

$$E_c = (16\text{–}24) \text{ MeV} \ . \tag{26}$$

The regeneration factor, f_{reg}, is given by

$$f_{\text{reg}} \equiv (P_{e2} - |U_{e2}|^2) = -(P_{1e} - |U_{e1}|^2) \ . \tag{27}$$

Let us comment on the features of the ratio R.

From eq. (24) it follows that if the adiabaticity in the high density (H) resonance inside the star is fulfilled, $P_H \rightarrow 0$, the Earth matter effect disappears. The reason is that in the adiabatic case the original electron neutrinos convert almost completely into ν_μ and ν_τ fluxes in the H resonance. Then the electron neutrinos detected at Earth result from the conversion of the original ν_μ and ν_τ fluxes. Since these fluxes are equal, eq. (12), no oscillation effect will be observed due to conversion in the low density resonance.

The Earth matter effect is maximal in the limit of strong violation of the adiabaticity in the H-resonance: $P_H \rightarrow 1$, when the dynamics is reduced to a two neutrino problem with oscillation parameters of the L resonance.

The jump probability P_H is determined by the density profile of the star and the oscillation parameters $|U_{e3}|^2 \approx \tan^2\theta_{13}$ and Δm^2_{atm}. As $|U_{e3}|^2$ decreases in the range allowed by the bound (9) the transition in the H resonance varies from perfectly adiabatic ($P_H \simeq 0$), for $|U_{e3}|^2 \gtrsim 5 \cdot 10^{-4}$, to strongly non-adiabatic ($P_H \simeq 1$), for $|U_{e3}|^2 \lesssim 10^{-6}$. The intervals of adiabaticity and strong adiabaticity violation change only mildly as Δm^2_{atm} varies in the presently allowed range.

The regeneration factor f_{reg}, eq. (27), and therefore R, have similar dependence on θ_n and $\Delta m^2_\odot$ as in the case of antineutrinos. These dependences are illustrated in fig. 5,

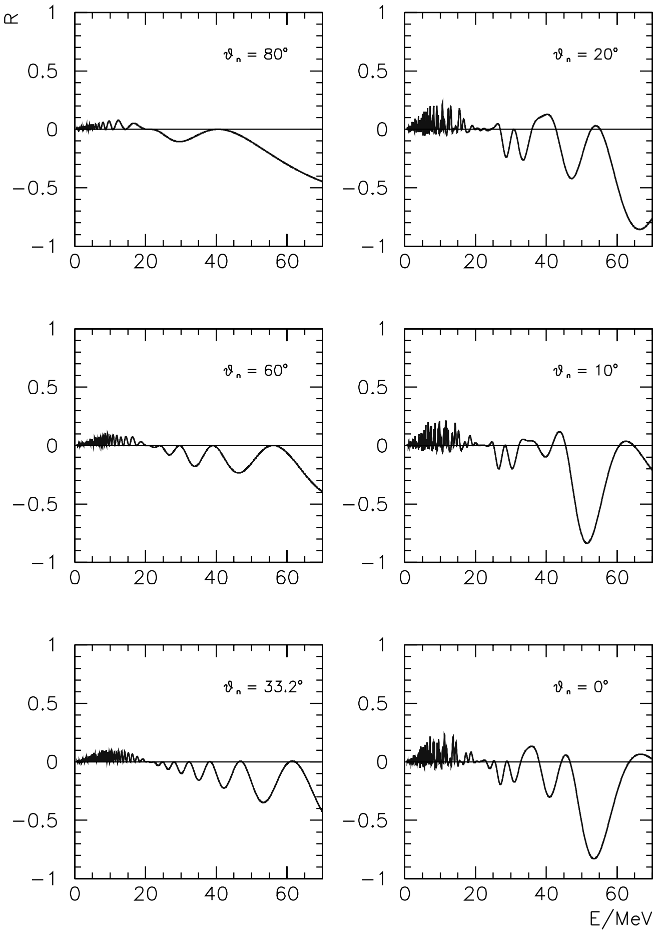


Fig. 5. – The relative Earth matter effect in ν_e channel, R, as a function of the neutrino energy for LMA oscillation parameters and various values of the nadir angle θ_n. We have taken $\Delta m^2_\odot = 5 \cdot 10^{-5}$ eV^2, $\sin^2 2\theta_\odot = 0.75$; $T_e = 3.5$ MeV, $T_x = 8$ MeV; $P_H = 1$ (or inverted hierarchy).

where $P_H = 1$ was taken [28]. The oscillation length and the period of the modulations in the energy spectrum increase with the increase of the energy and the decrease of $\Delta m^2_\odot$. The depth of the oscillations of the regeneration factor $f_{\rm reg}$ is larger than for antineutrinos since (if the L resonance is in the neutrino channel) matter enhances the ν_e mixing and suppresses the mixing of $\bar\nu_e$:

$$\sin^2 2\theta_m(\bar\nu) < \sin^2 2\theta_\odot < \sin^2 2\theta_m(\nu)\ . \tag{28}$$

The depth of oscillations has a resonant character, increasing as the resonance energy is approached. According to eq. (21) the depth gets larger for smaller $\Delta m^2_\odot$.

4˙2. *Schemes with inverted mass hierarchy.* – If the hierarchy of the mass spectrum is inverted the high density resonance is in the antineutrino channel and the Earth matter effect for $\bar{\nu}_e$ depends on the jump probability P_H. Expressions (14), (15) for the $\bar{\nu}_e$ fluxes are immediately generalized to

$$F_{\bar{e}}^D = F_{\bar{e}} + (F_{\bar{e}}^0 - F_x^0) P_H (\bar{P}_{1e} - |U_{e1}|^2) \ , \tag{29}$$

$$F_{\bar{e}} \approx F_{\bar{e}}^0 - (F_{\bar{e}}^0 - F_x^0)(1 - P_H |U_{e1}|^2) \ , \tag{30}$$

in analogy with eqs. (22), (23).

The relative deviation, $\bar{R}$, and the reduced flux factor:

$$\bar{R} = \bar{r} P_H \bar{f}_{\text{reg}} \ , \tag{31}$$

$$\bar{r} = \frac{F_{\bar{e}}^0 - F_x^0}{F_{\bar{e}}^0 P_H |U_{e1}|^2 + F_x^0 (1 - P_H |U_{e1}|^2)} \ . \tag{32}$$

If the hierarchy is inverted the Earth matter effect on $\bar{\nu}_e$ is affected by the adiabaticity in the high density resonance.

Now the conversion of ν_e is independent of P_H. The expressions of the neutrino fluxes F_e^D and F_e can be obtained from eqs. (22), (23) by the replacement $P_H \to 1$; they become analogous to eqs. (14), (15). With the same prescription, from eqs. (24), (25) one gets the expressions of the ratios R and r.

The results for inverted hierarchy of the spectrum are obtained from the description given for normal hierarchy by the replacement $P_H \to 1$. Therefore the results shown in fig. 5, in which $P_H = 1$ was used, apply to the case of inverted hierarchy.
Summarizing we can say that the mass hierarchy and the adiabaticity in the H density resonance (and thus value of U_{e3}) determine the channel (ν_e or $\bar{\nu}_e$) in which the Earth matter effects appear, which is

– both the ν_e and $\bar{\nu}_e$ channels if the H resonance is strongly non-adiabatic, $P_H = 1$, regardless to the hierarchy.

