Acta Physica Hungarica

VOLUME 64 · NUMBERS 1–3, 1988

Akadémiai Kiadó, Budapest

ACTA PHYS. HUNG. APAHAQ 64 (1–3) 1–326 (1988) HU ISSN 0231–4428

ACTA PHYSICA HUNGARICA

A JOURNAL OF THE HUNGARIAN ACADEMY OF SCIENCES

EDITED BY
I. KOVÁCS

Acta Physica publishes original papers on subjects in physics. Papers are accepted in English, French, German and Russian.

Acta Physica is published in two yearly volumes (4 issues each) by

AKADÉMIAI KIADÓ
Publishing House of the Hungarian Academy of Sciences
H-1054 Budapest, Alkotmány u. 21

Subscription information

Orders should be addressed to

KULTURA Foreign Trading Company
1389 Budapest P.O. Box 149

or to its representatives abroad.

Acta Physica Hungarica is abstracted/indexed in Chemical Abstracts, Current Contents–Physical, Chemical and Earth Sciences, Mathematical Reviews, Science Abstracts, Physics Briefs, Risk Abstracts

ACTA PHYSICA HUNGARICA

VOLUME 64

AKADÉMIAI KIADÓ, BUDAPEST
1988

CONTENTS
Volume 64

GENERAL

STATISTICAL PHYSICS

ELEMENTARY PARTICLES AND FIELDS

NUCLEAR PHYSICS

ATOMIC AND MOLECULAR PHYSICS

CONDENSED MATTER

ASTROPHYSICS

CROSS-DISCIPLINARY

CONTENTS

GENERAL

STATISTICAL PHYSICS

ELEMENTARY PARTICLES AND FIELDS

NUCLEAR PHYSICS

ATOMIC PHYSICS

ASTROPHYSICS

CROSS-DISCIPLINARY

Acta Physica Hungarica 64 (1-3), pp. 3-5 (1988)

EINE KLEINE NACHTPHYSIK*

W. THIRRING

Institute for Theoretical Physics, Vienna University
Vienna, Austria

(Received 8 January 1987)

Some thoughts are given to the anthropic principle and its traces in physics.

The non-informative title of this essay stems from the fact that it was written on a night flight when I was stirring up my memories. The reader who finds its scientific substance somewhat thin should blame that on the environment where it was scribbled.

In the post-war period George Marx and I struggled together, each of us in his country, to raise theoretical physics in central Europe to a level fitting for the intellectual potentialities of this area. Whereas I then turned to the hard mathematical facts he has reflected a great deal about the functioning of the universe, a subject rich in philosophical speculations. In this contribution I will also indulge in some related thoughts because it seems to me that some old ideas of men appear now in a new form. Theologians always wanted for men, the coronation of the creation, an appropriately distinguished place in the universe. Though we are admittedly at a rather cosy distance to the sun, our position in the galaxy or galaxy cluster seems to be nothing special, we did not get a first class ticket for our journey. In contradistinction the universe we are living in does indeed seem to be very distinguished. If we change its laws or other conditions just a little bit we do not only erase a subtle phenomenon like life from its surface but the scene changes far more drastically. I shall illustrate this claim by going through the history of the universe but making some slight deviations from its course.

1) At present one believes that the universe was created by a kind of quantum mechanical tunnel effect where the energy for the matter was paid for by the negative gravitational energy, the total energy being zero. The relevant time scale for such a spontaneous process is the Planck time 10^{-43} s and the remarkable fact is that this fluctuation was so successful that it lasted 10^{18} s. Starting an expanding universe is like launching a satellite, if you do not give it enough thrust it falls back to the ground, if you give it too much thrust it soon disappears into space. Only for very precisely selected initial conditions it will stay in orbit for a long time and this is what happened with our universe. Indeed, we have seen that on the relevant time scale its age is 10^{62}.

*Dedicated to Prof. G. Marx on his 60th birthday

2) We now think that the laws of nature have undergone an evolution and what we see now is the result of some freezing out process. Therefore the masses and coupling constants may be the results of historical accidents, as incalculable as the thickness of the ice on a lake after a cold night. In any case if they had turned out slightly different the appearance of the universe would change completely. Just consider the following possibilities.

a) The mass difference between the proton and the neutron is a sensitive effect, the small difference between various contributions. Changing the coupling constant a little bit it could easily come out the other way and make the proton heavier than the neutron. In this case the proton would be unstable and there would be no hydrogen, the basic raw material for cooking heavier elements. Thus the dark landscape of the universe would then be marked only by neutron stars.

b) The ratio between the mass of pions and electrons, being the lightest charged particles of families at best remotely related, is not calculable at present. A theory where the pion has a mass of 5 MeV but the electron 140 MeV seems equally good to the elementary particle physicist but would put the solid state physicist out of job. In this case the pion would be stable and matter consisted not of e^- and nuclei but of π^- and nuclei. The π^- would destroy heavier nuclei by the reaction $p + p + \pi^- \to p + n + 5$ MeV but even the fate of hydrogen would be wild. Since the pions are bosons the the binding energy of hydrogen would not increase $\sim N =$ number of hydrogen atoms but first like $Ry \cdot N^{5/3}$ until the kinetic energy of the proton stabilizes the system. This happens at densities $(\hbar^2/M_p e^2)^{-3} = 10^{10}$ times the density of water. At that stage it becomes a jellium of protons in a negative charge background which practically has no kinetic energy, its binding energy per particle is $(M_p/m_e)Ry \sim 2000\, Ry$. Thus in this scenario we have a potent and space saving fuel but there would be no rockets which could use it.

c) Universe with only hydrogen and perhaps a little helium would be dull, for life we need some heavier elements. To produce them there is the bottle neck that two α-particles do not stick together for any length of time and unless a third one happens to be around to make C^{12} the further evolution of elements is blocked. How one gets through this eye of the needle depends sensitively on the properties of the excited states of Be^8 and C^{12} and thus is essentially influenced by the strength and range of the nuclear forces. Yet the evolution of life as we know it depends on the existence of heavier elements. This list of the hurdles to the creation of life can easily be continued and experts from many fields can throw in some more facts which would impede the existence of life.

In view of all that some people thought that it should be a guiding principle for the laws of nature to be such that they eventually lead to life. It is called the anthropic principle and as physicist one has the duty to see how one could find a reason for that. There are several attitudes possible.

A) One can accept it by an act of faith and say that when the good Lord created the world he did it exceedingly cleverly so that it works so well. This seems to be the most reasonable answer in our present state of knowledge but it may not satisfy people who like to play with their imagination.

B) One can adopt a Darwinistic point of view and assume that there have been

many universes created but most of them were no good, either they collapsed too soon or the separation of the various interactions was all messed up e.t.c. Once there was finally the lucky strike where everything worked. Then intelligent creatures evolved and wondered why everything conspires to make their existence possible. Some people may like this idea but unless we see some traces of all these abortive universes there is not much scientific substance to such a hypothesis.

C) Since people now find self-organization for many systems without paying much attention to a realistic representation of matter, it may be that self-organization is a feature as general as the tendency to equilibrium and some sort of life could also evolve under quite different circumstances. For instance, even in our first scenario where there are only neutron stars, it could be that at its surface some highly organized structures come into being. As the time scale in nuclear matter is speeded up compared to ordinary matter by about a million they would evolve much faster and after a few thousand years higher civilizations may populate this otherwise dull scenario.

Obviously by such considerations one easily drifts into science fiction. To describe the situation in more scientific terms one might say that the present situation in physics is somewhat like in mathematics in the post-Gödel area. Hilbert had thought that one could cast mathematics in a closed rational system where every truth is provable, but Gödel had shown that this is not so. Similarly one could have hoped that physics became a closed rational system where every important fact could be deduced from the fundamental laws. It turned out that the facts important for us appear to be more of an accidental nature and beyond scientific deduction. What is engraved in the fundamental laws does not seem to care too much about us. Maybe our present understanding of nature just reflects the fact that our mind can readily grasp simplicity and symmetry, but has a hard time seeing through complexity.

Acta Physica Hungarica 64 (1–3), pp. 7–13 (1988)

PROGRESS IN OUR NOTION OF MASS* **

H. PIETSCHMANN

Institute for Theoretical Physics, Vienna University
Vienna, Austria

(Received 8 January 1987)

The development of our concepts of quantum field theory especially in the version of local gauge theories has brought about a tremendous success in predicting and explaining experimental facts. As a sideeffect, our understanding of the notion of mass has undergone dramatic changes. Since this is not accompanied by a change in experimental predictions or empirical notions, it is progress in a field which might be considered on the borderline of physics and philosophy of nature.

Introduction

My first encounter with George Marx was through his paper on the electron–muon mass difference [1]. I was working on my thesis, which dealt with a very similar problem [2] and this coincidence was the beginning of a long-lasting friendship as well as common scientific interests.

Our notion of mass has totally and dramatically changed since these days and in the following, I shall try to work out the various phases of this change; in each case, I shall try to explain the notion of the time as well as its origins and problems.

Mass is interaction energy

Around 1960 it was considered quite obvious that the small mass differences between neutral and charged partners of an iso-multiplet were entirely due to electromagnetic self-energy. The fact, that computations within quantum electrodynamics led to divergent integrals, was attributed to various reasons, depending on the point of view of the author; common to all was the opinion that the real origin of the divergent integrals lay in some physical phenomenon which was not properly described by theory at that time.

Straightforward calculations of the pion mass difference, the most typical example for an electromagnetic mass difference, led to [3]

$$m_{\pi^+}^2 - m_{\pi^0}^2 = \frac{3\alpha}{4\pi}\lambda^2, \tag{1}$$

*Dedicated to Prof. G. Marx on his 60th birthday

**Supported in part by "Fonds zur Förderung der wissenschaftlichen Forschung in Österreich", Project Nr. 5444.

where λ is a cut-off parameter of the dimension mass. The fact that the mass difference was proportional to the fine-structure constant α seemed to indicate the right order of magnitude.

The crucial test for this mass philosophy was, of course, the mass difference between neutrino and electron, since the neutrino was considered massless. For fermions, an electromagnetic mass shift is always proportional to the bare mass due to γ_5-invariance. Thus, in order to create an electromagnetic mass for a particle without "mechanical mass" (the term used in those years), γ_5-invariance had to be broken.

In 1958 the V–A theory was born, independently by Marshak and Sudarshan [4] and Feynman and Gell-Mann [5]. Feynman has based his arguments on a two-component version of quantum electrodynamics. The success of this ansatz was probably reason enough to prompt Feynman to suggest a mechanism for the breaking of γ_5-invariance and creating a mass for the electron, based on the same two-component wave equation for Dirac particles [6]. Whether or not this particular mechanism was assumed, the notion that the electron mass was of electromagnetic origin was quite generally accepted.

This, however, led to the consequence, that by analogy also the masses of the heavy particles should be created by strong interactions alone, and it seemed to be a natural observation that the mass of a particle was heavier the stronger its interactions were.

An exception of this rule of thumb was the muon. Thus it was quite obvious to attribute some kind of strong interaction to the muon, which should distinguish it from the electron. The solution of George Marx [1] was quite elegant: electromagnetic masses and mass differences were of the order of eV, masses (or mass differences) of the order of 1 GeV were generated by pionic interactions, and masses (and mass differences) of the order of 100 MeV were created by kaon interactions. Thus, the muon was assumed to interact with pairs of kaons (in order to conserve strangeness).

In a different picture, J. Schwinger in a famous paper [7] proposed an interaction of the muon with a new type of scalar field. The crucial test of all these hypotheses was the rather well-known magnetic moment of the muon [8].

Another crucial question in this scheme was the electromagnetic mass difference of the nucleons, i.e. the question why the neutron is heavier than the proton. In a famous paper, Feynman and Speisman [9] attributed this fact to the anomalously strong magnetic interactions of the nucleons (though this calculation is cut-off dependent). In other schemes, the "wrong sign" of the electromagnetic mass difference of nucleons (which is paralleled by that of K-mesons) was connected to larger multiplet structures. All these attempts are historic by now, but they show, that a quarter of a century ago, it seemed to be one of the main problems to "explain" the fact that the proton is the lightest (and therefore stable) baryon.

Before I close this Section, let me mention some attempts at finite calculations of electromagnetic mass differences. For hadrons, a promising scheme seemed the inclusion of form factors [10] or other damping methods due to strong interactions [11].

When current algebra provided a means to compute a finite mass difference for pions, it drew rather intensive attention from the particle physics community. The result was [12]

$$m_{\pi^+}^2 - m_{\pi^0}^2 = \frac{3\alpha}{2\pi} m_\rho^2 \ln 2. \tag{2}$$

However, it became clear quite soon [13] that the finiteness of the result rested heavily on the assumption of massless pions and stable ρ-mesons. A similar (equally problematic) finite calculation was possible for ρ -mesons [14] with the result

$$m_{\rho^+} - m_{\rho^0} = \frac{\alpha}{2} m_{\rho^0} \left\{ \frac{3}{8\pi} + \frac{4\pi f_\pi^2}{m_\rho^2} \right\}. \tag{3}$$

Heavy particles without strong interactions?

When the g–2 experiment reached higher and higher accuracy, it became more and more plausible that the muon was just a heavy electron without extra interactions. The first — still vague — ideas of something like a generation structure of elementary particles was published as early as 1958 [15]. This question was, however, completely settled only when the third generation was discovered.

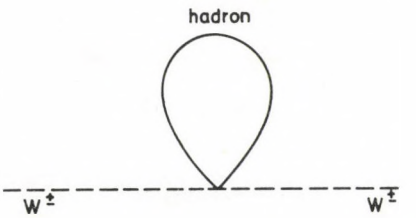


Fig. 1. Strong self-energy of intermediate bosons

In the next Section, we shall discuss how the notion of gauge particles emerged together with the problem of their masses. Before this notion was cleared up, the old idea of mass as interaction energy still persisted and it was therefore natural to assume that the intermediate bosons followed the same scheme. In the mid-sixties it was clear that their mass was larger than about 2 GeV. Since their main purpose was to mediate weak interactions, they were not allowed to participate in a strong Yukawa type interaction. In order to solve this puzzle, a strong quartic interaction of type $W^+W^-\overline{H}H$ (where H is any hadron) was proposed [16]. This could easily lead to strong self-energies of the type shown in Fig. 1.

Masses of gauge bosons

The general notion of gauge symmetry as the origin for interactions between fermions is rather old. With respect to our problem of masses, the question became virulent when Sakurai [17] suggested that strong interactions are based on the gauge principle with the vector mesons ρ, ω and ϕ as the gauge bosons. The masses of these vector mesons — well determined by experiments — became the main obstacle of new theoretical interest.

The idea that spontaneous symmetry breaking may be a cause for mass generation is also rather old [18]. With the renewed interest of vector mesons, various mechanisms were suggested in the literature [19]. A turning point in the history of this question was reached when the famous papers of Higgs appeared [20]. In spite of the fact that these considerations were immediately extended [21], its value in explaining the masses of gauge particles was only fully recognized in the famous papers by Salam and Weinberg [22].

Let us recall the basic Lagrangian of the Higgs model for one gauge boson:

$$\mathcal{L}_H = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + (D_\mu\Phi)(D^\mu\Phi)^* - \mu^2|\Phi|^2 - h|\Phi|^4. \tag{4}$$

With the abbreviations

$$D_\mu = \partial_\mu - ieA_\mu \ , \tag{5}$$

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \ . \tag{6}$$

The well-known trick is to allow μ^2 to become negative and Φ to obtain a non-vanishing vacuum expectation value

$$< 0|\Phi|0 >= \frac{\lambda}{\sqrt{2}} = \frac{1}{2}\sqrt{-\mu^2/h} \ . \tag{7}$$

Together with a very successful SU(2) × U(1) unification of weak and electromagnetic interactions, this led to a complete dramatic change in our notion of masses. Masses were no longer the result of interaction energy but a consequence of the structure of the vacuum. Not only were the gauge boson masses produced by spontaneous symmetry breaking but *all* masses of elementary particles (with a possible exception of the Higgs itself). Table I shows the relation of particle masses to the parameters of the Higgs Lagrangian (presuming the extension of the Lagrangian Eq. (4) to a Higgs doublet and the gauge boson structure of SU(2) × U(1)).

Notice that the masses of fermions and Higgs particles are not *really* predicted since the parameters f and h do not appear in any other experimentally accessible relation. But Table I shows the new philosophy: masses are created by the vacuum expectation value of the Higgs field.

Table I

Relation of particle masses to Higgs parameters and the weak mixing angle Θ_W

$\lambda^2 = 1/\sqrt{2}G_F$	
$m_W = \frac{e\lambda}{2\sin\Theta_W}$	$m_f = \frac{\lambda f}{2}$
$m_Z = \frac{e\lambda}{\sin 2\Theta_W}$	$m_H = \lambda\sqrt{2h}$

Through this change of paradigm the old question of why the neutron is heavier than the proton ceased to be a burning puzzle! It was now shifted to the fact that fermion masses were essentially free parameters and that the down quark is heavier than the up quark. But at the same time, for the first time in the history of particle physics, the problem of masses was directly connected to an experimental prediction: the prediction of the existence of a scalar particle — the Higgs boson.

Since there are absolutely no experimental indications as to the existence of a Higgs boson, several attempts have been made to circumvent this prediction [23].

Higgs boson masses

It has been mentioned in the last Section that the mass of the Higgs boson itself is essentially a free parameter. Yet there are interesting effects, which bind its value both from above and below. The effective lower bound of about 4 GeV stems from the fact that loop corrections create an effective potential which may spoil the spontaneous symmetry breaking effect if h is too small [24]. The upper bound stems from the observation that perturbation theory will cease to be meaningful if m_H exceeds the critical value [25]

$$m_H < m_{\text{crit}} = \left(\frac{8\pi\sqrt{2}}{3G_F}\right)^{1/2}. \tag{8}$$

Conclusion

Gauge theories have brought about a change of paradigm with respect to our notion of mass. Table II tries to represent this change.

It is quite clear that Table II is only a qualitative indication. However, it tries to show that some problems (like the neutron–proton mass difference) have become less burning, while it seems to me that the most important problem, which

Table II

Change of paradigm with respect to some problems concerning masses

Old view	New view	Search for solution
e–μ–puzzle	generation problem	horizontal symmetry
mass = interaction energy	Higgs phenomenon	find the Higgs or replace it by other mechanism
$m_n > m_p$	$m_d > m_u$	

is faced by elementary particle physics today, is the question of whether or not a Higgs boson exists and what kind of mechanism is the true cause for generating masses.

References

1. G. Marx and K.L. Nagy, Nucl.Phys., *12*, 125, 1959.
2. H. Pietschmann, Zum Massenspektrum der Leptonen, Ph.D. thesis, Univ. Vienna, unpublished. 1960.
3. A. Petermann, Helv. Phys. Acta, *27*, 441, 1954.
4. R.E. Marshak and E.C.G. Sudarshan, Phys. Rev., *109*, 1860, 1958.
5. R.P. Feynman and M. Gell-Mann, Phys.Rev., *109*, 193, 1958.
6. R.P. Feynman, Ann. Int. Conf. High En. Phys. CERN, p.216, 1958.
7. J. Schwinger, Ann. Phys. (N.Y.), *2*, 407, 1957.
8. H. Pietschmann, Acta Phys. Austr., *13*, 315, 1960.
9. R.P. Feynman and G. Speisman, Phys. Rev., *94*, 500, 1954.
10. S.K. Bose and R.E. Marshal, Nuovo Cim., *25*, 529, 1962.
11. L.M. Brown and H. Munczek, Phys. Rev. Lett., *20*, 680, 1968.
12. T. Das, G.S. Guralnik, V.S. Mathur, F.E. Low, J.E. Young, Phys. Rev. Lett., *18*, 759, 1967.
13. G.C. Wick and B. Zumino, Phys. Lett., *25B*, 479, 1967;
    M.B. Halpern and G. Segrè, Phys. Rev. Lett., *19*, 611, 1967;
    T. Kokott, H. Pietschmann, H. Rollnik, Z. f. Phys., *216*, 65, 1968.
14. H. Pietschmann, Acta Phys. Austr., *29*, 91, 1969.
15. M. Goldhaber, Phys. Rev. Lett., *1*, 467, 1958.
16. S. Pepper, C. Ryan, S. Okubo, R.E. Marshak, Phys. Rev., *137B*, 1259, 1965.
17. J.J. Sakurai, Ann. Phys. (N.Y.), *11*, 1, 1960.
18. Y. Nambu, Phys.Rev. Lett., *4*, 380, 1960.
19. For example J. Schwinger, in "Theoretical Physics", IAEA publication, Vienna, p.89, 1963.
20. P.W. Higgs, Phys. Lett., *12*, 132, 1964; Phys. Rev. Lett., *13*, 508, 1964; Phys. Lett., *145*, 1156, 1966.
21. T.W.B. Kibble, Phys. Rev., *155*, 1554, 1967.
22. A. Salam, Elementary Particle Theory, ed. N. Svartholm, Almqvist and Wiksell, Stockholm, 1968;

S. Weinberg, Phys. Rev. Lett., *19*, 1264, 1967.
23. For example, S. Coleman and E. Weinberg, Phys. Rev., *D7*, 1888, 1973.
24. S. Weinberg, Phys. Rev. Lett., *36*, 294, 1976;
A.D. Linde, JETP Lett., *23*, 64, 1976;
M. Veltman, Acta Phys. Polonica, *B8*, 475, 1977.
25. B.W. Lee, C. Quigg, H.B. Thacker, Phys. Rev., *D16*, 1519, 1977; Phys. Rev. Lett., *38*, 883, 1977.

Acta Physica Hungarica 64 (1–3), pp. 15–19 (1988)

IS THE MAGNETIC MOMENT OF THE NEUTRON A DIPOLE OR AN AMPERIAN CURRENT LOOP?*

F. MEZEI**

Central Research Institute for Physics
1525 Budapest, Hungary

(Received 8 January 1987)

In classical electrodynamics the two model representations of magnetic moments, viz. the dipole and the Amperian current loop models, give equivalent results. In contrast, for particles with overlapping wave functions, this equivalence breaks down. In the case of the neutron a heated theoretical debate in the 1930's on the choice of the correct model has been experimentally settled in the 1950's in favour of the current loop model. The question became of down-to-earth technical relevance in neutron spin echo spectroscopy.

It is a widespread belief that this question makes no sense and the two alternatives mentioned in the title are equivalent. This is perfectly wrong, and actually the current loop model has been proven to be the true one. [1].

This question was a hotly debated controversy from 1936 until it was experimentally settled in 1951, but all this has been mostly forgotten since then.

Let us first state the problem by considering the force acting on a magnetic moment in a magnetic field, as predicted by the two models of classical microscopic dipole or a classical microscopic Amperian current loop, as illustrated in Fig. 1. We will consider the forces given by ordinary electrostatics and magnetostatics in the limit of making the size of the object going to zero with the magnetic moment ($\boldsymbol{\mu} = e\mathbf{d}$ and $\boldsymbol{\mu} = I\mathbf{f}/c$ in the two cases, respectively, where $\mathbf{f}$ is the surface element vector of the current loop) kept constant. Surprisingly, the two results are different:

$$\begin{aligned}\mathbf{F}_d &= \mathbf{F}^+ + \mathbf{F}^- = (\boldsymbol{\mu}\cdot\operatorname{grad})\mathbf{H} \\ &= \operatorname{grad}(\boldsymbol{\mu}\cdot\mathbf{H}) - \boldsymbol{\mu}\times\operatorname{curl}\mathbf{H}\ ,\end{aligned} \tag{1}$$

$$\begin{aligned}\mathbf{F}_c &= \oint d\mathbf{F} = \frac{I}{c}\oint d\mathbf{l}\times\mathbf{H} \\ &= \frac{I}{c}\int \operatorname{grad}\,(\mathbf{H}\cdot d\mathbf{f}) - \frac{I}{c}\int(\operatorname{div}\mathbf{H})\,d\mathbf{f} \\ &= \operatorname{grad}(\boldsymbol{\mu}\cdot\mathbf{H})\ .\end{aligned} \tag{2}$$

*Dedicated to Prof. G. Marx on his 60th birthday

**Present address: Hahn-Meitner-Institute and Technical University Berlin, POB 390128 D-1000 Berlin /West/ 39.

The difference between the force acting on the dipole $\mathbf{F}_d$ and the one acting on the current loop $\mathbf{F}_c$ is nevertheless only non–zero, if the current density at the position of the magnetic moment is non–zero. Thus the well–known equivalence of these two models in electromagnetism of continuous media is fully compatible with this difference, since materials do not penetrate each other. We can also observe that Eq. (2) suggests that the energy of the magnetic moment in a field has the simple form $-\mu \cdot H$, while the situation is more complex for the dipole model. (Knowing nature's penchant for simplicity and aesthetics, we could just discard the dipole model at this point. In fact this would not be all that arbitrary: it just turns out to be logically difficult to accommodate in a microscopic theory both dipoles and currents as sources of fields.)

If we consider a magnetic moment inside a medium, we have to think of H as the microscopic field. Since we attribute this internal field to the Amperian microcurrents of the electrons, the average of the internal microfields on a macroscopic scale gives what is customarily called the magnetic induction of the medium B_m. (On the microscopic level it does not make sense to distinguish between B and H, and except otherwise stated I will write H. In some textbooks B is used in this context, i.e. as the real physical field in vacuum.) On the other hand, if we would postulate that magnetism in media is due to microscopic dipoles, then the internal microscopic field would average to what is usually referred to as the magnetic intensity in a medium, $H_m = B_m - 4\pi M$, with M being the macroscopic magnetic moment density. Thus there will be two simple limiting cases considering the average force exerted by a magnetic medium on our particle:

a) If both the particle moment and the magnetism of the media is described by Amperian currents:

$$F_c = \text{grad}\ (\mu \cdot B_m)\ ; \tag{3}$$

b) If both of them are described by dipoles:

$$F_d = \text{grad}(\mu \cdot H_m)\ . \tag{4}$$

I have listed these two cases not only because they are the ones which obviously correspond to simple Hamiltonians, but also because the actual debate on the nature of the neutron magnetic moment in 1930's and 40's ultimately boiled down to the choice between these two Hamiltonians [2].

Sure enough, the original controversy never appeared as such a simple matter at that time and there was quite some confusion. In his prophetic paper Bloch [3] introduced in 1936 what we today call magnetic neutron scattering, but unfortunately he used the dipole model. Schwinger reexamined the problem using a formal quantum mechanical treatment and arrived at a different expression for the cross section [4]. Incorrectly, he believed that this was due to the use of quantum mechanics from the outset in contrast to the classical character of some of Bloch's arguments. In his answer [5]. Bloch has correctly demonstrated that the difference has nothing to do with the use of quantum mechanics (which would violate the fundamental principle of correspondence), but with the dipole vs current loop assumption. He has shown that if we assume kind of a Lorentz hole in the electron

density around the neutron, the result will depend on the shape of this exclusion hole. This actually demonstrates the old difficulty in magnetostatics with the description of interpenetrating magnetizations, viz. what Bloch was trying to avoid was a not curl–free magnetic field, in contrast to our Eq.(1). In a little known paper from 1938 [6], Migdal claimed to have demonstrated that the model character of the neutron moment is irrelevant, which would restore the equivalence of the dipole and Amperian current pictures in magnetostatics. In view of the simple example of Eqs (1) and (2) this is of course incorrect, and on closer examination one finds that he used the Amperian model from the outset. Finally Ekstein [2] summarized the two types of Hamiltonians in Eqs (3) and (4), but he did not offer any interpretation for the two assumptions, and — incorrectly — he maintained the distinction between H and B even on the microscopic level. What this hid was a fact that nobody realized at that time, namely that the dipole vs current loop character of the microscopic sources of magnetism of the medium itself is as much part of the problem as the nature of the neutron moment.

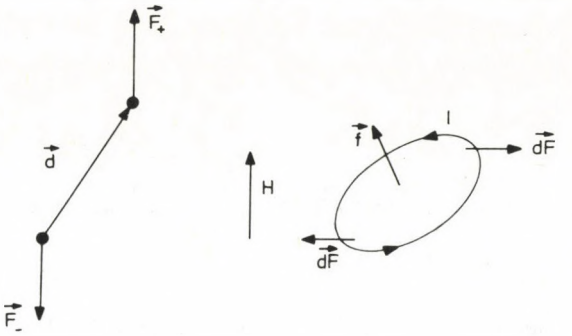


Fig. 1. Magnetic forces acting on a dipole and on a current loop

The forces illustrated in Fig. 1 can be actually measured by observing the deflection of the neutron beam traversing a superconducting foil in which high density surface currents can be induced. The result confirms the current loop model and it rules out the dipole model. Furthermore a ferromagnetic foil produces exactly the same deflections as an equivalent superconducting foil, showing that the spin magnetic moment of the electrons is also equivalent to a current loop, and not to a dipole.

Beyond the purely intellectual curiosity, a clear understanding of the problem of magnetic forces acting on neutron beams also became an important technical issue in high resolution NSE spectroscopy [7]. The resolution of the next instrument to be completed by 1988 at ILL will attain a fraction of a neV. In this method the neutron energy change in the sample scattering process is determined by directly comparing the incoming and outgoing velocities for each neutron individually. This is achieved by making the neutron spins perform large numbers of Larmor precessions in the

opposite sense before and after the scattering, as illustrated in Fig. 2. The Larmor precessions are initiated by turning the neutron spins perpendicular to the magnetic field direction with the help of the first $\pi/2$ flipper coil. The π spin flipper next to the sample inverts the apparent sense of the precessions and the second $\pi/2$ coil together with the polarization analyser serves for the observation of the precessing polarization of the scattered beam. The problem with the extreme high resolution now planned is that the flipper coils used to initiate and handle the precessions produce about 10 Oe fields, i.e. Zeeman energies of the same magnitude as the energy resolution of the spectrometer. Furthermore, in the flipper coils the neutrons traverse the current carrying wires, i.e. regions with curl $\mathbf{H} \neq 0$. Thus it became inevitable to analyse the influence of the flipper coils on the velocity of the neutron. It can be shown in detail [1] that with the Amperian current loop model found to be valid, the flipper coils produce no disturbing effects, essentially because of the existence of the bona fide potential energy $-\boldsymbol{\mu} \cdot \mathbf{H}$.

We will conclude by examining the physical origin of the Hamiltonian $-\boldsymbol{\mu} \cdot \mathbf{H}$. It is commonly believed that it corresponds to the mixed term in the energy of the total magnetic field $\mathbf{H} + \mathbf{H}_n$, where $\mathbf{H}_n$ is the magnetic field of the neutron:

$$\begin{aligned} E_{\text{field}} &= \frac{1}{4\pi} \int \mathbf{H} \cdot \mathbf{H}_n \, dV = \frac{1}{4\pi} \int \mathbf{H}_n \cdot \operatorname{curl} \mathbf{A} \, dV \\ &= \frac{1}{4\pi} \int \mathbf{A} \cdot \operatorname{curl} \mathbf{H}_n \, dV = \frac{1}{c} \int \mathbf{A} \cdot j_n \, dV \\ &= \int \mathbf{A} \cdot \operatorname{curl} \mathbf{M}_n \, dV = \int \mathbf{M}_n \cdot \operatorname{curl} \mathbf{A} \, dV \\ &= \boldsymbol{\mu} \cdot \mathbf{H} \,, \end{aligned} \tag{5}$$

where $\mathbf{A}$ is the vector potential for $\mathbf{H}$, and $\mathbf{j}_n = c \operatorname{curl} \mathbf{M}_n$ the current density for $\mathbf{H}_n$. This argument gives the correct result in electrostatics, but as we see, in magnetism it gives the negative of the correct answer. (This wrong sign has been overlooked in some otherwise valuable textbooks). The explanation of this apparent discrepancy is that if magnetic moments are moving with respect to each other there is a mutual induction effect which tends to change the currents of which the fields originate. It can be shown [8] that the electric energy required in order to keep these currents constant (e.g. provided by the power supply if the field is produced by a coil) is exactly equal to the change of $-\boldsymbol{\mu} \cdot \mathbf{H}$ for both the neutron and the source of the external field $\mathbf{H}$. Thus the total energy balance is indeed:

$$\Delta E_{\text{field}} + \Delta E^{\text{neutron}}_{\text{induction}} + \Delta E^{\text{field}}_{\text{induction}} = \Delta(-\boldsymbol{\mu} \cdot \mathbf{H}) \,. \tag{6}$$

It might appear somewhat surprising that the apparently local interaction term $-\boldsymbol{\mu} \cdot \mathbf{H}$ in reality implies the action by distant objects, which might just stress the fundamental nonlocality of quantum mechanics.

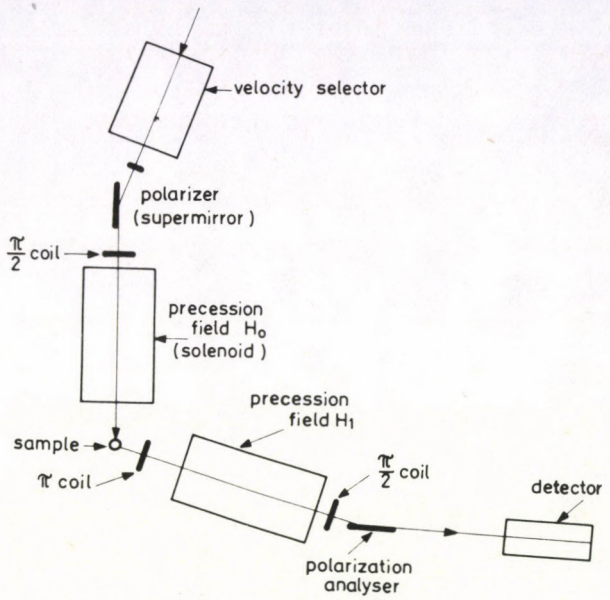


Fig. 2. Schematic lay-out of a Neutron Spin Echo spectrometer

References

1. F. Mezei, Physica, *137B*, 295, 1986.
2. H. Eckstein, Phys. Rev., *76*, 1328, 1949.
3. F. Bloch, Phys. Rev., *50*, 259, 1936.
4. J. Schwinger, Phys. Rev., *51*, 544, 1937.
5. F. Bloch, Phys. Rev., *51*, 994, 1937.
6. A. Migdal, Comptes Rendus /Doklady/ Acad. Sci. URSS, *20*, 551, 1938.
7. F. Mezei, Z. Phys., *255*, 146, 1972.
8. L.D. Landau and E.M. Lifshitz, Electrodynamics of Continuous Media, Pergamon, Oxford, 1984.

Acta Physica Hungarica 64 (1–3), pp. 21–26 (1988)

MECHANICAL EQUATION OF MOTION FOR ENERGY AND ENERGY-MOMENTUM TENSOR OF ELECTROMAGNETIC FIELD IN TRANSPARENT MEDIUM*

K. NAGY

Institute for Theoretical Physics, Roland Eötvös University
1088 Budapest, Hungary

(Received 8 January 1987)

With reference to one of K. F. Novobátzky's papers we deduce a mechanical equation of motion for energy of a non-isolated classical field. By way of application we consider the stress tensor of an electromagnetic plane-wave propagating in a transparent medium. It turns out from this consideration that Abraham's energy-momentum tensor describes the energy and momentum of electromagnetic field in an isotropic transparent medium correctly.

In one of his papers [1] K. F. Novobátzky posed a question whether one could find a dynamical theory of energy on the ground of equivalence of mass and energy. The answer was yes. It was shown that the equation of motion for energy of an isolated system is very similar to the equation of motion for a continuum. The mechanical equation of motion can be expressed by divergence of the stress tensor of the system, therefore it is related with the energy-momentum tensor.

In this context I consider the dynamical description of electromagnetic radiation in an isotropic and transparent medium.

Consider in general a closed system which consists of two interacting subsystems. A system consisting of a transparent medium and electromagnetic field is a typical example.

The energy-momentum tensor of the whole system is

$$\Theta_{\alpha\beta} = T_{\alpha\beta} + t_{\alpha\beta}, \tag{1}$$

where $T_{\alpha\beta}$ and $t_{\alpha\beta}$ denote the energy-momentum tensor of subsystems.

The system is isolated, therefore we have

$$\frac{\partial \Theta_{\alpha\beta}}{\partial x_\beta} = 0, \tag{2}$$

from where it follows that

$$\frac{\partial T_{\alpha\beta}}{\partial x_\beta} = -\frac{\partial t_{\alpha\beta}}{\partial x_\beta} = -f_\alpha, \qquad (\alpha = 1,2,3,4), \tag{3}$$

*Dedicated to Prof. G. Marx on his 60th birthday.

where f_α denotes the 4-force density between the interacting subsystems.

Let us now study a subsystem characterized by $T_{\alpha\beta}$. According to the above mentioned example $T_{\alpha\beta}$ is the energy-momentum tensor of electromagnetic field in a transparent medium which is supposed to be homogeneous and isotropic. One can derive $T_{\alpha\beta}$ from a Lagrangian of the subsystem by variation of metric tensor $g^{\alpha\beta}$. It follows from this definition that $T_{\alpha\beta}$ is symmetric. In consequence of conservation laws $\Theta_{\alpha\beta}$ is symmetric, therefore $t_{\alpha\beta}$ is symmetric as well. The well known physical meaning of components $T_{\alpha\beta}$ is the following:

$$T_{4k} = \frac{i}{c} S_k, \qquad T_{k4} = icg_k, \qquad (k = 1,2,3) \tag{4}$$

and

$$T_{44} = -u, \tag{5}$$

where **S** is the energy current density, **g** is the momentum current density and u is the energy density. Components T_{ij} are related with the local stresses.

The symmetry of $T_{\alpha\beta}$ yields the Planck's formula

$$\mathbf{g} = \frac{\mathbf{S}}{c^2} \tag{6}$$

which expresses the inertness of energy of all kinds.

From Eq. (3) we have

$$\frac{\partial T_{i\beta}}{\partial x_\beta} = \frac{\partial T_{ik}}{\partial x_k} + \frac{\partial T_{i4}}{\partial x_4} = -f_i, \qquad (i = 1,2,3). \tag{7}$$

$T_{i4} = T_{4i}$, consequently it follows from (4) that

$$\frac{\partial T_{ik}}{\partial x_k} + \frac{1}{c^2}\frac{\partial S_i}{\partial t} = -f_i, \qquad (i = 1,2,3). \tag{8}$$

Consider now Eq. (3) for the case α=4:

$$\frac{\partial T_{4\beta}}{\partial x_\beta} = \frac{\partial T_{4k}}{\partial x_k} + \frac{\partial T_{44}}{\partial x_4} = -f_4.$$

In the light of the meaning of T_{4k} and T_{44} and from equations (4) and (5) we have the following differential conservation law:

$$\frac{\partial S_k}{\partial x_k} + \frac{\partial u}{\partial t} = icf_4. \tag{9}$$

It turns out that icf_4 plays the role of energy-source-density. Let us define the velocity of energy-propagation as

$$v_k = \frac{S_k}{u}, \qquad (k = 1,2,3) \tag{10}$$

and let the fourth component be $v_4 = ic$. From the conservation law (9) we have

$$\frac{\partial(uv_k)}{\partial x_k} + \frac{\partial u}{\partial t} = icf_4. \tag{9'}$$

Consider now the divergence of tensor $uv_\alpha v_\beta$:

$$\frac{\partial(uv_\alpha v_\beta)}{\partial x_\beta} = v_\alpha \frac{\partial(uv_\beta)}{\partial x_\beta} + uv_\beta \frac{\partial v_\alpha}{\partial x_\beta}.$$

The first term on the right hand side can be written (see (9')) as

$$v_\alpha \frac{\partial(uv_\beta)}{\partial x_\beta} = icv_\alpha f_4$$

and the second term is equal to $u\dfrac{dv_\alpha}{dt}$.

Thus we have

$$u\frac{dv_\alpha}{dt} = \frac{\partial(uv_\alpha v_\beta)}{\partial x_\beta} - icv_\alpha f_4. \tag{11}$$

On the other hand

$$\frac{\partial(uv_i v_\beta)}{\partial x_\beta} = \frac{\partial(uv_i v_k)}{\partial x_k} + \frac{\partial S_i}{\partial t}, \qquad (i = 1,2,3). \tag{12}$$

From the comparison of (11) and (12) and from (8) one has the following mechanical equation of motion for energy:

$$u\frac{dv_i}{dt} = \frac{\partial}{\partial x_k}(uv_i v_k) - c^2 T_{ik}) - c^2 f_i - icv_i f_4, \qquad (i = 1,2,3). \tag{13}$$

This is a generalization of Novobátzky's equation for the case when the system interacts with enviroment, in consequence of which energy-source terms are arising. Let us notice that dividing Eq. (13) by c^2 one gets an equation of motion for density $\varrho = u/c^2$. This equation is completed with a source term and describes the mass flow accompanying the energy flow in general.

It follows from the above derivation that (13) holds within the framework of an arbitrary classical field theory. We only made use of the fact that energy-momentum tensor of the whole isolated closed system is divergenceless and that energy-momentum tensor $T_{\alpha\beta}$ is symmetric.

Dividing Eq. (13) by c^2 we get an equation which is very similar to the equation of motion for a deformable material body in continuum mechanics:

$$\varrho \frac{dv_i}{dt} = \frac{\partial}{\partial x_k}(\varrho v_i v_k - T_{ik}) - f_i - \frac{i}{c} v_i f_4, \qquad (i = 1,2,3). \tag{13'}$$

The expression in brackets plays the role of stress tensor σ_{ik}, the second term is the exterior force density coming from the enviroment and the third one is a material- (or energy-) source term. This last term is usually absent in continuum mechanics, since generally we examine the velocity field in a large distance from sources of mass (or energy). (13') has the form

$$\varrho \frac{dv_i}{dt} = \frac{\partial \sigma'_{ik}}{\partial x_k} + f'_i + q v_i, \qquad (i = 1,2,3), \tag{14}$$

where we have introduced the following notations:

$$\begin{aligned} \sigma'_{ik} &= \varrho v_i v_k - T_{ik}; & \varrho &= u/c^2, \\ f'_i &= -f_i, & q &= -\frac{i}{c} f_4. \end{aligned} \tag{15}$$

In case of a perfect fluid we know that

$$T_{ik} = p\delta_{ik} + (\varrho + \frac{p}{c^2 - v^2}) v_i v_k.$$

In gravitation field the exterior force density is $-f_i = g_i \varrho$. From Eq. (14) in classical $(v/c \ll 1)$ limit we have

$$\varrho \frac{dv_i}{dt} = \varrho g_i - \frac{\partial p}{\partial x_i}, \qquad (i = 1,2,3). \tag{16}$$

These equations are the well known Euler equations.

Let us study the stress tensor defined in (15). It can be derived from the energy-momentum tensor as follows (see (10), (4) and (5)):

$$\sigma'_{ik} = -\left(T_{ik} - \frac{T_{4i} T_{4k}}{T_{44}}\right). \tag{17}$$

To determine the stress tensor σ'_{ik} one has to solve the field equations and express the energy-momentum tensor through the field variables. The mechanical equation of motion (13) does not simplify the description of energy propagation, since the Poynting vector provides a more simple picture at kinematical level. It has a rather theoretical importance for dynamical description of energy propagation. For example the above equation of motion explains the diffraction in a medium of

inhomogeneous refractive coefficient: a deviating force arises which is proportional to the gradient of refractive coefficient.

Let us apply the equation of motion of energy for the situation mentioned above: an electromagnetic plane-wave is propagating in an isotropic and homogeneous insulating substance. Suppose that the plane-wave is propagating along the x-axis of a coordinate system which is fixed to the medium. The solutions of the Maxwell equations are

$$\begin{aligned} E_x &= 0, \qquad E_y = \frac{1}{n} f(t - \frac{nx}{c}), \qquad E_z = \frac{1}{n} g(t - \frac{nx}{c}), \\ H_x &= 0, \qquad H_y = -g(t - \frac{nx}{c}), \qquad H_z = f(t - \frac{nx}{c}), \end{aligned} \tag{18}$$

where $n = \dfrac{1}{\sqrt{\varepsilon\mu}}$ is the refractive coefficient and $v = c/n$ is the speed of propagation. f and g are arbitrary functions. The Poynting vector $\mathbf{S} = \dfrac{c}{4\pi}(\mathbf{E} \times \mathbf{H})$ aiso points towards the x-axis:

$$S_x = \frac{c}{4\pi n}(f^2 + g^2), \qquad S_y = S_z = 0. \tag{19}$$

The energy-density is

$$u = \frac{1}{4\pi}(f^2 + g^2). \tag{20}$$

It is clear from (10), (19) and (20) that the electromagnetic energy propagates along the x-axis with a velocity equal to the phase velocity of the wave. The force density on the right hand side of (13) must be equal to zero, since this velocity is constant in a homogeneous medium. Let us calculate the right hand side of (13). First we consider the energy-momentum tensor $T_{\alpha\beta}$. As is well known, there are various definitions for $T_{\alpha\beta}$. The two most common definitions are that of Abraham's and Minkowski's [2].