– the $\bar{\nu}_e$ channel for adiabatic H resonance, $P_H = 0$, and normal hierarchy.

– the ν_e channel for adiabatic H resonance, $P_H = 0$, and inverted hierarchy.

5. – Observations of the Earth matter effect

The observation of the Earth matter effect requires: i) separate detection of neutrinos of different flavours, ii) separate detection of neutrinos and antineutrinos, iii) the reconstruction of the neutrino energy spectrum.

We consider:

1. The detection of $\bar{\nu}_e$ at water Cherenkov detectors (SuperKamiokande and the outer volume of SNO) via the reaction

$$\bar{\nu}_e + p \to e^+ + n \ . \tag{33}$$

2. Heavy water detectors (the inner volume of SNO experiment) with the detection reactions:

$$\nu_e + d \to e + p + p \ , \tag{34}$$

$$\bar{\nu}_e + d \to e^+ + n + n \ , \tag{35}$$

which represent the dominant channel of CC detection. Events from the process (35) will be distinguished by those from (34) if neutrons are efficiently detected in correlation with the positron.

3. Liquid scintillator detectors (LVD), which are mostly sensitive to $\bar{\nu}_e$ via the reaction (33) with only little sensitivity to absorption processes on carbon nuclei.

The energy spectrum of the charged leptons reflects the spectrum of the neutrinos, with the following differences:

– the energy dependence of the cross section substantially enhances the high energy part of the spectrum.

– the integration over the neutrino energy and the convolution with the energy resolution function lead to averaging out the fast modulations in low energy part of the spectrum. Conversely, the large-period oscillations at high energies will appear in the lepton spectrum.

The Earth matter effects can be identified:

1. at a single detector, by the observation of deviations of the energy spectrum with respect to what expected from conversion in the star only;

2. by the comparison of energy spectra from different detectors.

In figs. 6-8 we show [28] examples of the spectra expected at SK, SNO and LVD for oscillation parameters from the LMA solution, $P_H = 1$ and various arrival times of the neutrino burst. We considered a supernova located in the direction of the galactic center (fig. 3 a)) at a distance $D = 10$ Kpc and releasing a total energy $E_{\rm B} = 3 \cdot 10^{53}$ ergs. The histograms represent the numbers of events from the reaction (33) for SK (panels a)) and LVD (panels b)); panels c) show the sum of the numbers of events from the reactions (35) and (33) at SNO. In d) we plot the numbers of events in the inner volume of SNO from the scattering (34).

Besides the present neutrino telescopes, the detection of supernova neutrinos is among the goals of future large volume detectors, like UNO [33] and NUSL [34]. We find that the comparison of the energy spectra observed by SK and by another detector with comparable or larger statistics could establish the Earth matter effects at more than $\sim 5\ \sigma$ level.

Besides the comparison of the spectra, more specific criteria of identification of the Earth effect can be elaborated if the location of the supernova and the solar neutrino oscillations parameters are known. For instance, for LMA parameters and rather superficial trajectory in the mantle the effect consists in a narrowing of the spectrum (see, *e.g.*, fig. 6). Thus the comparison of the widths of the spectra at different detectors may establish the Earth effect.

As a further illustration, in figs. 7 and 8 we show the expected spectra for the same parameters as in fig. 6 but different arrival times of the signal (see fig. 3 a)).

As we have mentioned already the very fact of establishing the Earth matter effect in the neutrino and/or in antineutrino channel will have important implications for the neutrino mass and flavor spectrum. For this it will be enough to study some integral effect of regeneration.

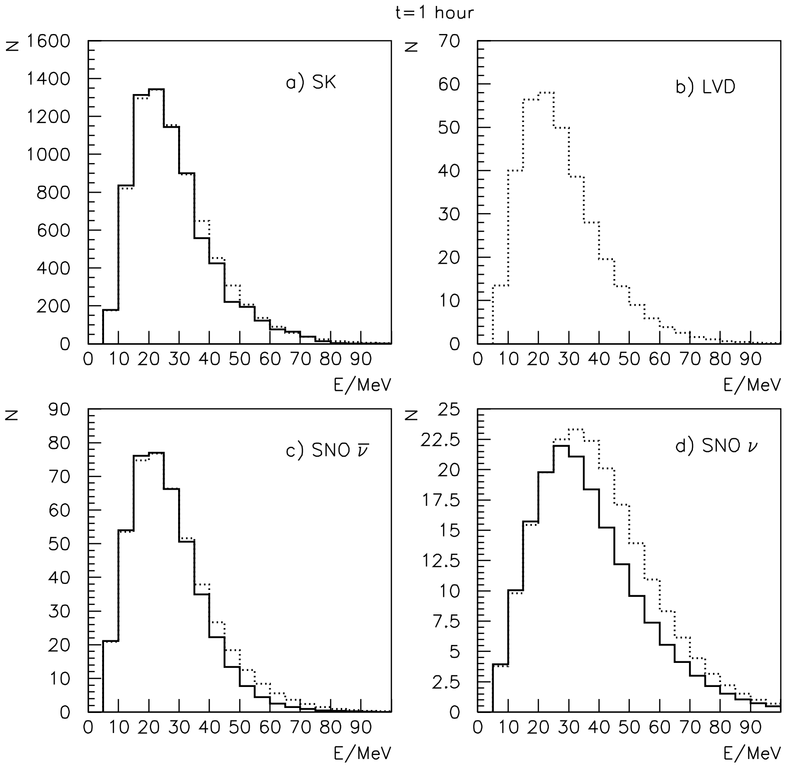


Fig. 6. – The energy spectra expected at SK, SNO and LVD with (solid lines) and without (dotted lines) Earth matter effect, for the same parameters as in figs. 4 and 5 and $t = 1$ hour of fig. 3 a). A distance $D = 10$ Kpc from the supernova and binding energy $E_B = 3 \cdot 10^{53}$ ergs have been taken. In this specific configuration LVD is not shielded by the Earth, thus observing undistorted spectrum. The histogram c) refers to the sum of events from $\bar{\nu}_e + p \to e^+ + n$ and $\bar{\nu}_e + d \to e^+ + n + n$ scatterings, while the panel d) shows the events from $\nu_e + d \to e + p + p$. In a) and b) only the events from $\bar{\nu}_e + p \to e^+ + n$ are shown.

6. – Neutrino mass spectrum and SN1987A

One of the unexpected features of the neutrino signals from SN1987A is the difference in the Kamiokande-2 (K2) and IMB spectra of events. Indeed, the data show

i) concentration of the IMB events in the energy interval $E \simeq 35$–40 MeV;

ii) absence of events at IMB above $E \simeq 40$ MeV (which looks like a sharp cut of the spectrum);

iii) Absence of events with $E \gtrsim 35$ MeV at K2.

Soon after the observation of SN1987A it was marked that the differences in the K2 and IMB spectra could be related to oscillations of $\bar{\nu}_e$ in the matter of the Earth and to the different positions of the detectors at the time of detection [10]. It was realized that, for this oscillations mechanism to work, one needs $\Delta m^2 \sim 10^{-5}$ eV^2 (*i.e.* in the region of the Earth regeneration effect) and large (close to maximal) mixing of $\bar{\nu}_e$.

We show that [23] the features i)-iii) can be explained by oscillations in the matter

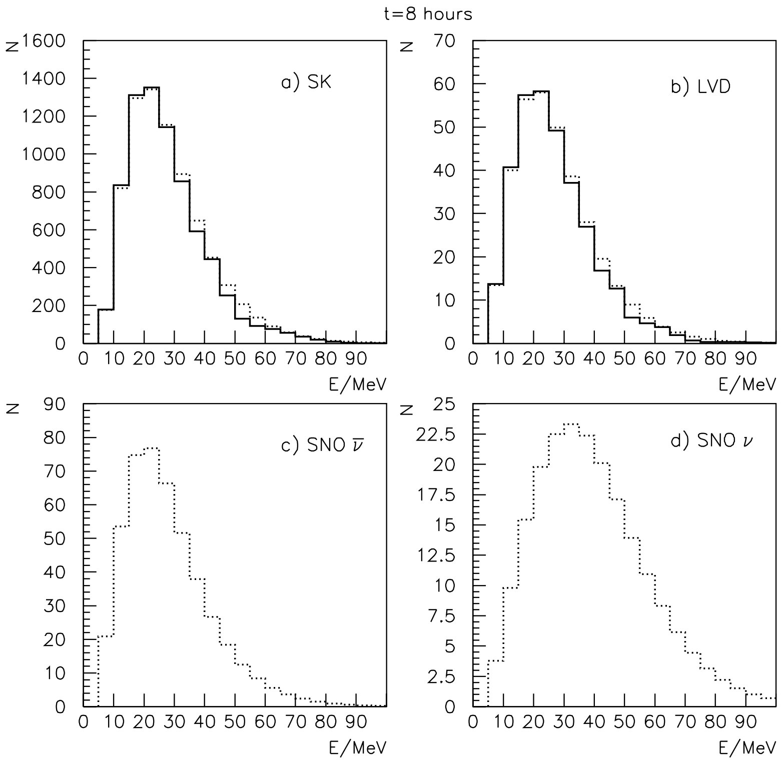


Fig. 7. – The same as fig. 6 for $t = 8$ hours of fig. 3 a). For this configuration SNO is unshielded by the Earth.

of the Earth. The difference of oscillation effects is related to the distances travelled by the neutrinos in the Earth: $d_{\rm IMB} \simeq 8535$ km for IMB, $d_{\rm K2} \simeq 4363$ km for K2, and to the average densities $\rho_{\rm IMB} \simeq 4.5\ {\rm g}\cdot{\rm cm}^{-3}$, $\rho_{\rm K2} \simeq 3.5\ {\rm g}\cdot{\rm cm}^{-3}$ along the trajectories. As a consequence, both the depths and the phases of oscillations at K2 and IMB are different. The explanation implies certain values of Δm^2 and $\sin^2 2\theta$.

To reproduce the characteristics described in i)-iii) we require [23]:

1) The oscillation maximum at IMB detector at $E \simeq 38$–42 MeV, that is, the phase of oscillations

$$\phi_{\rm IMB}(40) \equiv \frac{\pi d_{\rm IMB}}{l_m} = k\pi\ , \qquad k = 1, 2, 3, \ldots\ . \tag{36}$$

2) The oscillation minimum at IMB at $E \simeq 50$–60 MeV, so that the phase is semi-integer of π at these energies:

$$\phi_{\rm IMB}(60) = \pi\left(\frac{1}{2} + k\right)\ , \qquad k = 0, 1, 2, ..\ . \tag{37}$$

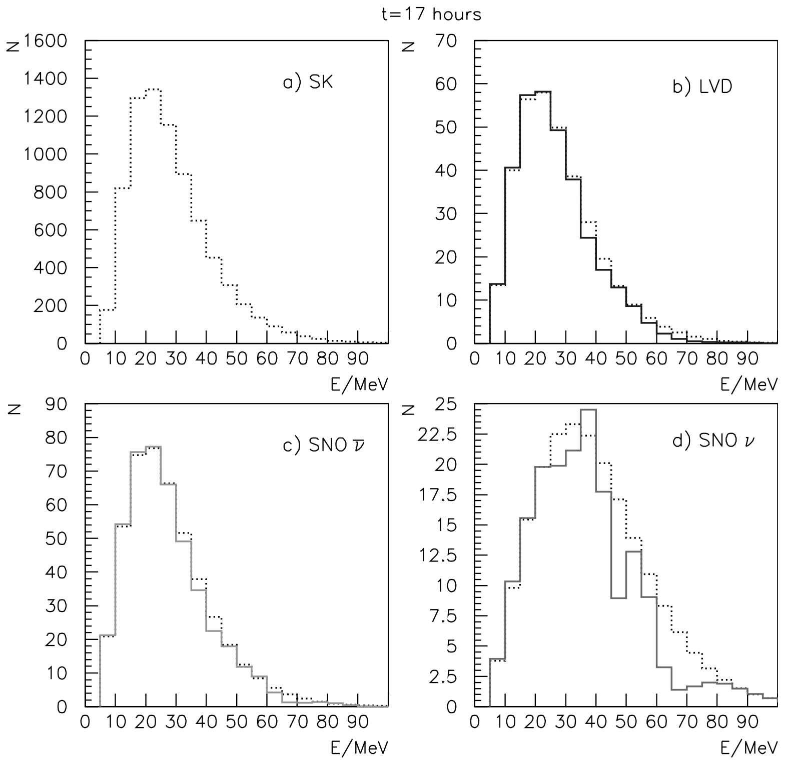


Fig. 8. – The same as fig. 6 for $t = 17$ hours of fig. 3 a). For this configuration SK is unshielded by the Earth.

3) The oscillation minimum at K2 at $E \simeq$ 38–42 MeV:

$$\phi_{\rm K2}(40) \equiv \frac{\pi d_{\rm K2}}{l_m} = \pi\left(\frac{1}{2}+k\right) , \qquad k = 0, 1, 2, ... \tag{38}$$

4) The Earth matter effect is maximal, $A_p \simeq A_p^{\rm max}$, at IMB at the energies $E \simeq$ 50–60 MeV, that is

$$E_R^{\rm IMB} \simeq 50\text{–}60 \text{ MeV} . \tag{39}$$

In fig. 9 from [23] we show the conditions (36), (38) and (39) in the Δm^2-$\cos 2\theta$ plane. As follows from the figure, there are bands in which the requirements (36) and (38) are satisfied simultaneously. They correspond to $\phi_{\rm IMB} \simeq 2\phi_{\rm K2} = 3\pi, 5\pi, 7\pi, ...$ The phase increases with Δm^2. Notice that the requirements (36)-(38) are satisfied in the whole relevant range of $\cos 2\theta$ if $\phi_{\rm IMB}$ equals odd multiples of π.

The band with $\Delta m^2 = (5\text{–}6) \cdot 10^{-5}$ eV^2 with $\cos 2\theta = 0.3$–0.5 covers the best fit point from recent analysis of solar neutrino data.

Notice that the region of parameters selected by the data is substantially narrower than that from the solar neutrino data themselves. Slight change of Δm^2, *e.g.*, increase lead to substantial worthening of the fit.