The Abraham's tensor $T^{(A)}_{\alpha\beta}$ is symmetric and for the above plane-wave it has the following simple form:

$$T^{(A)}_{\alpha\beta} = \begin{pmatrix} u & 0 & 0 & \frac{i}{n}u \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{i}{n}u & 0 & 0 & -u \end{pmatrix}. \tag{21}$$

This tensor is not divergenceless, which means that the electromagnetic wave exerts a force on the medium. The force density is

$$f_\alpha = -\frac{\partial T^{(A)}_{\alpha\beta}}{\partial x_\beta}. \tag{22}$$

From (21) one has

$$f_x = \frac{n^2-1}{cn}\frac{\partial u}{\partial t} = \frac{1-n^2}{n^2}\frac{\partial u}{\partial x}, \qquad f_y = f_z = 0. \tag{23}$$

The reaction-force of this acts on the electromagnetic wave (and in some sense on the energy) as an exterior force. Beyond this one the divergence of stress tensor σ'_{ik} has an effect as well. A straightforward calculation shows that these two forces equalize each other. The dielectricum is supposed to be at rest therefore f_4 is zero. Thus, in a homogeneous transparent medium there is not any force having an effect on the electromagnetic wave, according to the fact that the light propagates in a straight line with constant velocity.

If we choose the Abraham's energy-momentum tensor for electromagnetic field then the mechanical equation of motion (13) describes the wave-propagation in a homogeneous and isotropic transparent medium in accordance to the observations.

On the contrary, if we choose the non-symmetric Minkowski's expression, then the equation of motion is different from (13):

$$u\frac{dv_i}{dt} = \frac{\partial(uv_iv_k)}{\partial x_k} - ic\frac{\partial T_{4i}}{\partial t} - icv_i f_4. \tag{24}$$

Its form is different from that of Eq. (13), because the second term on the right hand side cannot be written as a divergence of a 3-dimensional tensor. Therefore one cannot define a stress tensor" the divergence of which would be a force density acting on energy (without any exterior force or source of energy). The dynamics of energy can be based on a symmetric energy-momentum tensor only. The Minkowski's energy-momentum tensor is not symmetric hence a dynamical description of the wave-propagation is impossible. The physical reason is that in the Minkowski's approach the dividing of the whole closed system into two subsystems (electromagnetic field and medium) is unlucky from the physical point of view because the momentum of the field and that of the medium are mixed [3]. A straightforward calculation shows that the right hand side of (24) is zero in accordance to the observations, but the physical meaning of the complete equation and the interpretation of several terms are insufficient.

References

1. K.F. Novobátzky, Magyar Fizikai Folyóirat, *1*, 5, 1953 (in Hungarian).
2. M.Abraham, Rend. Circ. Matem. Palermo, *28*, 1, 1909; *30*, 5, 1910;
   H. Minkowski, Nachr. Ges. Wiss. Göttingen, *53*, 1908;
   G. Marx and G. Györgyi, Acta Phys. Hung., *3*, 213, 1954;
   Annalen d.Phys. 6. Folge, *16*, 241, 1955.
3. G. Györgyi, Magyar Fizikai Folyóirat, *2*, 255, 1954 (in Hungarian)
   G. Marx, G.Györgyi, Annalen d. Phys. 6. Folge, *16*, 241, 1955.

Acta Physica Hungarica 64 (1-3), pp. 27-32 (1988)

KINEMATICS AND THE SPECIAL THEORY OF RELATIVITY*

Z. BAY and J.A. WHITE

The American University, Washington D.C. 20016, USA

(Received 8 January 1987)

Consistent use of the principles of classical kinematics leads to interesting theorems, connecting the coefficients of linear transformations between systems of inertia. Differential equations expressing such connections diminish the number of independent postulates required for the establishment of the Lorentz group.

For the measurement of length in moving systems an old classical principle of kinematics is used which states that length has to be measured as the difference of *simultaneous* coordinates. This old rule of kinematics should also be applied in a theory of special relativity (STR) where simultaneity is different in systems in relative motion to each other.

Thus, H. Poincaré [1] and A. Einstein [2] could show in the early history of relativity that the Fitzgerald–Lorentz contraction of moving bodies can easily be explained on the basis of transformation equations. If, in general (in one dimensional space) the transformation is [3]

$$\begin{aligned} t' &= a(v)t + b(v)x \ , \\ x' &= \frac{x - vt}{r(v)} \end{aligned} \tag{1}$$

and the inverse transformation is

$$\begin{aligned} t &= a'(v')t' + b(v')x' \ , \\ x &= \frac{x' - v't'}{r'(v')} \ , \end{aligned} \tag{1a}$$

where

$$v' = -\frac{v}{ar} \ , \quad a' = \frac{1}{f(v)} \ , \quad b' = -\frac{rb}{f} \ , \quad r' = \frac{f}{ar} \ , \quad f(v) = a(v) + vb(v)$$

and a rod of length, $L' = x'_2 - y'_1$ is at rest in the primed system B, then in the unprimed system A it has a length for the same values of t, which is

*Dedicated to Prof. G. Marx on his 60th birthday

$$x_2 - x_1 = L$$

and

$$L = r(x_2' - x_1') = rL \ . \tag{2}$$

Thus, the coefficient, $r(v)$, of the transformation (1) expresses the "contraction" and therefore it should be called the "contraction function" of (1). In the same way $r'(v')$ is the contraction function of the inverse transformation (1a).

In a similar way, a time interval, T', measured at a fixed point in B, as related to the time interval, T, for the point, moving with v in A is

$$T' = (a + vb)T = f(v)T \ , \tag{3}$$

thus, $f(v)$ expresses the "time dilatation function". Similarly, $f' = a' + v'b' = \frac{1}{a}$ is the time dilatation function of the inverse transformation (1a).

It is interesting to note that in the very extended literature (references listed in [4–7]) of STR apparently no attempt was taken to follow up further the consequences of kinematics in order to establish general connections among the coefficients of the transformation leading to STR. In particular, the functions $f(v)$, $r(v)$, $b(v)$ (and also the coefficients, $g(v)$ connecting the coordinates perpendicular to v in the three dimensional case) have been considered to be mathematically independent quantities for the determination of which several physical arguments (such as group character, reciprocity, invariance of the speed of light, etc.) should be used.

We will show below that consistent use of the requirements of kinematics restricts the freedom of choice of the afore-said coefficients, in that it establishes differential equations, valid for changes of relative velocities which connect the coefficients determining the transformations.

Before deriving those differential equations we investigate the physical meaning of the $b(v)$ and $b'(v')$ coefficients of Eqs (1) and (1a).

We take two events, (t_2, x_2) and (t_1, x_1) in A which are simultaneous in A, thus $t_2 = t_1$.

The corresponding events, (t_2', x_2'), (t_1', x_1') are not simultaneous in B. Their time difference is

$$t_2' - t_1' = b(v)(x_2 - x_1) = b(v)L \ ; \tag{4}$$

thus the

$$b(v) = \left(\frac{\partial t'}{\partial x}\right)^{t=\text{const.}} \tag{5}$$

function determines the time distribution in B for events simultaneous in A.

Therefore we may call the $b(v)$ function the "time distribution" function for the transformed system.

In a similar way: to simultaneity in system B corresponds the time distribution

$$t_2 - t_1 = b'(v')(x_2' - x_1') = b'(v')L' \tag{6}$$

in A. Thus,

$$b'(v') = \left(\frac{\partial t}{\partial x'}\right)^{t'=\text{const.}} \tag{7}$$

is the coefficient of the time distribution for system A for events simultaneous in B.

For inverse transformations

$$\begin{aligned} b' &= -rb/f \quad = -b/D \,, \\ b &= -r'b'/f' = b'/D' \,, \end{aligned} \tag{8}$$

where $D = f/r$ and $D' = 1/D$ are the determinants of the transformations.

We also have

$$b(v) = \left(\frac{\partial t'}{\partial x}\right)^{t} = -D\left(\frac{\partial t}{\partial x'}\right)^{t'} = -Db' \,; \tag{9}$$

thus, Eqs (8) and (9) result in

$$-b/b' = D = 1/D' \,. \tag{10}$$

The negative ratio of the inverse (or reciprocal) time distribution functions is equal to the determinant of the transformation.

The equations (8), (9) and (10) can also be interpreted by saying that the $b(v)$ and $b'(v')$ coefficients express the *dissynchrony* or *difference of synchronies* between two systems of inertia, *each in their respective measuring systems.*

The application of kinematics for changes of relative velocity between two systems of inertia

We take two simultaneous events in B, where B is moving with v with respect to A.

Suppose two signals are initiated at those events, for which

$$x'_2 - x'_1 = L' \qquad \text{and} \qquad t'_2 = t'_1 .$$

To those signals correspond two events in A : (t_2, x_2) and (t_1, x_1) which are not simultaneous in A. Therefore $x_2 - x_1 = \Delta x$ is not a length in A, since $t_2 - t_1 = \Delta t = -(b/a)\Delta x \neq 0$.

To obtain the length, L in A, corresponding to L' (at rest in B), we have to reduce Δx to simultaneity in A. Thus,

$$L = \Delta x - v\Delta t \tag{11}$$

$$= \Delta x \left[1 + \frac{b}{a}v\right] = \Delta x \frac{f}{a} \tag{12}$$

and

$$\Delta t = -(b/f)L \ . \tag{13}$$

Now, suppose that besides emitting the signals at $t'_2 = t'_1$ the same velocity increment (du' as measured in B, or dv as measured in A) is also imparted simultaneously in B to the endpoints x'_2 and x'_1.

Classical kinematics requires that the distance between those endpoints after acceleration does not change as measured in system B, because the difference of simultaneous coordinates remains L' in B. But a careful consideration should be given to the fact that those points are now not at rest in B; they are at rest in a system (we call it B_2) which moves with the velocity $v + dv$ in A.

We know that to two simultaneous events in A, separated by L, a time difference of $b(v)L$ corresponds in B. The time difference in B_2 differs from that in B by the increment, $d(bL)$ of the bL product for the velocity increment dv. That increment is proportional to dv and thus the difference of time distributions between B and B_2 is also proportional to du', where du' is of the same order of magnitude as dv when v/c is not close to unity.

Therefore, similar to Eq. (11), the relationship of L'' at rest in B_2, to L' at rest in B, can be written as

$$L'' = L' - du'd(bL) \ ,$$

and since du' and $d(bL)$ are both proportional to dv, L'' can differ from L' at the most to second order of the velocity increment. This result depends only on the condition that the coefficients of the transformation are continuous and differentiable functions of v. This important finding can be formulated as

Theorem I. If an equal velocity increment is simultaneously imparted to the points of a system of inertia, then distances in that system do not change to first order in that velocity increment.

While there is no first order change of distances in B, when the velocity increment is imparted simultaneously (according to synchrony in B) to points of the system, there will be first order changes of those distances, as measured in A. This can be seen from Eq. (11) where we obtain now for $v + dv$

$$L(v + dv) = \Delta x - (v + dv)\Delta t$$

and therefore

$$dL = -dv\Delta t \ . \tag{14}$$

(It is easy to visualize how the change, dL, is brought about: If the increase of velocity appears earlier at point x_1 and later by Δt at point x_2, as seen in A, then the distance between the two points diminishes by $dv\Delta t$ as measured in A.)

Eqs (13) and (14) give

$$dL = (b/f)Ldv$$

and

$$dL/L = (b/f)dv \ . \tag{15}$$

If $dv \to 0$, $L'' \to L'$ (Theorem I) and

$$dL/L = dr/r$$

and therefore

$$\frac{dr}{dv} = \frac{r}{f}b \ . \tag{16}$$

In the derivation of Eq. (16) we used classical kinematics and linearity of transformations. Therefore a similar differential equation should be valid for the inverse transformation. Thus,

$$\frac{dr'}{dv'} = \frac{r'}{f'}b' \tag{17}$$

and due to Eqs (6),(7),(8)

$$dr/dv = -b' \ , \tag{18}$$

$$dr'/dv' = -b \ . \tag{19}$$

These findings are summarized in

Theorem II. An increment of the relative velocity between two systems of inertia results in each system in an increment of the contraction function which is proportional to the difference of synchronies between the two systems.

Eqs (16) and (17) express Theorems I and II in mathematical form.

They introduce two conditions, required by the basic principles of kinematics, into relativity theory, valid for the coefficients of a homogeneous linear transformation.

It was known already since the early investigations of linear transformations, pertinent to relativity theory, that three conditions, applied to the aforesaid coefficients, have to be invoked to result in transformation of the type of the *Lorentz group*. In particular, conditions listed by W. Ignatowsky and P. Frank and H. Rothe [8] are:

a) the transformations must form a one-parameter homogeneous linear group;

b) $-v' = -v \ ; \quad r' = r \ .$

This already suffices to lead to

$$f(v) = r(v) = \sqrt{1 - \alpha v^2} \ , \quad b(v) = \frac{-\alpha v}{\sqrt{1 - \alpha v^2}} \ .$$

The constant of integration, α, can be inferred by physical arguments to be the reciprocal square of the invariant speed of light, c.

Now it can be shown that, owing to Eqs (16) and (17), a single condition, added to the general requirements, is sufficient to lead to the *Lorentz group*. This condition can either be one of the a), b), c), listed above, or another one (eg. the invariance of c, etc.).

The consequence of *Theorem I* is that if to the points of a system of inertia a continuous and equal acceleration is applied, always simultaneously according to the instantaneous synchrony of the system, distances in that system do not change. Distances as measured in other systems of inertia change according to *Theorem II*. This may give the answer to puzzling problems raised by E.Dewan and M.Beran, "On Stress Effects Due to Relativistic Contraction" [9].

Observance of the kinematic requirements can also be used to eliminate the supposition of a hypothetical rest system A in the empirical foundations of STR.

The kinematic equations are also useful in the criticism of some of the recommended "versions" of STR: for example of those based on the unisotropic definition of synchrony, advocated by Reichenbach [10] and pursued by the conventionalists, [11], [12]. They also are useful in the criticism of the so-called *RC–CR–ether* theories [13], involving physical rod contraction and clock retardation effects dependent on the velocity relative to a standing ether.

The detailed discussion of these questions will be published subsequently.

Acknowledgements

We are indebted to Drs. M. Danos of the National Bureau of Standards, Washington, D. C., and P. Farago of the University of Edinburgh, United Kingdom, for useful discussions.

References

1. H. Poincaré, C.R. Acad.Sci., *140*, 1504, 1905.
2. A. Einstein, Ann. Phys., *17*, 891, 1905.
3. Z. Bay and J.A. White, Acta Phys. Hung., *51*, 273, 1981.
4. W. Pauli, Theory of Relativity, Pergamon Press, New York 1958.
5. V. Berzi and V. Gorini, Journ. Math. Phys., *10*, 1518, 1969.
6. E. Breitenberger, Nuovo Cimento, *1B*, 1, 1971.
7. A.I. Miller, Einstein's Special Theory of Relativity, Addison-Wesley Publ. Co. 1981.
8. Ref.[4] p.11.
9. E. Dewan and M. Beran, Am. Journ.Phys., *27*, 517, 1959.
10. H. Reichenbach, The Philosophy of Space and Time, New York, Dover 1958.
11. J.A. Winnie, Philos. of Sci., *37*, 81 and 223, 1970.
12. C. Giannoni, Philos. of Sci., *45*, 17, 1978.
13. H. Erlichson, Am. Journ. Phys., *41*, 1068, 1978.

Acta Physica Hungarica 64 (1-3), pp. 33–48 (1988)

RELAXATION PROCESSES IN CHAOTIC STATES OF ONE DIMENSIONAL MAPS*

G. GYÖRGYI**

Institute for Theoretical Physics, Roland Eötvös University
1088 Budapest, Hungary

and

P. SZÉPFALUSY

Institute for Solid State Physics, Roland Eötvös University
1088 Budapest, Hungary
and Central Research Institute for Physics
1525 Budapest, Hungary

(Received 8 January 1987)

Relaxation of probability distributions towards their stationary state is studied in one dimensional fully developed chaotic maps. Exponential relaxation is described by solving the eigenvalue problem of the Frobenius–Perron operator for families of maps. This happens via perturbation theory, for which a straightforward formalism is presented. Critical maps with power law decay are also considered.

1. Introduction

In recent years there has been a rising interest in low dimensional nonlinear systems. Inspite of the simple deterministic laws of motion they obey, they may exhibit strikingly complex, chaotic, behavior. Chaotic phenomena have been detected in a wide variety of experimental situations ranging from electronic circuits and Josephson junctions to optical and hydrodynamical instabilities to autocatalytic chemical reactions [1-4].

In dissipative systems the trajectories are driven towards a limiting set, called the attractor. Attractors may be regular, where the system reaches equilibrium or is (quasi) periodic in time. Other attractors are explored in an erratic way, they are the limiting sets of chaotic trajectories, called the strange attractors. Chaotic motion is unpredictable in the sense that uncertainty in the trajectory at one moment yields exponentially increasing error in predictions about the motion. In other words, there is a sensitivity to initial conditions, nearby trajectories diverge exponentially

*Dedicated to Prof. G. Marx on his 60th birthday.

**Present address: Physics Department, Stevens Institute of Technology, Hoboken, NJ, 07030, USA

fast. The rate of the divergence, the Lyapunov exponent, is an important indicator of the degree of chaoticity [5,6]. Unpredictability also implies a loss of memory: in chaotic systems correlations decay with time.

Unlike regular trajectories in linear systems, chaotic curves typically cannot be expressed by closed formulae. As a matter of fact, the sensitivity to initial condition would render such formulae useless anyhow. Instead, statistical means are to be used. The long time behavior is basically described by the probability of visiting different regions in the phase space, that is, by the stationary probability density. Chaotic systems are also ergodic. This implies that time averages can be calculated as median values over the stationary probability density. The knowledge of the probability density, therefore, is essential. Probability densities themselves exhibit relaxation. No matter what the initial probability density is like, it usually approaches the stationary probability density. Whereas relaxation can be observed in regular systems as they relax to equilibrium, in the case of chaotic systems the phenomenon is richer. Namely, even if the initial density is concentrated on the attractor there is a relaxation process characterizing chaoticity. This process has a relaxation rate, which is by and large independent of the starting density and thus characteristic for the system itself. It also governs the decay of typical correlation functions. In this paper we study in detail this relaxation in the simplest chaotic systems, in one dimensional maps. Such maps correspond to flows in a three dimensional phase space in the limit of large dissipation [7]. They also emerge out of higher dimensional systems via slaving.

The paper is organized as follows. In Section 2 we review basic properties of chaotic maps and stress the significance of the Frobenius –Perron operator. Section 3 is devoted to the properties of the tent map. Next, in Section 4, a perturbation theory for the eigenvalue problem of the integral version of the Frobenius–Perron operator is described. This is applied in Section 5 for the calculation of probabilistic measures, relaxation rates and amplitudes. Critical maps with power law relaxation are discussed in the last Section.

2. Fully developed chaotic maps

We focus our attention on one dimensional maps of an interval [0,1]

$$x_{n+1} = f(x_n). \tag{2.1}$$

Here $f(0) = f(1) = 0$ and $f(x)$ has a single maximum $\hat{x}$, where $f(\hat{x}) = 1$. Further specifications understood in the definition of $f(x)$ are that $f(x)$ is smooth, except possibly at the maximum, and it is monotonically increasing and decreasing for $x < \hat{x}$ and $x > \hat{x}$, respectively. That is, every point has two preimages, except $x = 1$, which has only one, $\hat{x}$. We assume that the map does not have stable periodic orbits, but exhibits chaotic behavior. Chaos is fully developed in the sense that the iterations starting from almost all initial values fill in the whole interval

(0,1). The trajectory is ergodic, which means that for almost all initial values x_1 the time average of a function $A(x_n)$ is

$$\lim_{N\to\infty}(1/N)\sum_{n=1}^{N}A(x_n)=\int P(x)A(x)dx\equiv<A(x)>, \tag{2.2}$$

where $P(x)$ is the stationary probability density function associated with the iteration (2.1). A condition of ergodicity is given in [8]. The integration goes from 0 to 1, unless otherwise indicated. Because of the stationarity, it should not matter whether the initial condition in (2.2) is x_1 or $x_2 = f(x_1)$. Hence

$$<A(x)>=<A(f(x))> \tag{2.3}$$

follows for arbitrary $A(x)$. This is the condition of stationarity, or invariance, for the probability density $P(x)$. It is easy to see that the invariance condition is equivalent to

$$P(x)=\int\delta(x-f(x'))P(x')dx'\equiv\hat{F}P(x), \tag{2.4}$$

where $\delta(x)$ is the Dirac delta function, and $\hat{F}$ is called the Frobenius–Perron operator. Sometimes it is useful to write the action of $\hat{F}$ on an arbitrary function $A(x)$ in the form

$$\hat{F}A(x)=A(y)/|f'(y)|+A(z)/|f'(z)|, \tag{2.5}$$

where y and z are the preimages of x, and where elementary properties of the delta function have been used. The Frobenius–Perron operator determines the evolution of an initial probability density $P_0(x)$ as

$$P_n(x)=\hat{F}P_{n-1}(x)\equiv\hat{F}^nP_0(x). \tag{2.6}$$

In order to follow the iteration of the density we consider the spectral properties of $\hat{F}$. A function $R(x)$ is its (right) eigenfunction with eigenvalue m if

$$\hat{F}R(x)=mR(x). \tag{2.7}$$

No eigenvalue $m > 1$ may exist, if it did, the iterated density might grow without bounds. Furthermore, $P(x)$ is the unique eigenfunction with eigenvalue $m = 1$. All eigenfunctions with $m < 1$ satisfy $\int R(x)dx = 0$, which follows from integrating (2.7). Let us consider an initial density $P_0(x)$ which can be expanded in terms of eigenfunctions $R^k(x)$ as

$$P_0(x)=\sum_{k=0}^{\infty}a^kR^k(x). \tag{2.8}$$

Take $R^0(x) = P(x)$, and let m^1 be the eigenvalue with the second largest absolute value, belonging to $R^1(x)$. Since $\int R^k(x)dx = \delta_{k0}$, where δ_{k0} is the Kronecker

symbol, we have $a^0 = \int P_0(x)dx = 1$. After many iterations by $\hat{F}$, the dominant terms in $P_n(x)$ are

$$P_n(x) \sim P(x) + a^1(m^1)^{-n}R^1(x). \tag{2.9}$$

This shows the importance of the second largest eigenvalue: it governs relaxation towards the stationary density.

An alternative form for the invariance condition can be obtained by integrating both sides of $P(x) = \hat{F}P(x)$. This is conveniently written in terms of the invariant measure $\nu(x)$, defined as $\nu(x) \equiv \int_0^x P(x')dx'$. Using Eq. (2.5) and the conditions $\nu(0) = 0$, $\nu(1) = 1$, we obtain

$$\nu(x) = 1 + \hat{M}\nu(x). \tag{2.10}$$

Here the operator $\hat{M}$ is given by

$$\hat{M}\eta(x) \equiv \eta(f_l^{-1}(x)) - \eta(f_u^{-1}(x)), \tag{2.11}$$

and $f_l^{-1} < \hat{x}$ and $f_u^{-1} > \hat{x}$ denote the lower and upper branches of the inverse of f, respectively. Since $\hat{M}$ is an integrated version of $\hat{F}$ the spectrum of these operators is the same. Eq. (2.10) shall be our basic relation for calculating stationary properties of the iteration (2.1).

The speed of separation of two nearby points x and $x + dx$ is determined by the derivative of the mapping $f'(x)$. On the average, the divergence rate is the Lyapunov exponent

$$\lambda = \lim_{N\to\infty}(1/N)\sum_{n=1}^{N}\ln|f'(x_n)| = <\ln|f'(x)|> . \tag{2.12}$$

Ergodicity implies that the Lyapunov exponent λ does not depend on the initial condition for almost all x_1 in the interval.

Another characteristic of the map is the correlation function

$$C(n) = < xf^n(x) > - < x >^2, \tag{2.13}$$

where $f^n(x)$ is the n-th iterate of x, $f^0(x) = x$. For a chaotic map the correlation function decays, $C(n) \to 0$ if $n \to \infty$. The correlation function of arbitrary functions of x as of $A(x)$ and $B(x)$, is given by

$$C_{AB}(n) = < A(x)B(f^n(x)) > - < A(x) >< B(x) > . \tag{2.14}$$

Again $C_{AB}(n) \to 0$ is expected for large n. The way $C_{AB}(n)$ decays can conveniently be approached by rewriting its first term as

$$\begin{aligned} < A(x)B(f^n(x)) > &\equiv \int B(f^n(x))A(x)P(x)dx = \\ &= \int B(x')\delta(x' - f^n(x))A(x)P(x)dxdx' = \\ &= \int B(x')\hat{F}^n AP(x')dx', \end{aligned} \tag{2.15}$$

where the definition (2.4) of the Frobenius–Perron operator has been used. $\hat{F}^n AP$ is the n-th iterate of the function $A(x)P(x)$, which for large n has the leading terms

$$\hat{F}^n AP(x) \sim a^0 P(x) + a^1 (m^1)^{-n} R^1(x), \tag{2.16}$$

where $a^0 = \int A(x)P(x)dx \equiv < A(x) >$, and $R^1(x)$ is the eigenfunction with the second largest eigenvalue appearing in the expansion of $A(x)P(x)$. Therefore for large n we have

$$C_{AB}(n) \sim a^1 (m^1)^{-n} \int B(x')R^1(x')dx', \tag{2.17}$$

thus the second largest eigenvalue is the factor of correlation decay. (Here we disregarded the nontypical case when the integral in (2.17) or a^1 is accidentally zero.)

In summarizing, the second largest eigenvalue of the Frobenius–Perron operator, or, equivalently, that of $\hat{M}$, determines the relaxation of both the probability density and the correlation function.

In addition to the eigenfunctions a set of null functions emerges as well. Null functions may appear if the operator in question is not self-adjoint: $\hat{F}$ and $\hat{M}$ are not. We shall consider here only $\hat{M}$. A null function of order q, $\zeta^q(x)$ is defined by the property that

$$\hat{M}^{q+1}\zeta^q(x) = 0 \qquad \text{and} \qquad \hat{M}^q\zeta^q(x) \neq 0, \tag{2.18}$$

i. e. it is transformed into zero after exactly $q+1$ applications by $\hat{M}$. Note that a null function $\zeta^0(x)$ is an eigenfunction with eigenvalue zero. Null functions of order q form the null space Z^q. We also require that a linear combination of null functions in Z^q should belong to Z^q again, otherwise lower order null functions could be added to a $\zeta^q(x)$ causing an ambiguous situation. In the following we give a method for constructing null functions in the case of symmetric maps $f(x) = f(1-x)$. For such maps the branches of the inverse map are related by $f_l^{-1}(x) = 1 - f_u^{-1}(x)$. From (2.11) it is clear that all symmetric functions $s(x) = s(1-x)$ belong to the null space Z^0. Next, we construct the reverse operation of $\hat{M}$. It follows from (2.11) that the solution of

$$\hat{M}\eta(x) = \xi(x) \tag{2.19}$$

is

$$\eta(x) = 1/2 \text{ sgn } (1/2 - x)\,\xi(f(x)) + \text{ [symmetric function]}. \tag{2.20}$$

For the sake of uniqueness we set the symmetric function to zero, which yields the reverse operation

$$\hat{N}\zeta(x) = 1/2 \text{ sign } (1/2 - x)\,\zeta(f(x)). \tag{2.21}$$

It satisfies

$$\hat{M}\hat{N} = \hat{1} \tag{2.22}$$

where $\hat{1}$ is the identity. This enables us to reconstruct the hierarchy of null functions starting out from a suitably chosen base $\{\zeta^{0,j}(x)\}_j$ of Z^0, the null space of symmetric

functions. An analytic base is formed, for instance, by the functions $\sin[(2j-1)\pi x]$. Null functions of higher order can then be obtained as

$$\zeta^{q,j}(x) = \hat{N}^q \zeta^{0,j}(x) = 1/2^j \zeta^{0,j}(f^j(x)) \prod_{j=0}^{q-1} \text{sign}\,(1/2 - f^j(x)) \tag{2.23}$$

and they satisfy

$$\hat{M}\zeta^{q,j}(x) = \zeta^{q-1,j}(x). \tag{2.24}$$

These null functions are, in general, nonanalytic functions. As we shall see, however, the base $\{\zeta^{0,j}(x)\}_j$ can sometimes be chosen so that the resulting null functions $\zeta^{q,j}(x)$ are analytic.

The above procedure yields null spaces Z^q for all indices q. In such a case eigenfunctions can be constructed as infinite series of null functions. Such a construction is analogous to the way coherent states build up an eigenstate in a quantum oscillator. There, however, the operator is self-adjoint, thus the Hilbert space can be spanned out by either the eigenstates or the null functions, i. e. coherent states. In our case, the eigenfunctions do not form a complete basis; to complete the basis null functions should also be used at the same time. Those null functions, however, do not influence the asymptotics of the relaxation.

3. The linear map and its conjugates

As a simple example for fully developed chaotic one dimensional maps, consider the linear map

$$f_L(x) = 1 - |1 - 2x|, \tag{3.1}$$

which is sometimes called the tent map. From Eq. (2.5) it is clear that $P_L(x) \equiv 1$ is the stationary density, $\hat{F}_L P_L(x) = P_L(x)$, that is, the invariant measure is $\nu_L(x) = x$. Thus the Lyapunov exponent is $\lambda_L = \ln 2$, as calculated from (2.12).

The spectrum of the operator $\hat{M}_L$ can easily be given, provided we restrict ourselves to analytic functions [9,10]. Denoting the right eigenfunctions of $\hat{M}_L$ by ϕ_L^k and the eigenvalues by m_L^k we have

$$m_L^k = 1/4^k \qquad \text{and} \qquad \phi_L^k(x) = B_{2k+1}(x/2), \tag{3.2}$$

where $B_p(x)$ is the Bernoulli polynomial of order p. This can easily be seen by using (2.11) together with (3.1), and the identities

$$B_n(px) = p^{n-1} \sum_{j=0}^{p-1} B_n(x + j/p) \quad \text{and} \quad B_n(1-x) = (-1)^n B_n(x), \tag{3.3a-b}$$

with $p = 2$.

The analytic null functions can also be constructed [10]. Starting out from the zeroth order null functions $\zeta_L^{0,j}(x) = \sin[(2j-1)\pi x]$, $j \geq 1$, we get

$$\zeta_L^{q,j}(x) = 2^{-q} \sin[2^q(2j-1)\pi x]. \tag{3.4}$$

Note that the construction (2.23) also leads to these functions. It is worth mentioning that there is a family of polynomial null functions of index $q = 0$ as

$$b^0(x) \equiv \text{const.}, \tag{3.5a}$$

$$b^j(x) \equiv B_{2j}(x) - B_{2j}, \qquad j \geq 1, \tag{3.5b}$$

where the B_n's are the Bernoulli numbers. They are symmetric polynomials, $b^j(x) = b^j(1-x)$.

We shall also need a projection technique onto the eigen- and null functions. It can be shown that the operator

$$\bar{\phi}_L^j(x) \equiv \{2^{2j+1}/(2j+1)!\}\{d^{2j+1}/dx^{2j+1}\} \tag{3.6}$$

satisfies

$$\int \bar{\phi}_L^j(x)\phi_L^{j'}(x)dx = \delta_{jj'}, \tag{3.7a}$$

$$\int \bar{\phi}_L^j(x)\, b^{j'}(x)dx = 0. \tag{3.7b}$$

We also have the orthogonality property

$$\int \bar{b}^j(x)\, b^{j'}(x)dx = \delta_{jj'}, \tag{3.8}$$

where $\bar{b}^j(x)$ is defined by

$$\bar{b}^j(x) \equiv \{1/(2j)!\}\{d^{2j}/dx^{2j}\}. \tag{3.9}$$

In conclusion, an initial measure $\nu_0(x)$ can be expanded in terms of the eigen- and null functions using the projections above. As a rule of thumb, one first determines the coefficients a^k of the eigenfunctions ϕ_L^k, then subtracts the resulting series from the original function as $\nu_0(x) - \sum a^k \phi_L^k(x)$, and finally determines the projections of that difference onto the null functions.

The second largest eigenvalue turned out to be $m^1 = 1/4$. This governs the relaxation of correlation functions and probability densities, provided we restrict ourselves to analytic functions. In fact, it can be shown that functions which become analytic after a finite number of steps can be included as well.

Maps related by smooth coordinate transformation [11] to the tent map possess the same spectrum. Therefore, the relaxation factor is again 1/4. Suppose that

$$f(x) = u(f_L(u^{-1}(x))), \tag{3.10}$$

where $u(x)$ is a smooth function. Then the invariant measure of the map $f(x)$ is

$$\nu(x) = \nu_L(u^{-1}(x)) \equiv u^{-1}(x), \tag{3.11}$$

and all other eigen- and null functions of $\hat{M}$ are related in the same way to those of the tent map. A well-known example is the quadratic map [12]

$$f_Q(x) \equiv 4x(1-x) = 1 - (1-2x)^2, \tag{3.12}$$

related to the tent map through

$$u(x) = \sin^2(\pi x/2). \tag{3.13}$$

Consequently, the invariant measure is $\nu_Q(x) = (2/\pi)\arcsin\sqrt{x}$, and the probability density $P_Q(x) = (x(1-x))^{-1/2}/\pi$.

Note that in the case of the tent and quadratic maps, both the map and probability density are symmetric functions. It can be shown for such double symmetric maps, that the correlation function decays immediately as [11]

$$C(n) = C(0)\delta_{n0}. \tag{3.14}$$

Nevertheless, this is a speciality of the double symmetry, and the correlation $C_{AB}(n)$ between general analytic functions $A(x)$ and $B(x)$ decays exponentially with the factor 1/4.

4. Perturbation theory

Apart from the linear map and its conjugate family the eigenvalue problem of fully developed chaotic maps is, in general, not solved. For maps close to another map with known properties, a perturbative method seems to be a promising approach. Let us suppose, that the non-degenerate positive eigenvalues m_0^j and the associated eigenfunctions ϕ_0^j of the operator $\hat{M}_0$ are known. This operator corresponds to the fully developed chaotic map $f_0(x)$. Consider then the perturbed map

$$f(\varepsilon, x) = f_0(x) + \varepsilon f_1(x), \tag{4.1}$$

which is a fully developed chaotic map again. We assume that f_0 and f_1 are analytic functions apart from isolated points. Then $\hat{M}(\varepsilon)$ can be expanded in terms of ε as

$$\hat{M}(\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i \hat{M}_i. \tag{4.2}$$

The operators $\hat{M}_i$ can be expressed in terms of $f_0(x)$ and $f_1(x)$. For symmetric maps, $f(\varepsilon, x) = f(\varepsilon, 1-x)$, we give them explicitly in the Appendix. A further assumption is that the eigenvalues $m^j(\varepsilon)$ and eigenfunctions $\phi^j(\varepsilon)$ are analytic in ε,

$$m^j(\varepsilon) = \sum_{i=0}^{\infty} m_i^j \varepsilon^i, \tag{4.3a}$$

$$\phi^j(\varepsilon, x) = \sum_{i=0}^{\infty} \phi_i^j(x) \varepsilon^i. \tag{4.3b}$$

Then one can go along the usual lines to set up a perturbation theory for calculating the coefficients in the above series.

Let us expand both sides of

$$m(\varepsilon)\phi(\varepsilon, x) = \hat{M}(\varepsilon)\phi(\varepsilon, x), \tag{4.4}$$

where the superscript has been omitted. In the i-th order we obtain

$$\begin{aligned}(m_0 - \hat{M}_0)\phi_i(x) =& (\hat{M}_i - m_i)\phi_0(x) + (\hat{M}_{i-1} - m_{i-1})\phi_1(x) + \\ &+ \cdots + (\hat{M}_1 - m_1)\phi_{i-1}(x).\end{aligned} \tag{4.5}$$

If $m_1, \ldots, m_{i-1}$ and $\phi_1(x), \ldots, \phi_{i-1}(x)$ are known, the number m_i can be determined from the condition that the left hand side does not contain any terms proportional to the zeroth order eigenfunction $\phi_0(x)$, and neither does the right hand side. This prescription yields (we omit the x dependence)

$$\begin{aligned}m_i\phi_0 = \prod_0 \{\hat{M}_i\phi_0 + \sum\nolimits'_{\{n_j\}} (\hat{M}_{n_1} - m_{n_1})(m_0 - \hat{M}_0)^{-1}(\hat{M}_{n_2} - m_{n_2}) \ldots \\ \ldots (m_0 - \hat{M}_0)^{-1}(\hat{M}_{n_k} - m_{n_k})\phi_0\},\end{aligned} \tag{4.6}$$

where $\prod_0$ denotes the projection onto the eigenfunction ϕ_0 and the summation $\sum'$ goes over $n_1, n_2, \ldots, n_k$ with the constraints $\sum_{j=1}^{k} n_j = i$, $1 \leq n_j \leq i$, with all possible k's. Note that the insertions $(m_0 - \hat{M}_0)^{-1}$ do not cause singularities here. This is due to the fact that the numbers $m_1, m_2 \ldots, m_{i-1}$ have been chosen previously so that the dangerous eigenfunction disappears from all those functions $(m_0 - \hat{M}_0)^{-1}$ is acting upon. Having found m_i, we can also calculate $\phi_i(x)$ as

$$\phi_i = \sum\nolimits'_{\{n_j\}} (m_0 - \hat{M}_0)^{-1}(\hat{M}_{n_1} - m_{n_1})(m_0 - \hat{M}_0)^{-1} \ldots (m_0 - \hat{M}_0)^{-1}(\hat{M}_{n_k} - m_{n_k})\phi_0, \tag{4.7}$$

with $\sum_{j=1}^{k} n_j = i$, $n_j \geq 1$. Note that Eq. (4.5) leaves the projection onto $\phi_0, \prod_0\{\phi_i\}$, undetermined. This ambiguity is eliminated by setting it to zero in Eq. (4.7).

In the previous Section a method was given for constructing null functions for symmetric maps. Since any fully developed chaotic map is smoothly conjugate to a symmetric map, we can assume that $f(\varepsilon, x)$ is symmetric as well, thus by using $\hat{M}(\varepsilon)$ null functions can be constructed explicitly. Therefore, while a perturbation theory for null functions can also be set up, there is no immediate need for it.

Eq. (4.7) for the eigenfunction with eigenvalue unity gives the corrections for the perturbed invariant measure. The density can be obtained by derivation, and also the corrections to the Lyapunov exponent can be calculated straightforwardly. The second largest eigenvalue is the relaxation factor of an initial measure or, equivalently, of a density, and it also governs decay of correlation. The amplitude of relaxation can be calculated from the corresponding eigenfunction. In the following Section we show examples of maps for which these quantities are calculated explicitly.

5. Relaxation in the bilinear and biquadratic maps

In this Section we apply our results to two families of maps,

$$f_{BL}(\varepsilon, x) = 1 - (1-\varepsilon)|1-2x| - \varepsilon(1-2x)^2, \tag{5.1a}$$

$$f_{BQ}(\varepsilon, x) = 1 - (1-\varepsilon)(1-2x)^2 - \varepsilon(1-2x)^4, \tag{5.1b}$$

called bilinear and biquadratic, respectively. By means of perturbation theory their stationary probability density functions have been calculated [13] which has been put later on a rigorous basis [14].

We are interested in the spectrum and the eigenfunctions of the operators $\hat{M}_{BL}(\varepsilon)$ and $\hat{M}_{BQ}(\varepsilon)$. The null functions, as described in Section 2, can be calculated directly, so we will not deal with them here. In order to avoid too many indices we introduce the notation

$$m_{BL}(\varepsilon) = n(\varepsilon), \qquad m_{BQ}(\varepsilon) = p(\varepsilon), \tag{5.2a}$$

$$\phi_{BL}(\varepsilon, x) = \chi(\varepsilon, x), \qquad \phi_{BQ}(\varepsilon, x) = \psi(\varepsilon, x). \tag{5.2b}$$

Consider first the bilinear map. Seeking the perturbed eigenvalues $n^j(\varepsilon) = \sum_{i=0}^{\infty} n_i^j \varepsilon^i$ and eigenfunctions $\chi^j(\varepsilon, x) = \sum_{i=0}^{\infty} \chi_i^j(x)\varepsilon^i$ it turns out that the corrections in all orders of ε are polynomials. This follows from the fact that $(\hat{M}_{BL})_i \chi_0^j(x)$ is a finite sum of the polynomials $\chi_0^j(x) \equiv \phi_L^j(x)$ and $b^k(x)$, which have been described in Section 3. Furthermore, it can be shown that $(\hat{M}_{BL})_i b^k(x) = 0$ for all i and k. We calculated the first correction to the eigenvalues using Eq. (4.6) as

$$n^j(\varepsilon) = 1/4^j(1 - \varepsilon j(2j+1) + \ldots). \tag{5.3}$$

The largest eigenvalue $n^0(\varepsilon)$ is still one, as it should be. The stationary measure reads up to third order as

$$\begin{aligned} \nu_{BL}(\varepsilon, x) = & 1 + \chi^0(\varepsilon, x) = x + \varepsilon x(1-x) + \varepsilon^2/3x(8x^2 - 15x + 7) + \\ & + \varepsilon^3/3(21x^3 - 50x^2 + 40x - 11) + \ldots. \end{aligned} \tag{5.4}$$

The first correction to the eigenfunctions can be determined using Eq. (4.7) as

$$\chi_1^j(x) = \sum_{k=1}^{j-1} s_k^j \chi_0^k(x) + \sum_{k=1}^{j+1} t_k^j b^k(x), \tag{5.5a}$$

where

$$s_k^j = \frac{B_{2j-2k+2}}{2^{2j-2k}-1}\frac{2k-1}{2k+1}\binom{2j+1}{2k-1}\frac{4^{j-k+1}-1}{4^{j-k}-1}, \tag{5.5b}$$

$$t_{j+1}^j = \frac{2j+1}{2^{2j+1}}, \tag{5.5c}$$

$$t_k^j = \frac{2j+1}{k2^{2j+1}}B_{2j-2k+2}\left[\binom{2j}{2k-2} - \binom{2j+1}{2k-1}j(4^{j-k+1}-1)\right] - \frac{1}{4^j}\sum_{l=k}^{j-1}\frac{2l-1}{2l+1}(4^{j-l+1}-1)\binom{2j+1}{2l-1}\binom{2l+1}{2k-1}. \tag{5.5d}$$

Note that the sum in (5.5d) appears only for $k < j$. Whereas higher order corrections for a general j become more complicated, the second largest eigenvalue $n^1(\varepsilon)$ and the eigenfunction $\chi^1(\varepsilon, x)$ are relatively easy to calculate

$$n^1(\varepsilon) = 1/4(1 - 3\varepsilon + 4\varepsilon^2 - 5\varepsilon^3 + \dots), \tag{5.6}$$

$$\begin{aligned}\chi^1(\varepsilon, x) = {} & \chi_0^1(x) + \varepsilon/8\{3b^2(x) - 4b^1(x)\} + \\ & + \varepsilon^2/4\{192\chi_0^2(x) + 12b^2(x) - 15b^1(x)\} + \\ & + \varepsilon^3/16\{86b^3(x) - 1280\chi_0^2(x) - 75b^2(x) + 92b^1(x)\} + \\ & + \dots .\end{aligned} \tag{5.7}$$

We are facing a different situation in the case of the biquadratic map $f_{BQ}(\varepsilon, x)$ given by Eq. (5.1b). As has been pointed out in Section 3, the quadratic map is related through a coordinate transformation $u(x) = sin^2(\pi x/2)$ to the linear map. As a consequence, using the notation (5.2) we have

$$p_0^j = m_L^j, \tag{5.8a}$$

$$\psi_0^j(x) = \phi_L^j(u^{-1}(x)), \tag{5.8b}$$

where m_L^i and ϕ_L^i are given by Eq. (3.2). It is useful to introduce the functions

$$c^j(x) = b^j(u^{-1}(x)), \tag{5.9}$$

with b^j's defined in Eq. (3.5). The stationary measure $\nu_{BQ}(\varepsilon, x)$ can be calculated by realizing that $(\hat{M}_{BQ})_i\psi_0^0(x)$ is a sum of a finite number of null functions $\zeta_Q^{q,k}(x) = \zeta_L^{q,k}(u^{-1}(x))$. As a consequence, each correction to the measure is a finite

sum of null functions again, that is, it is of the form $(x(1-x))^{1/2}$ times a polynomial of x. Concerning smaller eigenvalues, $p^j(\varepsilon) < 1$, one ascertains that $(\hat{M}_{BQ})_i\psi_0^j(x)$ is an infinite series of the eigenfunctions $\psi_0^k(x)$ and $c^k(x)$. In this way we arrive at complicated expressions for the eigenfunctions, yet we can calculate the eigenvalues in relatively simple manner. This is due to the feature that $(\hat{M}_{BQ})_i\psi_0^j(x)$ does not contain $\psi_0^k(x)$, $k < j$, components. The projection prescribed in (4.6) is thus simplified. Hence

$$p^j(\varepsilon) = 1/4^j\{1 - \varepsilon j + \varepsilon^2 j(j+1)/2 - \varepsilon^3 j(2j^2+6j+1)/12 + \ldots\}. \tag{5.10}$$

As expected, the largest eigenvalue with $j = 0$ gets no correction. The terms above indicate that the convergence radius shrinks to zero for small eigenvalues, i. e. for $j \to \infty$. We have also calculated the eigenfunction $\psi^1(\varepsilon, x)$ to first order

$$\psi_1^1(x) = \sum_{j=1}^{\infty} t^j \psi_0^j(x) - \frac{(x(1-x))^{1/2}}{4\pi^3}[6 \arcsin^2\sqrt{x} - 6\pi \arcsin\sqrt{x} + 1], \tag{5.11}$$

where

$$t^1 = 1 \tag{5.12a}$$

$$t^j = (-1)^j \frac{3j\pi^{2j-2}4^{j-2}}{2(2j+1)!(4^{j-1}-1)}, \qquad j > 1. \tag{5.12b}$$

The expansions above enable us to present explicit formulae for the asymptotic relaxation rate and amplitude of probabilistic measures. An initial measure $\nu_0(x)$ evolves in time as $\nu_n(\varepsilon, x) = (1+\hat{M}(\varepsilon))^n\nu_0(x)$, and approaches the invariant measure $\nu(\varepsilon, x)$. We only consider analytic measure. It relaxes as

$$\nu_n(\varepsilon, x) - \nu(\varepsilon, x) \sim e^{-nr(\varepsilon)}w(\varepsilon)\phi^1(\varepsilon, x), \tag{5.13}$$

where we introduced the relaxation rate $r(\varepsilon) \equiv \ln(1/m^1(\varepsilon))$ and the amplitude of relaxation $w(\varepsilon)$. Note that $w(\varepsilon)$ depends on the initial measure, while the rate $r(\varepsilon)$ is invariant for a large class of initial measures. Using Eqs (5.6) and (5.10) the decay rate for the maps (5.1a-b) can be given as

$$r_{BL}(\varepsilon) = \ln 4 + 3\varepsilon + \varepsilon^2/2 + 2\varepsilon^3 + \ldots, \tag{5.14a}$$

$$r_{BQ}(\varepsilon) = \ln 4 + \varepsilon - \varepsilon^2/2 + \varepsilon^3/12 + \ldots. \tag{5.14b}$$

It is interesting to compare the rate $r(\varepsilon)$ with the Lyapunov exponent $\lambda(\varepsilon)$, which can be calculated from its definition (2.12) and by using the invariant measure. We only quote the first nontrivial correction as [13,15]

$$2\lambda_{BL}(\varepsilon) = \ln 4 - \varepsilon^2/3 + \ldots, \tag{5.15a}$$

$$2\lambda_{BQ}(\varepsilon) = \ln 4 - \varepsilon^2/8 + \ldots. \tag{5.15b}$$

The accidental coincidence between $2\lambda(\varepsilon)$ and $r(\varepsilon)$ at $\varepsilon = 0$ no longer holds in the perturbed cases. The Lyapunov exponent shows a maximum at $\varepsilon = 0$, which we interpret as a most chaotic state. The maximal sensitivity to initial condition, however, is not reflected by the decay rate $r(\varepsilon)$.

As for the relaxation amplitude $w(\varepsilon)$, we have to determine the projection of the initial measure onto the first eigenfunction. This we did in case of the Lebesgue measure $\nu_0(x) \equiv x$ as the initial measure, using the orthogonality properties (3.7-8). We obtained for the bilinear and biquadratic map respectively

$$w(\varepsilon) = 64/3\,\varepsilon^2(-1 + \varepsilon/3\ldots), \tag{5.16a}$$

$$w(\varepsilon) = 4\pi^2/3. \tag{5.16b}$$

Since the Lebesgue measure is the invariant one for the linear map, the amplitude $w(\varepsilon)$ for the bilinear map should go to zero if $\varepsilon \to 0$. Note that the amplitude in the second case is independent of ε. The relaxation formula (5.13) together with (5.14) and (5.16) has been studied numerically as well, yielding results in agreement with the theory.

6. Critical maps

So far we have focused on maps with exponential divergence of trajectories, i.e. a positive Lyapunov exponent, and with exponentially relaxing measure and correlation function. In some chaotic maps with positive Lyapunov exponent, however, the measure and correlation functions exhibit power law decay. This can happen if the initial slope of the map $f'(0) = 1$ and at the same time $f(x) \gtrsim x$ for small x, because then every time the iteration reaches the neighborhood of $x = 0$, it stays there for a long while [16]. An example for critical maps is

$$f_c(x) = 1 - |1 - 2x|^{1/2}, \tag{6.1}$$

the stationary probability density of which is given by [13,17,18]

$$P_c(x) = 2(1 - x). \tag{6.2}$$

Note that the map has a peak rather than a smooth maximum. This is needed if $f'(0) = 1$ in order to diminish the probability of driving the iteration near $x = 0$. Otherwise the asymptotic density would be $P(x) = \delta(x)$, i. e. the iteration would stay mostly near the origin, and the map would not generate fully developed chaos.

In critical maps correlations decay algebraically [19]. Consider the correlation function $C_{AB}(n)$ as defined in Eq. (2.14). A conveniently modified version of the result of [19]reads as

$$C_{AB}(n) \propto n^{-t}, \qquad t = (1 + l + l')/(z - 1). \tag{6.3}$$

The indices l, l', and z are defined by

$$f(x) = x + ax^z + \ldots, \qquad z > 1, \tag{6.4a}$$

$$P(x) \propto x^l, \qquad x << 1, \tag{6.4b}$$

$$A(x) \propto x^{l'}, \qquad x << 1. \tag{6.4c}$$

The power law (6.3), as pointed out in [19], represents the contribution of the neighborhood of the origin.

The iteration of the measure can be studied directly by writing

$$\begin{aligned} \nu_n(x) &= (1 + \hat{M})^n \nu_0(x) \\ &\equiv 1 + \sum_{a_1,\ldots,a_n = l,u} (-1)^D \nu_0(f_{a_n}^{-1}(f_{a_{n-1}}^{-1}(\ldots(f_{a_1}^{-1}(x))\ldots))), \end{aligned} \tag{6.5}$$

where $D = \sum_{i=1}^{n} \delta_{ua_i}$, $\delta_{uu} = 1$, $\delta_{ul} = 0$, and u, l refer to the upper and lower branches of the inverse map $f^{-1}(x)$. We can identify here a term exhibiting power law decay, in fact it is the term with the sequence $l, l, \ldots, l$. This can be shown as follows. From Eq. (6.4a) follows that the lower branch of the inverse is asymptotically

$$f_l^{-1}(x) = x - ax^z + \ldots. \tag{6.6}$$

In order to determine the n-th iterate of $f_l^{-1}(x)$ we rewrite the backward iteration in a differential form [16]

$$x' = x - ax^z \quad \rightarrow \quad x' - x = dx/dn = -ax^z. \tag{6.7}$$

This trick can be applied, since a backward iteration step causes little change in x. It immediately follows

$$f_l^{-n}(x) \propto n^{-1/(z-1)}, \tag{6.8}$$

whence

$$\nu_0(f_l^{-n}(x)) \propto n^{-u}, \qquad u = p_0/(z-1). \tag{6.9}$$

Here p_0 is the exponent of the initial measure, $\nu_0(x) \propto x^{p_0}$, $x \ll 1$. In summarizing, we demonstrated that in a critical chaotic map the relaxation of the measure exhibits power decay. Using (2.15) it can be shown that the exponent in (6.9) is in accord with that of the correlation function in Eq (6.3).

Appendix

We show a way to calculate the coefficients $\hat{M}_i$ in the ε expansion of the operator $\hat{M}(\varepsilon)$. We study symmetric maps $f(\varepsilon, x) = f(\varepsilon, 1-x)$ as given in Eq. (4.1). Let us introduce the operator $\hat{F}(\varepsilon)$

$$\hat{F}(\varepsilon)\phi(x) \equiv \phi(f(\varepsilon, x)). \tag{A.1}$$

It can easily be seen that

$$\hat{F}(\varepsilon)\hat{M}(\varepsilon) = \hat{F}_0\hat{M}_0, \tag{A.2}$$

where $\hat{F}_0 \equiv \hat{F}(0)$. Furthermore, provided it acts on an analytic function the operator $\hat{F}(\varepsilon)$ can be expanded as

$$\hat{F}(\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i \hat{F}_i = \sum_{i=0}^{\infty} f_1^i(x)\varepsilon^i \frac{1}{i!}\frac{d^i}{dx^i}\bigg|_{x=f_0(x)}. \tag{A.3}$$

We shall also use the left inverse of $\hat{F}_0$ given by

$$\hat{F}_{0L}^{-1}\phi(x) = 1/2\{\phi(f_{0l}^{-1}(x)) + \phi(f_{0u}^{-1}(x))\}. \tag{A.4}$$

The operators $\hat{M}_i$ can now be obtained in terms of $\hat{F}_j$ and $\hat{F}_{0L}^{-1}$ using the fact that the left hand side of Eq. (A.2) does not depend on ε:

$$\hat{M}_i = \hat{F}_{0L}^{-1} \sum_{\{n_j\}} (-1)^l \hat{F}_{n_1}\hat{F}_{n_2}\ldots\hat{F}_{n_l}\hat{M}_0 \tag{A.5}$$

with the constraint $\sum_{j=1}^{l} n_j = i$ and $n_j \geq 1$. Finally, we list the first three operators

$$\hat{M}_1\phi(x) = -G(x)Q'(x), \tag{A.6a}$$

$$\hat{M}_2\phi(x) = 1/2G(x)^2Q''(x) + G(x)G'(x)Q'(x), \tag{A.6b}$$

$$\hat{M}_3\phi(x) = -1/6Q'''(x)G(x)^3 - G(x)^2G'(x)Q''(x) - 1/2G(x)^2G''(x)Q'(x) - G(x)G'(x)^2Q(x)', \tag{A.6c}$$

where $G(x) \equiv f_1(f_0^{-1}(x))$, a definition which is unambiguous because $f_1(x)$ is symmetric, and $Q(x) \equiv \hat{M}_0\phi(x)$.

References

1. H.L. Swinney and J.P. Gollub (ed.) Chaotic Behaviour and Fluid Dynamics, Springer, Berlin, Heidelberg, New York, 1981.
2. H. Haken (ed.), Evolution of Order and Chaos, Springer, Berlin, Heidelberg, New York, 1982.
3. L. Garrido (ed.) Dynamical Systems and Chaos, Springer, Lecture Notes in Physics, *179*, 1983.
4. W. Horstehmke and D.K. Kondepudi (ed.), Fluctuations and Sensitivity in Nonequilibrium Systems, Springer Proceedings in Physics *1*, 1984.
5. V.I. Oseledec, Trans. Moscow. Math. Sov., *19*, 197, 1968.
6. Y.B. Pesin, Russian Math. Surveys, *32*, 55, 1977.
7. P. Collet and J.-P. Eckmann: Iterated Maps on an Interval as Dynamical Systems, Birkhäuser, Basel, Boston, 1980.
8. M. Misiurewicz, Publ. Math IHES, *53*, 17, 1981.

9. P. Grassberger, Relaxation in 1-Dimensional Chaotic Maps, preprint, WU B 83-9.
10. M. Dörfle, J. Stat. Phys., *40*, 93,1985.
11. S. Grossmann and S. Thomae, Z. Naturforschung, *32a*, 1353, 1977.
12. S.M. Ulam and J. Neumann, Bull. Am. Math. Soc., *53*, 1120, 1947.
13. G. Györgyi and P. Szépfalusy, J. Stat. Phys., *34*, 451, 1984.
14. G. Keller, Analytic Perturbations of N-fold Interval Covering Mappings, unpublished.
15. P. Szépfalusy and G. Györgyi, Phys. Rev., *A33*, 2852, 1986.
16. P. Manneville and Y. Pomeau, Commun. Math. Phys., *74*, 189, 1980.
17. G. Györgyi and P. Szépfalusy, Z. Phys., *B55*, 179, 1984.
18. P.C. Hemmer, J. Phys., *A17*, L247, 1984.
19. S. Grossmann and H. Horner, Z. Phys., *B60*, 79, 1985.

Acta Physica Hungarica 64 (1-3), pp. 49-57 (1988)

A PHASE TRANSITION IN "LIFE"*

N. MENYHÁRD

Central Research Institute for Physics
1525 Budapest, Hungary

(Received 8 January 1987)

A stochastic cellular automaton rule is considered which gives in two limiting cases Conway's "Life" and a Class 2 (according to Wolfram) cellular automaton exhibiting domain structure. By changing the difference between these two rules with concentration c, a phase transition between phases of strongly different densities is found experimentally. A generalized inhomogeneous mean field approximation is proposed to account for the findings at least semi-quantitatively.

Introduction

Cellular automata (CA) are simple mathematical models for describing cooperative phenomena far from thermal equilibrium. They consist of a lattice with k states on each site. The time evolution of the sites is synchronous, governed by local transition rules. The value of a given site is determined by previous values of the sites in a certain neighbourhood around it. CA's were introduced in the fifties by Neumann and have been studied extensively ever since. Wolfram [1–3] has made a fundamental contribution to CA theory recently by systematically investigating and classifying one and two-dimensional CA with deterministic local rules. He found essentially four types of behaviour when starting from a random initial configuration: Class 1 CA evolve to homogeneous final states, Class 2 CA are characterized by separated periodic structures (in space and time). By Class 3 CA chaotic behaviour is exhibited. Most interesting are those CA which belong to Class 4: they show complex localized and propagating structures. The well-known example is Conway's "Game of Life" [4]. It has been shown that the structures of "Life" can be combined to perform arbitrary processing of information so that "Life" supports universal computation [4].

Stochastic cellular automata show a variety of new phenomena which may be of great importance in the study of phase transitions. The CA rules can be made stochastic in different ways. Grassberger et al [5] studied two "elementary" one-dimensional CA, with very specific added noise and obtained a phase-transition from Class 2 to Class 3 behaviour via instability of kinks between ordered states. For a more general class of stochastic one-dimensional CA, using transfer matrix scaling, phase diagrams and critical properties of transitions between Class 1 – Class 3

*Dedicated to Prof. G. Marx on his 60th birthday

and Class 2 – Class 3 deterministic behaviour have been calculated and compared with numerical results by Kinzel [6].

Studies of phase transition in two-dimensional stochastic CA have been presented recently by Kaneko and Akutsu [7]. The simplest procedure to involve noise has been used by these authors, namely at each time step the value obtained by the deterministic rule is reversed with a probability p. None of these investigations has been, however, aimed at a Class 4 system.

Schulman and Seiden [8], on the other hand, studied Conway's "Life " with a stochastic component denoted by them as "temperature". "Temperature" acts to assign a finite probability for $k = 1$ and $k = 0$ (birth and death) independently from the states of the neighbourhood and the centre cell. By starting from an approximation akin to mean field theory of statistical physics and calculating the correlations to first order they have arrived at a phase transition from "Life"-like states to a disordered state characterized by strongly different equilibrium densities. The theory has been unsuccessful, however, in predicting the density of the "Life"-like equilibrium state.

The aim of the present work is to follow the transition between two 2D automata, namely "Life" and an automaton showing domain structure (Class 2), when evolving from a random start and with rule differing from "Life" in a single component of the transition rule (rule 736 according to Wolfram [3]).

The computer experiment reveals a phase transition at a critical concentration c_{crit} of this perturbing term, separating a "Life"-type stable phase of very low density ($\rho_\infty \approx 0$) and a chaotic-looking phase of finite density $\rho_\infty(c_{\text{crit}}) \approx 0.28$. $\rho_\infty(c)$ then increases with c up to about $\rho_\infty \approx 0.49$ at $c = 1.0$. A generalized mean-field-like approximation in which the average density is supposed to be inhomogeneous is proposed. In this framework cluster-surface cell- states are found to play a dominant role in bringing about the above mentioned transition.

The model

For the 8-neighbour outer-totalistic [1–3] two-dimensional CA the transition rule is of the general form

$$a_{ij}(t+1) = f(a_{ij}(t), a_{i,j+1}(t) + a_{i+1,j}(t) + \ldots\ldots + a_{i-1,j-1}(t)), \tag{1}$$

where the a_{ij} can take k values. The different automata have been specified by Wolfram through the code number defined as

$$C = \sum_n f(a,n)k^{kn+a}, \tag{2}$$

where a is the state of the centre cell and n are the possible values of the sum over the 8 neighbours ($n = 0, \ldots\ldots, 8$).

Here we shall consider the following automaton ($k = 2$):

$$
\begin{aligned}
&f(0,3) = f(1,2) = f(1,3) = 1;\\
&f(1,4) = 1 \qquad \text{with probability} \quad c,\\
&f(1,4) = 0 \qquad \text{with probability} \quad 1-c,\\
&f(a,n) = 0 \qquad \text{otherwise.}
\end{aligned}
$$

For $c = 0$ we get Conway's "Life" (code number 224),while for $c = 1$ a CA having domain structure results (code number 736). Our goal is to study the transition between these two systems by changing c.

Let us first review briefly some basic facts about these two automata.

1. "Life"

The possible configurations of "Life" have been widely studied and discussed in the literature. We will be interested here in those stable states only which develop from starting with a randomly filled lattice after about N^2 time steps [1].

We are using a 64 x 64 lattice here. The stable pattern which has eventually evolved is seen in Fig.1. These are some of the simplest stable or period 2 configurations of "Life".

2. "Domain"

Domain is a structure made up of four basic phases, namely two with vertical stripes (stripes either on odd or on even sites) and two phases with horizontal

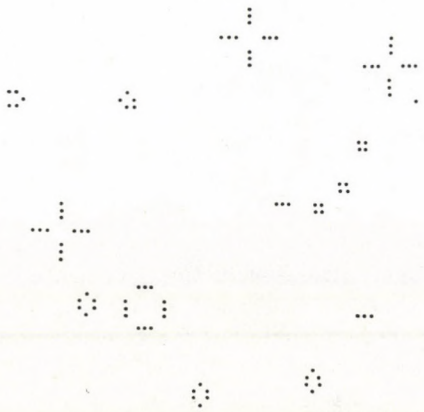


Fig. 1. Stable state of "Life" obtained in a lattice of 80 x 80 sites after 230 steps

stripes (again on odd or on even sites). Regions formed of one of these phases are invariant under the CA rule. Starting from a disordered configuration different phases evolve on different parts of the lattice and are separated by domain walls. The result is seen in Fig. 2.

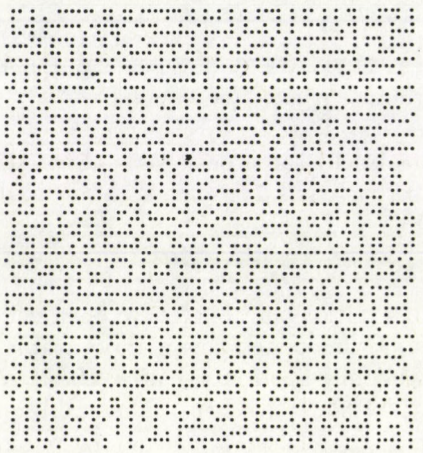


Fig. 2. "Domain" structure. Lattice size 80 x 80. Iteration steps: 75

Fig. 3. Some of the most common stable and oscillatory patterns of "Life"

Now turning to the case of general c's, let us first discuss the case of small c. Among the stable and small-period-oscillatory patterns of "Life" many are invariant against $f(1,4)$ as perturbation. Some of these are shown in Fig. 3.

If in the presence of some percent of $f(1,4)$ such configurations can evolve and far enough from each other, so that no effect of one on the other is exerted, and all other patterns die out, the stable phase can be called of type "Life".

The simplest quantity with which to characterize a CA configuration is the average density of sites with value 1, denoted by ρ. If the final state is of "Life"-type $\rho(\infty)$ is very small, effectively zero. With increasing c, however, the growth of

non-invariant and interacting (nearby) "Life" patterns hinder the lattice to become empty enough and finally some intermediate, chaotic-looking state sets in with finite density, which then increases with c. A theoretical expression of ρ as a function of t cannot be deduced in general.

The next Section is devoted to a kind of mean field approximation to this problem and to the tendencies which can be seen from such a simple-minded theory.

MF approximation

The average density of 1's will be denoted by $\rho(t)$, where time t changes in discrete steps. If correlations between values at different sites are neglected, an analog of the MFA of statistical physics can be given as follows [8,1]. All configurations are parametrized by their density ρ or (equivalently) by probabilities p and $q(= 1 - p)$ for each site to have value 1 and 0, respectively. The probability that there are n 1's in the 8-neighbourhood is

$$\binom{8}{n} p^n q^{8-n}. \tag{1}$$

The time evolution of the average density for the CA studied here is given by ($\rho \equiv p$)

$$\rho(t+1) = \binom{8}{2}\rho^2(1-\rho)^6 \rho f(1,2) + \binom{8}{3}\rho^3(1-\rho)^5 *$$

$$[\,\rho f(1,3) + (1-\rho) f(0,3)] + c\binom{8}{4}\rho^4(1-\rho)^4 \rho f(1,4) =$$

$$= 14\rho^3(t)(1-\rho(t))^4[2(1-\rho(t))(3-\rho(t)) + 5c\rho^2(t)]. \tag{2}$$

In equilibrium ($t \to \infty$) $\rho(t+1) = \rho(t)$. Eq. (2) gives a polynomial equation for $\rho(\infty) \equiv \rho_\infty$

$$1 = 14\rho_\infty^2(1-\rho_\infty)^4[2(1-\rho_\infty)(3-\rho_\infty) + 5c\rho_\infty^2], \tag{3}$$

which has the real solutions $\rho_\infty = 0$ and $\rho_\infty(c)$ as shown in Fig. 4.

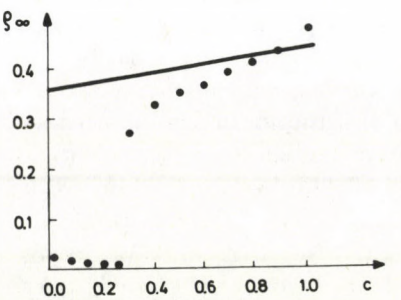


Fig. 4. Equilibrium value of the average density ρ_∞ shown as a function of the concentration c as given by experiment (dotted line) and the simple mean field approximation (full line)

The experimental data also shown in Fig. 4 with a dotted line have been obtained starting from a random initial configuration with $p_0 = 0.5$ and in a lattice of 32×32 sites, with periodic boundary conditions. Some of the data points have been obtained by adding a small "temperature" component of Schulman and Seiden [8] to the basic rule. Noise makes transient effects die out more quickly than at "T"$=0$ and the equilibrium state sets in typically in a few hundred steps making the computer experiment feasible on an IBM PC. The phase transition shown by the experiment at about $c \approx 0.3$ will be discussed later on.

The regions of stability of the fixed points $\rho_\infty = 0$ and $\rho_\infty(c)$ can easily be inferred from the curve of $\rho(t+1)$ as a function of $\rho(t)$ as shown in Fig. 5 for different values of c.

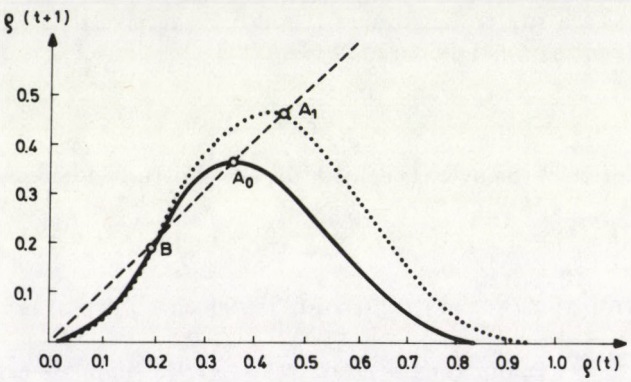


Fig. 5. MFA for the average density ρ for c=0 (full line) and c=1 (dotted line)

The points of intersection A_0 and A_1 with the straight line $\rho(t+1) = \rho(t)$ are the finite fixed points $\rho_\infty(c=0)$ and $\rho_\infty(c=1)$ (for inbetween values of c the curves lie between these two). Point B separates the domains of attraction of the fixed points $\rho_\infty = 0$ and A_i $(i = 0, 1)$. (The weak c-dependence of B has been neglected not to confuse the Figure unnecessarily).

For some fixed c, starting from a random initial configuration with $\rho(0) = 0.5$, ρ should approach $\rho_\infty(0)$ if the mean field approximation were applicable, i.e. the effect of correlations were negligible. This is clearly not the case which manifests itself most definitely in the stable state of "Life" with $\rho_\infty \approx 0$.

"Life"-like correlations are inclined to produce empty patches in some tens of iteration steps already. (See Fig. 6.) We propose here to treat the clusters which are formed in this way, surrounded by empty sites, in the spirit of the MFA as above. Let us examine a stripe of 1's which is two-sites wide as the simplest structure modelling a cluster in the sea of zero sites.

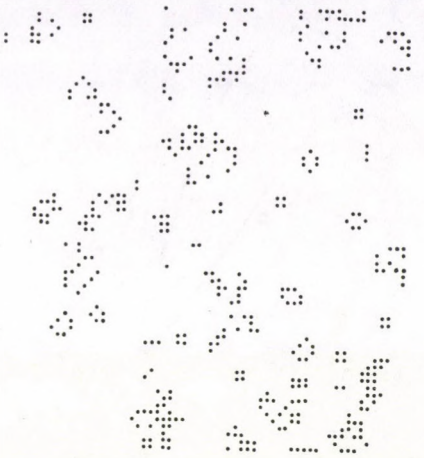


Fig. 6. Example of clustering, obtained with c=0.01 after 80 steps. Lattice size 80 x 80 with periodic boundary conditions

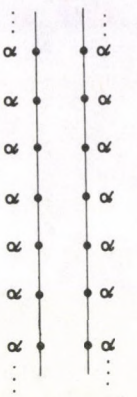


Fig. 7. Simplified cluster structure

Let us take the α-type surface sites and formulate a MF expression for the average density of 1's over them. Three of the 8 neighbours are now empty with probability 1. Thus the effective number of neighbours which can be "alive" is 5 and correspondingly

$$\rho_\alpha(t+1) = \binom{5}{3}\rho_\alpha^3(1-\rho_\alpha)^3 \cdot f(0,3) + \binom{5}{2}\rho_\alpha^3(1-\rho_\alpha)^3 f(1,2) + \\ + \binom{5}{3}\rho_\alpha^4(1-\rho_\alpha)^2 f(1,3) + \binom{5}{4}c\rho_\alpha^5(1-\rho_\alpha) f(1,4) = \\ = 10\rho_\alpha^3(1-\rho_\alpha)^2(2-\rho_\alpha) + 5c\rho_\alpha^5(1-\rho_\alpha). \tag{4}$$

For the α-sites Fig. 8 shows $\rho_\alpha(t+1)$ as a function of $\rho_\alpha(t)$ for different c-values.

For $c < c_{\text{crit}} = 0.39$ only the $\rho_\infty = 0$ fixed point is stable for the average density of α-type surface states. At $c = c_{\text{crit}}$ a new fixed point appears with finite density ($\rho_\infty(c_{\text{crit}}) = 0.5$). The magnitude of this new stable density increases steadily with increasing c reaching the value of $\rho_\infty = 0.66$ at $c = 1$.

It is possible to construct and examine wider cluster-mimicking stripes, as well. Extension to 3 layers is simple. In this case there are two different densities of structure ($0\alpha\beta\alpha 0$) and the result is the same as above only with a slightly different c_{crit}. Wider stripes fail to exhibit the depletion effect.

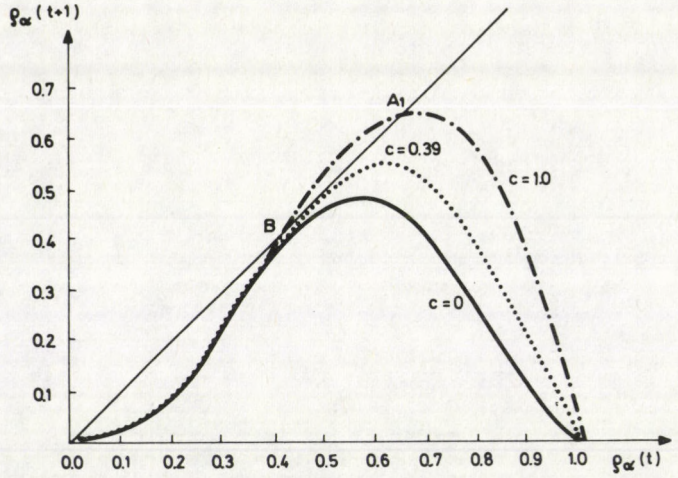


Fig. 8. $\rho_\alpha(t+1)$ as a function of $\rho_\alpha(t)$ for three values of c: full line c=0, dotted line c=0.39 and dashed line c=1. Point $B(\rho = 0.41)$ separates the regions of attraction of the fixed points ρ_∞=0 and $\rho_\infty(c=1) = 0.66(A_1)$

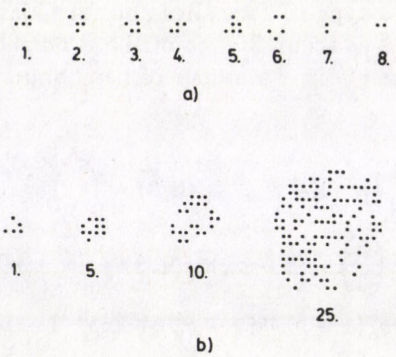


Fig. 9. "Glider" of "Life" perturbed by $f(1,4)$ with concentration c=0.2 (a) and c=0.9 (b). In case a) the "glider" dies out in 9 steps while in case b) growth sets in. For c=0, of course, it propagates forever (in an infinite medium and without any collision)

In this way the following oversimplified picture of the phase transition "Life-domain", exhibited in Fig. 4, arises. For small values of the perturbation $f(1,4)$

i.e. for low c, "Life"-like correlations lead to cluster-like objects after a random start (with $p_0 = 0.5$) in about, say, 50–100 iterations (depending on lattice-size, noise). For clusters which are thin enough the above surface-MF approximation-like treatment applies and the overall result is that surface effects will pull the system towards smaller and smaller densities. Eventually the density $\rho_\infty \approx 0$ of the "Life"-like stable state results. This is opposed by the volume term which favours a stable state at some finite density, depending on c ($0.37 \leq \rho_\infty(c) \leq 0.45$) and which will become dominating from some critical value of c on. In the oversimplified model treated above this critical value of the concentration was obtained as $c_{\text{crit}} \approx 0.4$. The experimental value is significantly lower, which is plausible if we consider that the whole picture of cluster-formation which should work in a self-consistent way may lose its validity for c-values which are not small. Namely, with increasing c growth of some simple Life-objects sets in which is very well demonstrated e.g. by the fate of the "glider" of "Life" when perturbed by f (1,4). (See Fig. 9.). On the other hand, the theoretical and experimental values of $\rho_\infty(c)$ are in a surprisingly good agreement.

A more systematic investigation of surface effects and their combination with volume terms for determining the density and also other relevant quantities like entropy would be needed to put the results presented above on a more firm basis. Also we hope to give a detailed study of the nature of the intermediate state which sets in at c_{crit} and of the formation of the "domain"-phase near $c=1$ in future work.

References

1. S. Wolfram, Rev. Mod. Phys., *55*, 601, 1983.
2. S. Wolfram, Physica, *D10*, 1, 1984.
3. N. Packard and S. Wolfram, J. Stat. Phys., *38*, 901, 1985.
4. E. R. Berlekamp, J. O. Conway and R. K. Guy, Winning ways for your mathematical plays, Academic Press, Vol. 2, Chap. 25, 1982;
   M. Gardner, Wheels, Life and other mathematical amusements, Freeman, 1983.
5. P. Grassberger, F. Krause and T.v.d. Twer, Phys., *A17*, L105, 1984.
6. W. Kinzel, Z. Phys., *B58*, 229, 1985.
7. K. Kaneko and Y. Akutsu, J. Phys., *A19*, L69, 1986.
8. L. S. Schulman and P. E. Seiden, J. Stat. Phys., *19*, 293. 1978.

Acta Physica Hungarica 64 (1–3), pp. 59–66 (1988)

OPTIMAL SIZE OF NEURAL NETWORKS WITH MULTIPLICATIVE LEARNING*

R. NÉMETH and T. GESZTI

Department of Atomic Physics, Roland Eötvös University
1088 Budapest, Hungary

(Received 8 January 1987)

A multiplicative model of learning by synaptic plasticity is reconsidered in order to relax an undesired symmetry restriction. This is found to leave essential previous results unchanged. In particular the network has an optimal size of about 700 interconnected neurons, tentatively identified with cortical columns. Bounds are given for the capacity of the multiplicative model for noiseless information storage.

1. Introduction

A great uprise of activity in studying physically motivated models of neural networks [1] has been motivated by Hopfields's observation [2] that such models are related to the spin-glass problem [3]. The connection is this: firing or non-firing neurons correspond to up or down spins [4]; excitatory or inhibitory synapses (connections between neurons) correspond to ferromagnetic or antiferromagnetic couplings between the spins. A spin system with a sufficiently random mixture of ferro- and antiferromagnetic interactions at low temperatures freezes into one of its several, apparently disordered, metastable states: this is called a spin-glass [3].

Despite the apparent disorder, the metastable states are rigorously determined by the actual values of the coupling strengths, and this is the feature connected to more specific neurological information: metastable states that attract thinking within some "basin of attraction" may correspond to learned memories, and the way they are learned is by appropriate modifications of the couplings or synaptic strengths, as postulated by Hebb [5].

The relaxation of the system towards one of its metastable states is governed by the firing signals each neuron receives from the others. If the state of the i-th neuron is characterized by the "quasi-spin" variable $S_i = \pm 1$ ($+1$ if fires, -1 if not), then it receives an input signal

$$h_i = \sum_{j \neq i} J_{ij} S_j , \tag{1}$$

where the (positive or negative) coupling constant J_{ij} represents the sign and strength of the synapse through which pulses from the j-th neuron are transmitted

*Dedicated to Prof. G. Marx on his 60th birthday

to the i-th. In the spin language, this input signal acts as an effective magnetic field, forcing the i-th neuron to alter its state (starting or ceasing to fire) if

$$S_i h_i < 0, \tag{2}$$

because then the "flip" $S_i \to -S_i$ reduces the "potential energy" $-S_i h_i$. Inequality [2] is apparently a sharp threshold condition that controls flipping or non-flipping.

Model systems whose dynamics are controlled by similar threshold conditions belong to the well-known family of cellular automata [6]. It is a decisive feature of Hopfield's model [2] that neurons for which condition (2) is satisfied are still allowed to flip but asynchronously, one at a time. This eliminates some nonrealistic periodicities, plaguing some applications of cellular automata.

Another simplifying feature of the model is the symmetry restriction

$$J_{ij} = J_{ji}, \tag{3}$$

which — although biologically unrealistic [1] — makes the approach to metastable states monotonous [2].

A central issue of the discussion is the memory storage capacity of the network, i.e. how many different memories can be simultaneously stored in the form of metastable states and recalled with an acceptable fraction of errors. Closely related is the problem of various learning algorithms, i.e. various modifications of the couplings J_{ij} in order to make given states of the spin system approximately metastable ones [7–10]. Even in the simplest cases the precise connection with the original spin-glass problem is rather nontrivial [11].

In Section 2 we summarize two possible learning algorithms [2, 10]. The second of them, called the multiplicative algorithm [10], is slightly modified here to allow obtaining new analytical results. A principal feature of the modification is to get rid of the symmetry requirement, Eq. (3). Section 3 describes a novel method to analyse the accuracy of memory recalling, taking advantage of this enhanced independence of the coupling constants. The results are used in Section 4 to discuss the cases of storage with no error and with a given level of relative error. The possible significance of the results in connection with the columnar structure of the cerebral cortex is discussed in Section 5.

2. Learning algorithms

If we want to store a given information encoded in the form of a vector $\{S_i^{(0)}\}(i = 1, ..., N$ where N is the number of neurons), then the choice

$$J_{ij} = S_i^{(0)} S_j^{(0)} \tag{4}$$

would apparently do the job: if the system is put into the same state $\{S_i^{(0)}\}$ then using Eq. (1) it is obvious that condition (2) is never satisfied, i.e. none of the

spins (neurons) would flip: $\{S_i^{(0)}\}$ is a stable (or, at least, metastable) state of the system, and if the neural network "thinks" about something sufficiently similar then the system relaxes into this state: the stored information is recalled (that is why this kind of memory is called associative).

The situation is changed if $p > 1$ vectors $\{S_i^{(a)}\}(a = 1, \ldots, p;\ i = 1, \ldots, N)$ have to be stored in the same network. The formula [12]

$$J_{ij} = \sum_{a=1}^{p} S_i^{(a)} S_j^{(a)} \tag{5}$$

acts in the right direction, however, in recalling anyone of the stored vectors, all the others play the role of quasi-random noise.

This noise has been analysed by Hopfield [2] in the following way. If we put the system into one of its memorized states, say $\{S_i^{(b)}\}$, noise makes the system relax into another state that differs from $\{S_i^{(b)}\}$ in some fraction of its components (bits). According to Hopfield's criterion [2], memory recalling is acceptable if it has probability less than 1/2 that this fraction is over 5%. This fraction is, however, hard to calculate and depends on dynamical details. Assuming that those details are not very relevant, one chooses therefore a closely analogous but simpler and more robust characteristics: the fraction of unstable neurons, i.e. those for which condition (2) is satisfied in the given initial state $\{S_i^{(b)}\}$. If furthermore $h_i S_i^{(b)}$ and $h_j S_j^{(b)}$ are regarded as independent for $i \neq j$ [10] then for a large network ($N \to \infty$) Hopfield's criterion is satisfied for a given critical value $\varepsilon^* = 0.052$ of the probability ε of instability (error) in a single bit. Therefore various learning algorithms can be tested through this quantity:

$$\varepsilon =: P(B_i^{(b)} < 0), \tag{6}$$

where

$$B_i^{(b)} = S_i^{(b)} \sum_{j \neq i} J_{ij} S_j^{(b)}. \tag{7}$$

Since all neurons are equivalent and so are all memorized vectors (see Eqs (5) or (8) below), ε does not depend on either i or b.

As a model obeying Eq. (3) and thereby defining a spin-glass with well-defined correlations built into the distribution of coupling constants, Eq. (5) is a convenient starting point for statistical physical studies [2,11]. On the other hand, it has several drawbacks as a model of real neural networks. In particular, synapses have their native signs that do not change on learning, and their non-zero native strengths even if nothing has been learned. Both these features are missed by the additive algorithm, Eq. (5), and this was the main motivation in [10] to introduce and analyse by the same approximate error counting method the *multiplicative* one, defined as

$$J_{ij} = T_{ij} \exp\left(x \operatorname{sgn} T_{ij} \sum_{a=1}^{p} S_i^{(a)} S_j^{(a)}\right), \tag{8}$$

where T_{ij} is the native strength of the coupling J_{ij}. This means that on learning a new vector $\{S_i^{(a)}\}$ each J_{ij} is either multiplied or divided by e^x, according to whether this coupling is "satisfied" or "frustrated" in the memorized state ($J_{ij}S_i^{(a)}S_j^{(a)} > 0$ or < 0).

In [10] we chose randomly $T_{ij} = T_{ji} = \pm 1$. Now we regard T_{ij} and T_{ji} as independent, continuous random variables of the same probability distribution for all i, j, characterized by the moments

$$< |T_{ij}| >= t \tag{9}$$

and

$$< T_{ij}^2 >= \sigma^2. \tag{10}$$

Allowing $T_{ij} \neq T_{ji}$ has serious consequences for the dynamics, however, it does not change the semiquantitative counting of recalling errors based on Eq. (6).

3. Noise analysis

In this Section we calculate the probability ε of error in a single bit (see Eq. (6)) as a function of the learning parameter x, the number of interconnected neurons N and the number of memorized vectors p. First let us rewrite Eq.(8) as

$$J_{ij} = T_{ij}(1 - v^2)^{-p/2} \prod_a (1 + vX_{ij}^{(a)}), \tag{11}$$

where the notations

$$v = \tanh x \tag{12}$$

and

$$X_{ij}^{(a)} = S_i^{(a)} S_j^{(a)} \operatorname{sgn} T_{ij} \tag{13}$$

have been introduced. Then Eq. (7) takes the form

$$B_i^{(b)} = (1 - v^2)^{-p/2} \sum_{j \neq i} |T_{ij}| Y_{ij}^{(b)} \tag{14}$$

with

$$Y_{ij}^{(b)} = (X_{ij}^{(b)} + v) \prod_{a \neq b} (1 + vX_{ij}^{(a)}). \tag{15}$$

Now we want to exploit the fact that since the random variables $X_{ij}^{(a)}$ are products of random factors of values ± 1, any two of them are uncorrelated if not equal. This is the principal novelty of the present approach with respect to [10]. As a consequence, the terms of the sum in Eq. (14) are independent random variables, and we can safely use the central limit theorem to approximate the probability

distribution function of $B_i^{(b)}$ through a Gaussian of mean value $\bar{B}$ and dispersion S. Then the desired probability of error in a single bit is obtained in the form

$$\varepsilon = 1 - \phi(Z) = 1/2 \operatorname{erfc}(Z/\sqrt{2}), \tag{16}$$

where

$$\phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-y^2/2} dy, \tag{17}$$

erfc is the complementary error function [13], and

$$Z = \bar{B}/S. \tag{18}$$

To evaluate B and S, we observe that $Y_{ij}^{(b)}$ is the product of a number of factors assuming randomly and independently one of the values $1 + v$ and $1 - v$. Then simple combinatorics give

$$\langle Y_{ij}^{(b)} \rangle = v, \tag{19}$$

$$\langle (Y_{ij}^{(b)})^2 \rangle = (1 + v^2)^p. \tag{20}$$

Finally, using the definitions (9), (10) and (14), we obtain

$$\bar{B} = (N - 1)tv, \tag{21}$$

$$S^2 = (N - 1)(\sigma^2(1 + v^2)^p - t^2 v^2), \tag{22}$$

which — neglecting 1 with respect to N — gives the desired result

$$Z = (N/[v^{-2}(1 + v^2)^p - 1])^{1/2}. \tag{23}$$

In writing Eq. (23) we have restricted ourselves to choosing a discrete set of synaptic strengths, $T_{ij} = \pm t$ which entails $\sigma^2 = t^2$. It is easy to see that any other choice reduces Z and therefore increases the error probability. For the present choice, which is retained in the remaining part of this paper, Eq. (23) should be compared to the result obtained* [10] for the symmetric model obeying Eq. (3),

$$Z = (N/[(1 + v^{-2}) \exp(x^2(p - 1)) - 1])^{1/2}. \tag{23a}$$

For the relevant limiting case $N >> 1$, $p >> 1$, $v << 1$ both Eqs (23) and (23a) approach the asymptotic form

$$Z^2 = Nx^2 \exp(-x^2 p), \tag{24}$$

*In [10] the last term -1 in the square bracket was missing due to a calculational error. Eq. (23a) corrects this error. Use of the correct Eq. (23a) reduces the discrepancy between simulations and approximate theory ([10], Fig. 1) by a factor of two.

which, introducing the notations

$$\alpha = p/N, \tag{25}$$

$$\tilde{N} = Nx^2, \tag{26}$$

can be rewritten as

$$Z^2 = \tilde{N}e^{-\tilde{N}\alpha}. \tag{27}$$

The requirement of no error needs $Z \to \infty$. On the other hand, as pointed out in [10], Hopfield's criterion is satisfied for $Z > Z^* \approx 1.6$. These two cases are investigated in the next Section.