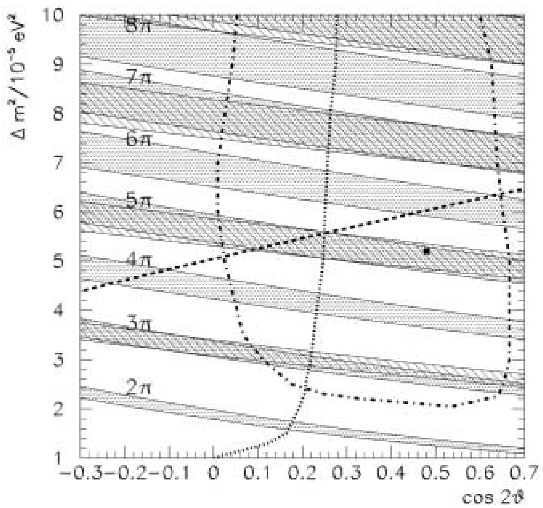


Fig. 9. – Bands of equal phases $\phi_{\rm IMB}(40) = k\pi$ (dotted regions) and $\phi_{\rm K2}(40) = \pi\,(1/2 + k)$ (dashed regions) in the Δm^2-$\cos 2\theta$ plane. The widths of the bands are determined by the requirement that the conditions (38) and (36) are satisfied in the energy interval E = 38–42 MeV. The region below the dashed line represents the band of strong Earth matter effect, where $A_p \gtrsim 0.7 A_p^{\rm max}$ (see eq. (39)). For comparison we show the 99% C.L. allowed region of the LMA solution of the solar neutrino deficit (dot-dashed contour). The dotted line represents the 99% C.L. exclusion curve from fig. 3a of ref. [20].

7. – Determination of type of hierarchy

The features of the Earth matter effects depend on the value of U_{e3} and on the type of mass hierarchy. For normal mass hierarchy and U_{e3} in the adiabatic range (which appears as the most plausible scenario) we expect regeneration effects in the antineutrino channel and no effect in the neutrino channel. In the supernova data further confirmations of such a possibility are i) the absence of the neutronization peak in ν_e and appearance of the ν_μ/ν_τ neutronization peak, ii) hard spectrum of ν_e during the cooling stage: $\langle E_e\rangle > \langle E_{\bar e}\rangle$. In the case of inverted mass hierarchy the Earth matter effect should be observed in the neutrino channel and no effect is expected in the antineutrino channel if $|U_{e3}|^2 > 10^{-5}$. This possibility will be confirmed by the observation of the ν_e-neutronization peak and of an hard spectrum of the $\bar\nu_e$ during the cooling stage.
In the limit $|U_{e3}|^2 \ll 10^{-5}$ the high density resonance is inoperative, so that the result is insensitive to the mass hierarchy. Oscillations appear in both the neutrino and antineutrino channels.

In practice, the observation of the Earth matter effect in the $\bar\nu_e$ channel and absence of the effect in ν_e channel will testify for normal mass hierarchy and $|U_{e3}|^2 > 10^{-5}$. In the opposite situation, effect in the ν_e channel and absence of the effect in $\bar\nu_e$ channel, the inverted hierarchy will be identified with $|U_{e3}|^2 > 10^{-5}$. However the present experiments have lower sensitivity to ν_e fluxes with respect to the fluxes of $\bar\nu_e$, so that it may be difficult to establish "zero" regeneration effect with high enough accuracy.

If the Earth matter effect is observed in both channels, one should compare the size of the effect with that predicted in the absence of the high resonance in a given channel. Thus, if the observed signal in the neutrino channel is smaller than what is predicted in the assumption of $P_H = 1$, whereas in the antineutrino channel prediction and observation coincide, we will conclude that the hierarchy is normal and the ratio of the observed to predicted signals in the neutrino channel can give the value of P_H. The opposite case of coincidence of the predicted and observed signals in the neutrino channel and suppressed observed signal in the antineutrino channel will testify for the inverted mass hierarchy.

Besides the probing of the neutrino mass spectrum and mixing, a study of the properties of the original neutrino fluxes can be done with Earth matter effects. In principle, a detailed study of the observed energy spectra will allow to reconstruct the flux factor as well as to determine the critical energy E_c.

8. – Conclusions

Neutrino astrophysics and in particular studies of solar, and supernova neutrinos play important role in realization of program of reconstruction of the neutrino mass and mixing spectrum.

As far as solar neutrinos are concerned, the LMA MSW solution of the solar neutrino problem looks very appealing. This solution

– gives the best fit of all available data;

– may lead to observable excess of the e-like events in the atmospheric neutrinos;

– can explain some features of the SN1987A neutrino signals;

– gives a hope to measure the CP-violation in lepton sector;

– can be tested soon.

Studies of supernova neutrinos can give very rich and probably unique information about neutrino mass spectrum. The studies

– can identify or confirm the solution of the solar neutrino problem;

– measure or restrict value of U_{e3};

– identify the type of mass hierarchy (ordering).

Here studies of the Earth matter effect on the supernova neutrinos are very important giving the Supernova model independent information. From theoretical point of view it is rather plausible that value of U_{e3} is not too far from the present upper bound. In this case the supernova neutrino studies will identify the mass hierarchy and put lower bound on U_{e3}.

If the LMA is the true solution of the solar neutrino problem, then a significant part of the $\bar{\nu}_e$ events detected from SN1987A were produced by the converted muon and tau antineutrinos. This means that in 1987 we observed the first appearance signal of neutrino conversion!

The observable effect of neutrino mass and mixing are based on neutrino transformations in matter. The Fermi coupling constant plays crucial role determining immediately the scale of phenomena and applications of effects.

* * *

The author is grateful to C. Lunardini and P. de Holanda. Substantial part of this talk was based on papers written in collaboration with Cecilia.

REFERENCES

[1] Wolfenstein L., *Phys. Rev. D*, **17** (1987) 2369; Mikheev S. P. and Smirnov A. Y., *Yad. Fiz.*, **42** (1985) 1441.

[2] See for recent reviews Bilenky S. M., Giunti C. and Grimus W., *Prog. Part. Nucl. Phys.*, **43** (1999) 1, Gonzalez-Garcia M. C. and Yosef Nir, hep-ph/0202058 and references therein.

[3] Lunardini C. and Smirnov A. Yu., *Nucl. Phys. B*, **583** (2000) 260.

[4] The CHOOZ Collaboration (Apollonio M. *et al.*), *Phys. Lett. B*, **466** (1999) 415 [hep-ex/9907037]; **420** (1998) 397 [hep-ex/9711002].

[5] Boehm F. *et al.*, *Phys. Rev. D*, **62** (2000) 072002 [hep-ex/0003022].

[6] SNO Collaboration (Ahmad Q. R. *et al.* nucl-ex/0204008, nucl-ex/0204009.

[7] de Holanda P. and Smirnov A. Yu., hep-ph/0205241.

[8] Mikheev S. P. and Smirnov A. Y., *Sov. Phys. JETP*, **64** (1986) 4.

[9] Mikheev S. P. and Smirnov A. Y., *Proceedings of the 6th Moriond Workshop on Massive Neutrinos in Astrophysics and in Particle Physics*, 1986, edited by O. Fackler and J. Tran Than Van (Editions Frontières, Gif-sur-Yvette, France) 1986, p. 355.

[10] Smirnov A. Y., talk given at the *Twentieth International Cosmic Ray Conference, Moscow, 1987.*

[11] Arafune J. and Fukugita M., *Phys. Rev. Lett.*, **59** (1987) 367.

[12] Arafune J. *et al.*, *Phys. Rev. Lett.*, **59** (1987) 1864.

[13] Minakata H. *et al.*, *Mod. Phys. Lett. A*, **2** (1987) 827.