4. Storage capacity

4.1 Noiseless storage

If the appearance of errors in different neurons (spins, bits) can be taken as independent [10] then neurons with $B_i^{(b)} > 0$ and < 0 are distributed binomially, and the mean number of stressed neurons (erroneous bits) is just $N\varepsilon$. Then ε itself is the fractional error, which vanishes if $Z \to \infty$, as seen from Eq. (16). With a fixed learning parameter x, it is easy to see from Eq. (24) that for $N \to \infty$, this requirement can be satisfied even with an infinite number of different words stored i.e. $p \to \infty$, if the rather restrictive upper bound

$$p(N) \leq (1 - \delta)x^{-2} \ln N \tag{28}$$

is obeyed with a small positive constant δ. However x can be tuned with growing N and p to achieve a more advantageous performance. For given N and P, the maximum of Z and therefore the minimum error is reached for

$$x_{id}^2 = p^{-1}, \tag{29}$$

which gives

$$Z_{id}^2 = (e\alpha)^{-1} \tag{30}$$

and, using the asymptotic expansion of the error function [13] for large Z,

$$\varepsilon_{id} = \left(\frac{\alpha}{e\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{e\alpha}\right). \tag{31}$$

This is closely analogous to the result obtained by Amit et al [15] by the replica method for the additive model, with only a slight variation of numerical parameters. Here, like in that case, now only $\alpha \to 0$, which is much less restrictive than (28), is needed for vanishing relative noise. Following the reasoning of [15], the vanishing of the absolute error $N\varepsilon$ now requires

$$p \leq N(e \ln N)^{-1}. \tag{32}$$

4.2. Storage with finite noise

Since Eq. (24) is equivalent to the result obtained in [10], the main results of that paper are recovered here. In particular, for a given learning parameter x and error fraction ε^* with a corresponding value $Z^*(\approx 1.6$ for Hopfield's criterion [2]), there is an upper bound for the relative storage capacity:

$$\alpha_{\max} = (Z^{*2}e)^{-1}, \tag{33}$$

which however can be reached for an optimal *size* of the network,

$$N_{\text{opt}} = Z^{*2}ex^{-2}. \tag{34}$$

For $Z^* = 1.6$ and $x = 0.1$, this gives $N_{\text{opt}} \approx 700$ [10].

5. Conclusions

We have proven that the main results of [10], namely the existence of an optimal network size assuring a maximum relative capacity (stored vector per neuron), does not depend upon the unrealistic assumption of symmetric synaptic strengths. The numerical estimate of the optimal size, $N_{\text{opt}} \approx 700$, also remains valid.

We still think it reasonable to interpret this result in terms of a tendency to partition a larger neural network into loosely interconnected blocks of the optimal size. In [10] these blocks were tentatively identified with peripherial ganglia. In the meantime we have learned [16] that those ganglia consist of functionally unrelated neurons. On the other hand, the numerical value estimated for N_{opt} recalls the similar size of columns in the cortex, contrary to the disclaim in [10]. However, current interpretation of the columnar structure of the cortex is developmental rather than functional [16], and much further work is needed to reach any sound conclusion about the possible connection between cortical columns and the optimalization described in the present work.

Acknowledgments

It is a pleasure to acknowledge that it was Professor G. Marx who turned our attention to the growing interest in spin-glass-like models of neural networks. We also thank enlightening discussions with Prof. J. Szentágothai and Drs. J. Kertész, P. Érdi and G. Barna.

References

1. J.W. Clark, J. Rafelski and J.V. Winston, Phys. Rep., *123*, 215, 1985.
2. J.J. Hopfield, Proc. Natl. Acad. Sci. USA., *79*, 2554, 1982.

3. Proc. Heidelberg Colloquium on Spin Glasses, Eds. J.L. Van Hemmen and I. Morgenstern, Lecture Notes in Physics, *192*, 1983.
4. W.S. McCulloch and W. Pitts, Bull. Math. Biophys., *5*, 115, 1943.
5. D.O. Hebb: The organization of behavior, Wiley, New York, 1949.
6. Proc. Interdisciplinary Workshop on Cellular Automata, Eds. D. Farmer, T. Toffoli and S. Wolfram, Physica., *10D*, 1-274, 1984.
7. W. Kinzel, Z. Phys., *B60*, 205, 1985.
8. Vik. S. Dotsenko, J. Phys., *C 18*, L1017, 1985.
9. G. Parisi, J. Phys., *A 19*, L617, 1986.
10. T. Geszti, Phys. Lett., *114A*, 334, 1986.
11. D.J. Amit, H. Gutfreund and H. Sompolinsky, Phys. Rev., A *32*, 1007, 1985.
12. L.N. Cooper, F. Liberman and E. Oja, Biol Cybern., *33*, 9, 1979.
13. M. Abramowitz and I.A. Stegun, Handbook of Mathematical Functions, Dover Publications, New York,Chapter 7. 1965.
14. G. Weisbuch and F. Fogelman-Soulie', J. Physique Lett., *46*, L623, 1985.
15. D.J. Amit, H. Gutfreund and H. Sompolinsky, Phys. Rev. Lett., *55*, 1530, 1985.
16. J. Szentágothai, private communication.

Acta Physica Hungarica 64 (1-3), pp. 67-72 (1988)

AN EXACT COMBINATORIAL TREATMENT OF THE ONE-DIMENSIONAL HARD-CORE FLUID*

A. BARANYAI and I. RUFF

Laboratory of Theoretical Chemistry, Department of Chemistry, Roland Eötvös University
1088 Budapest, Hungary

(Received 8 January 1987)

A novel solution is given of the problem of the one-dimensional hard-core fluid by distinguishing the nth neighbour of a particle. The pair correlation functions can be calculated by a very simple algorithm based on this solution. The equation of state is also given.

It is well known that the short-range order in dense fluids is mainly due to the steep repulsion wing of the pair potential, i.e. to the excluded volume effect, and this is only slightly modified by the long-range forces. The pair correlation function, $g(r)$, for a fluid that has a given density and consists of uniform hard spheres free of interactions is a good approximation of the pair correlation function of a one-component real fluid with the same density. It is also well known that, after reaching a definite number density, the hard-core model shows the same kind of phase transitions of freezing or melting [1].

It is thus quite obvious why so much effort has been made in the last decades to derive the exact pair correlation function and equation of state of the hard-core fluid. The most important achievement was the analytical solution of the Percus–Yevick equation [2]. This (mathematically complicated) solution yields a simple, closed, formula for the equation of state of the hard-core fluid, which is, however, not exact, since the final results differ depending on the choice whether the virial or the compressibility equation is used for closure:

$$\frac{pV}{Nk_BT_c} = \frac{1+\eta+\eta^2}{(1-\eta)^3}, \tag{1a}$$

or

$$\frac{pV}{Nk_BT_p} = \frac{1+\eta+\eta^2-3\eta^3}{(1-\eta)^3}. \tag{1b}$$

Still they proved to be good approximations, because they flank closely the function fitting computer simulation results that can be considered exact:

$$\frac{pV}{Nk_BT} = \frac{1+\eta+\eta^2-\eta^3}{(1-\eta)^3}, \tag{1c}$$

*Dedicated to Prof. G. Marx on his 60th birthday.

where $\eta = \pi\sigma^3\rho/6$ is the so-called packing density in which σ is the diameter of the rigid spheres and ρ the number density.

The equation of state of a three-dimensional hard-core fluid is the first quantitative and rigorous theoretical step from the perfect gas towards real dense fluids. While the two-dimensional gas or liquid is an adequate model of adsorbed layers on a plane surface, the one-dimensional fluid is only interesting from the theoretical point of view. From this latter aspect, however, both one- and two-dimensional fluids are much more attractive than three-dimensional ones. For instance the theory of phase transitions has been much more successful in two dimensions [3].

The $g(r)$ function of the one-component and one-dimensional hard-core fluid was first derived in the 1930's [4]. Since then several different derivations have been published, yet the simple approach given below has not appeared so far. In addition, its simplicity makes it useful for didactic purposes [5].

As an introduction to the problem, let us consider a section of straight line of length L along which a number N of point-like particles are distributed whose motions are independent of one another. In this case the distribution function of every particle will be the same constant function shown in Fig. 1.

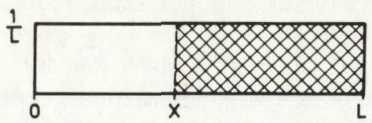


Fig.1. The even distribution of the point-like particles of the one-dimensional perfect gas (see text)

In a given configuration we can number the particles with respect to their serial number from one end of the section of line, i.e. we can name the nth closest particle to the origin. The occurrence of the nth particle at position x is proportional to the event that $n-1$ particles are before it and $N-n$ particles are behind it, i.e. since the probability of their occurrence is independent of one another, it is proportional to $(x/L)^{n-1}$, to $|(L-x)/L|^{N-n}$, and to the one-particle distribution function at position x which is $1/L$ as well as to the combinatorial number of realizations of the given arrangement, since the particles are undistinguishable. The nth particle can be selected in N different ways, while the number $n-1$ before the nth can be selected in $\binom{N-1}{n-1}$ different ways. (The selection of those behind it are already determined unequivocally by these quantities.)

Thus, the probability to find the nth particle at position x is:

$$P_n(x) = \frac{N}{L}\binom{N-1}{n-1}\left(\frac{x}{L}\right)^{n-1}\left(\frac{L-x}{L}\right)^{N-n}. \tag{2}$$

The sum of the $P_n(x)$'s for all n's from 1 to N must reproduce the sum of the constant distribution functions in Fig. 1. Indeed, the sum:

$$P(x) = N\left(\frac{1}{L}\right)^N \sum_{n=1}^{N} \binom{N-1}{n-1} x^{n-1}(L-x)^{N-n} \tag{3}$$

is the binomial decomposition of the constant N/L.

The generalization of this case to one in which the one-dimensional space L is replaced by a circle or any kind of closed curve of length L, is quite straightforward. Any point of such a curve may be chosen as the origin, and any of the two directions leading from the origin may be chosen as the direction of the positive coordinate. The results would be symmetrical for these choices. In this case the "radial distribution function in one dimension" – which is the corresponding pair correlation function – is:

$$g_L(r) = g_L(|x|) = [P(x) + P(-x)]/2\rho = P(x)/\rho. \tag{4}$$

We can follow this procedure for the case when the particles are not point-like but their "hard cores", impenetrable for other particles, occupy small sections of straight lines of length σ. A common model of this is a thread of length L with N pearls of diameter σ on it forming a necklace. Let us choose the origin at the centre of a particle. As seen in Fig. 2 the positions $x = 0$ and $x = L$ are equivalent.

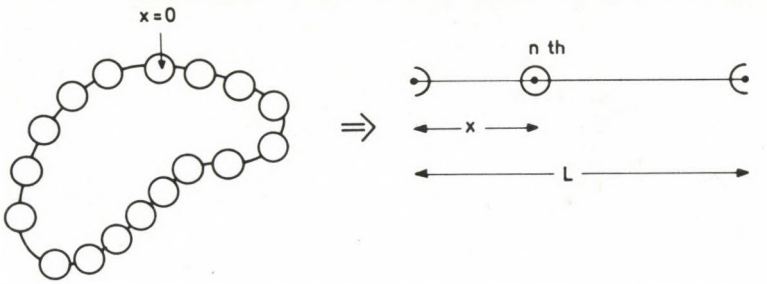


Fig.2. To the derivation of the pair correlation function of a one-dimensional hard-core fluid: choice of reference particle and the periodic occurrence of its nth neighbour

There is $L - N\sigma$ free space for the motion of the particles. Since the particle chosen to be at the origin "cannot move", only $N-1$ particles are to be considered. The probability of occurrence at position x of the nth closest neighbour, say, to the "right hand side" from the particle carrying the origin, is proportional to the $(n-1)$th power of the free space "before" the nth particle and to the $(N-n-1)$th power of the free space "behind" it. The combinatorial factors are the same as in Eq. (2), while the normalization factor is $1/(L - N\sigma)$ instead of $1/L$. Thus:

$$P_n^{\sigma}(x) = \frac{N-1}{L-N\sigma}\binom{N-2}{n-1}\left(\frac{x_n - n\sigma}{L-N\sigma}\right)^{n-1}\left(\frac{L-(N-n)^{\sigma}-x_n}{L-N\sigma}\right)^{N-n-1}. \tag{5}$$

The subscript n of x indicates that the distribution function of the nth particle can differ from zero only from the shortest possible distance $n\sigma$ up to the farthest possible distance $L-(N-n)\sigma$, since the rest of the space would be occupied by the closest packed particles before and behind it, respectively. Thus:

$$n\sigma \leq x_n \leq L-(N-n)\sigma, \tag{6}$$

while the sum of the distribution of all particles:

$$P^{\sigma}(x) = \frac{N-1}{L-N\sigma} \sum_{n=1}^{N-1} \binom{N-2}{n-1} \left(\frac{x_n - n\sigma}{L-N\sigma}\right)^{n-1} \left(\frac{L-(n-n)\sigma - x_n}{L-N\sigma}\right)^{N-n-1}. \tag{7}$$

Applying Eq. (4) again, one gets the pair correlation function:

$$g_L(r) = \frac{N}{L}\frac{N-1}{L-N\sigma} \sum_{n=1}^{N-1} \binom{N-2}{n-1} \left(\frac{x_n - n\sigma}{L-N\sigma}\right)^{n-1} \left(\frac{L-(N-n)\sigma - x_n}{L-N\sigma}\right)^{N-n-1}. \tag{8}$$

From this equation the thermodynamic limit can be obtained as the conditional limiting case when both N and L tend to infinity but the density, $\varrho = N/L$, is kept constant:

$$g(r) = \lim_{N\to\infty} \frac{N-1}{[N(1-\sigma\rho)]^{N-1}} \sum_{n=1}^{N-1} \binom{N-2}{n-1} (x_n\varrho - n\rho\sigma)^{n-1}[N(1-\sigma\rho) - (x_n\varrho - n\rho\sigma)]^{N-n-1}. \tag{9}$$

For hard-core fluids the virial equation of state includes only one value of the $g(r)$ function, viz. that assumed at distance $r=\sigma$:

$$\frac{p}{\rho k_B T} = 1 + \rho\sigma g(\sigma). \tag{10}$$

The value of $g(\sigma)$ can be given immediately on the basis of Eq. (8), since the contribution to the pair correlation function at contact can be given only by considering particle no. 1. Thus, for $x_1 = \sigma$ and $n=1$, we have from Eq.(8):

$$g(\sigma) = \frac{N-1}{N}\frac{1}{1-\rho\sigma}, \tag{11}$$

which, when $N \to \infty$, yields:

$$g(\sigma) = \frac{1}{1-\varrho\sigma}. \tag{12}$$

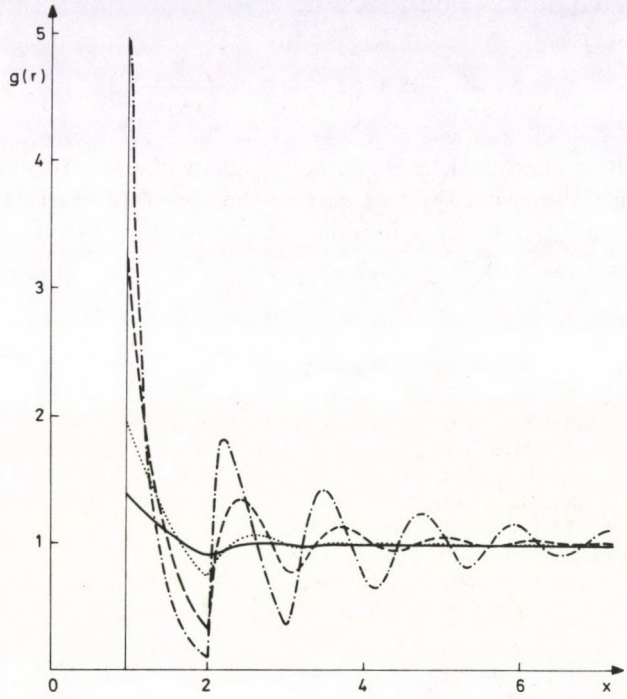


Fig.3. Pair correlation function of a one-dimensional hard-core fluid for $L = 100\sigma$ and various densities N/L (— : 0.3, · · · · · : 0.5, – – – – – : 0.7, –. – . – . : 0.8)

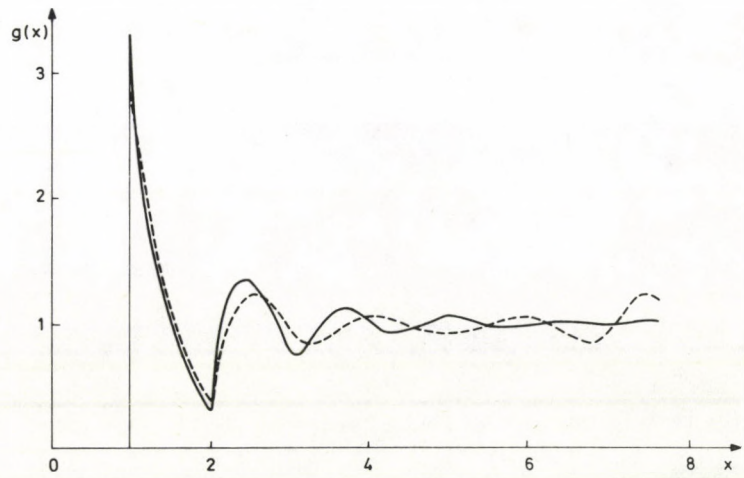


Fig.4. Illustration of the rapid convergence of the $g(x)$ function. For $\sigma = 1$ and $\rho = 0.7$, the dashed curve represents the case when $N = 7$ on $L = 10\sigma$ while the full curve corresponds to $N = 140$ on $L = 200\sigma$. The latter remains practically unaltered with the further simultaneous increase of N and L

This, when inserted into Eq. (10), leads to:

$$\frac{p}{\varrho k_B T} = 1 + \frac{\varrho\sigma}{1-\varrho\sigma} = \frac{1}{1-\varrho\sigma}. \tag{13}$$

This result is exact and, as seen, independent of the values of N and L. It depends only on the density $\rho = N/L$ and on the size of the particles.

In Figs 3 and 4, some $g(r)$ functions are shown which were calculated with various values of N, L, and σ.

References

1. C.A. Croxton, Liquid State Physics, Cambridge University Press, 1974;
J.P. Hansen, L.R. McDonald, Theory of Simple Liquids, Academic Press, New York, 1976.
2. E.Thiele, J. Chem. Phys., *39*, 474, 1963;
M.S. Wertheim, Phys.Rev. Letters, *10*, 321, 1963.
3. J.L. Tallon and R.M.J. Cotteril, Aust. J. Phys., *38*, 209, 1985.
4. L. Tonks, Phys. Rev., *50*, 955, 1934;
K.F. Herzfeld and M.G. Mayer, J. Chem. Phys., *24*, 38, 1934.
5. A. Baranyai and I. Ruff, J. Chem. Ed., to be published.

Acta Physica Hungarica 64 (1–3), pp. 73–90 (1988)

CONFORMAL STRUCTURES IN 2-DIMENSIONAL ISING-TYPE MODELS*

A. PATKÓS

Department of Atomic Physics, Roland Eötvös University
1088 Budapest, Hungary

(Received 8 January 1987)

The two-dimensional Ising-model is inexhaustible. Its role in discovery and demonstration of the importance of conformal invariance of two dimensional critical systems is reviewed. Numerical solution of an altered Ising model is described, which suggests unconventional new conformal structures.

1. Introduction

The concept of scale invariance is of central importance in understanding critical phenomena. The scale invariant fixed point Hamiltonian and the scaling fields describing deviations from it provide the framework, where universality of critical exponents follows a most natural way.

Although many of the critical exponents were found to be simple rational numbers, the general fluctuational theory of continuous phase transitions does not account for this feature properly. None of the calculational procedures (high temperature series, ε- expansion, etc.) is able to deal with this circumstance adequately.

Most of the known results refer to exponents characterizing the behaviour of the order parameter or its simplest (two-point) correlations. An elaborate theory of correlations of arbitrary complex composites built out of the order parameter is missing. In other words no general statement is known on the number of independent scaling dimensions (anomalous dimensions).

Several features of the explicit solution of the Ising model deviate substantially from the results of the mean field theory of Landau. This was one major motivation for developing the modern theory of critical phenomena.

A careful investigation of this model led L. P. Kadanoff in 1969 to the formulation of the so-called reducibility hypothesis [1], which is equivalent to the short distance expansion proposed by K. Wilson [2] in quantum field theory about the same time. According to this proposition the product of two scaling fields $O_\alpha(x)$ and $O_\beta(y)$ at nearby points x and y can be replaced by the linear combination of a complete set of scaling fields, with (maybe singular) coefficients:

$$\lim_{x\to y} O_\alpha(x)O_\beta(y) \sim \sum_\gamma C_{\alpha\beta\gamma}(x-y)O_\gamma(x). \tag{1}$$

*Dedicated to Prof. G.Marx on his 60th birthday

A. M. Poljakov has suggested in 1970 [3] that the algebraic structure of the coefficients $C_{\alpha\beta\gamma}$ should reflect the conformal symmetry of critical systems. The algebra of the generators of this symmetry, which contains scale transformation as special case, was found by M. Virasoro for 2-dimensional systems [4]. This is an infinite-dimensional structure (in contradistinction to higher-dimensions , where the conformal group is finite dimensional). The generators $L_n (n = ...-2,-1,0,1,2,...)$ fulfil the relation

$$[L_n, L_m] = (n-m)L_{n+m} + \frac{c}{12}\delta_{n,-m} n(n^2-1), \tag{2}$$

where the number c is called the central charge of the Virasoro algebra.

At the end of the seventies mathematicians took up the idea and the representation theory of the two-dimensional conformal group has been worked out by Kac [5] and Feigin and Fuks [6]. This development has initiated a renewed interest in the general theory of two-dimensional critical systems starting with the fundamental paper of Belavin, Poljakov and Zamolodchikov in 1984 [7]. This theory leads naturally to rational critical exponents and finds a class (the so-called minimal degenerate models) which is characterized by finite number of independent scaling fields. In particular, it turns out that the Ising model is the simplest unitary representative of conformal field theories, possessing only 3 independent scaling fields. The exponents derived from representation theory reproduce the values found in the explicit solution. In an important development Friedan, Qiu and Shenker [8] have found all those unitary representations with central charge smaller than unity which might be realized by two-dimensional critical systems.

What remained an art is just the way, how to identify a certain model with a member of the catalogue. In a series of papers J. L. Cardy [9] has suggested that exploiting the conformal mapping of the infinite system onto a finite width, infinite length strip allows the direct computation of the scaling exponents as the spectrum of the transfer operator of the strip. The correctness of this idea was demonstrated in analytic studies of the Ising [10] and numerical investigations of the 3-state Potts [11] models, where the identification was guessed already in the pioneering works [7], [19]. Recently, it has been understood which piece of the spectrum of the finite width strip transfer provides information also on the actual value of the central charge [12], [13].

Currently much attention is being payed to the effect of boundary conditions on the actual representation content appearing in the spectrum of the strip [14], [15], [16]. It is again the Ising model where the most detailed test of conformal invariance of systems with continuously tuned boundary conditions has been performed [17], [18]. A careful finite size analysis has revealed original conformal structures [36], which appear at first sight to be incompatible with the classification given in [8]. It is to be seen, whether this structure is a peculiarity of the Ising case or once more a more general new phenomenon is signalled when reanalyzing this classic system from a new point of view.

In Section 2 the analysis is summarized which led to the formulation of the reducibility hypothesis. Section 3 introduces the most important elements of the

2-dimensional conformal algebra and its representation theory. Correspondence between the spectra of scaling dimensions and levels of Hamilton operators of finite width systems is established in Section 4. Analytic results for finite quantum Ising chains serve for illustration, also of the effect of boundary conditions on the representation content.

The altered Ising model is introduced in Section 5. It represents a realisation of continuously tunable boundary conditions. Modifications of the Kadanoff analysis of Section 1 are discussed. Numerical results for the corresponding strip Hamiltonian are presented in Section 6, providing evidence for conformal invariance in this altered model, too. The appearing conformal structures show some puzzling features. Our conclusions are presented in Section 7.

2. The reducibility hypothesis

An analytic expression for the thermodynamical averages of products of variables $D_\gamma(y)$, ordered along a straight infinite line has been derived by Kadanoff and Ceva [20]. They have considered the "elementary" order parameter field (σ) defined on lattice sites, the dual disorder field (μ) on sites in the dual lattice. The slight displacements of these fields are of no importance in the critical point, when the correlation length is infinite. In addition to these, the fermion (ψ) and antifermion ($\overline{\psi}$) fields introduced by Kaufman [21] were considered and for completeness a separate notation for the unit operator has been introduced. The identification of these fields proceeds through a purposeful choice of the γ indices:

$$D_\gamma(y), \quad \gamma = \left\{ \begin{array}{c} 0 \\ 1/2 \\ -1/2 \\ 1 \\ -1 \end{array} \right\} \quad <=> \quad \left\{ \begin{array}{l} \text{unit operator} \\ \text{order parameter} \\ \text{disorder parameter} \\ \text{fermion} \\ \text{antifermion} \end{array} \right\}. \tag{3}$$

A quantity Γ_N is associated with each ordered product of N factors:

$$O_N = \prod_{j=1}^{N} D_{\gamma_j}(y_j), \quad y_j > y_{j-1} \tag{4}$$

which can be computed recursively

$$\Gamma_{j+1} = \Gamma_j + (-1)^{2\Gamma_j}\gamma_{j+1}, \quad \Gamma_1 = \gamma_1 \, . \tag{5}$$

The second term on the right hand side of Eq. (5) is called the "charge" of the variable at site y_{j+1}

$$q_{j+1} = (-1)^{2\Gamma_j}\gamma_{j+1}, \tag{6}$$

which gives for Γ_N

$$\Gamma_N = \sum_{j=1}^{N} q_j . \tag{7}$$

The correlation functions $< O_N >$ vanish unless $\Gamma_N = 0$. If the system is "neutral" a Coulomb-gas expression is valid:

$$< O_N >= e^{\Lambda}, \quad \Lambda = \sum_{j>i} q_i q_j \ln(y_j - y_i), \quad \Gamma_N = 0. \tag{8}$$

Consider specifically the nonzero two-point functions:

$$\begin{aligned} < D_{1/2}(y_2) D_{1/2}(y_1) > &=< D_{-1/2}(y_2) D_{-1/2}(y_1) > = (y_2 - y_1)^{-q_{1/2}^2} , \\ < D_1(y_2) D_{-1}(y_1) > &=< D_{-1}(y_2) D_1(y_1) > \quad = (y_2 - y_1)^{-q_1^2} . \end{aligned} \tag{9}$$

These expressions are compared to the general form of the critical correlation functions in two dimensions:

$$< \varphi(y_2)\varphi(y_1) > \sim (y_2 - y_1)^{-2x_\varphi}, \tag{10}$$

where x_φ is the scaling (or anomalous) dimension of the field φ. One concludes that

$$x_0 = 0, \quad x_{1/2} = x_{-1/2} = 1/8, \quad x_1 = x_{-1} = 1/2. \tag{11}$$

The next question is: Could one build new independent scaling fields out of those appearing in the list (3). The answer is given by the "fusion rule" of Kadanoff [1]. According to it the product of two D_γ's at infinitesimally separated points is to be replaced in all correlation functions by a single field whose index and charge are completely determined by the corresponding data of the factors:

$$\begin{aligned} D_\alpha(y + O) D_\beta(y - O) &\sim D_\gamma(y), \\ \gamma &= \alpha + (-1)^{2\alpha}\beta, \\ q_\gamma &= q_\alpha + (-1)^{2\alpha} q_\beta . \end{aligned} \tag{12}$$

This is equivalent to the short-distance expansion of Wilson [2]. Kadanoff has conjectured the existence of analogous expansion and operator algebra in general. He called this suggestion the reducibility hypothesis.

The complete list of the fields in case of the Ising model can be given by considering first the product of n fermions (antifermions). The composite field will have charge $n(-n)$. Adding to them an order or disorder field, operators with all kinds of halfinteger charges will occur. Any other combination leads to one of the above composite operators. Their scaling dimensions are found by the formula $x_\gamma = 1/2q_\gamma^2$ as before. Three infinite families are found, whose entries differ by an integer from the primary values.

The idea that it is the conformal symmetry of the critical system which is behind this tremendous simplification was raised by Polyakov [3]. It took however about 13 years until the infinite families were shown to be the simplest irreducible representatives of the two-dimensional conformal group.

3. Scaling fields form conformal blocks

The extension of the scale invariance of the fixed point Hamiltonian governing continuous phase transitions is the concept of its conformal invariance [22]. Its variation under a general (non-conformal) coordinate transformation $\alpha_\mu(x)$ in the two-dimensional plane is expressed through the weighted integral of the corresponding energy-momentum tensor:

$$\delta H = -\frac{1}{2\pi}\int d^2x \partial^\mu \alpha^\nu(x) T_{\mu\nu}(x), \qquad \delta x^\mu = \alpha^\mu(x). \tag{14}$$

Let us consider an $\alpha^\mu(x)$ which coincides with a conformal transformation $\varepsilon(z)$ within a circle $C(z = x + iy)$ and is identically zero outside C. The variation of the correlation functions of the theory due to the change (14) is expressed through a contour integral over $T_{\mu\nu}(x)$ as follows:

$$-\int_c n^\mu \varepsilon^\nu(z_s) < T_{\mu\nu}(x_s)\varphi_1(x_1)....\varphi_n(x_n) > ds =$$

$$= \sum_j \left\{ \varepsilon'(z_j)\Delta_j + \varepsilon(z_j)\frac{\partial}{\partial z_j} + \overline{\varepsilon'(z_j)}\overline{\Delta}_j + \overline{\varepsilon(z_j)}\frac{\partial}{\partial \overline{z}_j} \right\} < \varphi_1(x_1)....\varphi_n(x_n) > . \tag{15}$$

Here n^μ is a unit vector orthogonal to the line-element of the contour, x_s denotes points on the contour, $z = x + iy, \varepsilon'(z) = d\varepsilon/dz, \overline{\varepsilon(z)}$ is complex conjugate to $\varepsilon(z)$. Δ_j and $\overline{\Delta}_j$ are characteristic constants for the field φ_j. Note that $\overline{\varepsilon(z)} \neq \varepsilon(\overline{z})$, that is in general not only real analytic transformations are allowed.

The relation (15) is the conformal Ward identity. The equations (14) and (15) emphasize the role of $T_{\mu\nu}$ in determination of conformal properties of a critical system. This symmetric, traceless 2x2 matrix, which has vanishing divergence becomes particularly simple in the complexified coordinates $z = x + iy$ and $\overline{z} = x - iy$:

$$T = \begin{pmatrix} T_{zz} & 0 \\ 0 & T_{\overline{z}\overline{z}}, \end{pmatrix} \quad \partial_{\overline{z}} T_{zz} = \partial_z T_{\overline{z}\overline{z}} = 0. \tag{16}$$

An immediate consequence of Eq. (16) is that the components of the energy-momentum tensor are analytic functions

$$T_{zz} = T(z), \quad T_{\overline{z}\overline{z}} = \overline{T}(\overline{z}). \tag{17}$$

Therefore the product $T(z)\varphi(z_1, \overline{z}_1)$ (and analogously $\overline{T}(\overline{z})\varphi(z_1, \overline{z}_1)$) can be expanded into Laurent-series with respect to z around the point z_1:

$$T(z)\varphi(z_1, \overline{z}_1) = \sum_{n=-\infty}^{\infty} (z - z_1)^{-2-n}(\hat{L}_n \varphi(z_1, \overline{z}_1)). \tag{18}$$

This expansion defines the operators $\hat{L}_n(-\infty \leq n \leq \infty)$ through the inverse transformation

$$\hat{L}_n\varphi(z_1,\overline{z}_1) = \frac{1}{2\pi i}\oint_{C(z_1)} dz(z-z_1)^{n+1}T(z)\varphi(z_1,\overline{z}_1). \tag{19}$$

The conformal Ward identity restricts strongly the non-zero coefficient functions $\hat{L}_n\varphi(z_1,\overline{z}_1)$, when Eq.(15) is rewritten into a local form. For a conformal transformation of the z-coordinate it is equivalent to the short distance expansion

$$T(z)\varphi(z_1,\overline{z}_1) = \frac{\Delta}{(z-z_1)^2}\varphi(z_1,\overline{z}_1) + \frac{1}{z-z_1}\frac{\partial}{\partial z_1}\varphi(z_1,\overline{z}_1) + \text{regular terms.} \tag{20}$$

Similar relations are derived for $\overline{T}\cdot\varphi$, with $\overline{\Delta}$ replacing Δ, in variable $\overline{z}$. From the comparison of (18) and (20) follows that

$$\begin{aligned} &\hat{L}_n\varphi(z,\overline{z}) = 0, && \hat{\overline{L}}_n\varphi(z,\overline{z}) = 0, && \text{if } n>0,\\ &\hat{L}_0\varphi(z,\overline{z}) = \Delta\cdot\varphi(z,\overline{z}), && \hat{\overline{L}}_0\varphi(z,\overline{z}) = \overline{\Delta}\cdot\varphi(z,\overline{z}), \end{aligned} \tag{21}$$

where $\hat{\overline{L}}_n$ are the operators defined in the Laurent expansion of $\overline{T}(\overline{z})$.

A field with property (21) is called primary. An infinite family can be constructed starting from it, by the application of all types of combinations of the operators $\hat{L}_{-n}$ and $\hat{\overline{L}}_{-n}$ $(n>0)$. The evaluation of the subsequent applications of $\hat{L}$-operators to φ requires the information on the short distance expansion of the product $T(z)T(z'),(\overline{T}\cdot\overline{T})$:

$$T(z)T(z') = \frac{2}{(z-z')^2}T(z') + \frac{1}{z-z'}\frac{\partial}{\partial z'}T(z') + \frac{c/2}{(z-z')^4} + \text{regular terms.} \tag{22}$$

Relative to Eq. (21) the new feature is the presence of the inverse quartic power proportional to the conformal anomaly number or central charge (c). Then a double use of Eq. (19) followed by a lengthy algebra leads to the algebra of the $\hat{L}_n$ operators [4]:

$$[\hat{L}_n,\hat{L}_m] = (n-m)\hat{L}_{n+m} + \frac{c}{12}\delta_{n,-m}n(n^2-1). \tag{23}$$

The algebra determines the $\hat{L}_0(\hat{\overline{L}}_0)$ eigenvalues of the secondary fields:

$$\hat{L}_0(\hat{L}_{-n}\varphi(z_1,\overline{z}_1)) = (\Delta+n)\varphi(z_1,\overline{z}_1). \tag{24}$$

A whole family of secondary fields is then found associated with the primary field φ

$$\left\{\varphi,\hat{L}_{-1}\varphi,\hat{\overline{L}}_{-1}\varphi,\hat{L}_{-2}\varphi,\hat{\overline{L}}_{-2}\varphi,\hat{L}^2_{-1}\varphi,\hat{\overline{L}}^2_{-1}\varphi,\hat{L}_{-1}\hat{\overline{L}}_{-1}\varphi,\ldots.\right\} \tag{25}$$

which have $\hat{L}_0$ and $\hat{\overline{L}}_0$ eigenvalues differing by integers from Δ and $\overline{\Delta}$. This conformal block realizes an infinite dimensional irreducible representation of the direct

product group $O \times \overline{O}$ where O stands for conformal transformations in variable z and $\overline{O}$ for transformations in $\overline{z}$.

The meaning of Δ and $\overline{\Delta}$ becomes clear when studying the correlation function of a primary field. The differential operator representation of $\hat{L}_{\pm 1}$, $\hat{L}_0$ [7] determines fully its functional form:

$$< \varphi(z_1, \overline{z}_1)\varphi(z_2, \overline{z}_2) >= (z_1 - z_2)^{-2\Delta}(\overline{z}_1 - \overline{z}_2)^{-2\overline{\Delta}}. \tag{26}$$

For secondary fields the powers are replaced by $2(\Delta + n)$ and $2(\overline{\Delta} + n)$.

When looking back to our Ising example, one has to ask what are the conditions under which the conformal group admits just three irreducible representations to be realised in a continuum theory? The values of the highest weights should be chosen to reproduce the correlations appearing in (9):

$$\begin{aligned} &\Delta_\sigma = \overline{\Delta}_\sigma = 1/16, \\ &\Delta_\psi = 1/2, \qquad\qquad \overline{\Delta}_\psi = 0, \\ &\Delta_{\overline{\psi}} = 0, \qquad\qquad \overline{\Delta}_{\overline{\psi}} = 1/2. \end{aligned} \tag{27}$$

The representation theory of the Virasoro algebra was developed by mathematicians to the end of the seventies. In particular, Kac [5], and Feigin and Fuks [6] have found a series of representations with $c < 1$, where the number of primary fields is finite:

$$\begin{aligned} c &= 1 - \frac{6}{m(m+1)}, \\ \Delta_{p,q} &= \frac{[(m+1)p - mq]^2 - 1}{4m(m+1)}, \qquad 1 \le p \le m-1, \quad 1 \le q \le p. \end{aligned} \tag{28}$$

In 1984 Friedan et al proved [8] that the series parametrized by the integer, m, provides all unitary representations with $c < 1$.

If the representations of the conformal algebra are of any relevance for the critical Ising model, one has to find the specific m value yielding by (28) just those dimensions, which appear in (27). Choosing $m = 3$, one has indeed

$$c = 1/2, \quad \Delta_{1,1} = 0, \quad \Delta_{2,1} = 1/2, \quad \Delta_{2,2} = 1/16. \tag{29}$$

This means that the conformal block of the unit operator realizes the $(0, 0)$, that of the order-parameter the $(1/16, 1/16)$, the fermion the $(1/2, 0)$, while the antifermion the $(0, 1/2)$ representation of the Virasoro algebra with central charge $c = 1/2$.

A few other models have been also identified with representations characterized by low values of m, but it is still a question of "phenomenological" investigation to find out for an explicit lattice model what its conformal representation content at criticality is.

4. Lattice phenomenology of the scaling dimension spectra

Finite conformal maps can be used to relate the correlation functions in the infinite plane to those measurable in restricted geometries. The most convenient among them is the logarithmic map

$$w(z) = \frac{L}{2\pi} \ln z \tag{30}$$

($w = u + iv$), which relates the original correlations to those measurable in an infinitely long strip of width L (Fig. 1).

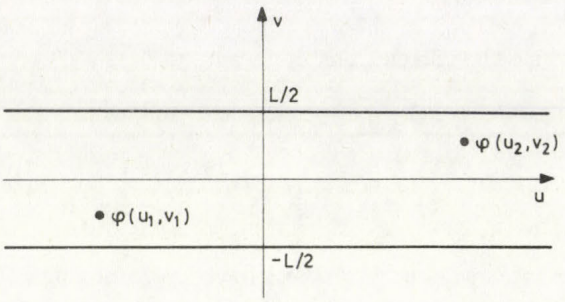


Fig. 1. Correlation function in the finite width strip

The transformation law for the two-point function is the following [9]:

$$< \varphi(z_1, \overline{z}_1)\varphi(z_2, \overline{z}_2) >= |w'(z_1)|^x \, |w'(z_2)|^x < \varphi(w_1, \overline{w}_1)\varphi(w_2, \overline{w}_2) > \tag{31}$$

($w_i = w(z_i)$). Assuming that φ is primary, one makes use of Eqs (26) and (30) to derive

$$<\varphi(u_1, v_1)\varphi(u_2, v_2) >= \left(\frac{2\pi}{L}\right)^{2x} \left[2 \cos h \frac{2\pi}{L}(u_1 - u_2) - 2 \cos \frac{2\pi}{L}(v_1 - v_2)\right]^{-x}$$

$$= \left(\frac{2\pi}{L}\right)^{2x} \sum_{m,\overline{m}} a_m a_{\overline{m}}$$

$$\exp\left[-\frac{2\pi}{L}(x + m + \overline{m})(u_1 - u_2) + i\frac{2\pi}{L}(s + m - \overline{m})(v_1 - v_2)\right], \tag{32}$$

where $x = \Delta + \overline{\Delta}$, $s = \Delta - \overline{\Delta}$.

The strip correlation function can be represented directly in terms of the eigenstates of its transfer operator:

$$< \varphi(u_1, v_1)\varphi(u_2, v_2) >=$$

$$= \sum_n \sum_k < 0 \,|\hat{\varphi}(v_1)|\, n,k > e^{-(E_n - E_0)(u_1 - u_2)} < n,k \,|\hat{\varphi}(v_2)|\, 0 > . \qquad (33)$$

In this expression $\{E_n\}$ denotes the energy eigenvalues of the eigenstate $|n,k>$ of the transfer operator. A degeneracy index k is also introduced. The eigenstate belonging to the largest eigenvalue is denoted by $|0>$. The operator $\hat{\varphi}(v)$ coincides with the field $\varphi(0,v)$ in the representation where the eigenvectors are functionals of this field.

Comparing (32) and (33) one finds the correspondence between the following sets of quantities:

$$\begin{aligned} \{E_n - E_0\} \quad &\Longleftrightarrow \left\{\frac{2\pi}{L}(x + m + \overline{m})\right\}, \\ \{< 0 \,|\hat{\varphi}(v_1)|\, n,k >\} \quad &\Longleftrightarrow \left\{a_1 \exp\left(i\frac{2\pi}{L}(s + m - \overline{m})v_1\right)\right\}. \end{aligned} \qquad (34)$$

Then the degeneracy index k is equated to $s + m - \overline{m}$, but it is more important that the first equality provides a general foundation for the famous gap-amplitude relation occasionally found earlier in specific case studies [23].

The extremely interesting conclusion is that there is a one-to-one correspondence between the states of the Hilbert space contributing to the 2-point function of a primary field and the fields in the conformal block of that field. This correspondence allows to find the spectrum of the scaling dimensions of a model by computing the eigenvalue spectrum of the transfer operator of the continuum field theory on a strip.

The transfer operator for infinitesimal time intervals is simply related to the Hamilton operator of the quantum field theory defined on a line of length L

$$\hat{T}(\Delta t) = 1 - \Delta t \hat{H}. \qquad (35)$$

So one is led to the problem of finding the spectrum of this Hamilton operator.

A practical procedure is to consider first the lattice regulated version of $\hat{H}$, where N discrete points will represent a section of the strip along the v-direction in Fig. 1. As $\hat{H}$ belongs to a conformal invariant theory the only dimensional scale for it is provided by the lattice constant a_s:

$$\hat{H} = \frac{1}{a_s}\hat{W}, \qquad (36)$$

where $\hat{W}$ is some dimensionless operator.

For concreteness we return to the Ising case, where the Hamiltonian of the critical lattice system is given by

$$\hat{W} = -\sum_{i=1}^{N} \sigma_i^x - \sum_{i=1}^{N} \sigma_i^z \sigma_{i+1}^z, \qquad (37)$$

supplemented with a periodic boundary condition. From finite size scaling theory [24] it is known that in the critical point all energy gaps of this finite system are inversely proportional to N:

$$w_n - w_0 = \frac{2\pi}{N} H_n. \tag{38}$$

Substituting (38) back into (36) and using $L = N \cdot a_s$ one understands that amplitudes A_n, which can be determined by diagonalizing (37) give an estimate for the spectra of the scaling dimensions which in the limit $N \to \infty, L = N \cdot a_s =$ fixed should provide us with their true values if the lattice regularisation makes sense.

Fortunately the spectrum of (37) is known [25] in analytic terms, so the Ising system is perfect for illustrative purposes. Performing a Jordan–Wigner [26] transformation one arrives at a fermionic Hamiltonian whose solution is found following the strategy of Lieb et al [27]

$$H_F = \sum_{n=1}^{N} (c_n^+ c_n - 1/2) - 1/2 \sum_{n=1}^{N-1} (c_n^+ - c_n)(c_{n+1}^+ + c_{n+1}) + \left(\frac{1}{2} - Q\right)(c_N^+ - c_N)(c_1^+ + c_1). \tag{39}$$

$Q = 0$ denotes the energy sector (excitations built of an even number of fermionic states), while $Q = 1$ is for the spin sector with an odd number of filled fermionic levels. The eigenvalues for periodic chains were found by Pfeuty [25]

$$\Lambda_k = 2 \sin k/2, \qquad k = \frac{2m}{N}\pi, \qquad \text{if } Q = 1,$$
$$, \qquad k = \frac{2m+1}{N}\pi, \qquad \text{if } Q = 0 \tag{40}$$

$(m = 0, 1 \ldots . N - 1)$. The lowest excited state in the $Q = 1$ sector is given by filling the $k = 0$, $\Lambda_0 = 0$ level. Its energy is given purely by the difference of zero-point fluctuation energies in the two sectors:

$$E_0(Q = \underline{1}) = -\frac{1}{2} \sum_{m=0}^{N-1} \Lambda(k = \frac{2\pi}{N} m) = -\cot \frac{\pi}{2N},$$
$$E_0(Q = 0) = -\frac{1}{2} \sum_{m=0}^{N-1} \Lambda(k = \frac{2m+1}{N}\pi) = -(\sin \frac{\pi}{2N})^{-1}, \tag{41}$$
$$A_{\text{gap}} \cdot \frac{2\pi}{N} = \lim_{N\to\infty} (E_0(Q = 1) - E_0(Q = 0)) = \lim_{N\to\infty} \tan \frac{\pi}{4N}.$$

The lowest gap in the $Q = 0$ sector is given by

$$A_{\text{energy}} \cdot \frac{2\pi}{N} = \Lambda(k = \frac{\pi}{N}) + \Lambda(k = \frac{2N-1}{N}\pi) = 4 \sin \frac{\pi}{2N}. \tag{42}$$

When one performs the continuum limit $N \to \infty$

$$A_{\text{gap}} = 1/8, \qquad A_{\text{energy}} = 1 \tag{43}$$

is found. The first reproduces the result for $\Delta_{2,2}$ in (29), if taking into account that $A_{\text{gap}} = \Delta_{2,2} + \overline{\Delta}_{2,2}$, while the second result is consistent with the fermion scaling dimensions given in (29), when one notices that the energy operator is a fermion-antifermion composite: $\overline{\psi}\psi$.

Another important consequence of conformal invariance concerns the leading finite size correction to the ground state energy. It follows from the anomalous transformation of the energy-momentum tensor-components under conformal maps:

$$T(z) \longrightarrow |\, w'(z)\,|^2\; T(w) + \frac{c}{2}\left[\frac{d^3w}{dz^3}\Big/\frac{dw}{dz} - \frac{3}{2}\left(\frac{d^2w}{dz^2}\Big/\frac{dw}{dz}\right)^2\right].$$

That is under the logarithmic map (30) one finds

$$T \longrightarrow \left(\frac{L}{2\pi z}\right)^2 T + \frac{c}{24z^2}, \qquad \overline{T} \longrightarrow \left(\frac{L}{2\pi \overline{z}}\right)^2 \overline{T} + \frac{c}{24\overline{z}^2}. \tag{44}$$

Using the expression of the L_0 and $\overline{L}_0$ generators as Cauchy integrals over $T(z)$ and $\overline{T}(\overline{z})$, one finds for the generators in the strip

$$L_0 + \overline{L}_0 \longrightarrow \frac{L}{2\pi}\left[\frac{1}{2\pi}\int_0^L (T + \overline{T})dv\right] + \frac{c}{12}. \tag{45}$$

The quantity in the square bracket on the right-hand side is the Hamiltonian. Therefore the application of the Hamilton operator on the ground state

$$H \mid 0 >= \frac{2\pi}{L}(\overline{L}_0 + L_0) \mid 0 > -\frac{\pi c}{6L} \mid 0 >= -\frac{\pi c}{6L} \mid 0 > . \tag{46}$$

Using the expression for $E(Q = 0)$ (Eq. (42)) one checks the validity of (46) in the lattice regularised Ising field theory:

$$\frac{\pi c}{6L} \stackrel{?}{=} \lim_{N\to\infty} \frac{1}{a_s}\frac{1}{\sin\frac{\pi}{2N}} = \frac{2N}{\pi a_s} + \frac{1}{2}\cdot\frac{\pi}{6L} + \mathrm{O}(L^{-2}). \tag{47}$$

The first term on the right hand side diverges when $a \to 0$. It has to be subtracted in the procedure of the regularisation. The equality of the $\mathrm{O}(L^{-1})$ terms reproduces $c = 1/2$ (c.f. Eq.(29)).

The strategy described in this Section for the identification of the irreducible representations of the conformal group appearing in lattice models has been applied extensively to some famous models of statistical physics, like the 3-states Potts [28] or the Ashkin–Teller model [29]. It was suggested also that the evidence for or against conformal invariance will be decisive in disputes about the nature of the phase transition in specific models, like the 6-state cubic model [30].

5. Algebra of exponents in a modified Ising model

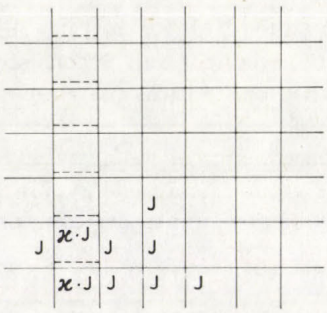


Fig. 2. The altered Ising model in the infinite plane

Consider the critical Ising model in the infinite plane. Let us assume that couplings along a straight ladder are changed from J_c to $\kappa \cdot J_c$ (Fig.2). This change is a surface effect, which allows to introduce exponents characterizing correlations along the defect line [31, 32]. The corresponding exponents vary continuously with κ. The situation was understood to be a consequence of the marginal nature of the surface-energy field $(\sigma(\mathbf{x})\sigma(\mathbf{x}+\mathbf{e}a_s))$ [33]. The modifications are restricted to the spin sector exclusively, where they can be stated the most compact way using the multipoint correlations introduced in Section 2 [34]. Relations (5) and (8) expressing the operator algebra and the correlations in terms of the charges are unchanged. The charges however are "renormalized":

$$|q_\gamma| = \begin{cases} \gamma, & \text{if} \quad \gamma = \text{integer}, \\ \gamma + f_\gamma(\kappa), & \text{if} \quad \gamma = \text{halfinteger}. \end{cases} \tag{48}$$

Only 1 independent function $f_\gamma(\kappa)$ appears because the operator algebra relates the others to a single one. For example the product of an order and a disorder operator being a fermion one has

$$|q_{1/2}| + |q_{-1/2}| = 1, \tag{49}$$

which means by the relation $x = q^2/2$

$$\sqrt{2x_\sigma(\kappa)} + \sqrt{2x_\mu(\kappa)} = 1. \tag{50}$$

This relation was first recognized by Brown [37]. Also the series of halfinteger charged composites is seen to possess modified charges:

$$\begin{aligned} \sqrt{2x_\sigma(n,\kappa)} &= \sqrt{2x_\sigma(\kappa)} + n = n + 1 - \sqrt{2x_\mu(\kappa)}, \\ \sqrt{2x_\mu(n,\kappa)} &= \sqrt{2x_\mu(\kappa)} + n. \end{aligned} \tag{51}$$

As $f_\gamma(\kappa)$ takes values continuously depending on κ, the exponents appearing in (51) are all independent, so it is suggestive to assume, that one is dealing with infinite number of primary fields. Before that stage, however, one has to answer very basic questions concerning the validity of conformal invariance in this situation.

Does one find the same spectrum of scaling exponents from the Hamiltonian obtained for the system arising from a logarithmic map, for instance? Do the above independent dimensions possess their own conformal blocks? Are the level degeneracies those which follow from character formulae for irreducible representations [35]?

Affirmative answers to some of these questions have been given in partial investigations [17, 18]. A complete numerical study of scaling dimensions followed by the construction of the spectrum generating algebra was attempted in [36] and [37]. It led to firmly established conformal structures, which are related in a very original way to the characterisation presented above for the perfect Ising system.

6. Conformal strip phenomenology for the altered Ising model

The logarithmic map of the infinite plane with a straight defect line leads to a strip containing two parallel defects. The physical spectrum is built of fermion levels of the following Hamiltonian:

$$\begin{aligned} H_F = &\sum_{n=1}^{n}(c_n^+ c_n - 1/2) - 1/2\sum_{n=1}^{N-1}(c_n^+ - c_n)(c_{n+1}^+ + c_{n+1}) + \\ &+ \left(\frac{1}{2} - Q\right)(c_N^+ - c_N)(c_1^+ + c_1) + \\ &+ \frac{1}{2}(1-\kappa)\left[(c_{\frac{N}{4}}^+ - c_{\frac{N}{4}})(c_{\frac{N}{4}+1}^+ + c_{\frac{N}{4}+1}) + (c_{\frac{3N}{4}}^+ - c_{\frac{3N}{4}})(c_{\frac{3N}{4}+1}^+ + c_{\frac{3N}{4}+1})\right]. \end{aligned} \tag{52}$$

The analytic, $N \to \infty$ solution of (52) is not yet known. Its numerical diagonalisation can be achieved following steps of the procedure outlined by Lieb et al [27]. The eigenvalues were found for chains consisting of up to 210 sites. With the help of an extrapolation technique proposed by Bulirsch and Stoer [38] 8 digit stable values of the first 12 eigenvalues were obtained to follow diagonal representation

$$\frac{N}{2\pi}H_F = \sum_{r=0}^{\infty}\left\{\Lambda_r^{(1)}\eta_{r(1)}^+\eta_{r(1)} + \Lambda_r^{(2)}\eta_{r(2)}^+\eta_{r(2)}\right\},$$

$$\Lambda_r^{(1)} = \frac{1}{2} - \Delta(\kappa) + r + O(1/N), \qquad \Lambda_r^{(2)} = \frac{1}{2} + \Delta(\kappa) + r + O(1/N). \tag{53}$$

In the $Q = 0$ sector $\Delta(\kappa) = 1/2$ independently of the value of κ. In the spin sector ($Q = 1$) $\Delta(\kappa)$ was found to fit extremely well the functional form

$$\Delta(\kappa) = 1 - \frac{2}{\pi}\arctan\frac{1}{\kappa}. \tag{54}$$

Primary fields in the $Q = 0$ sector remain what they are in the perfect Ising model, implying the unchanged form (47) for the leading ground state correction.

In the sector $Q = 1$ infinitely many primary fields appear. Their gap values are given as

$$\frac{N}{2\pi} G_p^{(i)} = \sum_{r=0}^{2p} \Lambda_r^{(i)} + \Delta E. \tag{55}$$

ΔE is the difference in the zero point fluctuation energies in the sectors $Q = 1$ and $Q = 0$ (c.f. Eq.(41)). This quantity was found to fulfil the relation

$$\Delta E = \frac{1}{2}\Delta^2(\kappa). \tag{56}$$

The gaps in (55) are rewritten with help of (54) and (56) in a form which exactly reproduces the known Hamiltonian limit [39] of Eq.(48). This completes the evidence for conformal invariance present in the altered model.

One constructs with help of the single fermionic levels also the secondary states in the $Q = 1$ sector. Their degeneracy at level m is given by the number of partitions of this integer. That is the generating function has the simple expression:

$$P(q) = \sum_m d(m) q^m = \prod_{n=1}^{\infty} (1 - q^n)^{-1}. \tag{57}$$

A half-infinite defect line is mapped into a strip problem containing a single defect only. Its numerical solution was achieved along the same lines [37] as presented above. The main difference is that the diagonal form of the Hamiltonian is of the form (53) now in both ($Q = 0, 1$) sectors. The characteristic functions $\Delta_0(\kappa)$ and $\Delta_1(\kappa)$ were found analytically and are different. It is an interesting question for future investigations to see whether the form (53) and (57) is the most general structure which can be realized in Ising type systems. It is the construction of the spectrum generating algebra in terms of the eigenmodes $\eta_r(i)$, which determines the type of representation of the conformal group which appears in the defected Ising model.

Consider the following operators:

$$\mathcal{L}_n(\Delta) = \frac{1}{2} \sum_{p \in Z+1/2} (p - \Delta) : \eta_{n-p+\Delta} \eta_{p-\Delta} : \quad , \quad \eta_\alpha^+ \equiv \eta_{-\alpha}, \tag{58}$$

where : : denotes normal ordering, that is

$$: \eta_\alpha \eta_\beta := \begin{cases} \eta_\alpha \eta_\beta, & \alpha < 0, \\ -\eta_\beta \eta_\alpha, & \alpha > 0, \\ \frac{1}{2}[\eta_\alpha \eta_\beta], & \alpha = 0. \end{cases} \tag{59}$$

With help of standard anticommutation rules one shows that $\mathcal{L}_n(\Delta)$ and $\mathcal{L}_n(-\Delta)$ ($[\mathcal{L}_n(\Delta), \mathcal{L}_{n'}(\Delta')] = 0$, $\Delta' \neq \pm\Delta$) form in the special points $\Delta = 0,\ 1/2$ two separate $c = 1/2$ Virasoro algebras. The value $\Delta = 1/4$ is also of peculiar nature, because then the two infinite oscillator series of (53) collapse into a unique series of spacing π/N on absolute scale. The corresponding single Virasoro algebra has again $c = 1/2$.

Away from the specific values $\Delta = 0,\ 1/2,\ 1/4$ the algebra of the operators (58) does not close:

$$
\begin{aligned}
[\mathcal{L}_n(\Delta), \mathcal{L}_m(\Delta)] &= \frac{n-m}{2}\mathcal{L}_{n+m}(\Delta) - \frac{1}{24}\delta_{n,-m}\left(n^3 + n\left(\frac{1}{2} + 6\Delta^2(\kappa)\right)\right), \\
[\mathcal{L}_n(\Delta), \mathcal{L}_m(-\Delta)] &= \frac{n-m}{2}\mathcal{L}_{n+m}(-\Delta) - \frac{n^2}{4}\sum_{p=1/2+Z} :\eta_{n-p-\Delta}\eta_{p+\Delta}: + \\
&\quad + \frac{1}{4}\delta_{n,-m}\left(\frac{n^3}{3} + n^2\Delta + n\left(\Delta^2 - \frac{1}{12}\right)\right).
\end{aligned}
\tag{60}
$$

Introducing

$$
K_n = \mathcal{L}_n(\Delta) + \mathcal{L}_n(-\Delta) \tag{61}
$$

the two extra terms of the computing structures cancel and a single Virasoro algebra is found

$$
[K_n, K_m] = (n-m)K_{n+m} + \delta_{n,-m}\frac{1}{12}(n^3 - n(1 - 12\Delta^2)). \tag{62}
$$

It is known from the representation theory of the conformal group [40], that the central charge of the algebra is uniquely fixed by the coefficient in front of n^3 on the right hand side of (62), while the coefficient of n in the same term is fixed only by convention.

Indeed, the modified generators

$$
K'_0 = K_0 + \frac{1}{2}\Delta^2(\kappa), \qquad K'_{n\neq 0} = K_{n\neq 0} \tag{63}
$$

are seen to fulfil the standard form of a $c = 1$(!) Virasoro algebra. The representation in terms of the fermion fields (58) is unitary, as one easily convinces himself that

$$
K'^{+}_n = K'_{-n}. \tag{64}
$$

Also it is known [5] that the degeneracy of all unitary representations of $c \geq 1$ Virasoro algebras follows the pattern (57) and there are no restrictions on the value of the highest weights.

It is still not clear at this stage of the analysis what dictates the relations between the highest weights of the different conformal blocks as given by Eqs (50) and (51)? An important insight is provided by comparing the critical defected Ising system with a real boson theory supplemented with specific boundary conditions.

Following [41] we consider the free field $\Phi(x)$ defined on the interval $O \leq x \leq L$ and taking its values on a circle of radius β^{-1}:

$$0 \leq \Phi \leq 2\pi\beta^{-1}, \qquad \mathrm{mod}(2\pi\beta^{-1}). \tag{65}$$

Boundary conditions are chosen to be:

$$\Phi(0) = 0, \qquad \Phi(L) = \frac{\omega}{\beta}, \qquad \mathrm{mod}(2\pi\beta^{-1}). \tag{66}$$

The field can be expanded into Fourier series, separately in sectors labeled by the number of windings of the Φ-field on its configurational circle:

$$\Phi^{(m)}(x) = \frac{2\pi}{\beta}\left[\frac{x}{L}\left(m + \frac{\omega}{\pi}\right) + \sum_{m=1}^{\infty} a_n \sin\frac{\pi x n}{L}\right]. \tag{67}$$

The partition function of this boson theory, $Z_\beta(L,\omega,\beta)$ is compared to the partition function of the critical, defected Ising model , which is defined by the general expression

$$Z_{\mathrm{ISING}}(L,T) = \sum_{\text{primary}} \sum_{m} d_{pr}(m) \exp\left[\left(-G_{pr} - \frac{2\pi}{L}m\right)T\right]. \tag{68}$$

Using (55), (56) and (57) one finds ($q \equiv \exp(-2\pi T/L)$)

$$Z_{\mathrm{ISING}}(Q = 1, 2 - defects) = Z_B\left(\frac{L}{2}, \pi(1 + \Delta_1(\kappa)), \sqrt{\pi}\right). \tag{69}$$

Different conformal blocks contribute to (68) in a way, which is equivalent to the contribution of the different winding number sectors to the bosonic partition function. A further extension of the equivalence for fermionic fields proceeds along the lines outlined in [41].

In conclusion of this Section we state that the defected critical Ising model in the $Q = 1$ sector provides unitary representations of a single $c = 1$ Virasoro algebra (except in the specific points $\Delta = 0$, 1/4, 1/2). A posteriori it is very natural that two commuting Virasoro algebras of the periodic Ising model fuse into a single one under more general boundary conditions.

7. Conclusions

Application of conformal invariant quantum field theory to critical phenomena in 2-dimensional statistical systems is one of the brightest achievements of the theoretical physics of the eighties. Critical scaling exponents form irreducible

highest weight representations of the Virasoro algebra. Many well-known systems possessing finite number of conformal blocks appear naturally to be identifiable with unitary representations with central charge smaller than unity.

In the Ising model the operator algebra which follows from what one calls today the representation content of the $c = 1/2$ Virasoro algebra, had been discovered in very explicit terms long before the general group theoretical framework was worked out. Also, it allows an explicit analytic check of those features which reflect conformal invariance in quantities calculated for finite size systems. The extensive study of other models which led to confirmation of the representation theory is one of the most undisputable successes of lattice regularized field theories in defining continuum field theories at their critical points.

Finally, it was again the Ising model, where one has demonstrated how the boundary conditions can change even the central charge of the model. It will be interesting to see, whether this observation is extendable to other models.

The present understanding is still not fully satisfactory. One cannot foresee which representation of the conformal group will appear in the continuum limit of a certain lattice system. This question might become the focus of research in the immediate future.

References

1. L.P. Kadanoff, Phys. Rev.,*188*, 859, 1969.
2. K.G. Wilson, Phys. Rev.,*179*, 1499, 1969.
3. A.M. Poljakov, ZhETF Letters.,*12*, 538, 1970.
4. M. Virasoro, Phys. Rev.,*D1*, 2933, 1970.
5. V.G. Kac, Lecture Notes in Physics.,*94*, 441, 1979.
6. B.L. Feigin and D.B. Fuks, Funk. Anal.,*16*, 47, 1982.
7. A.A. Belavin, A.M. Poljakov and A.B. Zamolodchikov, Nucl. Phys., *B241*, 333, 1984.
8. D. Friedan, Z. Qiu and S. Shenker, Phys. Rev. Lett.,*52*, 1575, 1984.
9. J.L. Cardy, J. Phys.*A17*, L385, 1984; ibid, L961, 1984.
10. T.W.B. Burkhardt and I. Guim, J. Phys.,*A18*, L33, 1985.
11. G. von Gehlen and V. Rittenberg, J. Phys.,*A19*, L625, 1986.
12. H.W.J. Blöte, J.L. Cardy and M.P. Nightingale, Phys. Rev. Lett., *56*, 742, 1986.
13. I. Affleck, Phys. Rev. Lett.,*56*, 746, 1986.
14. J.L. Cardy, Nucl. Phys.,*B270, FS16*, 186, 1986.
15. J.-B. Zuber, Phys. Lett.,*176B*, 127, 1986.
16. J.L. Cardy, Nucl. Phys.,*B275, FS17*, 200, 1986.
17. L. Turban, J. Phys.*A18*, L325, 1985.
18. L.G. Guimaraes and J.R. Drugowich de Felicio, J. Phys.,*A19* , L341, 1986.
19. Vl. S. Dotsenko, Nucl. Phys.,*B235, FS11*, 54, 1984,
20. L.P. Kadanoff and H. Ceva, Phys. Rev.,*B3*, 3918, 1971.
21. B. Kaufman, Phys, Rev.,*76*, 1232, 1949.
22. A. Z. Patashinskii and V. L. Pokrovskii, Fluctuation Theory of Phase Transitions, Pergamon Press, Oxford., 1979.
23. J.M. Luck, J. Phys.,*A15*, 169, 1982.
24. M. E. Fisher, in Proc. of Varenna Summer School, Academic Press, New York, 1970.
25. P. Pfeuty, Ann. Phys. (N. Y.),*57*, 79, 1970.
26. E. Jordan and E. Wigner, Z. Phys.,*47*, 631, 1928.
27. E. Lieb, T. Schultz and D. Mattis, Ann. Phys. (N. Y.),*16*, 407, 1961.

28. G. von Gehlen and V. Rittenberg, J. Phys.,*A19*, L625, 1986.
29. G. von Gehlen and V. Rittenberg, J. Phys.,*A19*, L631, 1986.
30. G. von Gehlen and V. Rittenberg, J. Phys.,*A19*, 2439, 1986.
31. R.V. Bariev, ZhETF.,*77*, 1217, 1979.
32. B.M. McCoy and J.H.H. Perk, Phys. Rev.,*44*, 840, 1980.
33. A.C. Brown, Phys. Rev.,*B25*, 331, 1982.
34. L.P. Kadanoff, Phys. Rev.,*B24*, 5382, 1981.
35. A. Rocha-Cariddi, in Vertex Operators in Mathematics and Physics, ed. J. Lepovsky, Springer, Berlin, 1985.
36. M. Henkel and A. Patkós, J. Phys.*A20*, 2199, 1987.
37. M. Henkel and A. Patkós, Nucl. Phys.*B285*, 29, 1987.
38. R. Bulirsch and J. Stoer, Num. Math.,*6*, 413, 1964.
39. I. Peschel and K.D. Schotte, Z. Phys.,*54*, 305, 1984.
40. G. Segal, Comm. Math. Phys.,*80*, 301, 1981.
41. N. Craigie, W. Nahm and S. Narain, Ann. Phys. (N.Y.),*174*, 78, 1987.

Acta Physica Hungarica 64 (1–3), pp. 91–109 (1988)

THE QUARK–GLUON PHASE OF MATTER*

J. KUTI**

Department of Physics, University of California
at San Diego, La Jolla, California 92093, USA

Some important physical properties of the quark–gluon phase of matter are discussed. The transition temperature and the latent heat per unit volume for the creation of quark–gluon plasma from ordinary hadron matter are calculated from a very large scale computer simulation of Quantum Chromodynamics at finite temperatures.

1. Introduction

I dedicate this work to Professor George Marx on his 60th birthday. I have had the good fortune and privilege to be his student, coworker and close associate. What I have learned from him about physics and human curiosity will guide me for the rest of my life.

The work reported here is based on some recent results of a large scale collaboration spread over two continents. We have studied the physical conditions for the transition of ordinary hadron matter into the quark–gluon phase of matter within the framework of Quantum Chromodynamics at finite temperatures. My collaborators on the project are S.A. Gottlieb, A.D. Kennedy, S. Meyer, B.J. Pendleton, R.L. Sugar and D. Toussaint. Without their contributions I could not discuss here some interesting results and I am very thankful to them.

2. Quark matter in nature

Quarks and gluons are the known fundamental building blocks of strongly interacting hadron matter. They are organized into nucleons, and mesons which hold the nucleons together in the nucleus. Quarks and gluons first were believed to be confined to the inside of hadrons permanently. Later its was conjectured that nuclear matter at large densities or high temperatures where nucleons overlap may transform into a deconfined quark–gluon plasma phase (quark matter) where quarks are free to move around. As the quark density or the temperature of the plasma grows the interaction among quarks and gluons is expected to become negligible (asymptotic freedom) and quark matter may become asymptotically an ideal gas of quark and gluon quanta.

*Dedicated to Prof. G. Marx on his 60th birthday

**On leave from the Central Research Institute for Physics, Budapest

Acta Physica Hungarica 64, 1988
Akadémiai Kiadó, Budapest

Quark matter could only be observed under some exotic conditions. The most spectacular scenario is provided by the early universe, a few microseconds after the big bang. At that time, in the last phase transition of the early universe, at a few hundred MeV temperature, quarks and gluons became permanently bound inside hadrons, they mainly formed nucleons and the lowest mass mesons.

We have no direct access to quark matter in the early universe, because the big bang cannot be replayed. Fortunately, a little bang can be created under laboratory conditions in ultra-relativistic heavy ion collisions. The density of matter inside the nucleus is about $3 \times 10^{14} \mathrm{g/cm}^3$. The density of matter inside the nucleon is only about three times larger than ordinary nuclear density. It is, therefore, concievable that if nuclei are compressed to five or ten times their normal density, a transition may occur into the quark matter phase. In ultrarelativistic nuclear collisions the required densities, and, therefore, quark matter in bulk may be created. The high temperature, measured in hundred MeV in the central region of the collision, will significantly enhance the possibility of making quark matter in the little bang. The observation of the created quark–gluon plasma will require very clever plasma diagnostic tools.

It is interesting to note that the cores of neutron stars under very large pressure may consist of quark matter. However, to distinguish between quark matter and ordinary nuclear matter in the core of neutron stars is a very difficult astrophysical problem.

3. Quantum chromodynamics

Quantum Chromodynamics (QCD) provides a rigorous theoretical framework for the study of hadron matter under extreme conditions. The physical conditions which are required to create quark matter have to be derived from QCD.

During the last few years quark–gluon thermodynamics has been studied extensively within the framework of lattice Quantum Chromodynamics. The lattice provides a nonperturbative regularization scheme for the theory. Computer simulation (Monte Carlo methods) is the only available practical tool for calculations from first principles in lattice QCD outside the framework of perturbation theory. Perturbation theory is, of course, not applicable to the study of phase transitions.

The first computer simulation of the deconfining phase transition [1,2] and the investigation of gluon thermodynamics with SU(2) color [3] opened the way to detailed quantitative results for the realistic SU(3) color group. The computer simulations substantially strengthened the arguments of early theoretical work [4,5] on the deconfining phase transition and motivated further research with new ideas and rigorous results [6,7,8,9] .

In lattice QCD calculations one of the difficult problems is to remove the lattice cutoff effects from physical quantities. This can only be accomplished with confidence in the scaling regime of the theory where the renormalization group β function is universal and known in perturbation theory. The removal of the cutoff is

important, if we want to determine the transition temperature and the latent heat of the transition in measurable physical units, that is in units of MeV.

We believe that the deconfining phase transition with its first-order character may also be the best tool to study the continuum limit of lattice QCD on more general grounds. The determination of the transition temperature T_c is a unique test of the onset of scaling behavior, since T_c is free of cut-off dependent ultraviolet divergences. Locating the transition temperature is relatively easy because the system undergoes a sharp first-order phase transition where rounding effects are small and controlled by finite size scaling theory.

At present these calculations can be carried out only when the effects of quark vacuum polarization are neglected. The hope is that the experience gained from these studies of the pure gauge theory will be useful when improved techniques and larger computing power make inclusion of dynamical fermions practical. In fact, computer simulations in the presence of dynamical quark effects are becoming feasible now on medium size lattices.

The partition function for the Euclidean Wilson action $S_E(U)$ on the lattice is defined by the functional integral

$$Z = \int \prod_{x,t,\mu} dU \ \exp\left[-\beta S_E(U)\right], \tag{3.1}$$

where the integral in Eq. (3.1) is with respect to the Haar measure and the lattice coupling is $\beta = \frac{6}{g^2}$. The Wilson action S_E is defined as a sum over all unoriented plaquettes,

$$S_E(U) = \sum_{\text{plaquettes}} (1 - \frac{1}{3} \text{ Re Tr } UUUU). \tag{3.2}$$

The lattice spacing a is a known function of the bare coupling g in the continuum limit,

$$a = \Lambda^{-1} \left(\frac{16\pi^2}{11g^2}\right)^{\frac{51}{121}} e^{\frac{-8\pi^2}{11g^2}}, \tag{3.3}$$

where Λ is the lattice scale parameter. In the continuum limit Λ is related to the string tension σ of the quenched theory by the relation

$$\Lambda = c \cdot \sqrt{\sigma}, \tag{3.4}$$

with a string tension estimated to be $\sigma \approx (400 \text{ MeV})^2$ from charmonium spectroscopy. The constant c can be determined from Monte Carlo calculations of the static quark antiquark potential. Its estimated value is about $c \approx \frac{1}{80}$ from recent very large scale computer simulations [10] .

Since the transition temperature T_c of the deconfining phase transition is a measurable physical quantity in the continuum limit, it is renormalization group invariant; that is $\frac{d}{da} T_c\left(g(a)\right) = 0$, where a is the lattice cutoff. Using the known two loop perturbative β function

$$-\beta(g) = b_0 g^3 + b_1 g^5 + \ldots \tag{3.5}$$

with

$$b_0 = \frac{11}{16\pi^2},$$

and

$$b_1 = \frac{102}{(16\pi^2)^2}.$$

T_c depends on the lattice coupling g as

$$T_c = \text{const}\,\frac{1}{a}\,\exp\left[-\int_0^g \frac{dg'}{\beta(g')}\right]. \tag{3.6}$$

The constant in Eq.(3.6) must be determined from non-perturbative calculations of T_c in the scaling regime.

We use a Monte Carlo measurement of the function $aT(g)$ to determine the β function. The results are to be compared with the two-loop form of Eq. (3.5) to verify asymptotic scaling and to determine the physical value of T_c in MeV.

4. Global Z(3) symmetry and the deconfinement transition

Gluon thermodynamics for finite temperatures is realized by lattices with spatial volume n_s^3 and temporal size n_t. The temporal size n_t is identified with the inverse of the temperature of the quantum field system in lattice spacing units. Strictly speaking the spatial size n_s should be taken to infinity for fixed n_t in the thermodynamic limit. This is only approximately realized in Monte Carlo calculations, and finite size effects become an important issue in the discussion.

The order parameter of the deconfinement phase transition is the Polyakov loop $P(\mathbf{x})$ which is defined as

$$P(\mathbf{x}) = \text{Tr}\prod_{t=1}^{n_t} U_t(\mathbf{x}, t), \tag{4.1}$$

where U_t is an SU(3) matrix along a time-like link at spatial location $\mathbf{x}$ at time t. We shall denote the spatial average of $P(\mathbf{x})$ by P and use it as the order parameter of the transition. The free energy F_Q of an isolated external quark source is related to the order parameter P by the relation

$$P \approx e^{-n_t a F_Q}, \tag{4.2}$$

where the proportionality constant is a Z(3) symmetric phase factor.

Lattice QCD at finite temperature has a global Z(3) symmetry, in addition to the local gauge symmetry associated with the color SU(3) group. The action under the Z(3) symmetry is invariant under multiplication of all time-like links on a single time slice by the same element of Z(3). Under this symmetry transformation the order parameter transforms as

$$P \longrightarrow z \cdot P, \tag{4.3}$$

where z is a group element from Z(3). In the low-temperature confined phase $P = 0$ and the Z(3) symmetry is unbroken. In the high-temperature deconfined phase $P \neq 0$ and the symmetry is broken with a three-fold degeneracy. The high temperature broken symmetry phase is identified as the deconfined phase because the free energy F_Q of an isolated external quark source is finite on the lattice in the deconfined phase.

The dynamics of the Polyakov loop is determined by a three-dimensional effective theory: by local gauge invariance all time-like links can be set to the unit SU(3) matrix except on one time slice. By integrating out the spatial link variables, one can derive an effective SU(3) spin model in three dimensions to describe the interaction of the time-like link variables in the time-slice [6,7,8] . If this effective SU(3) spin model has short range interactions, it is in the same universality class as the three-state Potts model in three dimensions and therefore a first order transition is expected.

Following the method of Polonyi and Szlachanyi [7] , we develop a qualitative physical picture for the phase transition and the breakdown of the Z(3) symmetry at the transition point. This picture will guide our analysis of the computer simulation results. In strong coupling we can derive an effective action in three dimensions for the partition function of Eq. (3.1) by integrating out the spatial link variables,

$$Z = \int [dP] \ \exp\ [-S_{\text{eff}}(P)]. \tag{4.4}$$

The effective action is given by

$$S_{\text{eff}}(P) = \beta_{\text{eff}} \sum_{x,i} |P(\mathbf{x}) - P(\mathbf{x} + e_i)|^2 + \tag{4.5}$$

$$\sum_{x} (-\frac{1}{2} \ln[27 - 18|P(\mathbf{x})|^2 + 8 \ \text{Re} \ P(\mathbf{x})^3 - |P(\mathbf{x})|^4] - 6\beta_{\text{eff}}|P(\mathbf{x})|^2)$$

where $\beta_{\text{eff}} = (\frac{1}{3g^2})^{n_t}$ and the Polyakov loop $P(\mathbf{x})$ is defined above. The effective action S_{eff} in invariant under global Z_3 transformations given in Eq. (4.3).

Mean field calculations on the effective theory defined by Eqs (4.4) and (4.5) predict a first order transition at a critical value of β_{eff} above which the Z_3 symmetry is spontaneously broken. At criticality the three broken Z_3 phases coexist with the unbroken phase at the origin. Fig. 1 depicts the effective potential of the mean

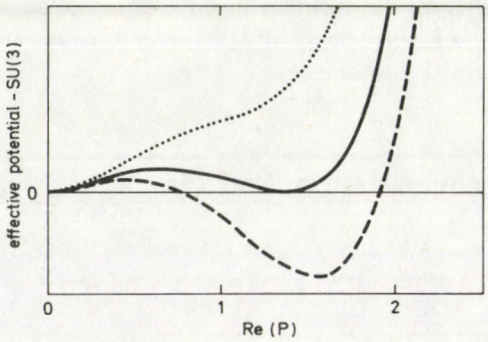


Fig. 1. The effective potential is shown in the mean field approximation to Eq. (3.5) for three different values of β_{eff} as we step through the transition point. The dotted line corresponds to the unbroken confined phase, the solid line represents coexistence, and the dashed line is calculated at a value of β_{eff} in the broken deconfined phase

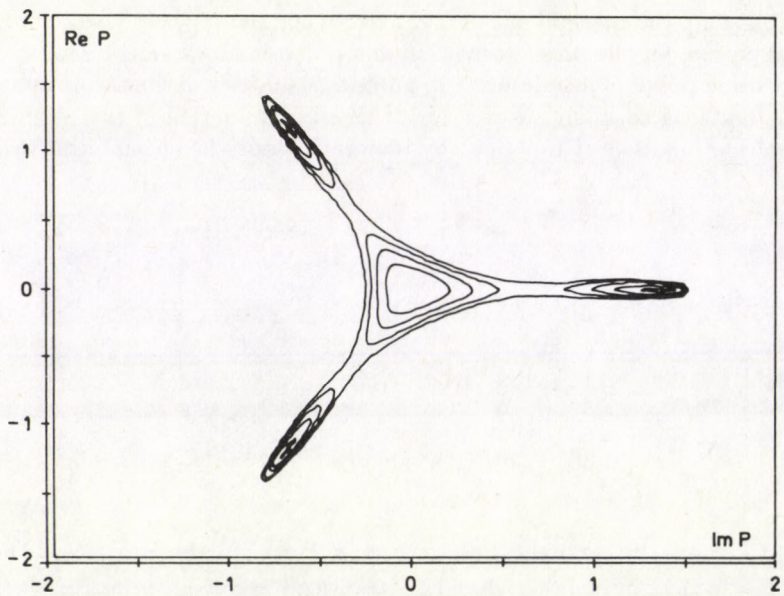


Fig. 2. The contour plot of the effective potential is shown at coexistence in the complex plane of the order parameter P. The analytic mean field calculation compares very well with the computer simulation at $n_t = 2$ in Fig. 3 at the coexistence point $\frac{6}{g^2} = 5.092$

field calculation as a function of the real part of the order parameter P for three different values of β_{eff} as we step through the first order transition point. Fig. 2 shows the contour plot of the effective potential in mean field approximation on the complex plane of the order parameter P at the coexistence point of the broken and unbroken phases. The effective potential has a minimum around the origin which corresponds to the unbroken confined phase. The Z(3) symmetric three other minima describe the broken deconfined phase. This qualitative picture has been made very convincing in computer simulations.

5. The transition temperature and computer simulation

In the computer simulation of the phase transition the number of time slices n_t is kept fixed (inverse temperature in lattice spacing units) while the coupling constant g is varied to search for the transition point. As g varies, the temperature changes in physical units. The transition temperature T_c corresponds to some critical coupling g_c for each value of n_t. In the scaling limit, for a sequence of n_t values, the transition temperature T_c has to follow the functional form of Eq. (3.6) as a function of g_c. The perturbative β function of Eq. (3.5) has to be used in the exponent of the expression for T_c in Eq. (3.6).

In an earlier paper [11] a Monte Carlo calculation of the critical coupling g_c was reported for n_t ranging from 2 to 10. Fig. 3 demonstrates how the first order phase transition is recognized in the computer simulation for $n_t = 2$ and $n_s = 9$. In the standard notation of $\beta = \frac{6}{g^2}$ the coexistence point of the two phases is at $\beta = 5.092$ which corresponds to fairly strong coupling. The similarity of the Polyakov loop distribution at coexistence in Fig. 3 is in striking qualitative agreement with the contour plot of Fig. 2 which was obtained in mean field approximation at strong coupling. At small n_t values we have good theoretical understanding of the detailed picture supported quantitatively by computer simulation. At larger values of n_t, however, we are loosing our analytic power to calculate things, but the computer simulation continues to provide the required quantitative results.

The most surprising result of this early calculation was that a pronounced non-scaling behavior was found in the coupling constant range $5.1 < 6/g^2 < 6.1$. The implications of this finding were twofold. First, an important upper bound was set on the lattice coupling constant g and on the lattice cutoff a for continuum physics results. Second, it became obvious that the determination of the transition temperature of the deconfinement phase transition in physical units (MeV) will require n_t values larger than 10, and therefore, given the condition $n_s >> n_t$, a very large scale computer simulation.

A new very large scale computer simulation was undertaken by our collaboration and results were reported for n_t from 8 to 14 [12]. The calculations were done on Cyber 205 supercomputers and on ST100 array processors. Different programs were used, one using the "quasi-heatbath" method [13] , and the other the Metropolis method. The Cyber 205 code was running in 32 bit precision with 19 μs/link update time using the quasi-heatbath algorithm. The ST100 code used 16

bit precision with 105 μs/link update time with 15 Metropolis hits/link including measurements at every fourth sweep and reunitarization after every second sweep.

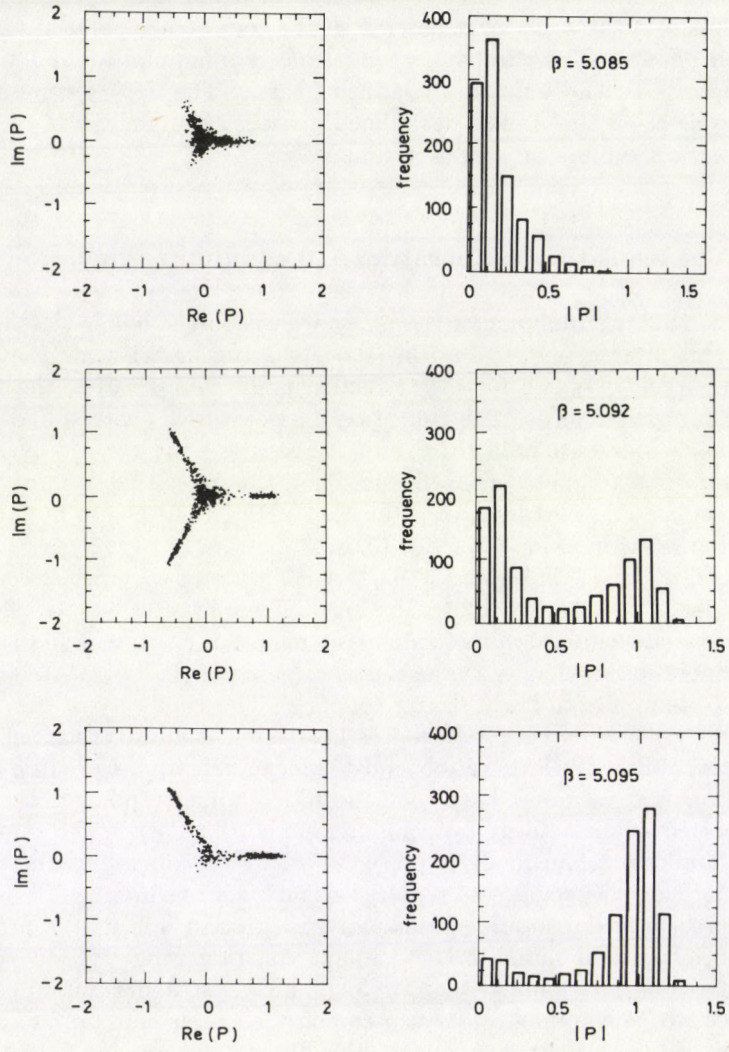


Fig. 3. The graphs on the left show the distribution of the Polyakov loop P in the complex plane for a 9^3 by 2 lattice. The top row correspond to $\beta < \beta_c$ (confined), the middle row is at β_c, and the bottom row has $\beta > \beta_c$ (deconfined). Each point corresponds to the value of the average Polyakov loop on a given configuration, and each measurement is separated by 20 sweeps at β_c and 5 sweeps away from the transition region. The histograms on the right show the distribution of points as a function of the radial distance (absolute magnitude of the Polyakov loop)

Runs were reported in length from 1×10^4 to 3×10^4 sweeps. Except for the first run at each lattice size, the lattice was initialized to the end result of a run at

a nearby value of β on the same size lattice. The first 2000 sweeps at each coupling were discarded. This number, 2000, represents a compromise between the ideal of a truly independent start and the relative scarcity of data.

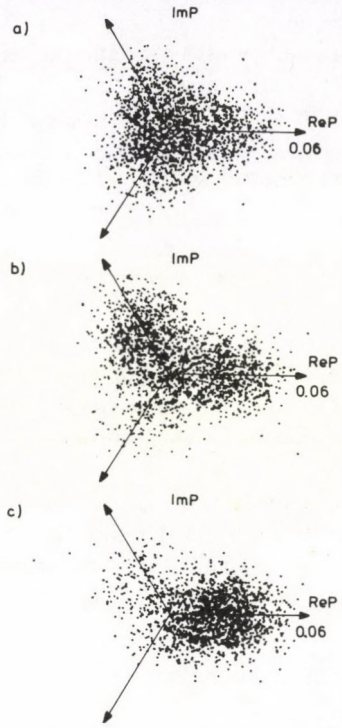


Fig. 4. Scatter plots of the Polyakov loop on $19^3 \times 14$ lattices at $6/g^2 = 6.45, 6.475,$ and 6.5 showing the transition from the confined phase to the deconfined phase. The runs contain 28500, 22000, and 23750 sweeps respectively, exclusive of warmups. The average Polyakov loop over the lattice is plotted at every tenth sweep

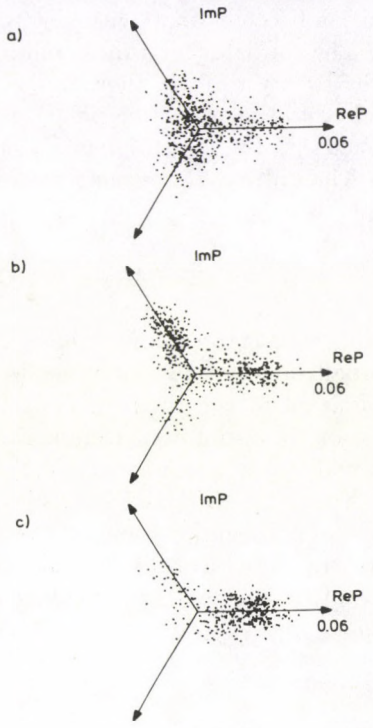


Fig. 5. The same data as in Fig. 1. The five successive measurements have been averaged to smooth out the high frequency fluctuations

Some of these new results are shown from three Monte Carlo runs in Fig. 4. These results are from $19^3 \times 14$ lattices, the largest size we studied at that time. In the first run we are below the transition point in the confined phase. The second run shows the system near the transition point with the coexisting confined and deconfined phases. In the third run we are above the transition point in the deconfined phase. Since then we cumulated more data on very large lattices up to the size of $23^3 \times 14$.