[14] Lagage P. O. *et al.*, *Phys. Lett. B*, **193** (1987) 127.

[15] Notzold D., *Phys. Lett. B*, **196** (1987) 315.

[16] Rosen S. P., *Phys. Rev. D*, **37** (1988) 1682.

[17] Kuo T. K. and Pantaleone J., *Phys. Rev. D*, **37** (1988) 298.

[18] Minakata H. and Nunokawa H., *Phys. Rev. D*, **38** (1988) 3605.

[19] Wolfenstein L., *Phys. Lett. B*, **194** (1987) 197.

[20] Smirnov A. Y., Spergel D. N. and Bahcall J. N., *Phys. Rev. D*, **49** (1994) 1389.

[21] Dighe A. S. and Smirnov A. Y., *Phys. Rev. D*, **62** (2000) 033007.

[22] Jegerlehner B., Neubig F. and Raffelt G., *Phys. Rev. D*, **54** (1996) 1194.

[23] Lunardini C. and Smirnov A. Y., *Phys. Rev. D*, **63** (2001) 073009.

[24] Minakata H. and Nunokawa H., *Phys. Lett. B*, **504** (2001) 301.

[25] Kachelriess M., Tomas R. and Valle J. W. F., *JHEP* **01** (2001) 030.

[26] Minakata H., *Nucl. Phys. Proc. Suppl.*, **100** (2001) 237.

[27] Takahashi K., Watanabe M. and Sato K., (2000), hep-ph/0012354.

[28] Lunardini C. and Smirnov A. Yu., *Nucl Phys. B*, **616** (2001) 307.

[29] For a review on supernova neutrino experiments see Scholberg K., *Nucl. Phys. Proc. Suppl.*, **91** (2000) 331 [hep-ex/0008044].

[30] See Fulgione W., for the LVD Collaboration, *Nucl. Phys. Proc. Suppl.*, **70** (1999) 469, and references therein.

[31] See Virtue C. J., for the SNO Collaboration, *Nucl. Phys. Proc. Suppl.*, **100** (2001) 326 [astro-ph/0103324], and references therein.

[32] The Super-Kamiokande Collaboration (Fukuda Y. *et al.*), *Phys. Rev. Lett.*, **81** (1998) 1158 [hep-ex/9805021].

[33] Jung C. K., *Feasibility of a next generation underground water Cherenkov detector: UNO*, hep-ex/0005046.

[34] The recent NUSL proposal is available at http://www.sns.ias.edu/˜ jnb/.

B.11 G.M. Zaslavsky: From the FPU-problem (LA–1940 report) to chaos

G.M. Zaslavsky: "From the FPU-problem (LA–1940 report) to chaos," *Nuovo Cimento B* **117**, 1275 (2002).

IL NUOVO CIMENTO Vol. 117 B, N. 9-11 Settembre-Novembre 2002

From the FPU-problem (LA-1940 Report) to chaos(*)

G. M. Zaslavsky

Courant Institute of Mathematical Sciences, New York University
251 Mercer St., New York, NY 10012, USA
Department of Physics, New York University
2-4 Washington Place, New York, NY 10003, USA

(ricevuto il 4 Novembre 2002; approvato il 4 Dicembre 2002)

Summary. — We discuss the FPU problem in the context of an attempt to find a transition from regular dynamics to the chaotic one. The Fermi mechanism of acceleration was a precursor of the FPU problem. The FPU problem has inspired scientific activities in Hamiltonian integrability, chaos, and the validity of discretization of differential equations. We discuss briefly the latter two issues as well as some new achievements in the theory of chaos.

PACS 05.45.-a – Nonlinear dynamics and nonlinear dynamical systems.
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

FPU is an abbreviation for E. Fermi, J. Pasta and S. Ulam. The FPU-problem is less specific and related to the questions that have been raised in a fairly short preprint [1], now famous and widely known as LA-1940. This paper was first published in [2], it appears in the Institute of Nuclear Physics (INP) at Novosibirsk (former Soviet Union) after ten years of its "publication" in Los Alamos, and the main information about the problems formulated in [1] and their discussion, the scientists from Novosibirsk extracted from other publications [3,4]. At that time the interest to the nonlinear problems at INP was fairly explicit, especially due to the leading scientists Roald Sagdeev, one of the creators of the so-called quasilinear theory [5] which will be discussed later, and Boris Chirikov who proposed the resonance overlapping criteria [6] as a condition of transition to chaos.

(*) Paper presented at the IX ICRA Network Workshop "Fermi and Astrophysics" (Rome, Pescara, September 2001) held under the auspices of the Italian Committee for the Celebration of the Hundredth Anniversary of the birth of Enrico Fermi. Joint copyright SIF and World Scientific.

It may be very difficult to truck all the details to explain how the so self-efface publication initiates tremendous number of publications at least in three important contemporary directions: solitons and integrability; chaos and nonintegrability; discrete *vs.* continuous. The FPU paper proposed an intensive use of fast computers to study nonlinear problems, to simulate fundamental principles of physics and, particularly, origin of statistical laws and thermalization.

It seems that an interest of Fermi to the last problem had a long story after publication of a paper on the mechanism of acceleration of cosmic particles [7] known now as Fermi acceleration. The acceleration of particles can appear as a result of particles scattering on randomly moving magnetic clouds. The crucial point of this mechanism was the randomness of the particles wandering, which was not evident for the regularly moving clouds. The origin of the randomness was important for the Fermi acceleration and the FPU problem could be considered as a model for the dynamical origin of statistical laws.

In this article we briefly touch the problem of chaos and Fermi acceleration, and the problem of discretization regarding their relation to the FPU-problem. The influence of the FPU-paper [1] on the origin of the theory of solitons one can find in the article of David Campbell [8] in a special issue dedicated to Stanislav Ulam.

2. – FPU-model

The original paper [1] considers a nonlinear string

$$y_{tt} = y_{xx}(1 + 3\beta y_x^2) + \gamma y_{xxxx} \tag{1}$$

with a periodic boundary condition. It was supposed that different initially excited oscillations should be thermalized after a while due to their interaction through the nonlinear term. As a result of thermalization, one can expect equipartition of full energy beteween all oscillating modes. In this way a problem was posted on the appearance of statistical features in nonlinear dynamical systems. To answer the question, the authors of [1] decided to exploit the first fast computer MANIAC I at Los Alamos. For this goal (1) should be replaced by a set of difference equations

$$\begin{aligned} &\ddot{y}_n = y_{n-1} - 2y_n + y_{n+1} + \beta[(y_{n+1} - y_n)^2 - (y_n - y_{n-1})^2]\ , \\ &(n = 0, 1, \ldots, N-1) \end{aligned} \tag{2}$$

with a cyclic condition $y_0 = y_N$. The system (2) corresponds to the N coupled nonlinear oscillators. The expected thermalization, *i.e.* equipartition of energy between different oscillators or some oscillator clusters, did not appear and almost no transfer of energy between modes was observed in [1].

That is how the FPU problem arised and, in addition, the question of a possibility of the replacement of the continuous problem (1) by the discrete problem (2) was posted. Some later Ulam put the new problem in the following way: What are we losing when we replace differential equations by their difference approximation, and what is new in the difference equations that does not exist in their differential prototype?