In an attempt to clarify scatter plots such as the ones depicted in Fig. 4 we have made "blocked" scatter plots in which we averaged the Polyakov loop over several successive measurements. The idea is that we are averaging out the high frequency (in Monte Carlo sweeps) scatter in the data but leaving the much slower movement of the system among the different available phases. Fig. 5 shows the same data as in Fig. 4, where five successive points have been averaged. The transition becomes more visible with this technique.

The coexistence of the confined and deconfined phases over runs of extended length (twenty to forty thousand sweeps were typically required to study the system close to the transition point) and the jump in the order parameter P provide evidence that the transition is of first order.

The width ΔT of a temperature driven first order phase transition is expected to scale as

$$\frac{\Delta T}{T} \approx \frac{1}{sV}, \tag{5.1}$$

where $V = n_s^3$ is the spatial volume and s is the latent entropy of the system [14,15]. For small values of the jump in the "magnetization" P the latent entropy will be proportional to the jump in P at the transition point. Near a first-order phase transition we also anticipate a shift in T_c as a function of the volume V with the same scaling law given by Eq. (5.1).

Estimates of T_c can be made from visual inspection of scatter plots such as those in Figs 4 and 5 and plots of the magnitude of the order parameter versus sweep number. Naturally it would be useful to have a quantitative measure of the degree of confinement or deconfinement of a finite size lattice.

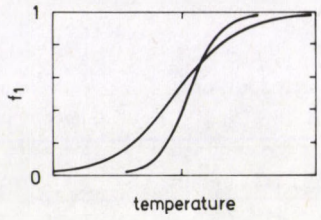


Fig. 6. The deconfinement fraction f_1 as a function of temperature on two different size spatial lattices at fixed value of n_t. The width of the rounding scales with $\frac{1}{V}$. The curves are expected to cross each other at $f \approx \frac{3}{4}$

In an attempt to do this we have studied the dimensionless quantities f_1 and f_2 which are the fractions of time spent in a long Monte Carlo run at fixed coupling and lattice size in the deconfined and confined phases, respectively. The fraction f_1 vanishes in the confined phase on an infinite spatial lattice. Fig. 6 shows f_1

schematically on two different spatial lattices at fixed n_t. The widths of the curves scale with the $\frac{1}{V}$ scaling law of Eq. (5.1) and the two curves cross at the best possible estimate for T_c in a finite volume.

The crossing is expected to occur at $f_1 = \frac{3}{4}$, if, at coexistence, the integrated Polyakov loop density in the complex phase space is equal in the confined phase around the origin to the integrated density in the three Z(3) symmetric peaks of the deconfined phase. In that case the four coexisting peaks of the Polyakov loop distribution will occur with equal probabilities in a long computer run. Mean field calculations at strong coupling support that picture as can be deduced from Figs 1 and 2 and from the details of the analytic mean field calculation.

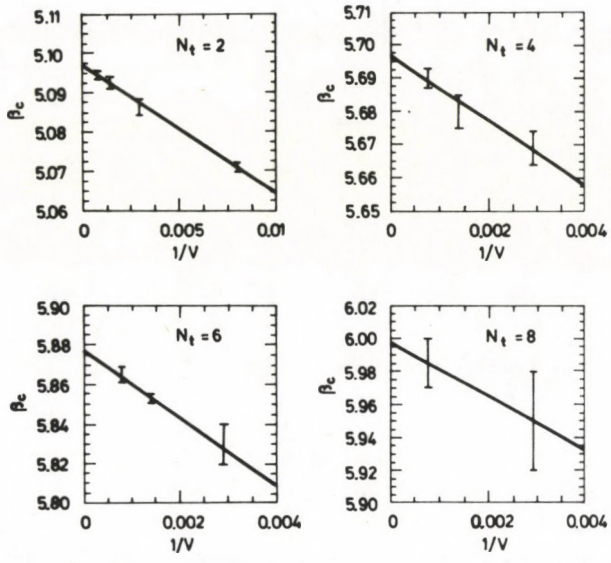


Fig. 7. Computer simulation measurement of β_c for different spatial volumes at $n_t = 2, 4, 6,$ and 8 The lines show the finite-size scaling extrapolations for the $f_1 = \frac{1}{2}$ criteria

In our earlier work [11] the value $f_1 = \frac{1}{2}$ was chosen using the somewhat more ad hoc argument that at coexistence the system would spend half time in the confined phase and half time in the deconfined phase ignoring the Z(3) structure of the broken (deconfined) phase. With this choice we expect a shift in the transition point proportional to $\frac{1}{V}$. Of course, we expect to extrapolate in the limit when V becomes very large to the transition point obtained from the $f_1 = \frac{3}{4}$ condition.

Fig. 7 shows the finite size scaling shift of β_c from the the condition $f_1 = \frac{1}{2}$. That has to be compared with Fig. 8 for $n_t = 2$ where the deconfinement fraction is plotted for $n_s = 5, 7$ and 9. The crossing point is at $f_1 = \frac{3}{4}$ at critical coupling

$\beta_c = 5.095$ which agrees to a very good approximation with the extrapolated value obtained from Fig. 7 for $n_t = 2$. We conclude that at $n_t = 2$ our theoretical picture and the computer simulation are in very good agreement. The same analysis can be used then for larger values of n_t.

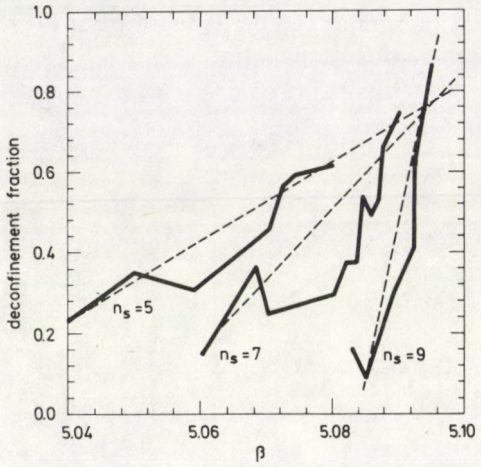


Fig. 8. The deconfinement fraction as a function of the coupling for $n_t = 2$ at spatial lattice sizes $n_s = 5, 7,$ and 9.
The zig-zag of the curves indicates the statistical error involved in the calculations. The curves appear to cross each other at $f_1 = \frac{3}{4}$ within statistical uncertainties

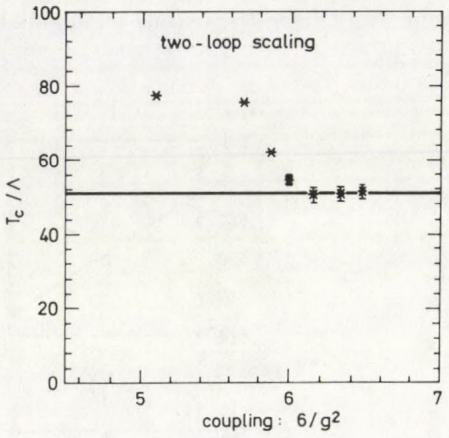


Fig. 9. The onset of scaling in the deconfinement temperature

Table I summarizes our results of a large number of Monte Carlo runs for different values of n_t.

Estimation of the statistical errors is difficult because of the limited amount of data. Even our longest runs contain only one or two handful of tunnelings among the different phases, so one cannot accurately divide the runs into many independent subsets as we would do if we had much more data. Thus our quoted errors represent the range of β over which the scatter plots and time histories appear to change from confined to deconfined using the $f_1 = \frac{3}{4}$ criteria.

The main physics result of our work is depicted in Fig. 9. The measured values of T_c/Λ are plotted there in the range from 2 to 14 for n_t. Where the relation between g_c and $a \cdot T_c$ is that predicted by two loop perturbation theory, this graph will be a horizontal line. The height of the line gives the constant of proportionality between T_c and Λ.

Table I
Critical couplings on various lattice sizes

Time slices	Largest spatial size n_s	Critical β	Statistical error
2	13	5.097	0.005
4	11	5.696	0.005
6	17	5.88	0.01
8	19	6.02	0.02
10	19	6.18	0.02
12	19	6.33	0.02
14	23	6.45	0.02

Table II
Data base

Time slices	Spatial size	β	Sweeps	Machine/algorithm
10	17	6.13	32000	ST100/Metropolis
10	17	6.17	36000	ST100/Metropolis
12	19	6.28	22000	ST100/Metropolis
12	19	6.30	24000	ST100/Metropolis
12	19	6.32	14000	Cyber/heatbath
14	21	6.42	34750	Cyber/heatbath
14	21	6.45	40440	Cyber/heatbath

Table III
Latent heat

N_t	β	Plaquette gap	$\Delta\epsilon/\Lambda^4 \times 10^{-7}$	Statistical error
10	6.13	0.000303(59)	3.18	0.93
	6.17	0.000275(52)	3.46	0.67
12	6.28	0.000080(51)	1.65	1.18
	6.30	0.000181(40)	4.08	0.91
	6.32	0.000228(83)	5.62	0.21
14	6.42	0.000080(28)	3.10	1.08
14	6.45	0.000013(34)	0.58	1.50

The following remarkable structure emerges: after apparent early scaling between $5.1 < 6/g^2 < 5.7$ there is strong scaling violation in the range $5.7 < 6/g^2 < 6.10$ and finally asymptotic scaling is observed for $6.15 < 6/g^2 < 6.50$. The onset of scaling is at much weaker coupling than early optimistic expectations. This means that either a large increase in computer power or a substantial improvement on Wilson's lattice action is needed for practical calculations of hadron properties. However, our work does provide evidence that Monte Carlo calculations with $\beta > 6.15$ on sufficiently large lattices can provide believable answers for continuum quantities in pure gauge QCD.

Recently a new computer simulation of the string tension $\sqrt{\sigma}$ was carried out on a very large lattice of the size $24 \times 24 \times 24 \times 48$ at the coupling $\beta = 6.3$ [10]. De Forcrand finds $\sqrt{\sigma} \approx 80\Lambda$ where Λ is the lattice scale parameter defined before. From Fig. 9 we find $T_c = (51 \pm 3)\Lambda$. The combination of the two calculations gives $\frac{T_c}{\sqrt{\sigma}} = 0.6$ which corresponds to $T_c = 240$ MeV in physical units if $\sqrt{\sigma} \approx 400$ MeV is chosen from heavy quark spectroscopy.

I should also note that a similar calculation of the deconfinement transition temperature was carried out on the Columbia special purpose machine with similar findings [16] .

6. The latent heat of the deconfinement transition

In the previous Section we have learned from our large scale computer simulation that deconfinement is associated with a first order phase transition. For physical applications it is important to know the latent heat per unit volume of the transition. It will determine the amount of energy per unit volume which has to be pumped into the system at the transition point to make deconfined quark matter.

In many of the Monte Carlo runs used in the study of the transition temperature we monitored the expectation value of the plaquette. In this Section we use these plaquette measurements to estimate the latent heat of the deconfinement transition [17]. To the extent that the deconfinement transition in pure gauge QCD is a good approximation to the real world including quarks, this number is important to the question of whether the deconfined phase will be observed in heavy ion collisions, as well as to studies of the early universe.

The theory of the latent heat in lattice QCD can be found in the literature. The particular formula that we apply is due to Svetitsky and Fucito [18], and relates the latent heat to the discontinuity of the average plaquette across the first order transition. We will briefly summarize here the theoretical basis for the latent heat calculation.

For any discrete statistical system with states $|j>$ of energy $E_j \equiv < j|H|j >$ at temperature $T = 1/k_B\underline{\beta}$ the partition function in the canonical ensemble is defined by $Z \equiv \sum_j e^{-\underline{\beta}E_j}$. Underlined $\underline{\beta}$ is used here to distinguish the physical temperature of the system from the unfortunate notation of $\beta = \frac{6}{g^2}$ in QCD.

The energies E_j depend upon an extensive parameter V, the volume of the "box". From this we immediately find that the internal energy is

$$U \equiv < H >= \frac{1}{Z} \sum_j e^{-\underline{\beta} E_j} E_j = - \left(\frac{\partial \ln Z}{\partial \beta} \right)_V \tag{6.1}$$

and the entropy is

$$S \equiv -k_B \sum_j p_j \ln p_j = -k_B \sum_j \frac{e^{-\underline{\beta} E_j}}{Z} [-\underline{\beta} E_j - \ln Z] = k_B \underline{\beta} U + k_b \ln Z,$$

thus we have the fundamental relation between the free energy F and the partition function $F \equiv U - TS = -\frac{1}{\underline{\beta}} \ln Z$.

The second law of thermodynamics states that $dU = TdS - PdV$, hence $dF = -SdT - PdV$. The pressure P may be defined by

$$P = - \left(\frac{\partial U}{\partial V} \right)_S = - \left(\frac{\partial F}{\partial V} \right)_T = \frac{1}{\underline{\beta}} \left(\frac{\partial \ln Z}{\partial V} \right)_{\underline{\beta}}. \tag{6.2}$$

On the lattice the partition function for finite temperature QCD may be written as

$$Z = \int [dU] \ \exp \left(\frac{6}{g^2} S_B[U] \right), \tag{6.3}$$

where the lattice action is

$$S_B[U] = \frac{a_s}{a_t} \sum_{\text{spacelike}} \frac{1}{3} \text{ Tr Re } U_{pl} + \frac{a_t}{a_s} \sum_{\text{timelike}} \frac{1}{3} \text{ Tr Re } U_{pl}, \tag{6.4}$$

and the lattice spacing in the spatial and thermal directions are a_s and a_t respectively. We assume the lattice volume to be $V = (a_s n_s)^3$, the temperature to be given by $\beta = 1/k_b T = a_t n_t$, and we define the energy density to be $\epsilon \equiv U/V$, so we obtain

$$\epsilon = - \left(\frac{1}{a_s^3 n_s^3 n_t} \right) \left(\frac{\partial \ln Z}{\partial a_t} \right)_{a_s}, \tag{6.5}$$

$$P = \left(\frac{1}{3 a_s^2 n_s^3 a_t n_t} \right) \left(\frac{\partial \ln Z}{\partial a_s} \right)_{a_t}. \tag{6.6}$$

Using the explicit form for the SU(3) partition function, together with the definition of the configuration-averaged plaquette operator

$$P_s = \frac{\frac{a_s}{a_t} \sum_{\text{spacelike}} \text{ Tr Re } U_{pl}}{3 n_s^3 n_t},$$

$$P_t = \frac{\frac{a_t}{a_s} \sum_{\text{timelike}} \text{ Tr Re } U_{pl}}{3 n_s^3 n_t},$$

(notice that our individual plaquettes are normalized so that their maximum value is three), we find

$$\epsilon = \left(\frac{1}{a_s^3 a_t}\right)\left[\frac{6}{g^2} < P_t - P_s > - a_t \left(\frac{\partial(6/g^2)}{\partial a_t}\right)_{a_s} < P_s + P_t >\right], \quad (6.7)$$

$$P = \left(\frac{1}{3a_s^3 a_t}\right)\left[\frac{6}{g^2} < P_t - P_s > + a_s \left(\frac{\partial(6/g^2)}{\partial a_s}\right)_{a_t} < P_s + P_t >\right]. \quad (6.8)$$

The averages are defined by $< \Omega >= \frac{1}{Z}\int [dU] \exp(6g^{-2} S_B[U]\Omega[U])$ and $S_B/n_s^3 n_t = P_s + P_t$. The two terms in each of the above expressions come from the explicit dependence of S_B upon a_s or a_t and from the implicit cutoff dependence of the coupling constant g.

The expressions just derived for the energy density and pressure hold for any temperature, but at T_c, the deconfinement temperature, we know in addition that the pressure is continuous (the transition is of first order, so the system changes between the confined and deconfined phases at constant T and P because the free energy F is degenerate for the two phases). At T_c we have, therefore,

$$\epsilon_c = -\frac{2}{a^4} a \frac{\partial(6/g^2)}{\partial a} < P >, \quad (6.9)$$

where we have also set $a_s = a_t = a$ and $P = (P_s + P_t)/2$.

If the coupling is sufficiently small that the system follows the asymptotic scaling laws we can use the weak-coupling form for the β-function (the dependence of g on a_s and a_t are the same in this limit), and also remove the cutoff dependence of our result by reexpressing it in terms of Λ:

$$\beta(g) \equiv -a\frac{\partial g}{\partial a} = -b_0 g^3 - b_1 g^5 + \ldots, \quad (6.10)$$

$$a\Lambda(g) = (b_0 g^2 + \ldots)^{-b_1/2b_0^2} \exp\left(\frac{-1}{2b_0 g^2}\right), \quad (6.11)$$

where the first two (universal) coefficients were given before: $b_0 = 11/(4\pi)^2$ and $b_1 = 102/(4\pi)^4$. Let the energy density difference between the two phases be $\Delta\epsilon$, then our final result is that the latent heat of the transition is

$$\frac{\Delta\epsilon}{\Lambda^4} = 24\left[\frac{11}{(4\pi)^2} + \frac{102g^2}{(4\pi)^4} + \ldots\right]\left(\frac{11g^2}{(4\pi)^2}\right)^{204/121} \exp\left(\frac{32\pi^2}{11g^2}\right) < \Delta P > \quad (6.12)$$

with $< \Delta P >$ being the change in the average plaquette between the two phases.

We will discuss now our measurement of the jump in the plaquette at the transition point and the related uncertainties. Because the gap in the plaquette is extremely small at large n_t and because the spatial volumes of our lattices are relatively small, it is not easy to measure the gap. We attempt to take a long run

at a single value of β, ideally exactly at the critical value, and separate the run into portions where our finite lattice is "confined" or "deconfined". On our finite size lattices there is a region of β in which we see phase coexistence, with occasional tunneling from one phase to another. The central assumption of our procedure is that the configurations of our lattices can be divided into confined and deconfined configurations and the average plaquette measured in each phase. This is obviously an idealization, and there will be a resulting systematic error on our results.

We used a total of seven Monte Carlo runs in this analysis. To be useful, a run had to show coexistence (i.e. tunneling), be rather long, and have the plaquette measurements recorded along with the Polyakov loops. Table II contains a summary of the runs used.

Table III contains a summary of the plaquette gaps in all the seven runs. This Table should be interpreted with some caution. It is clear that the gap measured from a particular Monte Carlo run is reasonably insensitive to the exact cutoff used. However, this alone does not tell us that another run would give the same gap.

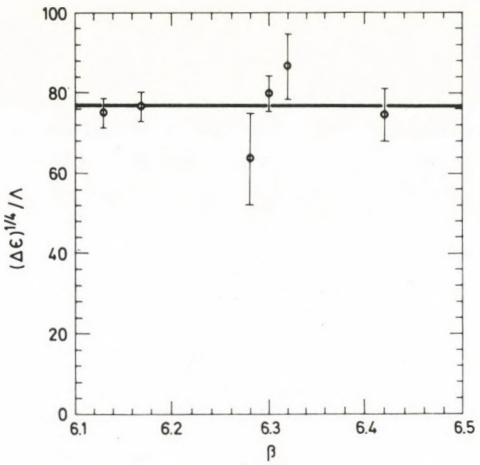


Fig. 10. The scaling behavior of the latent heat as measured in units of Λ

The errors quoted in the Table are statistical errors only. That is, they refer to the uncertainty in the average value of the plaquette under the assumptions that the configurations can be separated into deconfined and confined configurations, and that we have done the separation correctly. Both these assumptions are approximate, and both will introduce some systematic error. It is not clear to us whether our lattices go from one phase to the other as a whole or whether domain walls form. In the first case, we would expect the average energy of an "undecided" configuration to be between the deconfined and confined averages, while in the second case we would expect it to be higher. We don't know what direction the

systematic error caused by the inclusion of such configurations will be. The second systematic error comes from misclassifying configurations. This almost certainly causes our result to be smaller than the correct number.

It is not clear how to choose a best value for the gap from the above Table. Because we suspect that smearing reduces the gap, for the moment we will take the largest result from a magnitude cut in each run. Again, the differences as we vary the cut are fairly small.

We can now use the formula in Eq. (6.12) to calculate the physical latent heat. The resulting values are plotted in Fig 10. The weighted average of the seven runs is $(3.17 \pm 0.35) \times 10^7$, with a χ^2 of 7.2 for six degrees of freedom.

The result depicted in Fig. 10 can be translated into a latent heat per unit volume in $\frac{\text{GeV}}{\text{fermi}^3}$ units using the relation $\sqrt{\sigma} \approx 80\Lambda$ and the string tension $\sqrt{\sigma} \approx$ 400 MeV from heavy quark spectroscopy. This way we find the latent heat per unit volume to be about 2.5 $\frac{\text{GeV}}{\text{fermi}^3}$ which is a large energy density. This number is very sensitive (fourth power) to our input string tension value and the proportionality factor between $\sqrt{\sigma}$ and Λ.

7. Conclusion

In this report I demonstrated the power of very large scale computer simulations for the nonperturbative study of Quantum Chromodynamics. I presented evidence that the deconfinement phase transition is of first order and the transition temperature was determined in the scaling limit of the theory where cutoff effects can be eliminated. The transition temperature was estimated to be about 240 MeV. The latent heat of the first order transition was also calculated in the same very large scale computer simulation.

Our calculations were carried out in the sector of QCD neglecting quark vacuum polarization effects. These dynamical quark effects will be included in the next generation of computer simulations on medium size lattices in the very near future.

References

1. L.D. McLerran and B. Svetitsky, Phys. Letters, *98B*, 195,1981.
2. J. Kuti, J. Polonyi and K. Szlachanyi, Phys. Letters, *98B*, 199, 1981.
3. For a review of early work, see H. Satz, Phys. Rep., *88*, 439, 1982.
4. A.M. Polyakov, Phys. Lett., *72B*, 477, 1978.
5. L. Susskind. Phys. Rev. D., *20*, 2610, 1979.
6. N. Weiss, Phys. Rev D., *24*, 475, 1981; Phys. Rev. D, *25*, 2667, 1982.
7. J. Polonyi and K. Szlachanyi, Phys. Lett., *110B*, 963, 1982.
8. L.G. Yaffe and B. Svetitsky, Phys. Rev. D., *26*, 963, 1982; Nucl. Phys., *B210*, 423, 1982.
9. C. Borgs and E. Seiler, Nucl. Phys., *B215*, 125, 1983.
10. Ph. de Forcrand, private communication.
11. A.D. Kennedy, J. Kuti, S. Meyer and B.J. Pendleton, Phys. Rev. Lett., *54*, 87, 1985.
12. S.A. Gottlieb, J. Kuti, D. Toussaint, A.D. Kennedy, S. Meyer, B.J. Pendleton and R.L. Sugar, Phys. Rev. Letters, *55*, 1958, 1985.

13. N. Cabibbo and E. Marinari, Phys. Lett., *119B*, 387, 1982.
14. Y. Imry, Phys. Rev. B, *21*, 2042, 1980.
15. M.E. Fisher and A.N. Berker, Phys. Rev., B, *26*, 2507, 1982.
16. N.H. Christ and A.E. Terrano, Phys Rev. Lett., *56*, 111, 1985.
17. S.A. Gottlieb, J. Kuti, D. Toussaint, A.D. Kennedy, S. Meyer, B.J. Pendleton and R.L. Sugar, UC San Diego preprint, 1986.
18. B. Svetitsky and F. Fucito, Phys. Lett., *131B*, 165, 1983.

Acta Physica Hungarica 64 (1-3), pp. 111-122 (1988)

HYDRODYNAMICS OF REHADRONIZATION*

B. LUKÁCS and J. ZIMÁNYI

Central Research Institute for Physics
1525 Budapest, Hungary

N.L. BALAZS

State University of New York at Stony Brook, Stony Brook
NY 11794 USA

(Received 8 January 1987)

It has been shown that strange and non-strange quarks generally possess different hadronization rates in a quark–nucleon phase transition. Since the main process driving the rehadronization is the expansion and cooling of the fireball of quarks, this complicated phase transition is intimately connected with relativistic hydrodynamics. Here the consistent hydro+thermodynamical description of the transition is presented.

1. Introduction

In equilibrium thermodynamics it is a commonplace that the necessary condition for phase equilibrium is the equality of all the thermodynamic intensives between the phases. This can be visualized in two different ways. Either one can observe that in this case all the intensive gradients driving conductive fluxes are absent, or one can find that, with fixed total values of the extensives, the entropy is maximal at a homogeneous distribution of the intensives [1].

However, in a recent publication [2] it has been pointed out that to keep the zero value of strangeness in a heavy ion reaction one has to use different strange quark chemical potential, μ_s, in the quark and hadron phases. Therefore ref. [2] has suggested that μ_s should possess a jump at the phase boundary. But in this case the phase transition would be accompanied by entropy change, which is possible but not in the equilibrium (or quasistationary) limit, which is a physical limit when the hydrodynamical evolution is of moderate velocity.

Ref. [3] has eliminated this problem by showing that the quark–hadron phase transition has to proceed through a coexistence region where μ_s changes continuously from 0 to a finite value, being in each volume element the same for the two phases. However, obviously, the presence of such a coexistence region leads to serious problems in the hydrodynamical description of an expanding fireball undergoing phase transition from quarks into hadrons.

*Dedicated to Prof. G. Marx on his 60th birthday

The aim of this paper is to set up a new type of hydrodynamical formalism suited to the present problem. Section 2 briefly summarizes results of [3], needed here. Section 3 gives the new type of hydrodynamical equations, while Section 4 discusses the boundary conditions at the interfaces of different phases. Section 5 contains the explicit forms of the equations in CM system.

2. Phase transition in systems of neutral strangeness

First we recapitulate how the phase equilibrium conditions are obtained. Consider a thermodynamical system whose independent extensives are X^I; for the general statements the particular number and specification of these extensives is unnecessary. The entropy S is a homogeneous linear function of X^I. Now assume that the matter can be in two different phases between which the transition is free of constraints. Then the total entropy of states in a container keeping the values of the extensives fixed is

$$\begin{aligned} S &= S(X_1^I) + S(X_2^I), \\ X_1^I + X_2^I &= X^I = \text{fixed}, \end{aligned} \tag{2.1}$$

where the lower index labels the particular phase. Since equilibrium belongs to the entropy maximum [1], by a variation one obtains

$$\frac{\partial S}{\partial X_1^I} = \frac{\partial S}{\partial X_2^I}, \tag{2.2}$$

which is just the equality of intensives, called Gibbs criteria.

Now consider a quark plasma, produced in some previous compression phase of the heavy ion collision. Charmed, bottom and top quarks can be neglected in collisions of the order of GeV/nucleon beam energy, and gluons are not conserved particles, therefore their number is determined by the temperature. Furthermore in first approximation the neutron and proton numbers are equal, therefore in the quark plasma up and down quarks will appear symmetrically. So the quark plasma can be characterized by three independent extensive densities: the energy density ρ, the number density of light quarks n_q and the number density of strange quarks n_s. The conjugate entropic intensives are $1/T, -\mu_q/T$ and $-\mu_s/T$, respectively, where μ is the chemical potential and T is the temperature, while the fourth intensive, conjugate to the volume V is p/T, where p is the pressure. Now assume that this plasma begins to rehadronisate: then, in equilibrium, Eq. (2.2) requires the equality of quark and hadron intensives. However, still we have to specify the independent characteristic data of the hadronic matter. Because hadrons consist of quarks and can be mutually transformed into each other, the complete equilibrium can be characterised by only two chemical potentials also on the hadronic side, with e.g.

$$\begin{aligned} \mu_n &= 3\mu_q, \\ \mu_\Lambda &= 2\mu_q + \mu_s, \end{aligned} \tag{2.3}$$

etc. Then Eq. (2.2) leads to the equality of T, μ_q and μ_s between the phases, with the additional condition

$$p_1(T, \mu_q, \mu_s) = p_2(T, \mu_q, \mu_s). \tag{2.4}$$

Here we will not discuss the specific forms of the pressure functions.

Now, the initial condition of the transition is a quark plasma with zero strangeness; then in the plasma $\mu_s = 0$. Substituting this into Eq. (2.4) one obtains a relation between μ_q and T. However, in the hadronic matter generally the strangeness does not vanish at $\mu_s = 0$, so is seems that one has to choose between chemical equilibration and conservation of strangeness [2].

However, in the above discussion a tacit assumption was used, namely that the nuclear matter keeps the equilibrium with the original state of the quark plasma. This is not necessarily so, furthermore, the above paradox shows that it cannot be true. Consequently, with the appearance of the first hadronic droplets the quark plasma ceases to be neutral for strangeness; $\mu_s \neq 0$, and takes such a value that the actual quark–hadron mixture in the volume element investigated can be globally neutral at the same μ_s value in both phases.

This construction is mathematically possible. The physical processs behind it is a fractional distillation of quarks. Let us start from $\mu_s = 0$; then, because of

$$2m_K < m_\Lambda \tag{2.5}$$

first the s quarks go into the hadronic phase, forming mainly kaons. Then the quark plasma becomes s-dominated, μ_s starts to increase, until at some value the s quarks already possess the needed energy to build up the Λ hyperons. The final state of the rehadronisation is a hadronic matter neutral for strangeness but already some positive μ_s [3].

According to this, in a given volume element the evolution of the matter consists of three stages: quark plasma with $\mu_s = 0$, quark–hadron mixture with growing positive μ_s and the final hadronic matter. At fixed time the corresponding 3 regions can be found in a fireball. Consequently, the hydrodynamics of the expansion and transition will be complicated. In the next Section we will explicitly incorporate the above phase equilibrium picture into the hydrodynamical formalism.

3. The evolution equations of hydrodynamics

Our system is an expanding sphere of particle mixture, in some region with two coexisting phases. Nevertheless, in principle, the equations of motion and continuity are obtained in the same manner as for one component and one phase. Our starting equations are:

$$\dot{n}_\alpha + n_\alpha u^r{}_{;r} = 0, \tag{3.1}$$

$$T^{ir}{}_{;r} = 0, \tag{3.2}$$

where u^i is the four-velocity, n is a particle density, the subscript α labels the conserved components, and T^{ik} is the energy-momentum tensor of the fluid; the semicolon stands for covariant derivative, while the dot is the comoving derivative

$$\dot{n} = n_{,r}\, u^r. \tag{3.3}$$

Eqs (3.1–3) are valid in any coordinate system; the actual forms for CM system will be discussed later, but, obviously, Eqs (3.1) are evolution equations for the particle densities. Now we turn to Eq. (3.2).

T^{ik} can be decomposed by using the velocity field u^i; one obtains

$$\begin{aligned} T^{ik} &= \rho u^i u^k + q^i u^k + u^i q^k + p^{ik}, \\ u^r u_r &= -1, \\ q^r u_r &= p^{ir} u_r = 0. \end{aligned} \tag{3.4}$$

For the observer comoving with the matter ρ is the energy density, q^i is the conductive energy (heat) flux and p^{ik} is the spatial stress tensor [4]. Restricting ourselves to perfect fluids,

$$\begin{aligned} q^i &= 0, \\ p^{ik} &= p(g^{ik} + u^i u^k), \end{aligned} \tag{3.5}$$

where p is the thermodynamic pressure and g_{ik} is the metric tensor of the spacetime.

Substituting the form (3.4) into Eq. (3.2), two equations are obtained:

$$(\rho + p) u^i{}_{;r} u^r + p_{,r} (g^{ir} + u^i u^r) = 0, \tag{3.6}$$

$$\dot{\rho} + (\rho + p) u^r{}_{;r} = 0. \tag{3.7}$$

The first equation is the equation of motion, being $u^i{}_{;r} u^r$ the acceleration, while Eq. (3.7) is the balance equation for energy. Eqs (3.1), (3.7) result in an entropy equation too, because

$$s = s(n_{\alpha'} \rho). \tag{3.8}$$

(We, of course, assume that all the particle densities not restricted by conservation laws take their equilibrium values, i.e. those of the entropy maximum.) Now, for a one phase system the entropy production can be evaluated via Eqs (3.1) and (3.7), yielding

$$\dot{s} + s u^r{}_{;r} = 0 \tag{3.9}$$

and the same is true for two phases in equilibrium [5]. We will use therefore Eq. (3.9) instead of Eq. (3.7). So the system of evolution equation consists of Eqs (3.1), (3.6) and (3.9); since the independent extensive densities are n_α and either ρ or s, the evolution of p is determined and then Eq. (3.6) gives the velocity changes for any moment.

However, during phase transition the variables n_α and ρ cannot be directly used, because the conditions of phase equilibrium are the equalities of the intensives; similarly, now there are "chemical" (i.e. hadronization) reactions during the phase transition, whose equilibrium can also be formulated via intensives. Thus we have to rewrite the balance equations (3.1), (3.9) in terms of chemical potentials, temperature and pressure. It can easily be done for a pure phase as follows.

Consider a fluid characterised by two chemical potentials μ_1 and μ_2 and the temperature T. Then, due to thermodynamic relations given in the Appendix.

$$\begin{aligned} p &= p\,(\mu_1, \mu_2, T), \\ n_1 &= p_{,_1}\,, \\ n_2 &= p_{,_2}\,, \\ s &= p_{,_T}\,, \\ \rho &= Ts + \mu_1 n_1 + \mu_2 n_2 - p, \end{aligned} \tag{3.10}$$

where the comma and subscript stand for derivatives with respect to thermodynamic variables and 1 is a shorthand notation for n_1. Now Eqs. (3.1), (3.9) become linear in $\dot{\mu}_i$ and $\dot{T}$, while the coefficients are second derivatives of p. The equations so obtained can be written into a compact form by exploiting the Riemannian structure of the thermodynamic state space. Namely, let us introduce the independent intensive parameters as coordinates in this state space [6]

$$x^I = (T, \mu_1, \mu_2). \tag{3.11}$$

Then the second derivatives of p build up a metric tensor [7]

$$g_{IK} = \frac{\partial^2 p}{\partial x^I \partial x^K} \tag{3.12}$$

and the balance equations get the form

$$g_{IK}\dot{x}^R + p_{,_I}\, u^r{}_{;r} = 0. \tag{3.13}$$

By multiplying with g^{IK}

$$g_{IK} g^{KR} = \delta_I^K, \tag{3.14}$$

the result is

$$\dot{x}^I = -g^{IR} p_{,_R}\, u^r{}_{;r}. \tag{3.15}$$

So we have explicitly obtained the evolution equations for the intensives; substituting the ρ and p values into Eq. (3.6) the evolution of the velocity is obtained as well. One still has to show that the components n_1 and n_2 can be identified by the conserved baryon and strangeness numbers; this is done in the Appendix, here we only note that in equilibrium the number of independent chemical potentials is 2 in both phases [2].

Now let us turn to the coexistence region, where the matter is a mixture of quark and hadronic phases. Assuming that the droplets are sufficiently small for a continuous hydrodynamics, all the density-like quantities are to be volume averaged as, e.g.

$$s = \alpha s_q + (1-\alpha) s_h, \tag{3.16}$$

where the subscript labels the particular phase and

$$\alpha = V_q/(V_q + V_h). \tag{3.17}$$

The x^I state space coordinates take the same values for both phases as a consequence of equilibrium, therefore the local state is characterized by x^I and α. So now one more evolution equation is needed. This can be obtained from the condition of mechanical equilibrium

$$p_q(x^I) = p_h(x^I). \tag{3.18}$$

In order to obtain again a compact form here we introduce averaged and difference quantities as

$$\begin{aligned} D_I &\equiv p_{q,I} - p_{h,I}, \\ \bar{p}_I &= \alpha p_{q,I} + (1-\alpha) p_{h,I}, \\ \bar{g}_{IK} &= \alpha g_{qIK} + (1-\alpha) g_{hIK}. \end{aligned} \tag{3.19}$$

Then, from Eqs (3.1), (3.9):

$$\bar{g}_{IK}\dot{x}^R + D_I\dot{\alpha} + \bar{p}_I u^r{}_{;r} = 0, \tag{3.20}$$

while the dot derivative of Eq. (3.18) yields

$$D_R\dot{x}^R = 0. \tag{3.21}$$

By introducing the contravariant vector componens via

$$D^I = g^{IR} D_R, \tag{3.22}$$

Eqs (3.20–21) give

$$\begin{aligned} \dot{\alpha} &= -\frac{D^R \bar{p}_R}{D^S D_S} u^r{}_{;r}, \\ \dot{x}^I &= \left(\frac{D^R \bar{p}_R}{D^S D_S} D^I - \bar{p}^I \right) u^r{}_{;r}. \end{aligned} \tag{3.23}$$

Being the phase transition of first order, $D_I \neq 0$.

Now the evolution of the local thermodynamical quantities is determined in the mixed phase; by using volume averaged ρ and p values in Eq. (3.6) the equation of motion is known as well. Then, in principle, a particular coordinate system can be chosen and the evolution equations can be integrated from an appropriate initial condition. However, first the question of possible discontinuities is to be discussed.

4. The phase boundaries

Perfect fluid models always possess two kinds of possible instabilities. First, the Reynolds number, which is inversely proportional to the viscosity coefficient, is infinite, therefore turbulency is expected [8]. Second, in the lack of transport processes shock fronts may be built up. (If there is either viscosity or heat transfer, the front becomes a smooth transition region of finite width [9].) Now, these instabilities may or may not be physical depending on the actual values of transport coefficients in the actual system investigated; if necessary, the shock fronts can be incorporated into the evolution equations via the Rankine–Hugoniot equations [9], while the turbulency is at least formally excluded by the high symmetry of the flow. The development of shock fronts is highly sensitive on the initial conditions.

In this paper we are concentrating on effects arising from the phase transition and the existence of a mixed phase region. Therefore, here we assume that the initial conditions are sufficiently smooth in the sense that they could not lead to discontinuities without phase transition. However, the phase boundaries do represent some kinds of discontinuities, so they are to be discussed.

Our particular initial condition is a hot, compressed, static sphere. Therefore the pure quark phase occupies the central region, up to some r_1; here $\alpha=1$ identically. At r_1 there is the first phase boundary B_1 separating the pure quark phase from the mixture. In the second region α is continuously decreasing, and where it reaches the 0 value, there is the second phase boundary B_2 whence $\alpha = 0$ and a pure hadronic phase can be found.

Now, one may assume that both x^I and α are continuous in r, since the equality of the intensives is the condition for phase equilibrium. However, there is no sufficient reason to assume the continuity of the first derivatives, and generally at least α possesses jumps in the first derivative at phase boundaries (being the derivative identically 0 outside). But then the first derivatives of x^I must be discontinuous, too, in order to compensate the effect in p, otherwise a shock front will develop. Namely, consider Eq. (3.6) in both sides of B_1 (or B_2). The coefficient $\rho + p$ is continuous while crossing B_1. However, $p_{,i}$ can be written on the two sides as

$$\begin{aligned}(p_{,i})_{r_1-0} &= p_{q,R} x^R{}_{,i}\,,\\ (p_{,i})_{r_1+0} &= \bar{p}_R x^R{}_{,i} + (p_q - p_h)\alpha_{,i}\,.\end{aligned} \tag{4.1}$$

At the boundary the values of $p_{q,I}$ and $\bar{p}_I$ coincide due to the continuity of x^I and α. So, in order to get a smooth pressure gradient, the jumps of $x^I{}_{,i}$ and $\alpha_{,i}$ must be matched; if not, the acceleration will be different in the two sides of the boundary even in the static initial configuration, so a discontinuous flow pattern will develop. As it was mentioned above, here we restrict ourselves to smooth initial conditions; then the jumps are matched at $t=0$, and in principle the stability of the equation of motion against shock front is no new condition compared to the situation without phase transition.

Nevertheless, we still have to check if the phase equilibrium can be continuously maintained at B_1 during the expansion. The question is not trivial, because the evolution equations (3.15) and (3.23) clearly differ, therefore the comoving derivatives of intensives of matter elements just on different sides of B_1 differ too. However, this fact does not imply any discontinuity developing at B_1, since B_1 is not comoving with the matter. The situation can be visualized as follows.

Consider the moment t, when the boundary is at r_1; the three-velocity of the matter just at the boundary is v, while that of B_1 is w; for definiteness' sake assume that $w > v$. Then in the moment $t + dt$ the above mentioned matter element is at $r_1 + vdt$, while B_1 is at $r_1 + wdt$. Thus both matter elements just at the two sides of B_1 in $t + dt$, have been governed by the common evolution equation (3.23) (that of the mixed phase) up to $t + dt$, and therefore their x^I values are equal. Mutatis mutandis, this argumentation remains valid for $w < v$ too, while for $w = v$ B_1 is comoving with the matter, therefore $\dot{\alpha} = 0$, and then Eqs (3.15) and (3.23) yield the same evolutions.

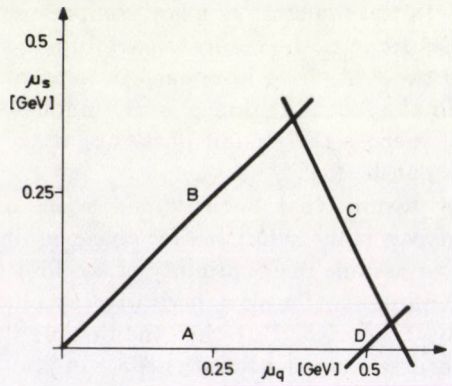


Fig. 1. Approximative picture of the phase coexistence region on the (μ_q, μ_s) plane at moderate temperatures. The possible region is the interior of the ABCD irregular quadrangle. Line A is the original neutral quark-plasma; beyond Line B $\mu_K < 0$, therefore $n_K < n_{\overline{K}}$ at low n_Λ, so both n_{sq} and n_{sh} would be positive. On Line C $\mu_\Lambda = m_\Lambda$, but $\mu_K < m_K$, so beyond that again both phases possess positive strangeness. Finally, beyond Line D there is a Bose condensate of kaons, therefore for substantial amount of hadronic phase the total strageness would be strongly negative.

The above discussion leads to the consequence that matching conditions are insufficient to govern the motion of B_1 and B_2. This is true, and these boundaries are determined by the evolutions of the intensives. There is a definite domain of x^I where mixed phase can exist, and at any moment t the boundaries are localised in the volume elements just leaving this domain. For moderate temperatures the coexistence region is roughly temperature-independent, and displayed on Fig. 1 as being approximately the interior of the irregular quadrangle ABCD; B_1 is roughly

located at A and D, while B_2 at B or C. (More explanation can be found in the Figure Caption.) Since in this way the motion of B's is completely determined by the equations discussed previously, and, on the other hand, its actual calculation may depend on the details of the evolution of the matter (e.g. whether it crosses line B or C when going into the pure hadronic phase), we leave the technical details of this question to the particular calculations.

5. The center of mass coordinate system

As we have seen, the variables of the problem are as follows: for the local description of the matter $x^I = T, \mu_1, \mu_2$, and α, which is 0 or 1 in pure phases; for the flow the components of u^i. The needed information is the forms of the potential functions $p_q(x^I)$ and $p_h(x^I)$, and the initial and boundary conditions. The initial conditions do not need more discussion, the boundary conditions are: $\mu_s = 0$ at $B_1, (n_s)_h = 0$ at B_2 and $p = 0$ at the surface [10]. Then the problem is completely defined, except that we are still in an arbitrary coordinate system, thus the form of the covariant derivative, denoted by a semicolon, is not specified.

Three different kinds of coordinate systems can be serious candidates: the first is comoving with the matter, the second is comoving with phase transition and the third is rigid, centered at the center of mass of the sphere. While the first possibility is simple enough in some sense [11], now the phase boundary is not comoving with the matter, so some part of the usual simplicity is lost. The second system is not uniquely defined outside the coexistence region. The third can be constructed in the following way.

Let us introduce spherical coordinates in the Minkowski spacetime

$$\xi^i = (t, r, \vartheta, \varphi) \tag{5.1}$$

in such a way that the metric tensor obtains the form

$$g_{ik} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2 \vartheta \end{bmatrix}. \tag{5.2}$$

In this coordinate system, due to the sphericity of the expansion and the normalization of u^i,

$$u^i = \left(\sqrt{1+u^2}, u, 0, 0\right). \tag{5.3}$$

Since u is the spatial component of the four-velocity, it is not bound from above. The three-velocity can be obtained as $v^1 = u^1/u^0$.

The covariant derivative of u^i is defined as

$$u^i{}_{;k} = u^i{}_{,ki} + \Gamma_{kr} u^r, \tag{5.4}$$

where $\Gamma_{ik}{}^{m}$ is the Christoffel symbol [12], calculable from g_{ik}. Since the acceleration is orthogonal to u^i, and the latter possesses only one nontrivial independent component, Eq. (3.6) reduces to one component, e.g. to the radial one. Its explicit form is

$$(\rho + p)(\sqrt{1+u^2}\, u_{,0} + u u_{,1}) + p_{,0}\sqrt{1+u^2}\, u + p_{,1}(1+u^2) = 0, \tag{5.5}$$

where, as it has been discussed, ρ and p depend on t and r via T, μ_1, μ_2 and α. For the evolution equations of these latter local variables, note that, according to the definition (3.3):

$$T = T_{,0}\sqrt{1+u^2} + T_{,1}\, u \tag{5.6}$$

and so on, while using Eq. (5.4),

$$u^r{}_{;r} = \frac{u u_{,0}}{\sqrt{1+u^2}} + u_{,1} + \frac{2}{r} u \tag{5.7}$$

and then all the needed particular equations can be directly constructed.

6. Conclusions

In this paper we have demonstrated that fractional destillation type processes in the phase transition of a multicomponent fluid do not hinder the consistent relativistic hydrodynamical description of the transition: in fact, the system of evolution equations is presented in Section 3. Of course, some simplifying assumptions have been used, as e.g. the equilibrium nature of the phase transition and the uniqueness of the velocity field. Nevertheless, it is well known that delayed phase transitions are consistent with hydrodynamics, too [13]. As for the uniqueness of u^i, it is worthwhile to note that no a priori guarancy exists for the proportionality of the particle current vectors n^i_α; if these vectors are not parallel, then diffusion effects appear [4]. However, for sufficiently strong interactions viscosity-type forces lead to the equilibration of velocities.

The presented formalism can be applied for calculating the strange particle yields in a heavy ion collision containing deconfinement stage of the evolution. We restricted the number of different quark flavors to 2, but this was done only for simplicity's sake: since each individual flavor possesses its own conserved quantum number, the equations have the same structure in the generic case.

Acknowledgements

The authors would like to thank Dr. I. Lovas for illuminating discussions. This work was partly supported by the International Exchange Program of the Hungarian Academy of Sciences and the National Science Foundation of the USA.

Appendix: The variables of the thermodynamic potential

Consider a multicomponent matter. Then it fulfils the thermodynamic relations

$$s = s(\rho, n_\alpha), \tag{A.1}$$

$$p = Ts - \rho + \sum_\alpha \mu_\alpha n_\alpha, \tag{A.2}$$

$$\frac{1}{T} = \frac{\partial s}{\partial \rho},$$

$$\frac{\mu_\alpha}{T} = -\frac{\partial s}{\partial n_\alpha}. \tag{A.3}$$

Hence, by introducing T and μ_α as new variables, one obtains [14], [15]

$$n_\alpha = \frac{\partial p}{\partial \mu_\alpha},$$

$$s = \frac{\partial p}{\partial T}, \tag{A.4}$$

while ρ can be obtained from Eq. (A.2).

Now, consider our p functions. In the quark phase clearly

$$p_q = p_q(\mu_q, \mu_s, T). \tag{A.5}$$

In the hadronic phase

$$p_h = p_h(\mu_h, T), \tag{A.6}$$

wher μ_h stands for all the different hadronic chemical potentials, but, due to combined chemical and phase equilibrium,

$$\mu_h = c_{hq}\mu_q + c_{hs}\mu_s, \tag{A.7}$$

where c_{hq} and c_{hs} are the numbers of normal and strange quarks building up the particular hadron. Therefore p_h is a function of μ_q, μ_s and T, too.

Now we show that, instead of μ_q and μ_s, one can introduce special combinations μ_1 and μ_2 so that the conserved number densities n_B and n_S be direct derivatives of p. In order to see this, let us write

$$n_B = \frac{1}{3}(n_q + n_s) = \frac{1}{3}\left(\frac{\partial p}{\partial \mu_q} + \frac{\partial p}{\partial \mu_s}\right),$$

$$n_S = n_s = \frac{\partial p}{\partial \mu_s}, \tag{A.8}$$

in both phases. Then, by introducing new combinations

$$\mu_1 = 3\mu_q,$$
$$\mu_2 = -\mu_q + \mu_s \tag{A.9}$$

into p, Eqs (A.8) give

$$n_B = \frac{\partial p}{\partial \mu_1},$$
$$n_S = \frac{\partial p}{\partial \mu_2}, \tag{A.10}$$

as used in this paper. For our present purposes the specification of the particular form of the pressure function is not necessary.

References

1. H.B. Callen, Thermodynamics. J. Wiley, New York, 1960.
2. Kang S. Lee, M.J. Rhoades-Brown and U. Heinz, Phys.Lett., *B174*, 123, 1986.
3. B. Lukács, J. Zimányi and N.L. Balazs, Phys. Lett., *B183*, 27, 1987.
4. J. Ehlers, in: Relativity, Astrophysics and Cosmology, ed. W. Israel, D. Reidel Publ. Co., Dordrecht, 1973.
5. B. Lukács, Acta Phys. Pol., *B14*, 33, 1983.
6. F. Weinhold, J. Chem. Phys., *63*, 2479, 1975.
7. L. Diósi, G. Forgács, B. Lukács and H.L. Frisch, Phys. Rev., *A29*, 3343, 1984.
8. G.K. Batchelor, An Introduction to Fluid Dynamics, Cambridge Univ. Press, 1970.
9. S. Flügge, Handbuch der Physik Vol. 8/1. Springer, Berlin-Göttingen-Heidelberg, 1959.
10. J.L. Synge, Relativity: the Generaal Theory. North-Holland, Amsterdam, 1960.
11. B. Kämpfer and B. Lukács, Acta Phys. Hung., *61*, 317, 1987.
12. L.P. Eisenhart, Riemannian Geometry, Princeton Univ. Press, 1950.
13. H.W. Barz, B. Kämpfer, L.P. Csernai and B. Lukács, Phys. Lett., *143B*, 334, 1984.
14. L.D. Landau and E.M. Lifshic, Teoreticheskaya fizika V. Statisticheskaya fizika, chast' I. Nauka, Moscow, 1976.
15. J.D. Walecka, Phys. Lett., *B59*, 109, 1975.

Acta Physica Hungarica 64 (1–3), pp. 123–142 (1988)

GAUGE-INVARIANT QHD LAGRANGIANS*

K. SAILER and I. LOVAS

Department of Theoretical Physics, Kossuth Lajos University
4010 Debrecen, Hungary

(Received 8 January 1987)

A possible way of the introduction of charged vector mesons as gauge fields has been analysed and applied to the gauge-invariant and renormalizable generalization of Walecka's model and that of the linear σ-model

1. Introduction

The charged constituents of the atoms, the atomic nucleus and the electrons, are bounded to each other by Coulomb force, while among the neutral atoms Van der Waals force takes place, which can be derived from the fundamental Coulomb force acting between the constituents of the atoms. Nevertheless, it is reasonable to consider the Van der Waals force as a fundamental one, if the single electron and nuclear degrees of freedom are not playing any significant role in the physical phenomenon being in the scope of our interest. (For example, the thermodynamical properties of a dilute monoatomic gas are not — or almost not — influenced by the internal dergrees of freedom of the atoms).

In nuclear physics, one confronts with a similar situation. The color-charged constituents of the hadrons, the quarks and the antiquarks, are bounded by strong interaction described by QCD, while the interaction between the colorless hadrons (e. g. the NN-and πN-interaction) reveals itself as a force of Van der Waals type. It is believed that this force can be derived from QCD. Nevertheless, it is clear that a lot of properties of finite nuclei, nuclear matter and nuclear reactions are rather slightly affected by the internal degrees of freedom of the nucleons. Thus, it seems reasonable to work out an effective theory, in which the baryon and meson degrees of freedom are the basic ingredients.

Relativistic quantum many-body systems, such as nuclear systems, can be treated in the framework of quantum field theory. The quantum field theory, pretending to describe hadron–hadron interactions and hadronic many-body systems is called Quantumhadrodynamics (QHD). The question is, how construct the Lagrangian of QHD. It has been established [1, 2, 3] that in addition to the nucleon field at least scalar-isoscalar field σ and vector-isoscalar field ω_μ should be included into the QHD Lagrangian in order to reproduce nuclear saturation. This simple version of QHD, Walecka's model is renormalizable [4].

*Dedicated to Prof. G. Marx on his 60th birthday.

It seems to be important to maintain renormalizability when one intends to generalize Walecka's model. First, only in a renormalizable theory is it possible to compute quantum corrections to the expectation values of observables in a consistent way. Second, a renormalizable theory has a finite number of parameters (masses and coupling constants) and having fitted them to experimental data all the observables are finite. Third, there is no need of introducing a high-momentum cut-off and therefore a renormalizable theory is highly insensitive to the details of the short-range hadron–hadron interactions.

Here it should be mentioned that although nonrenormalizable generalizations of Walecka's model exist, they have not made a success. Including for example, a $\sigma^2\omega_\mu\omega^\mu$ interaction term into the Lagrangian, the model is unable to describe the ground states of finite nuclei [5, 6]. Introducing the pion field with nonrenormalizable pseudovector NNπ-coupling, predictions have been made for the threshold density of the pion condensation [7, 8]. The predicted threshold densities are significantly smaller, than it would be expected [9].

The Lagrangian of QHD is strongly restricted by the requirement of renormalizability. In Walecka's model the Yukawa coupling of the neutral vector-meson field ω_μ to the conserved baryon current is renormalizable. More problematical is the introduction of the charged vector-isovecto r field $\boldsymbol{\rho}_\mu$, which is responsible for the asymmetry energy in systems with different numbers of neutrons and protons ($N \neq Z$). Its coupling to the conserved vector-isovector current is not renormalizable. The only way of including charged vector mesons without destroying renormalizability is to introduce them as massless gauge fields and to generate their masses by the Higgs mechanism.

In Section 2 of the present paper we shall analyse the procedure of including vector mesons in a renormalizable way. In Section 3 and Section 4 examples will be given on the gauge-invariant generalization of Walecka's model and that of the linear σ-model, respectively.

2. Vector mesons as gauge fields

In order to introduce gauge fields, one should decide, what global symmetries of the model will be raised to the level of local symmetries. It would seem natural to connect the inclusion of the vector fields ω_μ and $\boldsymbol{\rho}_\mu$ with making local the baryon number symmetry U(1) and the isospin symmetry SU(2), respectively. In order to get massive gauge fields, however, one should assume that the above mentioned symmetries are spontaneously broken by the vacuum state. But such an assumption is unacceptable, since it would mean that the vacuum state is an eigenstate neither of the baryon number nor of the isospin.

The idea, how to overcome this difficulty, has been proposed by de Wit [10], Fabricius and Fleischer [11] and Serot [12]. We shall outline their procedure in a more general way. Let us suppose, S is the Lie-group of unbroken physical symmetries leading to the conservation of the baryon number, isospin, etc. The idea is to start with a Lagrangian of the larger symmetry group $G \times H \supset S$, where G and

H are Lie-groups of global and local symmetries, respectively. Vector meson fields are introduced now as gauge fields of the local symmetry group H. Thereafter, by introducing selfinteracting Lorentz-scalar fields, the symmetry group $G \times H$ is spontaneously broken down to its subgroup S and, in the same time, the gauge fields become massive due to the Higgs mechanism.

This program can be realized by making use of the following (trivial) mathematical statement. Let G and H be two isomorphic Lie-groups, whose generators corresponded to each other by the isomorphism are $\hat{\mathcal{G}}^a$ and $\hat{\mathcal{H}}^a$ $(a = 1, 2, \ldots, N)$, respectively. Then $\hat{\mathcal{S}}^a = \hat{\mathcal{G}}^a + \hat{\mathcal{H}}^a$ $(a = 1, 2, \ldots, N)$ are generators of subgroup $S \subset G \times H$ and the subgroup S is isomorphic with both G and H.

The model with the required properties can be constructed making the following assumptions:

i/ Let S be the Lie-group of the unbroken global symmetries. Choose the Lie-groups G and H to be isomorphic with S. Let denote by $\hat{\mathcal{G}}^a$ and $\hat{\mathcal{H}}^a$ $(a = 1, 2, \ldots, N)$ the generators of the groups G and H, respectively, which are corresponded to each other by the isomorphism. Let identify the group S with the subgroup of $G \times H$, whose generators are $\hat{\mathcal{S}}^a = \hat{\mathcal{G}}^a + \hat{\mathcal{H}}^a$ $(a = 1, 2, \ldots, N)$.

ii/ Construct a Lagrangian $\mathcal{L}_0$ with the global symmetry $G \times H$ from the fields with the following transformation properties:

	G	$\times$	H
ψ	singlet		multiplet
ϕ	singlet		multiplet
Φ	multiplet		multiplet

Here ψ is a fermion field, ϕ and Φ denote Lorentz scalar boson fields. All the fields are assumed to be G singlets, except for the field Φ, which is meant to produce the spontaneous symmetry breaking, as we shall see below. (We use real representations for the Lorentz-scalar fields.)

iii/ Let the field $\Phi \in \mathbf{R}^n$ have a self-interaction

$$V(\Phi^T\Phi) = \mu^2\Phi^T\Phi + \lambda(\Phi^T\Phi)^2 \qquad (\mu^2 < 0, \quad \lambda > 0) \tag{1}$$

with the symmetry $O(n)$. The dimension of the representation is chosen to fulfil the relations $\mathcal{R}_n(G), \mathcal{R}_n(H) \subset O(n)$, where $\mathcal{R}_n(G)$ and $\mathcal{R}_n(H)$ are the $n \times n$ matrix representations of the groups G and H, respectively, according to which the field Φ transforms.

As a consequence of assumption iii/ the field Φ has a nonvanishing vacuum

expectation value [13]

$$< \Phi >= \begin{pmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ w \end{pmatrix} \neq 0 \qquad (w = \sqrt{-\mu^2/(2\lambda)}) \tag{2}$$

and its mass matrix is diagonal of the form

$$(\mathcal{M}_\Phi^2)_{ij} = \left.\frac{\delta^2 V}{\delta\Phi_i \delta\Phi_j}\right|_{\Phi=<\Phi>} = 8\lambda < \Phi >_i < \Phi >_j \quad (i,j = 1,2,\ldots,n) \tag{3}$$

i. e.

$$\mathcal{M}_\Phi^2 = \begin{pmatrix} 0 & \ldots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & 0 \\ 0 & \ldots & 0 & m_\eta^2 \end{pmatrix}, \qquad m_\eta^2 = 8\lambda w^2. \tag{4}$$

It means, that the field Φ has $(n-1)$ Goldstone components with zero mass and one Higgs-component with the mass m_η.

iv/ Choose the representations $\mathcal{R}_n(G)$ and $\mathcal{R}_n(H)$ in which

$$\hat{\mathcal{G}}^a < \Phi >= \Gamma_\Phi(\hat{\mathcal{G}}^a) < \Phi >\neq 0, \quad \hat{\mathcal{H}}^a < \Phi >= \Gamma_\Phi(\hat{\mathcal{H}}^a) < \Phi >\neq 0 \quad (a = 1,2,\ldots,N), \tag{5}$$

but

$$\hat{\mathcal{S}}^a < \Phi >= \Gamma_\Phi(\hat{\mathcal{S}}^a) < \Phi >= 0 \quad (a = 1,2\ldots,N). \tag{6}$$

As a consequence, the symmetry group $G \times H$ of dimension $2N$ breaks down spontaneously to its subgroup S of dimension N. A comparison with the consequences of assumption iii/ shows that $n \geq N+1$ should be fulfilled.

v/ In order to introduce a minimum number of Goldstone components of the field Φ, we choose

$$n = N + 1. \tag{7}$$

As we shall see later, assumption v/ enables us to transform away all the Goldstone components of the field Φ.

vi/ The Lagrangian with the global symmetry $G \times H$ has now the following form

$$\begin{aligned} \mathcal{L}_0 =& \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi - g_\phi \bar{\psi}\Gamma\psi\phi + \\ &+ \frac{1}{2}\partial_\mu \phi^T \partial^\mu \phi - \frac{1}{2} m_\phi^2 \phi^T \phi + \\ &+ \frac{1}{2}\partial_\mu \Phi^T \partial^\mu \Phi - V(\Phi^T \Phi). \end{aligned} \tag{8}$$

Let now raise the symmetry H to the level of local symmetry, introducing the gauge fields $b_\mu^a (a = 1, 2, \ldots, N)$ with minimal coupling to the fields being not H-singlet. This is done by replacing their derivatives with covariant ones:

$$\begin{aligned} D_\mu \psi &= (\partial_\mu - i g_H \Gamma_\psi(\hat{\mathcal{H}}^a) b_\mu^a)\psi, \\ D_\mu \phi &= (\partial_\mu - i g_H \Gamma_\phi(\hat{\mathcal{H}}^a) b_\mu^a)\phi, \\ D_\mu \Phi &= (\partial_\mu - i g_H \Gamma_\Phi(\hat{\mathcal{H}}^a) b_\mu^a)\Phi. \end{aligned} \tag{9}$$

Here Γ_ψ, Γ_ϕ and Γ_Φ denote the matrix representations of group H according to which the fields ψ, ϕ and Φ transform, respectively. The following Lagrangian can be obtained:

$$\begin{aligned} \mathcal{L} = &\bar{\psi}(i\gamma^\mu D_\mu - m)\psi - g_\phi \bar{\psi}\Gamma\psi\phi + \\ &+ \frac{1}{2}(D_\mu \phi)^T D^\mu \phi - \frac{1}{2} m_\phi^2 \phi^T \phi + \\ &+ \frac{1}{2}(D_\mu \Phi)^T D^\mu \Phi - V(\Phi^T \Phi) - \frac{1}{4} b_{\mu\nu}^a b^{\mu\nu a} \end{aligned} \tag{10}$$

with the field strength

$$b_{\mu\nu}^a = \partial_\mu b_\nu^a - \partial_\nu b_\mu^a + g_H f^{abc} b_\mu^b b_\nu^c. \tag{11}$$

(f^{abc} are the structure constants of the Lie-group H).

Due to assumptions v/ and vi/, all the $N = n - 1$ Goldstone components of the field Φ can be transformed away. In order to do this, the field Φ is parametrized as follows

$$\Phi = \exp\{i\Gamma_\Phi(\hat{\mathcal{H}}^a)\xi^a(x)\} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ w + \eta(x) \end{pmatrix}. \tag{12}$$

Performing now the local gauge transformation

$$\hat{U}_H(x) = \exp\{-i\hat{\mathcal{H}}^a \xi^a(x)\}, \tag{13}$$

only the Higgs component $\eta(x)$ of the field Φ is retained in the Lagrangian,

$$\Phi \to \Phi' = \hat{U}_H(x)\Phi = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ w + \eta(x) \end{pmatrix}. \tag{14}$$

(For simplicity, we shall not distinguish the transformed fields from the original ones in our notations, unless a distinction is unavoidable.)

As a consequence of the Higgs mechanism [13], in the Lagrangian appears a mass term quadratic in the gauge fields,

$$\frac{1}{2}(M^2)^{ab}b_\mu^a b^{\mu b} \tag{15}$$

with the symmetric and real mass matrix

$$(M^2)^{ab} = g_H^2 w^2 \Gamma_\Phi(\hat{\mathcal{H}}^a)_{ni}\Gamma_\Phi(\hat{\mathcal{H}}^b)_{in} \tag{16}$$

(no summation for n). Therefore exists an orthogonal matrix ϑ by means of which the mass matrix M^2 can be diagonalized

$$M_B^2 = \vartheta M^2 \vartheta^T = \text{diagonal}\,, \tag{17}$$

and the mass eigenstates are

$$B_\mu^a = \vartheta^{aa'} b_\mu^{a'}. \tag{18}$$

Disregarding the constant terms, the selfinteraction potential functional V can be rewritten in terms of the Higgs field

$$V = \frac{1}{2}m_\eta^2\eta^2 + 4\lambda w\eta^3 + \lambda\eta^4, \tag{19}$$

where

$$m_\eta^2 = -4\mu^2 = 8\lambda w^2 \tag{20}$$

is the mass of the Higgs boson η.

Finally, the Lagrangian takes the form:

$$\begin{aligned}
\mathcal{L} =& \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - g_\phi\bar{\psi}\Gamma\psi\phi + g_H\bar{\psi}\gamma^\mu\Gamma_\psi(\hat{\mathcal{H}}^a)(\vartheta^T)^{ab}B_\mu^b\psi + \\
&+ \frac{1}{2}(\partial_\mu\phi)^T\partial^\mu\phi - \frac{1}{2}m_\phi^2\phi^T\phi + \\
&+ \frac{1}{2}\partial_\mu\eta\partial^\mu\eta - \frac{1}{2}m_\eta^2\eta^2 - 4\lambda w\eta^3 - \lambda\eta^4 + \\
&+ ig_H\phi^T\Gamma_\phi(\hat{\mathcal{H}}^a)(\vartheta^T)^{ab}B_\mu^b\partial^\mu\phi + \\
&+ \frac{1}{2}g_H^2\vartheta^{ca}\phi^T\Gamma_\Phi(\hat{\mathcal{H}}^a)\Gamma_\phi(\hat{\mathcal{H}}^b)\phi(\vartheta^T)^{bd}B_\mu^c B^{\mu d} + \\
&+ \frac{1}{2}g_H^2\eta(\eta+2w)\vartheta^{ca}\Gamma_\Phi(\hat{\mathcal{H}}^a)_{ni}\Gamma_\Phi(\hat{\mathcal{H}}^b)_{in}(\vartheta^T)^{bd}B_\mu^c B^{\mu d} - \\
&- \frac{1}{4}B_{\mu\nu}^a B^{\mu\nu a} + \frac{1}{2}(M_B^2)^{ab}B_\mu^a B^{\mu b},
\end{aligned} \tag{21}$$

where the notation

$$B_{\mu\nu}^a = \partial_\mu B_\nu^a - \partial_\nu B_\mu^a + g_H f^{abc}B_\mu^b B_\nu^c \tag{22}$$

has been introduced. Use has been made of the equation

$$(\hat{U}_H\Phi)^T\Gamma_\Phi(\hat{\mathcal{H}}^a)\partial^\mu(\hat{U}_H\Phi) = 0. \tag{23}$$

Up to this point, the scalar-meson field ϕ has been supposed to have vanishing vacuum expectation value. Let us now examine the case, when it has a nonvanishing vacuum expectation value, due to which the group S of global symmetries is spontaneously broken down to its subgroup S_0. This can happen, if the following further assumptions are made:

vii/ Let the field ϕ have a selfinteraction of the form

$$\mathcal{V}(\phi^T\phi) = \mu_0^2\phi^T\phi + \lambda_0(\phi^T\phi)^2, \tag{24}$$
$$(\phi \in \mathbf{R}^{n_\phi}, \quad \mu_0^2 < 0, \quad \lambda_0 > 0),$$

with the symmetry $O(n_\phi)$, which is larger than the $n_\phi \times n_\phi$ matrix representations of the groups G and H, according to which the field ϕ transforms, $\mathcal{R}_{n_\phi}(G), \mathcal{R}_{n_\phi}(H) \subset O(n_\phi)$.

viii/ The existence of a subgroup $S_0 \subset S$ of unbroken symmetries with the generators $\hat{\mathcal{S}}^a (a = 1, 2, \ldots, N_0, N_0 < N)$ is supposed:

$$\hat{\mathcal{S}}^a < \phi' >= 0 \qquad (a = 1, 2, \ldots, N_0), \tag{25}$$

where $\phi' = \hat{U}_H(x)\phi$.

Denote the vacuum expectation value of the field ϕ by

$$< \phi' >= \begin{pmatrix} 0 \\ \vdots \\ 0 \\ w_0 \end{pmatrix} \qquad (w_0 = \sqrt{-\mu_0^2/(2\lambda_0)} \neq 0) \tag{26}$$

and make the following shift

$$\phi' =< \phi' > +\varphi, \qquad (\varphi \in \mathbf{R}^{n_\phi}). \tag{27}$$

The Goldstone components of the field Φ can be eliminated by the local gauge transformation $\hat{U}_H(x)$, as before, but the number $n-1$ Goldstone components of the field φ cannot be eliminated. They correspond to real Goldstone bosons with zero mass, $m_i^2 = 0$ $(i = 1, 2, \ldots, n_\phi - 1)$, while the n_ϕ-th component of the field $\varphi, \varphi_{n_\phi} = \sigma'$ is a Higgs boson with the mass

$$m_\sigma^2 = -4\mu_0^2 = 8\lambda_0 w_0^2, \tag{28}$$

as it can be read from the selfinteraction potential (24) replacing Eq. (27) into it and omitting the constant term:

$$\mathcal{V}(\phi^T\phi) = \frac{1}{2}m_\sigma^2\sigma'^2 + 4\lambda_0 w_0\sigma'\left(\sum_{i=1}^{n_\phi-1}\varphi_i\varphi_i + \sigma'^2\right) + \lambda_0\left(\sum_{i=1}^{n_\phi-1}\varphi_i\varphi_i + \sigma'^2\right)^2. \tag{29}$$

The mass matrix of the vector-meson fields b_μ^a is also modified as a consquence of the additional symmetry breaking

$$(M^2)^{ab} = g_H^2(w^2\Gamma_\Phi(\hat{\mathcal{H}}^a)_{ni}\Gamma_\Phi(\hat{\mathcal{H}}^b)_{in} + w_0^2\Gamma_\phi(\hat{\mathcal{H}}^a)_{n_\phi j}\Gamma_\phi(\hat{\mathcal{H}}^b)_{jn_\phi}),$$
$$(i = 1,2,\ldots,n; \quad j = 1,2,\ldots n_\phi), \tag{30}$$

but it can be diagonalized as before.

The Lagrangian takes now the following form:

$$\begin{aligned}
\mathcal{L} =& \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - g_\phi\bar{\psi}\Gamma\psi(<\phi'> + \varphi) + \\
&+ g_H\bar{\psi}\gamma^\mu\Gamma_\psi(\hat{\mathcal{H}}^a)(\vartheta^T)^{ab}B_\mu^b\psi + \\
&+ \frac{1}{2}\left(\partial_\mu\sigma'\partial^\mu\sigma' + \sum_{i=1}^{n_\phi-1}\partial_\mu\varphi_i\partial^\mu\varphi_i\right) - \frac{1}{2}m_\sigma^2\sigma'^2 - \\
&- 4\lambda_0 w_0\sigma'\left(\sigma'^2 + \sum_{i=1}^{n_\phi-1}\varphi_i\varphi_i\right) - \lambda_0\left(\sigma'^2 + \sum_{i=1}^{n_\phi-1}\varphi_i\varphi_i\right)^2 + \\
&+ ig_H\sum_{j=1}^{n_\phi-1}\left[(w_0+\sigma')\Gamma_\phi(\hat{\mathcal{H}}^a)_{n_\phi j} + \sum_{i=1}^{n_\phi-1}\varphi_i\Gamma_\phi(\hat{\mathcal{H}}^a)_{ij}\right]\partial^\mu\varphi_j\cdot b_\mu^a + \\
&+ g_H^2 w_0\vartheta^{ca}\Gamma_\phi(\hat{\mathcal{H}}^a)_{n_\phi j}\Gamma_\phi(\hat{\mathcal{H}}^b)_{jk}(\vartheta^T)^{bd}\varphi_k B_\mu^c B^{\mu d} + \\
&+ \frac{1}{2}g_H^2\varphi_i\vartheta^{ca}\Gamma_\phi(\hat{\mathcal{H}}^a)_{ij}\Gamma_\phi(\hat{\mathcal{H}}^b)_{jk}(\vartheta^T)^{bd}\varphi_k B_\mu^c B^{\mu d} + \\
&+ \frac{1}{2}\partial_\mu\eta\partial^\mu\eta - \frac{1}{2}m_\eta^2\eta^2 - 4\lambda w\eta^3 - \lambda\eta^4 + \\
&+ \frac{1}{2}g_H^2\eta(\eta+2w)\vartheta^{ca}\Gamma_\Phi(\hat{\mathcal{H}}^a)_{ni}\Gamma_\Phi(\hat{\mathcal{H}}^b)_{in}(\vartheta^T)^{bd}B_\mu^c B^{\mu d} - \\
&- \frac{1}{4}B_{\mu\nu}^a B^{\mu\nu a} + \frac{1}{2}(M_B^2)^{ab}B_\mu^a B^{\mu b}.
\end{aligned} \tag{31}$$

All what was said before, has a straighforward generalization for the case, when S is a direct product of several semisimple Lie-groups, $S = S_1 \times S_2 \times \ldots \times S_q$. We start now with the symmetry $G \times H$, where $G = G_1 \times G_2 \times \ldots \times G_q$ and $H = H_1 \times H_2 \times \ldots \times H_q$. Then we introduce q gauge fields and coupling constants raising the symmetry H to the level of local symmetry, and introduce the same number of multiplets of Lorentz-scalar fields, namely, one multiplet to each of the subgroups $G_r \times H_r$ which is singlet with respect to the other symmetries and which breaks the symmetry $G_r \times H_r$ to its subgroup S_r spontaneously. For this purpose, we prescribe appropriate selfinteractions of the form of Eq. (1) to these scalar fields ($\mu_r^2 < 0$, $\lambda_r > 0$, $r = 1,2,\ldots,q$).

3. Generalization of Walecka's model

As an example, the general scheme described in the previous Section has been applied to the generalization of Walecka's model. Let S be the direct product of

the isospin symmetry group SU(2) and that of the baryon number symmetry group U(1), S=SU(2)×U(1). Now we construct a Lagrangian with the symmetry $G \times H$, where G=(SU(2)× U(1))$_g$ and H=(SU(2)×U(1))$_l$. (lower case g and l indicate global and local symmetries, respectively). The transformation properties of the fields are given as follows:

	$(SU(2) \times U(1))_g$		$\times$ $(SU(2) \times U(1))_l$	
ψ	0	0	1/2	1
$\boldsymbol{\phi}$	0	0	1	0
σ	0	0	0	0
$\Phi_{(1)}$	1/2	0	1/2	0
$\Phi_{(2)}$	0	1	0	-1
	${}_g\hat{T}^a$	${}_g\hat{Y}$	${}_l\hat{T}^a$	${}_l\hat{Y}$
	—	—	g_1	g_2
	—	—	$B^{\mu a}_{(1)}$	$B^{\mu}_{(2)}$

The generators of the subgroups G_1=SU(2)$_g$, G_2=U(1)$_g$, H_1=SU(2)$_l$, H_2=U(1)$_l$, S_1=SU(2) and S_2=U(1) denoted by ${}_g\hat{T}^a$, ${}_g\hat{Y}$, ${}_l\hat{T}^a$, ${}_l\hat{Y}$, $\hat{T}^a$ and $\hat{Y}$ are represented as follows (a, j, k=1,2,3):

$$\Gamma_\psi({}_g\hat{T}^a) = 0, \quad \Gamma_\psi({}_l\hat{T}^a) = \Gamma_\psi(\hat{T}^a) = \frac{1}{2}\tau^a, \quad \Gamma_\psi({}_g\hat{Y}) = 0,$$

$$\Gamma_\psi({}_l\hat{Y}) = \Gamma_\psi(\hat{Y}) = 1;$$

$$\Gamma_\phi({}_g\hat{T}^a) = 0, \quad \Gamma_\phi({}_l\hat{T}^a)_{jk} = \Gamma_\phi(\hat{T}^a)_{jk} = -i\varepsilon_{ajk}, \quad \Gamma_\phi({}_g\hat{Y}) = 0,$$

$$\Gamma_\phi({}_l\hat{Y}) = \Gamma_\phi(\hat{Y}) = 0;$$

$$\left[\Gamma_{\Phi_{(1)}}({}_g\hat{T}^a)\Phi_{(1)}\right]_j = -\frac{i}{2}\Phi_{(1)0}\delta_{aj} + \frac{i}{2}\varepsilon_{jak}\Phi_{(1)k}, \quad \left[\Gamma_{\Phi_{(1)}}({}_g\hat{T}^a)\right]_0 = +\frac{i}{2}\Phi_{(1)a},$$

$$\left[\Gamma_{\Phi_{(1)}}({}_l\hat{T}^a)\Phi_{(1)}\right]_j = +\frac{i}{2}\Phi_{(1)0}\delta_{aj} + \frac{i}{2}\varepsilon_{jak}\Phi_{(1)k}, \quad \left[\Gamma_{\Phi_{(1)}}({}_l\hat{T}^a)\right]_0 = -\frac{i}{2}\Phi_{(1)a},$$

$$\left[\Gamma_{\Phi_{(1)}}(\hat{T}^a)\Phi_{(1)}\right]_j = i\varepsilon_{jak}\Phi_{(1)k}, \quad \left[\Gamma_{\Phi_{(1)}}(\hat{T}^a)\Phi_{(1)}\right]_0 = 0,$$

$$\Gamma_{\Phi_{(1)}}({}_g\hat{Y})\Phi_{(1)} = \Gamma_{\Phi_{(1)}}({}_l\hat{Y})\Phi_{(1)} = \Gamma_{\Phi_{(1)}}(\hat{Y})\Phi_{(1)} = 0;$$

$$\Gamma_{\Phi_{(2)}}({}_g\hat{T}^a) = \Gamma_{\Phi_{(2)}}({}_l\hat{T}^a) = \Gamma_{\Phi_{(2)}}(\hat{T}^a) = 0,$$

$$\Gamma_{\Phi_{(2)}}({}_g\hat{Y}) = Q, \quad \Gamma_{\Phi_{(2)}}({}_l\hat{Y}) = -Q, \quad \Gamma_{\Phi_{(2)}}(\hat{Y}) = 0, \tag{32}$$

where the matrix

$$Q = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} \tag{33}$$

has been introduced. By direct calculation one can check, that requirement iv/ is fulfilled.

The procedure described in Section 2 leads to the following expression for the Lagrangian:

$$\begin{aligned}\mathcal{L} =&\bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - g_\phi\bar{\psi}\mathbf{\Gamma}\psi\boldsymbol{\phi} - g_\sigma\bar{\psi}\psi\sigma + g_1\bar{\psi}\gamma^\mu\frac{\tau}{2}\mathbf{B}_{(1)\mu}\psi + g_2\bar{\psi}\gamma^\mu B_{(2)\mu}\psi + \\ &+\frac{1}{2}\partial_\mu\boldsymbol{\phi}\partial^\mu\boldsymbol{\phi} - \frac{1}{2}m_\phi^2\boldsymbol{\phi}\boldsymbol{\phi} + g_1(\boldsymbol{\phi}\times\partial^\mu\boldsymbol{\phi})\mathbf{B}_{(1)\mu} + \frac{1}{2}g_1^2\left(\boldsymbol{\phi}\times\mathbf{B}_{(1)\mu}\right)\left(\boldsymbol{\phi}\times\mathbf{B}_{(1)}^\mu\right) + \\ &+\frac{1}{2}\partial_\mu\sigma\partial^\mu\sigma - \frac{1}{2}m_\sigma^2\sigma^2 - U_{(\sigma)} + \\ &+\frac{1}{4}g_1^2w_1\eta\mathbf{B}_{(1)\mu}\mathbf{B}_{(1)}^\mu + \frac{1}{8}g_1^2\eta^2\mathbf{B}_{(1)\mu}\mathbf{B}_{(1)}^\mu + \\ &+\frac{1}{2}\partial_\mu\eta\partial_\eta^\mu - \frac{1}{2}m_\eta^2\eta^2 - \frac{1}{4}\frac{m_\eta^2}{M_1}g_1\eta^3 - \left(\frac{m_\eta}{M_1}\right)^2\frac{g_1^2}{32}\eta^4 + \\ &+g_2^2w_2\zeta B_{(2)\mu}B_{(2)}^\mu + \frac{1}{2}g_2^2\zeta^2B_{(2)\mu}B_{(2)}^\mu + \\ &+\frac{1}{2}\partial_\mu\zeta\partial^\mu\zeta - \frac{1}{2}m_\zeta^2\zeta^2 - \frac{1}{2}\frac{m_\zeta^2}{M_2}g_2\zeta^3 - \frac{1}{8}\left(\frac{m_\zeta}{M_2}\right)^2g_2^2\zeta^4 - \\ &-\frac{1}{4}\mathbf{B}_{(1)\mu\nu}\mathbf{B}_{(1)}^{\mu\nu} + \frac{1}{2}M_1^2\mathbf{B}_{(1)}^\mu\mathbf{B}_{(1)\mu} - \frac{1}{4}B_{(2)\mu\nu}B_{(2)}^{\mu\nu} + \frac{1}{2}M_2^2B_{(2)\mu}B_{(2)}^\mu, \end{aligned} \tag{34}$$

where

$$\mathbf{B}_{(1)\mu\nu} = \partial_\mu\mathbf{B}_{(1)\nu} - \partial_\nu\mathbf{B}_{(1)\mu} + g_1\mathbf{B}_{(1)\mu}\times\mathbf{B}_{(1)\nu}, \quad B_{(2)\mu\nu} = \partial_\mu B_{(2)\nu} - \partial_\nu B_{(2)\mu}. \tag{35}$$

Here $\mathrm{B}_{(r)\mu}$ and g_r denote the gauge fields and the coupling constants connected with the groups H_r of local symmetries, respectively; η and ζ denote the Higgs components of the fields $\Phi_{(1)}$ and $\Phi_{(2)}$ with the masses m_η^2 and m_ζ^2, respectively. The mass matrices of the gauge fields are diagonal ($\vartheta = 1$):

$$\left(M_{(1)}^2\right)^{ab} = (g_1w_1)^2\delta^{ab} = M_1^2\delta^{ab}, \quad M_2^2 = (g_2w_2)^2. \tag{36}$$

Making use of Eqs (20) and (36), the parameters λ_1 and λ_2 of the selfinteraction potentials for the fields $\Phi_{(1)}$ and $\Phi_{(2)}$ can be expressed in terms of the masses of the Higgs bosons and those of the gauge bosons:

$$\begin{aligned} w_1 &= 2M_1/g_1\,, \quad \lambda_1 = \tfrac{1}{32}(m_\eta g_1/M_1)^2\,, \\ w_2 &= M_2/g_2\,, \quad \lambda_2 = \tfrac{1}{8}(m_\zeta g_2/M_2)^2\,. \end{aligned} \tag{37}$$

The fields can now be identified with known particle fields:

$$\begin{aligned} \psi &\rightarrow \text{nucleon field}, \\ \boldsymbol{\phi} &\rightarrow \boldsymbol{\pi}\text{ : pion field}, \\ \mathbf{B}_{(1)\mu} &\rightarrow \boldsymbol{\rho}_\mu\text{ : rho-meson field}, \\ B_{(2)\mu} &\rightarrow \omega_\mu\text{ : omega-meson field}. \end{aligned}$$

Appropriately changing the notations

$$m_\rho = M_1, \quad m_\omega = M_2, \quad g_1 = g_\rho, \quad g_2 = g_\omega, \quad g_\phi = g_\pi, \quad m_\phi = m_\pi$$

and supposing a pseudoscalar $NN\pi$-coupling ($\Gamma = \gamma_5\tau$), the following gauge-invariant and renormalizable Lagrangian has been obtained:

$$\begin{aligned}
\mathcal{L} &= \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - g_\pi\bar{\psi}\gamma_5\boldsymbol{\tau}\psi\boldsymbol{\pi} - g_\sigma\bar{\psi}\psi\sigma + g_\rho\bar{\psi}\gamma^\mu\frac{\boldsymbol{\tau}}{2}\psi\boldsymbol{\rho}_\mu + g_\omega\bar{\psi}\gamma^\mu\psi\omega_\mu + \\
&+ \frac{1}{2}\partial_\mu\boldsymbol{\pi}\partial^\mu\boldsymbol{\pi} - \frac{1}{2}m_\pi^2\boldsymbol{\pi}^2 + g_\rho(\boldsymbol{\pi}\times\partial^\mu\boldsymbol{\pi})\boldsymbol{\rho}_\mu + \frac{1}{2}g_\rho^2(\boldsymbol{\pi}\times\boldsymbol{\rho}_\mu)(\boldsymbol{\pi}\times\boldsymbol{\rho}^\mu) + \\
&+ \frac{1}{2}\partial_\mu\sigma\partial^\mu\sigma - \frac{1}{2}m_\sigma^2\sigma^2 - U(\sigma) + \\
&+ \frac{1}{2}g_\rho m_\rho\eta\boldsymbol{\rho}_\mu\boldsymbol{\rho}^\mu + \frac{1}{8}g_\rho^2\eta^2\boldsymbol{\rho}_\mu\boldsymbol{\rho}^\mu + \frac{1}{2}\partial_\mu\eta\partial^\mu\eta - \frac{1}{2}m_\eta^2\eta^2 - \frac{1}{4}\frac{m_\eta^2}{m_\rho}g_\rho\eta^3 - \left(\frac{m_\eta g_\rho}{m_\rho}\right)^2\frac{1}{32}\eta^4 + \\
&+ g_\omega m_\omega\zeta\omega_\mu\omega^\mu + \frac{1}{2}g_\omega^2\zeta^2\omega_\mu\omega^\mu + \frac{1}{2}\partial_\mu\zeta\partial^\mu\zeta - \frac{1}{2}m_\zeta^2\zeta^2 - \frac{1}{2}\frac{m_\zeta^2}{m_\omega}g_\omega\zeta^3 - \frac{1}{8}\left(\frac{m_\zeta g_\omega}{m_\omega}\right)^2\zeta^4 - \\
&- \frac{1}{4}\boldsymbol{\rho}_{\mu\nu}\boldsymbol{\rho}^{\mu\nu} + \frac{1}{2}m_\rho^2\boldsymbol{\rho}_\mu\boldsymbol{\rho}^\mu - \frac{1}{4}\omega_{\mu\nu}\omega^{\mu\nu} + \frac{1}{2}m_\omega^2\omega_\mu\omega^\mu. \qquad (38)
\end{aligned}$$

Lagrangian (38) differs from that used by Serot [12] in the unique treatment of the isoscalar and isovector vectormesons as gauge fields. Transforming the pseudoscalar $NN\pi$-coupling in Eq. (38) into a pseudovector coupling by means of a nonlinear chiral transformation and adding a $\sigma\boldsymbol{\pi}^2$ interaction term with appropriately chosen coupling constant to the Lagrangian (38), a gauge-invariant and renormalizable Lagrangian can be obtained which is in accordance with low-energy πN-phenomenology [14].

4. Generalization of the linear σ-model

It has been established on the basis of low-energy πN-phenomenology that the chiral $\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R$ symmetry is almost an exact symmetry in hadron physics. Furthermore, QCD is also chiral invariant if the quarks are regarded as massless particles. Therefore, it is a question of interest, how gauge invariant and renormalizable Lagrangian with chiral symmetry can be constructed and used as a model Lagrangian of QHD.

Let S the chiral symmetry group $S = \mathrm{SU}(2)_L \times \mathrm{SU}(2)_R \times \mathrm{U}(1)$. We construct a Lagrangian with the symmetry $G \times H$, where $G = (\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R \times \mathrm{U}(1))_g$ and $H = (\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R \times \mathrm{U}(1))_l$. The transformation properties of the fields are given as follows:

	$(SU(2)_L$	×	$(SU(2)_R)_g$	×	$SU(2)_L$	×	$SU(2)_R)_l$	×	$U(1)_g$	×	$U(1)_l$
ψ_L	0		0		1/2		0		0		1
ψ_R	0		0		0		1/2		0		1
ϕ	0		0		1/2		1/2		0		0
χ_L	1/2		0		1/2		0		0		0
χ_R	0		1/2		0		1/2		0		0
ξ	0		0		0		0		1		-1
Generators	${}_g\hat{T}^a_L$		${}_g\hat{T}^a_R$		${}_l\hat{T}^a_L$		${}_l\hat{T}^a_R$		${}_g\hat{Y}$		${}_l\hat{Y}$
Coupling constants	—		—		G_x		G_Y		—		g_ω
Gauge fields	—		—		X^a_μ		Y^a_μ		—		ω_μ

Here ϕ is a chiral multiplet of scalar-pseudoscalar mesons. If its interactions with other fields are turned off, the Lagrangian of the free field ϕ is that of the linear σ-model with the selfinteraction (24) ($n_\phi = 4$). The field ϕ has a nonvanishing vacuum expectation value. Due to this fact the symmetry group S is spontaneously broken down to its subgroup $S_0 = \mathrm{SU}(2) \times \mathrm{U}(1)$ corresponding to isospin and baryon number conservation. This is just the case of the additional symmetry breaking.