The problem of thermalization appeared to have two subproblems: using the contemporary language it is occurrence of chaos in (2) or (1) and redistribution of energy between different degrees of freedom. The first answer on the condition of chaos was in [9]. Other papers of this volume will describe more these two subproblems.

3. – Fermi acceleration

Here we would like to return to the question that led Fermi to the FPU problem. It was clear for the FPU that the main issue is a nonlinearity of the system that should induce a random dynamics. Later Ulam specified this idea proposing a simple nonlinear model to study a randomization: a particle that bounces between two walls with periodic oscillatinos of one of the walls [10,11]. A physical motivation for the model known as the Ulam one was a possibility of the energy transform from astronomical bodies to particles or even to space vehicles. Stochastic acceleration in a field of the rotating double star can be considered as an example where the Fermi acceleration mechanism [7] should work. This Fermi's idea of the 1949 can be considered as a precursor to the FPU-problem and it is not surprising that after unsuccessful results with eqs. (2) Ulam returns to the much simpler model of the bouncing particle [10, 11]. The results of simulations were again negative [11].

The origin of failures in the FPU and Ulam models became clear after a theory and understanding of the phenomenon of chaos and its alternative theory of stability occurred to be more evident and explicit. A theory and simulation for the Ulam model appeared in [12]. More information and adjoint problems were published in [13], and some continuation in its investigation can be found in [14, 15].

Consider a map

$$p_{n+1} = p_n + Vf(x_n)$$

$$x_{n+1} = x_n + \frac{a}{p_{n+1}^{\nu}} + b \quad (\mathrm{mod}\ 2\pi)\,, \tag{3}$$

where all variables are dimensionless, (p, x) are generalized momentum and coordinate; $f(x)$ is periodic: $f(x+2\pi) = f(x)$; V, a, b, ν are constants, and n corresponds to a discrete time. For the Ulam model considered in [12] p is velocity, $f(x)$ is the sawtooth function that defines the periodic oscillations of a wall, n defines the time instant of the n-th collision between the particle and the oscillating wall, and $\nu = 1$.

The strong chaos occurs in the map (3) when

$$K_u \equiv (a\nu V/p^{\nu+1})|f'(x)| \gg 1 \tag{4}$$

and the condition $K \sim 1$ can be considered as a threshold for the chaos. In fact, system (3) always possesses chaotic trajectories located in some part of the phase space, which sometimes is difficult to find due to very small phase volume that they occupy. The role of chaos is fast mixing of phases x and slow diffusion along p following to the equation

$$\frac{\partial F(p,t)}{\partial t} = \frac{1}{2}\,\frac{\partial}{\partial p}\mathcal{D}(p)\frac{\partial F(p,t)}{\partial p}\,, \tag{5}$$

where the diffusion coefficient

$$\mathcal{D}(p) \equiv \langle\langle(\Delta p)^2\rangle\rangle/\Delta t \tag{6}$$

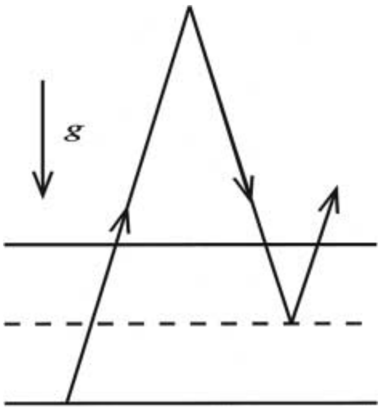


Fig. 1. – A model of Fermi acceleration in the field of gravity.

with

$$\Delta p = Vf(x) \ , \tag{7}$$

$\langle\langle \ldots \rangle\rangle$ is averaging over the phase x, and Δt is a time between two consequent bounces, which depends on the model.

The model (3) has strong limitations of the acceleration rate for $\nu \geq 1$, and another variant of the Fermi acceleration in a gravitational field with an oscillating horizontal plate as an energy source was proposed in [13] with $\nu = -1$:

$$p_{n+1} = p_n + 2V \ f(x_n)$$

$$x_{n+1} = x_n + \frac{V}{2ga} p_{n+1} \qquad (\text{mod } 1), \tag{8}$$

where a is the amplitude of the oscillating plate of the infinite mass, $2V$ is its velocity amplitude, g is the gravitational constant, and $f(x)$ was taken in the sawtooth form

$$f(x) = 1 - 2x \qquad (0 < x < 1) \tag{9}$$

(see fig. 1). Let us comment that (8) coincides with the so-called Chirikov-Taylor, or standard, map [16] if one takes $f(x) = \sin 2\pi x$. Equation (8) has solution that shows unlimited acceleration if

$$K = 2V^2/ga \gtrsim 1 \tag{10}$$

and provides the accleration of particles with [13]

$$\langle p^3 \rangle = \text{const} + 3gV^2 t \tag{11}$$

since $\Delta t = 2p/g$, *i.e.* $\langle |p| \rangle \sim t^{1/3}$.

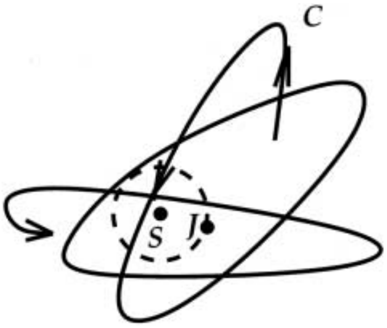


Fig. 2. – A comet (C) dynamics perturbed by the Jupiter (J) rotations around the Sun (S).

4. – Kepler Map

The Kepler Map [17-19] may be considered as one of the most interesting applications of the Fermi acceleration. For example, comets from the so-called Oort cloud, that can reach a visual zone, have very elongated orbits. These comets are perturbed by different planets from which the strongest perturbation is due to the Jupiter. Strongly elongated orbits permit to introduce an iterative map between the comet consequent entrances of the visual zone.

A simplified version of dynamics of comets with strongly elongated orbits is to consider the Sun, Jupiter orbit, and a comet orbit in one plane (see fig. 2). It is the so-called restricted 3-body problem. The Jupiter mass is of 10^{-3} of the Sun mass which makes Jupiter's perturbation fairly sensitive to comets. The corresponding map can be written in the form

$$E_{n+1} = E_n + \sigma\mu f(\vartheta_n)\,,$$

$$\vartheta_{n+1} = \vartheta_n + 2\pi/(-2E_{n+1})^{3/2}\,, \tag{12}$$

where E_n is the comet energy

$$E = \frac{1}{2}p^2 - \frac{1-\mu_{\rm J}}{|\mathbf{r}-\mathbf{r}_s|} + \frac{\mu_{\rm J}}{|\mathbf{r}-\mathbf{r}_{\rm J}|} \tag{13}$$

after the n-th passing near the Jupiter zone, ϑ is the corresponding angle between the Jupiter radius $\mathbf{r}_{\rm J}$ and the comet radius $\mathbf{r}$ when the comet is in the epicenter point. Other parametes are $\mu_{\rm J}$ is the Jupiter mass, the Sun mass is 1, $\sigma = \pm 1$ depending on direction of the comet rotation with respect to the Jupiter rotation, and (13) is written in the center-of-mass system. The form of the periodic function $f(\vartheta)$ as well as more precise map were obtained in [19]. The Kepler map (12) is similar to (3) with $\nu = 3/2$ and a strong chaos occurs for

$$\mu|f'(\vartheta)| \gtrsim |E|^{5/2}\,, \tag{14}$$

i.e. for fairly small energies and large semi-axis of comets. The Fermi mechanism of acceleration works, if the condition (14) is valid, generating diffusion of the comets and their possible escape from the solar system.

Other, similar to (3) and (12) maps but with different values of ν can be found in [20].

5. – Discretization and chaos

As was mentioned in the introduction, a validity of the descretization of differential equations and their replacement by difference equations is another important problem initiated by the LA-1940 report. The material of this section can be extended by important examples from the sect. 9.5 of [20].

Consider first an integrable dynamics of a pendulum

$$\ddot{x} + \omega_0^2 \sin x = 0\ . \tag{15}$$

Its discrete variant possesses chaotic orbit showing in this way how strong can be effects of discretization. Indeed, replace (15) by

$$x_{n+1} - 2x_n + x_{n-1} + \omega_0^2 (\Delta t)^2 \sin x_n = 0\ , \quad x_n \equiv x(n\Delta t) \tag{16}$$

and rewrite (16) in the form

$$p_{n+1} = p_n - \omega_0^2 \Delta t \cdot \sin x_n\ ,$$

$$x_{n+1} = x_n + \Delta t p_{n+1} \tag{17}$$

where the second equation defines momentum p. Equation (17) is just the standard map with the parameter

$$K = \omega_0^2 \Delta t^2 \ll 1 \tag{18}$$

since Δt is very small. That means the chaotic dynamics exists in some exponentially narrow stochastic layers.

One can state that the physical meaning of the replacement of differential by difference, called $\mathcal{D} \to \Delta$ transition, is an introducing into the system a high frequency artificial force. Particularly, for the pendulum (15) $\mathcal{D} \to \Delta$ transition is equivalent to the following replacement of the Hamiltonian:

$$H_0 = \frac{1}{2}p^2 - \omega_0^2 \cos x \to H_0 - 2\omega_0^2 \cos x \cdot \cos\left(\frac{2\pi}{\Delta t}t\right)\ . \tag{19}$$

The second term is just a potential of the discretization, which has frequency $2\pi/\Delta t \gg \omega_0$.

More serious changes can appear if the system is already chaotic or has more than one degrees of freedom [20]. This makes the FPU-problem more contemporary that it was expected at the beginning when the role of computing does not enter such sphere of possibilities as the computer-assisted proofs of theorems.

6. – Developments in the chaotic dynamics

Recurrences of the oscillating profiles observed in [1] and later in [3, 4] indicated an absence of thermalization or, using a more contemporary language, chaos. Recently the FPU recurrences were observed in modulationally unstable optical fibers [21]. It is an additional but not less important aspect raised by the paper [1]: One can diagnose the

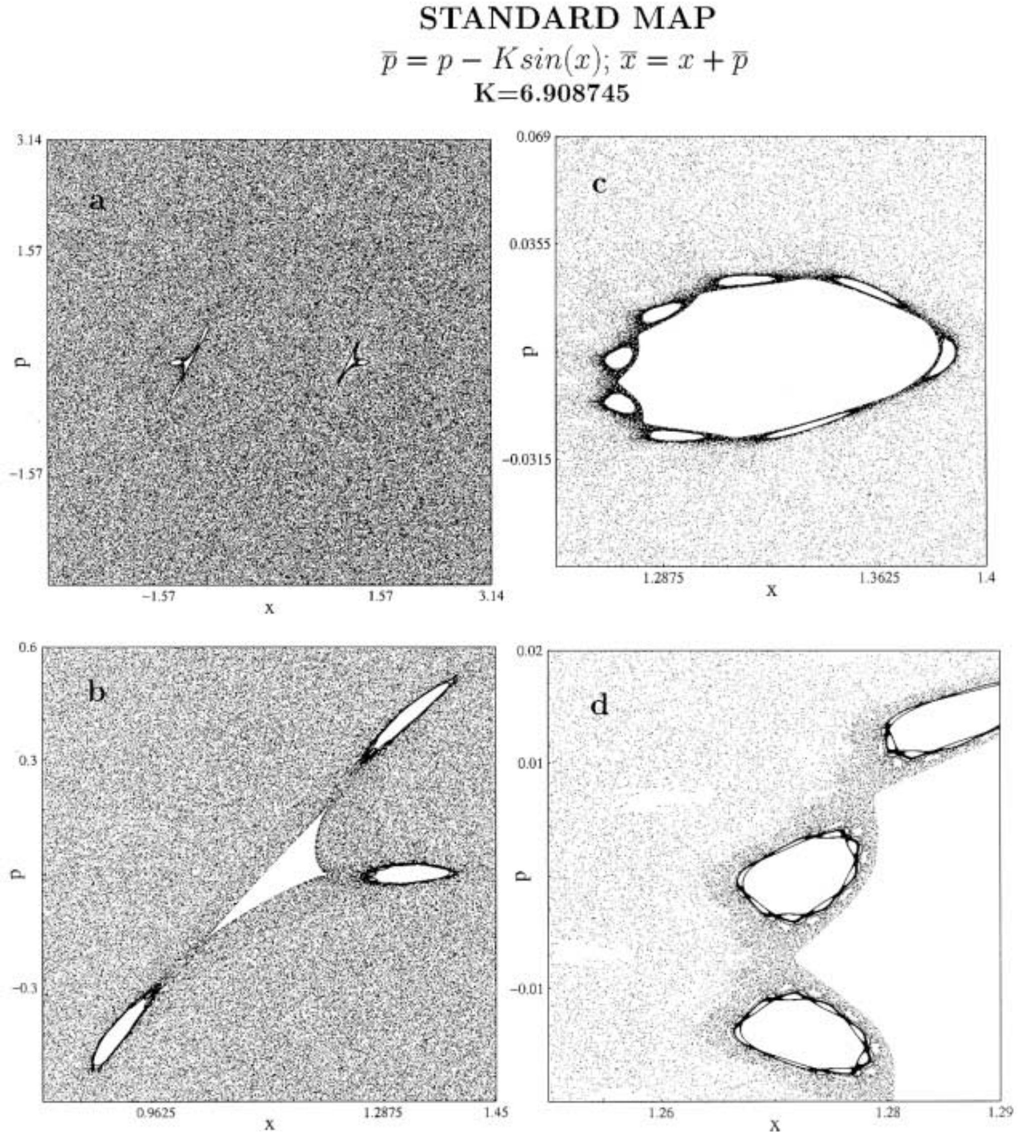


Fig. 3. – A trajectory of the standard map with a parameter $K = 6.908745$ reveals a hierarchy of sticky islands in the sequenced zooms a-d.

level of randomness of dynamics from the recurrences distribution. An effective use of the theory and simulation of the Poincaré recurrences helps to penetrate into some deeply hidden features of Hamiltonian chaos, namely to understand its “non-chaotic” elements, such as intermittancy, superdiffusion, and strange (or fractional) kinetics [22-24]. To understand the last statement, consider a classical problem of the so-called quasilinear theory [5]. Let a charged particle move in the field of a wave packet of electrostatic waves

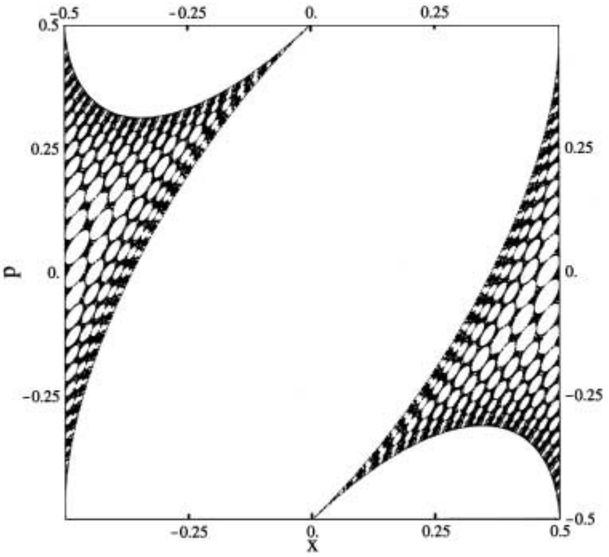


Fig. 4. – A trajectory for the sawtooth type of the standard map with zero Lyapunov exponent for $K = -1.893$.

in one dimension for simplicity:

$$\ddot{x} = \frac{e}{m} \sum_k E_k \cos(kx - \omega_k t) \ . \tag{20}$$

Under some conditions for the amplitudes E_k, and dispersion $\omega_k = \omega_k(k)$ dynamics of particles is chaotic and can be described by the quasilinear equation

$$\frac{\partial f(p,t)}{\partial t} = \frac{\pi}{2} \frac{\partial}{\partial p} \sum_k |E_k|^2 \delta \left(\omega_k - \frac{kp}{m} \right) \cdot \frac{\partial f(p,t)}{\partial p} \tag{21}$$

which is of the diffusional type (see more about its derivation in [14]). Fermi acceleration of particles can also be obtained from (21). This approach describes a normal diffusion when the mixing properties in the phase space are fairly uniform. The condition of the space-time uniformity is important.

In real Hamiltonian systems the condition of the type of (4), (10), or (14) do not mean yet the appearance of the real randomness in the way it forms a basic platform for statistical physics or in the way it was searching for in the FPU paper (see more discussion in [22]). The phase space of systems is not uniform and the presence of islands of stability leads to a kind of stickiness of trajectories to the island's boundaries. The level of stickiness and its characteristic times depend on control parameters and, sometimes, can be very strong and well observable in the experiments [25]. An examle in fig. 3 shows for the standard map how a hierasrchy of islands 2-3-8-8-8... can appear for a special value of the control parameter K, creating a strong stickiness of trajectories to the island's boundaries (dark strips in fig. 3) and leading to the superdiffusive anomalous transport with

$$\langle p^2 \rangle \sim t^{\mu} \tag{22}$$

with $1 < \mu < 2$.

The origin of (22) is not trivial. It follows from a very nontrivial generalization of the diffusional equation by the so-called fractional kinetic equation

$$\frac{\partial^{\beta} f(p,t)}{\partial t^{\beta}} = \mathcal{D} \frac{\partial^{\alpha} f(p,t)}{\partial |p|^{\alpha}} \tag{23}$$

where (α, β) are fractional and their values depend on some intimate features of chaotic dynamics with intermittancy [24]. Particularly from (23) follows

$$\langle |p|^{\alpha} \rangle \sim t^{\beta} \tag{24}$$

which gives in the case of full self-similarity that

$$\mu = 2\beta/\alpha \ . \tag{25}$$

This example opens new features for the Fermi acceleration and for the understanding which kinds of randomness one can expect from the dynamics.

Another example of nontrivial diffusion is for the sawtooth map (8) when the condition (10) is not valid and the Lyapunov exponent is zero. Figure 4 shows how complicated can the behavior of one trajectory. It is definitely random and it performs a kind of random walk but the description of this type of random dynamics needs a longer explanation.

These examples show how far we have stepped on the way of understanding of the origin of statistical physics laws after the FPU paper and how much more new questions are still to be answered.

* * *

I am very thankful to R. Ruffini and V. Gurzadyan for the invitation to this special issue in celebration of the 100th anniversary of the birth of Enrico Fermi. This work was supported by the U.S. Navy Grants N00014-96-1-00557 and N00014-97-1-0426, and the U.S. DOE Grant DE-FG02-92ER54184.

REFERENCES

[1] Fermi E., Pasta J. and Ulam S., Los Alamos Scientific Report LA-1940 (1955).

[2] Fermi E., Pasta J. and Ulam S., in *Collected Papers of Enrico Fermi*, Vol. **2**, edited by E. Segré (The University of Chicago, Chicago) 1965, p. 977.

[3] Ford J. and Waters J., *J. Math. Phys.*, **4** (1963) 1293.

[4] Zabusky N. J. and Kruskal M., *Phys. Rev. Lett.*, **15** (1965) 240.

[5] Vedenov A. A., Velikhov E. P. and Sagdeev R. Z., *Nucl. Fusion*, **1** (1961) 82.

[6] Chirikov B. V., *Atom. Energ. (Atomic Energy)*, **6** (1959) 630.

[7] Fermi E., *Phys. Rev.*, **75** (1949) 1169.

[8] Campbell D., in *Los Alamos Science*, **15**, Special Issue dedicated to Stanislav Ulam, p. 218 (1987).

[9] Izrailev F. M. and Chirikov B. V., *Dokl. Akad. Nauk*, **166** (1966) 57.

[10] Ulam S. M., Los Alamos, MS-2219 (1958).

[11] Ulam S. M., *Proc. 4th Berkeley Symp. on Math. Stat. and Probability*, **3** (Berkeley-Los Angeles) 1961, p. 315.

[12] Zaslavsky G. M. and Chirikov B. V., *Dokl. Akad. Nauk*, **159** (1964) 306 (translation: *Sov. Phys.-Dokl.*, **9** (1965) 989).

[13] Zaslavsky G. M., *Statistical Irreversibility in Nonlinear Systems* (Moscow, Nauka) 1970 (translation: Culham Laboratory, 1976).

[14] Sagdeev R. Z., Usikov D. A. and Zaslavsky G. M., *Nonlinear Physics* (Harwood Acad. Publ. Chur.) 1988.

[15] Lichtenberg A. J. and Liberman M. A., *Regular and Stochastic Motion* (Springer, Berlin) 1983.

[16] Chirikov B. V., *Phys. Rep.*, **52** (1979) 263.

[17] Sagdeev R. Z. and Zaslavsky G. M., *Nuovo Cimento*, **97** (1987) 119.

[18] Petrosky T. J., *Phys. Lett. A*, **117** (1986) 328.

[19] Natenzon M. Ya., Neishtadt A. I., Sagdeev R. Z., Seryakov G. K. and Zaslavsky G. M., *Phys. Lett. A*, **145** (1990) 255.

[20] Zaslavsky G. M., Sagdeev R. Z., Usikov D. A. and Chernikov A. A., *Weak Chaos and Quasiregular Patterns* (Cambridge Univ. Press, Cambridge) 1991.

[21] Van Simaeys G., Emplit Ph. and Haelterman M., *Phys. Rev. Lett.*, **87** (2001) 033902.

[22] Zaslavsky G. M., *Phys. Today*, **8** (1999) 39.

[23] Shlesinger M. F., Zaslavsky G. M. and Klafter J., *Nature*, **363** (1993) 31.

[24] Zaslavsky G. M., Edelman M. and Niyazov B., *Chaos*, **7** (1997) 159.

[25] Solomon T. H., Weeks E. R. and Swinney H. L., *Phys. Rev. Lett.*, **71** (1993) 3975; *Physica D*, **76** (1994) 70.