The matrix representations of the generators satisfying requirement iv/ are given as follows:

$$\Gamma_\psi({}_l\hat{T}^a_L) = \frac{1+\gamma_5}{2}\frac{\tau^a}{2}, \quad \Gamma_\psi({}_l\hat{T}^a_R) = \frac{1-\gamma_5}{2}\frac{\tau^a}{2}, \quad \Gamma_\psi({}_g\hat{T}^a_{L,R}) = 0,$$

$$\Gamma_\psi({}_l\hat{Y}) = 1, \quad \Gamma_\psi({}_g\hat{Y}) = 0,$$

$$\Gamma_\phi({}_l\hat{T}^a_L) = \Theta^a, \quad \Gamma_\phi({}_l\hat{T}^a_R) = \varrho^a, \quad \Gamma_\phi({}_g\hat{T}^a_{L,R}) = 0,$$

$$\Gamma_\phi({}_g\hat{T}^a_{L,R}) = 0, \quad \Gamma_\phi({}_g\hat{Y}) = 0,$$

$$\Gamma_{\chi_L}({}_g\hat{T}^a_L) = \varrho^a, \quad \Gamma_{\chi_L}({}_l\hat{T}^a_L) = \Theta^a, \qquad \Gamma_{\chi_L}({}_g\hat{Y}) = \Gamma_{\chi_L}({}_l\hat{Y}) = 0,$$

$$\Gamma_{\chi_R}({}_g\hat{T}^a_R) = \Theta^a, \quad \Gamma_{\chi_R}({}_l\hat{T}^a_R) = \varrho^a, \qquad \Gamma_{\chi_R}({}_g\hat{Y}) = \Gamma_{\chi_R}({}_l\hat{Y}) = 0,$$

$$\Gamma_\zeta({}_g\hat{Y}) = Q, \qquad \Gamma_\zeta({}_l\hat{Y}) = -Q, \quad \Gamma_\zeta({}_g\hat{T}^a_{L,R}) = \Gamma_\zeta({}_l\hat{T}^a_{L,R}) = 0, \tag{39}$$

where the following matrices are introduced

$$\Theta^1 = \frac{i}{2}\begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \Theta^2 = \frac{i}{2}\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix},$$

$$\vartheta^1 = \frac{i}{2}\begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \vartheta^2 = \frac{i}{2}\begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix},$$

$$\Theta^3 = \frac{i}{2}\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix},$$

$$\vartheta^3 = \frac{i}{2}\begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{40}$$

(The matrix Q is given by Eq. (33)).

The convariant derivative of the nucleon field

$$D_\mu\psi = (\partial_\mu - iG_x\frac{1+\gamma_5}{2}\frac{\boldsymbol{\tau}}{2}\mathbf{X}_\mu - iG_Y\frac{1-\gamma_5}{2}\frac{\boldsymbol{\tau}}{2}\mathbf{Y}_\mu - ig_\omega\omega_\mu)\psi =$$
$$= (\partial_\mu - ig_\rho\frac{\boldsymbol{\tau}}{2}\boldsymbol{\rho}_\mu - ig_A\gamma_5\frac{\boldsymbol{\tau}}{2}\mathbf{A}_\mu - ig_\omega\omega_\mu)\psi \tag{41}$$

can be expressed in terms of the vector and axialvector fields defined as follows:

$$g_\rho\boldsymbol{\rho}_\mu = \frac{1}{2}(G_x\mathbf{X}_\mu + G_Y\mathbf{Y}_\mu), \quad g_A A_\mu = \frac{1}{2}(g_x\mathbf{X}_\mu - G_Y\mathbf{Y}_\mu). \tag{42}$$

After some algebra we get for the convariant derivative the expression

$$D_\mu\phi = (\partial_\mu - iG_x\Theta\mathbf{X}_\mu - iG_Y\boldsymbol{\vartheta}\mathbf{Y}_\mu + ig_\omega Q\omega_\mu)\phi =$$
$$= \begin{pmatrix} \partial_\mu\phi_0 + g_A\mathbf{A}_\mu\boldsymbol{\phi} \\ \partial_\mu\boldsymbol{\phi} - g_A\mathbf{A}_\mu\phi_0 + g_\rho\boldsymbol{\rho}_\mu x\boldsymbol{\phi} \end{pmatrix}, \tag{43}$$

where ϕ is the four-component column vector

$$\phi = \begin{pmatrix} \phi_0 \\ \boldsymbol{\phi} \end{pmatrix}. \tag{44}$$

Performing the local gauge transformation $\hat{U}_H(x)$, we get

$$\chi_L = \begin{pmatrix} w_L+\eta_L(x) \\ 0 \end{pmatrix}, \quad \chi_R = \begin{pmatrix} w_R+\eta_R(x) \\ 0 \end{pmatrix}, \quad \xi = \begin{pmatrix} 0 \\ v+\zeta(x) \end{pmatrix}, \tag{45}$$

where η_L, η_R and ζ are the Higgs fields and w_L, w_R and v are the appropriate vacuum expectation values. The covariant derivatives of the fields χ_L, χ_R, and ζ are given by

$$D_\mu \chi_L = (\partial_\mu - iG_x \Theta^a X_\mu^a)\chi_L = \begin{pmatrix} \partial_\mu \eta_L \\ -\frac{1}{2} G_x \mathbf{X}_\mu (w_L + \eta_L) \end{pmatrix}, \tag{46}$$

$$D_\mu \chi_R = (\partial_\mu - iG_Y \boldsymbol{\vartheta}^a Y_\mu^a)\chi_R = \begin{pmatrix} \partial_\mu \eta_R \\ \frac{1}{2} G_Y \mathbf{Y}_\mu (w_R + \eta_R) \end{pmatrix}, \tag{47}$$

$$D_\mu \xi = (\partial_\mu + ig_\omega Q \omega_\mu)\xi = \begin{pmatrix} -g_\omega \omega_\mu (v+\zeta) \\ \partial_\mu \zeta \end{pmatrix}. \tag{48}$$

The Lagrangian with the symmetry $G_g \times H_l$ has the form

$$\begin{aligned} \mathcal{L} = & \bar{\psi} i\gamma^\mu D_\mu \psi - g_\phi \bar{\psi}(\phi_0 + i\gamma_5 \boldsymbol{\tau}\boldsymbol{\phi})\psi + \frac{1}{2}(D_\mu \phi)^T D^\mu \phi - \mathcal{V}(\phi^T \phi) + \\ & + \frac{1}{2}(D_\mu \chi_L)^T D^\mu \chi_L - V_L(\chi_L^T \chi_L) + \frac{1}{2}(D_\mu \chi_R)^T D^\mu \chi_R - V_R(\chi_R^T \chi_R) + \\ & + \frac{1}{2}(D_\mu \xi)^T D^\mu \xi - V(\xi^T \xi) - \frac{1}{4}\mathbf{X}_{\mu\nu}\mathbf{X}^{\mu\nu} - \frac{1}{4}\mathbf{Y}_{\mu\nu}\mathbf{Y}^{\mu\nu} - \frac{1}{4}\omega_{\mu\nu}\omega^{\mu\nu}, \end{aligned} \tag{49}$$

where the following field strengths have been introduced:

$$\mathbf{X}_{\mu\nu} = \partial_\mu \mathbf{X}_\nu - \partial_\nu \mathbf{X}_\mu + G_X \mathbf{X}_\mu \times \mathbf{X}_\nu, \qquad \mathbf{Y}_{\mu\nu} = \partial_\mu \mathbf{Y}_\nu - \partial_\nu \mathbf{Y}_\mu + G_Y \mathbf{Y}_\mu \times \mathbf{Y}_\nu, \tag{50}$$

$$\omega_{\mu\nu} = \partial_\mu \omega_\nu - \partial_\nu \omega_\mu. \tag{51}$$

Introducing the field strength of the vector and axialvector fields by

$$\mathbf{F}_{\mu\nu} = \partial_\mu \boldsymbol{\rho}_\nu - \partial_\nu \boldsymbol{\rho}_\mu + g_\rho \boldsymbol{\rho}_\mu \times \boldsymbol{\rho}_\nu + \frac{g_A^2}{g_\rho} \mathbf{A}_\mu \times \mathbf{A}_\nu \tag{52}$$

and

$$\mathbf{G}_{\mu\nu} = \partial_\mu \mathbf{A}_\nu - \partial_\nu \mathbf{A}_\mu + g_\rho (\mathbf{A}_\mu \times \boldsymbol{\rho}_\nu + \boldsymbol{\rho}_\mu \times \mathbf{A}_\nu), \tag{53}$$

we get

$$\mathbf{X}_{\mu\nu} = \frac{g_\rho}{G_X}\mathbf{F}_{\mu\nu} + \frac{g_A}{G_X}\mathbf{G}_{\mu\nu}, \quad \mathbf{Y}_{\mu\nu} = \frac{g_\rho}{G_Y}\mathbf{F}_{\mu\nu} - \frac{g_A}{G_Y}\mathbf{G}_{\mu\nu} \tag{54}$$

and

$$-\frac{1}{4}\mathbf{X}_{\mu\nu}\mathbf{X}^{\mu\nu} - \frac{1}{4}\mathbf{Y}_{\mu\nu}\mathbf{Y}^{\mu\nu} =$$
$$= -\frac{1}{4}\left[\left(\frac{1}{G_X^2} + \frac{1}{G_Y^2}\right)(g_\rho^2\mathbf{F}_{\mu\nu}\mathbf{F}^{\mu\nu} + g_A^2\mathbf{G}_{\mu\nu}\mathbf{G}^{\mu\nu}) + \left(\frac{1}{G_X^2} - \frac{1}{G_Y^2}\right)2g_\rho g_A\mathbf{F}_{\mu\nu}\mathbf{G}^{\mu\nu}\right]. \tag{55}$$

These kinetic energy terms are diagonal and have the standard normalization factor 1/4 if and only if

$$G_X = G_Y = G, \qquad g_A = g_\rho = \frac{G}{\sqrt{2}}. \tag{56}$$

The kinetic energy of the fields χ_L and χ_R is given by

$$\frac{1}{2}(D_\mu\chi_L)^T D^\mu\chi_L + \frac{1}{2}(D_\mu\chi_R)^T D^\mu\chi_R = \frac{1}{2}\partial_\mu\eta_L\partial^\mu\eta_L + \frac{1}{2}\partial_\mu\eta_R\partial^\mu\eta_R +$$
$$+\frac{1}{2}\left[(w_L^2 + w_R^2)\frac{1}{4}(g_\rho^2\boldsymbol{\rho}_\mu\boldsymbol{\rho}^\mu + g_A^2\mathbf{A}_\mu\mathbf{A}^\mu) + (w_L^2 - w_R^2)\frac{1}{2}g_\rho g_A\boldsymbol{\rho}_\mu\mathbf{A}^\mu\right] +$$
$$+\frac{1}{8}(2w_L\eta_L + \eta_L^2)(g_\rho\boldsymbol{\rho}_\mu + g_A\mathbf{A}_\mu)^2 + \frac{1}{8}(2w_R\eta_R + \eta_R^2)(g_\rho\boldsymbol{\rho}_\mu - g_A\mathbf{A}_\mu)^2. \tag{57}$$

The kinetic energy of the field ξ has the form

$$\frac{1}{2}(D_\mu\xi)^T D^\mu\xi = \frac{1}{2}\partial_\mu\zeta\partial^\mu\zeta + \frac{1}{2}g_\omega^2 v^2\omega_\mu\omega^\mu + \frac{1}{2}g_\omega^2\omega_\mu\omega^\mu(2v\zeta + \zeta^2). \tag{58}$$

Terms quadratic in the fields $\boldsymbol{\rho}_\mu$, and $\mathbf{A}_\mu$ appear in expression (57) and they are diagonal if and only if

$$w_L = w_R = w. \tag{59}$$

In order to guarantee Eq. (59), the appropriate parameters of the selfinteraction potentials are assumed to be equal for the left–right pair of the fields χ_L and χ_R:

$$\mu_L^2 = \mu_R^2 = \mu_\chi^2, \qquad \lambda_L = \lambda_R = \lambda_\chi. \tag{60}$$

Then the masses of the Higgs fields η_L and η_R are identical

$$m_L^2 = m_R^2 = m_\eta^2. \tag{61}$$

The kinetic energy of the field ϕ also contains terms quadratic and diagonal in the gauge fields:

$$\frac{1}{2}(D_\mu\phi)^T D^\mu\phi = \frac{1}{2}(\partial_\mu\varphi_0 + g_A\mathbf{A}_\mu\boldsymbol{\varphi})^2 + \frac{1}{2}(\partial_\mu\boldsymbol{\varphi} - g_A\mathbf{A}_\mu\boldsymbol{\varphi}_0 + g_\rho\boldsymbol{\rho}_\mu\times\boldsymbol{\varphi})^2 +$$
$$+\frac{1}{2}(g_A w_0)^2\mathbf{A}_\mu\mathbf{A}^\mu - g_A w_0\mathbf{A}_\mu(\partial^\mu\boldsymbol{\varphi} - g_A\mathbf{A}^\mu\varphi_0 + g_\rho\boldsymbol{\rho}^\mu\times\boldsymbol{\varphi}). \tag{62}$$

Here the shift (27) has been performed and the notations

$$\varphi = \begin{pmatrix} \varphi_0 \\ \boldsymbol{\varphi} \end{pmatrix}, \qquad <\phi> = \begin{pmatrix} w_0 \\ 0 \end{pmatrix} \tag{63}$$

introduced.

Since Eq. (62) contains mixed quadratic term of the form $\mathbf{A}_\mu \partial^\mu \boldsymbol{\varphi}$, the terms quadratic in $\mathbf{A}_\mu$ and $\partial_\mu \boldsymbol{\varphi}$ should be brought into a diagonal form, before asking for the particle content of the model. Diagonalization can be performed by means of the transformation.

$$\mathbf{A}_\mu \to \mathbf{A}'_\mu = \mathbf{A}_\mu - \kappa \partial_\mu \boldsymbol{\varphi} \qquad (\kappa = \text{const.}) \tag{64}$$

As a result, we get the following expression

$$\begin{aligned} &\frac{1}{2} g_A^2 (w_0^2 + \frac{1}{2} w^2) \mathbf{A}_\mu \mathbf{A}^\mu - g_A w_0 \mathbf{A}_\mu \partial^\mu \boldsymbol{\varphi} + \frac{1}{2} \partial_\mu \boldsymbol{\varphi} \partial^\mu \boldsymbol{\varphi} = \\ &= \frac{1}{2} g_A^2 (w_0^2 + \frac{1}{2} w^2) \mathbf{A}'_\mu \mathbf{A}^{\mu\prime} + \left[g_A^2 (w_0^2 + \frac{1}{2} w^2) \kappa - g_A w_0 \right] \mathbf{A}'_\mu \partial^\mu \boldsymbol{\varphi} + \\ &+ \frac{1}{2} \left[1 - 2 g_A w_0 \kappa + g_A^2 (w_0^2 + \frac{1}{2} w^2) \kappa^2 \right] \partial_\mu \boldsymbol{\varphi} \partial^\mu \boldsymbol{\varphi}. \end{aligned} \tag{65}$$

The terms with $\mathbf{A}'_\mu \partial^\mu \boldsymbol{\varphi}$ drop out in Eq. (65) if we choose

$$\kappa = w_0 \left[g_A (w_0^2 + \frac{1}{2} w^2) \right]^{-1}. \tag{66}$$

Renormalizing the pseudoscalar-isovector field $\boldsymbol{\varphi}$ as follows

$$\boldsymbol{\pi} = Z_\varphi^{-1/2} \boldsymbol{\varphi}, \qquad Z_\varphi^{-1} = 1 - g_\rho \kappa w_0, \tag{67}$$

the kinetic energy of the renormalized field $\boldsymbol{\pi}$ can be rewritten into the standard form $\frac{1}{2} \partial_\mu \boldsymbol{\pi} \partial^\mu \boldsymbol{\pi}$. Now the mass terms of the gauge fields are diagonal and the following masses can be obtained

$$m_\rho^2 = \frac{1}{2} g_\rho^2 w^2, \quad m_A^2 = m_\rho^2 + g_\rho^2 w_0^2, \quad m_\omega^2 = g_\omega^2 v^2. \tag{68}$$

The renormalization constant Z_φ^{-1} can be expressed in terms of the masses of the gauge fields:

$$Z_\varphi^{-1} = (m_\rho / m_A)^2. \tag{69}$$

Performing the transformation given by Eqs (64), (67) and introducing the notations $\varphi_0 = \sigma'$, $g_\phi = g_\sigma$, we get for the Lagrangian the following form

$$\mathcal{L} = \mathcal{L}_\psi + \mathcal{L}_{SC} + \mathcal{L}_H + \mathcal{L}_g, \tag{70}$$

where the expressions of the different terms will be given below.

The term $\mathcal{L}_\psi$ is given by

$$\mathcal{L}_\psi = \bar{\psi} i\gamma^\mu (\partial_\mu - ig_\rho \frac{\boldsymbol{\tau}}{2} \boldsymbol{\rho}_\mu - ig_\rho \gamma_5 \frac{\boldsymbol{\tau}}{2} \mathbf{A}'_\mu - ig_\omega \omega_\mu)\psi + \\ + \bar{\psi}\gamma^\mu g_\pi^{pv} \gamma_5 \boldsymbol{\tau} \partial_\mu \boldsymbol{\pi} \psi - ig_\pi^{ps} \bar{\psi} \gamma_5 \boldsymbol{\tau} \psi \boldsymbol{\pi} - g_\sigma \bar{\psi}(w_0 + \sigma')\psi. \tag{71}$$

Thus the mass of the nucleon can be identified with

$$m = g_\sigma w_0. \tag{72}$$

The following coupling constants have appeared in Eq. (71)

$$g_\pi^{ps} = g_\sigma \frac{m_A}{m_\rho}, \qquad g_\pi^{pv} = \frac{g_\pi^{ps}}{2m} \left(\frac{m}{m_A} \frac{g_\rho}{g_\sigma} \right)^2. \tag{73}$$

(The superscripts '*ps*' and '*pv*' indicate pseudoscalar and pseudovector coupling.) As a consequence of the transformation (64), pseudovector coupling of the pseudoscalar-isovector meson field to the nucleon field has appeared.

Due to the spontaneous breaking of the chiral symmetry S, the Higgs field σ' with the mass $m_\sigma^2 \neq 0$ and the Goldstone fields $\boldsymbol{\pi}$ with zero mass present themselves in the model. It is well known [13] that by adding to the selfinteraction potential $\mathcal{V}(\phi^T \phi)$ the explicit symmetry breaking term $-\varepsilon\sigma'$, the Goldstone fields become massive. Doing this, we get for the scalar meson term in Eq. (70) the expression:

$$\mathcal{L}_{SC} = \frac{1}{2}(\partial_\mu \sigma' + S_{0\mu})^2 + \frac{1}{2}(\partial_\mu \boldsymbol{\pi} + \mathbf{S}_\mu)^2 + \\ + \left(\frac{m_A}{m_\rho} - 1 + \kappa \frac{m_A}{m_\rho} \right) \mathbf{S}_\mu \partial^\mu \boldsymbol{\pi} - g_\rho w_0 \mathbf{S}_\mu \mathbf{A}^{\mu\prime} - \frac{1}{2} m_\sigma^2 \sigma'^2 - \frac{1}{2} m_\pi^2 \boldsymbol{\pi}^2 - \\ - 4\lambda_0 w_0 \sigma' \left(\sigma'^2 + \frac{m_A^2}{m_\rho^2} \boldsymbol{\pi}^2 \right) - \lambda_0 \left(\sigma'^2 + \frac{m_A^2}{m_\rho^2} \boldsymbol{\pi}^2 \right)^2. \tag{74}$$

Here the following notations have been introduced:

$$S_{0\mu} = g_\rho \frac{m_A}{m_\rho} \mathbf{A}'_\mu \boldsymbol{\pi} + g_\rho \kappa \left(\frac{m_A}{m_\rho} \right)^2 \boldsymbol{\pi} \partial_\mu \boldsymbol{\pi}, \tag{75}$$

$$\mathbf{S}_\mu = g_\rho \frac{m_A}{m_\rho} \boldsymbol{\rho}_\mu \times \boldsymbol{\pi} - g_\rho \kappa \frac{m_A}{m_\rho} \sigma \partial_\mu \boldsymbol{\pi} - g_\rho \sigma \mathbf{A}'_\mu. \tag{76}$$

The masses of the scalar and pseudoscalar mesons are given by

$$m_\sigma^2 = \frac{\varepsilon}{w_0} + 8\lambda_0 w_0^2, \qquad m_\pi^2 = \frac{\varepsilon}{w_0} Z_\varphi = \frac{\varepsilon}{w_0} \left(\frac{m_A}{m_\rho} \right)^2. \tag{77}$$

Thus we obtain

$$\varepsilon = m_\pi^2 \frac{m}{g_\sigma}\left(\frac{m_\rho}{m_A}\right)^2, \qquad \lambda_0 = \frac{g_\sigma^2}{8m^2}\left[m_\sigma^2 - m_\pi^2\left(\frac{m_\rho}{m_A}\right)^2\right]. \tag{78}$$

The Lagrangian of the Higgs fields η_L, η_R in Eq. (70) takes the following form

$$\begin{aligned}\mathcal{L}_H =& \frac{1}{2}\partial_\mu\eta_L\partial^\mu\eta_L + \frac{1}{2}\partial_\mu\eta_R\partial^\mu\eta_R + \frac{1}{2}\partial_\mu\zeta\partial^\mu\zeta + \\ &+ \frac{1}{8}g_\rho^2\left[(\boldsymbol{\rho}_\mu + \mathbf{A}_\mu' + \kappa Z_\varphi^{1/2}\partial_\mu\boldsymbol{\pi})^2(2w\eta_L + \eta_L^2) + \right. \\ &\left. + (\boldsymbol{\rho}_\mu - \mathbf{A}_\mu' - \kappa Z_\varphi^{1/2}\partial_\mu\boldsymbol{\pi})^2(2w\eta_R + \eta_R^2)\right] + \frac{1}{2}g_\omega\omega_\mu\omega^\mu(2v\zeta + \zeta^2) - \\ &- \frac{1}{2}m_\eta^2(\eta_L^2 + \eta_R^2) - 4\lambda_\eta w(\eta_L^3 + \eta_R^3) - \lambda_\eta(\eta_L^4 + \eta_R^4) - \frac{1}{2}m_\zeta^2\zeta^2 - 4\lambda_\zeta v\zeta^3 - \lambda_\zeta\zeta^4.\end{aligned} \tag{79}$$

Here the parameters are given by

$$w = \sqrt{2}\frac{m_\rho}{g_\rho}, \qquad \lambda_\eta = \frac{m_\eta^2}{8w^2} = \frac{1}{16}\left(\frac{m_\eta}{m_\rho}g_\rho\right)^2, \tag{80}$$

$$v = \frac{m_\omega}{g_\omega}, \qquad \lambda_\zeta = \frac{m_\zeta^2}{8v^2} = \frac{1}{8}\left(\frac{m_\zeta}{m_\omega}g_\omega\right)^2. \tag{81}$$

Finally the gauge term $\mathcal{L}_g$ in Eq. (70) has the form

$$\begin{aligned}\mathcal{L}_g =& -\frac{1}{4}\mathbf{F}_{\mu\nu}\mathbf{F}^{\mu\nu} - \frac{1}{4}\mathbf{G}_{\mu\nu}\mathbf{G}^{\mu\nu} - \frac{1}{4}\omega_{\mu\nu}\omega^{\mu\nu} + \\ &+ \frac{1}{2}m_\rho^2\boldsymbol{\rho}_\mu\boldsymbol{\rho}^\mu + \frac{1}{2}m_A^2\mathbf{A}_\mu'\mathbf{A}^{\mu\prime} + \frac{1}{2}m_\omega^2\omega_\mu\omega^\mu,\end{aligned} \tag{82}$$

where the explicit expressions of the fields strengths are as follows:

$$\mathbf{F}_{\mu\nu} = \partial_\mu\boldsymbol{\rho}_\nu - \partial_\nu\boldsymbol{\rho}_\mu + g_\rho\left[\boldsymbol{\rho}_\mu \times \boldsymbol{\rho}_\nu + (\mathbf{A}_\mu' + \kappa Z_\varphi^{1/2}\partial_\mu\boldsymbol{\pi}) \times (\mathbf{A}_\nu' + \kappa Z_\varphi^{1/2}\partial_\nu\boldsymbol{\pi})\right], \tag{83}$$

$$\mathbf{G}_{\mu\nu} = \partial_\mu\mathbf{A}_\nu' - \partial_\nu\mathbf{A}_\mu' + g_\rho\left[(\mathbf{A}_\mu' + \kappa Z_\varphi^{1/2}\partial_\mu\boldsymbol{\pi}) \times \boldsymbol{\rho}_\nu + \boldsymbol{\rho}_\mu \times (\mathbf{A}_\nu' + \kappa Z_\varphi^{1/2}\partial_\nu\boldsymbol{\pi})\right], \tag{84}$$

$$\omega_{\mu\nu} = \partial_\mu\omega_\nu - \partial_\nu\omega_\mu. \tag{85}$$

Taking into account Eqs (68) and (72), the mass of the axialvector field is given by

$$m_A^2 = m_\rho^2 + \left(\frac{g_\rho}{g_\sigma}m\right)^2. \tag{86}$$

The Lagrangian given by Eq. (70) leads to PCAC in the form

$$\partial_\mu j_5^{\mu a} = \varepsilon Z_\varphi^{1/2} \pi^a = m_\pi^2 f_\pi \pi^a, \tag{87}$$

where f_π denotes the decay constant of the pion.

Taking into account Eqs (67), (78) and (87) the relation

$$f_\pi = \frac{m}{g_\sigma} \cdot \frac{m_\rho}{m_A} \tag{88}$$

can be obtained. Using Eqs (73) and (88), we conclude that

$$\frac{f_\pi}{\sqrt{2}m} g_\pi^{ps} = \frac{1}{\sqrt{2}} \quad . \tag{89}$$

Eq. (89) contradicts the Goldberger–Treiman relation [15]:

$$\frac{f_\pi}{\sqrt{2}m} g_\pi^{ps} = \frac{G_A}{G_V} \approx 1.2 \quad .$$

The model, which has been obtained in the present Section, in the gauge-invariant and renormalizable generalization of the model with the symmetry SU(2) × SU(2) proposed by Gasiorowicz and Geffen [16].

5. Concluding remarks

Renormalizable, gauge invariant generalization of Walecka's model and that of the linear σ-model have been given. Both models might be used for the Lagrangian of QHD. Nevertheless, there is a significant conceptional difference between them, reflected by the role the scalar-isoscalar field plays in them.

In Walecka's model the scalar-isoscalar field σ is included in order to produce the intermediate range attraction between the nucleons. In this context the σ-exchange imitates the exchange of two interacting pions [17]. The vacuum expectation value of the field σ vanishes, but due to the attractive NN-interaction trough the σ exchange a nonvanishing ground state expectation value of the field σ can appear in a nuclear many-body system. This results in the decrease of the effective mass of the nucleon [18].

In the linear σ-model, the scalar-isoscalar field (ϕ_0 in our notation) plays a quite different role. In this case the field ϕ_0 is a Higgs boson and it has a non-vanishing vacuum expectation value, causing spontaneous symmetry breaking. As a consequence of this, the originally massless nucleons become massive and their mass is determined by the vacuum expectation value of the scalar-isoscalar field. The linear σ-model has been developed by Gell-Mann and Lévy [19] just to solve the paradox, how chiral symmetry (as an almost good symmetry) and the nucleon with

its finite mass can exist at the same time in nature. Clearly, the scalar-isoscalar field imitates now some dynamics, responsible for hadronic structure. It can hardly be expected that its coupling to the nucleon field would produce a realistic intermediate range attraction. Lagrangian (70) surely does not, as it contains pseudoscalar and pseudovector $NN\pi$-couplings at the same time, which does not allow the dominance of the p-wave interaction in low-energy πN-scattering, and it leads to a relation, contradicting the Goldberger–Treiman relationship. It is an open question, whether the gauge invariant generalization of the linear σ-model could be reinterpreted as an effective theory for QCD. Recently, similar attempts have been made on the basis of the nonlinear σ-model [20] .

References

1. G. Marx, Nucl. Phys., *1*, 660, 1956.
2. G. Marx and J. Németh, Acta Phys. Hung., *18*, 77, 1964.
3. J.D. Walecka, Ann. Phys. (N. Y.), *83*, 491, 1974.
4. S.A. Chin, Ann. Phys. (N. Y.), *108*, 301, 1977.
5. J. Boguta, Phys. Lett., *120B*, 34, 1983.
6. J. Kunz, J. Boguta, Proc. of the Int. Workshop on Gross Properties of Nuclei and Nuclear Excitations XIV. Hirschegg 1986, ISSN 0720-8715.
7. B. Banerjee, N. K. Glendenning and M. Gyulassy, Nucl. Phys., *A361*, 326, 1981.
8. I. Lovas, J. Németh and K. Sailer, Nucl. Phys., *A430*, 731, 1984.
9. A. Faessler, Quantum Electrodynamics of Strong Fields, Ed. W. Greiner, Plenum Press, N. Y., 1983, NATO Advanced Study Institutes Series, Series B, Physics, V. 80, p. 701.
10. B. de Wit, Nucl. Phys., *B51*, 237, 1973.
11. K. Fabricius and J. Fleischer, Phys. Rev., *D19*, 353, 1979.
12. B.D. Serot, Phys. Lett., *B86*, 146, 1979.
    Erratum, Phys. Lett. *B87*, 403, 1979.
13. E.S. Abers and B.W. Lee, Phys. Rep., *9*, 1, 1973.
14. T. Matsui and B.D. Serot, Ann. Phys. (N. Y.), *144*, 107, 1982.
15. Yu.V. Novozhilov, Vvedenie v teori'u elementarnyh chastic, Nauka, Moscow, 1972.
16. S. Gasiorowicz and D.A. Geffen, Rev. Mod. Phys., *41*, 531, 1969.
17. G.E. Brown and A.D. Jackson, The Nucleon–Nucleon Interaction, North-Holland, Amsterdam, 1976.
18. B.D. Serot and J.D. Walecka, Adv. Nucl. Phys., *16*, 1, 1986.
19. M. Gell-Mann and M. Lévy, Nuovo Cim., *16*, 53, 1960.
20. J. Balog and P. Vecsernyés. KFKI-1985-47, HU ISSN 0368 5330.

Acta Physica Hungarica 64 (1–3), pp. 143–146 (1988)

INFRARED ASYMPTOTICS OF THE QUARK PROPAGATOR IN NONABELIAN GAUGE THEORIES II*

G. PÓCSIK

Institute for Theoretical Physics, Roland Eötvös University
1088 Budapest, Hungary

Computing Centre, Roland Eötvös University
1088 Budapest, Hungary

and

T. TORMA

(Received 8 January 1987)

Assuming a k^{-4} singularity for the gluon propagator at $k \to 0$, it is shown that in Quantum Chromodynamics the quark propagator is finite on the mass shell in axial and covariant gauges, and in axial gauges it has an essential singularity in g^2 at $g^2 = 0$ around the mass shell.

In a recent paper [1] we have calculated the infrared behaviour of the quark propagator $S_F'(p)$ for the gauge group SU(2) in a vacuum saturation approximation using a confining gluon propagator. The purpose of this note is to extend the calculation to the case of SU(3). For SU(2) $S_F'(p)$ is finite in the mass shell limit and this remains true also for SU(3). Furthermore, in both cases the quark propagator has an essential singularity in g^2 at $g^2 = 0$ around the mass shell, but the result on the mass shell is independent of the coupling constant.

We start with the general result [1] in the infrared region

$$S_F'(p) = \frac{1}{(2\pi)^4} \int_{-\infty}^{\infty} (S^I(q) + S^{II}(q)) d^4q, \tag{1}$$

$$\frac{1}{(2\pi)^4} \int S^I(q) d^4q = -i(\gamma_\mu p^\mu + m) \int_0^{\infty} f(2b\nu^2) e^{i(p^2 - m^2 + i\varepsilon)\nu} d\nu, \tag{2}$$

$$\frac{1}{(2\pi)^4} \int S^{II}(q) d^4q = i\gamma^\mu p^\nu G_{\mu\nu}(2p\nu') \frac{g^2}{4b} \int_0^{\infty} \frac{df(2b\nu^2)}{d\nu} e^{i(p^2 - m^2 + i\varepsilon)\nu} d\nu, \tag{3}$$

*Dedicated to Prof. G. Marx on his 60th birthday.

where $G_{\mu\nu}(2p\nu')$ is the confining gluon propagator in coordinate-space including the confining term Ω^2/k^2. γGp is independent of ν [1]:

$$\frac{g^2}{2}\gamma^\mu p^\nu G_{\mu\nu}(2p\nu') = \begin{cases} ib\dfrac{\gamma p - (\gamma n)(pn)n^{-2}}{p^2-(pn)^2n^{-2}}, & \text{in axial gauges } n^2 \neq 0 \\ ib_{F,L}\dfrac{\gamma p}{p^2}, & \text{in covariant gauges} \end{cases} \tag{4}$$

and

$$b = \frac{\delta\Omega^2 g^2}{16\pi^2}(p^2-(pn)^2n^{-2}), \tag{5}$$

$$b_F = -\frac{\Omega^2 g^2}{16\pi^2}\Gamma(\varepsilon)p^2 = \frac{4}{3}b_L. \tag{6}$$

$\delta = -1$ corresponds to the usual axial gauge. The b_F (Feynman gauge) and b_L (Landau gauge) are written down in $n = 4+\varepsilon$ dimensions, thus for $n \to 4$, $b_{F,L} \to -\infty (p^2 \to 0)$. For SU($N$) the function f is defined by

$$f(z) = \sum_{n=0}^{\infty}\frac{z^n}{(2n)!}\left(\prod_{j=1}^{2n}\lambda_{a_j}\right)\sum_{d.p.}\delta_{a_1a_2}\ldots\delta_{a_{2n-1}a_{2n}}. \tag{7}$$

The $\sum_{d.p.}$ goes over all the distinct a_ia_j-pairings, λ_a's are the $N\times N$ colour matrices, for SU(N) $a = 1,\ldots,N^2-1$.

First of all $f(z)$ can be written as

$$f(z) = \sum_{n=0}^{\infty}\frac{(z/2)^n}{n!(2n)!}\Delta_{\underline{x}}^n(\underline{\lambda}\,\underline{x})^{2n} \tag{8}$$

here $\underline{x}$ means an N^2-1 dimensional arbitrary vector and $\Delta_{\underline{x}}$ is the corresponding Laplace-operator. $f(z)$ is analytic in z everywhere. $f(z)$ is unchanged if one adds an arbitrary polynomial without a $2n$-th order term to $(\underline{\lambda}\,\underline{x})^{2n}$ in (8) and takes $\underline{x}=0$. In this way

$$f(z) = \sum_{n=0}^{\infty}\frac{(z/2)^n}{n!}\Delta_{\underline{x}}^n \operatorname{ch}(\underline{\lambda}\,\underline{x})|_{\underline{x}=0}. \tag{9}$$

Now, Fourier-transform $ch(\underline{\lambda}\,\underline{x})$ and carry out the differentiation, then a Gauss-type integral emerges and we get for Re $z>0$

$$f(z) = \frac{(2\pi)^{\frac{1}{2}(N^2-1)-1}}{(z)^{\frac{1}{2}(N^2-1)}}\int d^{N^2-1}\underline{x}e^{-\frac{x^2}{2z}}ch(\underline{\lambda}\,\underline{x}) = f_+(z)+f_-(z), \tag{10}$$

where

$$f_{\pm}(z) = \frac{1}{2}(2\pi)^{\frac{1}{2}(N^2-1)-1} \int d^{N^2-1}\underline{x} e^{-\frac{x^2}{2} \pm \sqrt{z}\underline{\lambda}\underline{x}}. \tag{11}$$

It is easy to see that (10), (11) are valid also for abritrary z's.
Substitute (10), (11) into (2) and integrate over ν:

$$\frac{1}{(2\pi)^4} \int d^4qS^I = \frac{1}{2}(2\pi)^{\frac{1}{2}(N^2-1)-1}(\gamma p + m) \int d^{N^2-1}\underline{x} e^{-\frac{x^2}{2}} \cdot$$
$$\cdot [(p^2 - m^2 + i\sqrt{-2b}\underline{\lambda}\underline{x} + i\varepsilon)^{-1} + (p^2 - m^2 - i\sqrt{-2b}\underline{\lambda}\underline{x} + i\varepsilon)^{-1}]. \tag{12}$$

From now on we assume $N = 3$. Then, employing the properties of Gell-Mann matrices one can show that $\underline{\lambda}$ rules out in the integrand of (12) and

$$\frac{1}{(2\pi)^4} \int d^4qS^I = \frac{(2\pi)^3(a + i\varepsilon)}{2}(\gamma p + m) \int d^8\underline{x} e^{-\frac{x^2}{2}} [(a + i\varepsilon)^2 - \frac{4}{3}b\underline{x}^2]^{-1}, \tag{13}$$

$a = p^2 - m^2$. (13) can be expressed by the integral [2]

$$Ei(x) = P \int_{-\infty}^{x} \frac{e^y}{y} dy, \tag{14}$$

namely

$$\frac{1}{(2\pi)^4} \int d^4qS^I = \frac{(2\pi)^7 a}{8b}(\gamma p + m) \left\{ \left(\frac{3a^2}{8b}\right)^3 e^{-\frac{3a^2}{8b}} \cdot \right.$$
$$\left. \cdot \left[Ei\left(\frac{3a^2}{8b}\right) - i\pi\Theta(b)\ \mathrm{sgn}\ a \right] - \left(\frac{3a^2}{8b}\right)^2 - \frac{3a^2}{8b} - 2 \right\}. \tag{15}$$

In the same way we obtain

$$\frac{1}{(2\pi)^4} \int d^4qS^{II} = \frac{(2\pi)^7 ig^2}{6b} \gamma^\mu p^\nu G_{\mu\nu}(2p\nu') \cdot$$
$$\cdot \left\{ \left(\frac{3a^2}{8b}\right)^4 e^{-\frac{3a^2}{8b}} \left[Ei\left(\frac{3a^2}{8b}\right) - i\pi\Theta(b)\ \mathrm{sgn}\ a \right] - \right. \tag{16}$$
$$\left. - \left(\frac{3a^2}{8b}\right)^3 - \left(\frac{3a^2}{8b}\right)^2 - 2\frac{3a^2}{8b} + 6 \right\}.$$

For small x we use the expansion [2]

$$Ei(x) = \ln|x| + C + \frac{x}{1 \cdot 1!} + \frac{x^2}{2 \cdot 2!} + \ldots, \tag{17}$$

where the series beside ln $|x|$ defines an analytic function of x, C is the Euler constant.

In axial gauges ($b < 0$), and in covariant gauges in $4+\varepsilon$ dimensions ($b_{F,L} < 0$) we get

$$S_F'(p) \to \frac{(2\pi)^7 i g^2}{b} \gamma^\mu p^\nu G_{\mu\nu} \tag{18}$$

in the mass shell limit $a \to 0$, that is, $S_F'(p)$ is finite and independent of g. This may be a manifestation of the quark confinement. Similarly (18) holds also in covariant gauges in four dimensions.

Let us note that for small a, $a \neq 0$, (15) and (16) have the following properties:

$$\begin{aligned} (15) &= 0(a) \text{ analytic part } + 0(a^7 \ln a), \\ (16) &= (18) + 0(a^8 \ln a), \end{aligned} \tag{19}$$

so that in S_F' a branch cut appears which vanishes for $a \to 0$.

Being [2]

$$e^x Ei(-x) = -\frac{1}{x} F_1(x), \tag{20}$$

where $F_1(x)$ has an essential singularity at $x = 0$, it follows that in axial gauges $S_F'(p)$ has an essential singularity in g^2 at $g^2 = 0$ for $a \neq 0$. This remains true in $4 + \varepsilon$ dimensions in covariant gauges.

References

1. G. Pócsik and T. Torma, ITP Rep, *No.* 431, July 1985. Acta Phys. Hung., *62*, 107, 1987.
2. I. S. Gradshteyn and I. M. Ryzhik, Tables of Integrals, Sums, Series and Products, Phys.-Math. Lit., Moscow, 1963.

Acta Physica Hungarica 64 (1-3), pp. 147-155 (1988)

THE HIGGS MESON MASS AND THE SCALE OF NEW PHYSICS IN THE STANDARD MODEL* **

P. HASENFRATZ and J. NAGER

Institute for Theoretical Physics, University of Bern
CH-3012 Bern, Switzerland

(Received 8 January 1987)

There are indications for the Weinberg–Salam model being an effective theory only. This would imply a relation between the scale ($\Lambda^{\rm cut}$) before which new physics should be observed and the Higgs meson mass m_H. The relation $\Lambda^{\rm cut} = \Lambda^{\rm cut}(m_H)$ is studied in an approximate, but nonperturbative way. For $m_H/m_W \gtrsim 10$, $\Lambda^{\rm cut}$ is of the order of the Higgs meson mass itself (i.e. this is an upper bound), while for $m_H/m_W \lesssim 6$, $\Lambda^{\rm cut}$ can be very large and no practically interesting constraint emerges.

1. Introduction

The old Fermi theory of weak interactions gave an excellent prescription of many low energy weak phenomena. The Fermi model is, however, not renormalizable, the cut-off $\Lambda^{\rm cut}$ cannot be removed. The model had to be considered as an effective theory. At a certain scale the Fermi model breaks down and one could expect new physics to occur before this scale.

There are indications for the Weinberg–Salam model having the same problem: although it is perturbatively renormalizable, the cut-off cannot be removed from the full theory. This would not be a catastrophe. At scales much below the cut-off, the predictions remain unchanged. At high energies however, the model is expected to break down. Actually, this scale $\Lambda^{\rm cut}$, before which "new physics" (i.e. phenomena not described by the standard model) should be observed, can be predicted within the framework of the standard model.

The scale $\Lambda^{\rm cut}$ depends on the parameters of the theory, most notably on the value of the Higgs meson mass m_H. Since m_H is unknown, one can only predict a functional relation $\Lambda^{\rm cut} = \Lambda^{\rm cut}(m_H)$. The cut-off $\Lambda^{\rm cut}$ is expected to decrease as m_H is increased. When m_H is large, the cut-off becomes comparable to the Higgs mass itself, and the model loses its meaning. This gives an upper bound on the Higgs mass m_H.

The determination of the relation $\Lambda^{\rm cut} = \Lambda^{\rm cut}(m_H)$ is a fully nonperturbative problem. The upper bound on m_H it implies follows from a basic criterion (namely that cut-off should be (much) larger than a physical mass). There exist upper

*Dedicated to Prof. G. Marx on his 60th birthday

**Work supported partly by Schweizerischer Nationalfonds

bound estimates in the literature, which use ad hoc criteria and perturbation theory [1]. The present work is based on nonperturbative renormalization group (RG) considerations and closest in spirit to a proposal of Dashen and Neuberger [2].

Of course, the question addressed here is sensible only, if the Weinberg–Salam model is really an effective theory with an unremovable cut-off. The arguments come mainly from the theoretical and numerical investigations of an SU(2) gauge theory with a scalar doublet:

$$\begin{aligned} \mathcal{L}_E &= 1/4 F^a_{\mu\nu} F^a_{\mu\nu} + \sum_{i=1}^{2} (D_\mu \Phi)^*_i (D_\mu \Phi)_i + 1/2\tau \sum_{\alpha=1}^{4} \phi^2_\alpha + 1/4\lambda \left(\sum_{\alpha=1}^{4} \phi^2_\alpha \right)^2 , \\ \Phi &= 1/\sqrt{2} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}, \\ D_\mu &= \partial_\mu 1 - igA^a_\mu \frac{\tau^a}{2}. \end{aligned} \tag{1}$$

This is the part of the standard model, which is relevant for the Higgs mechanism. This model has been the subject of detailed numerical studies during the last few years [3]. They seem to indicate that the only point in the 3 dimensional coupling constant space (g^2, r, λ), where a cut-off independent field theory can be defined (without a priori losing the gauge or the scalar part the model) is the Gaussian point $(g^2 = 0, r = 0, \lambda = 0)$.

What kind of theory is obtained by tuning the parameters towards the Gaussian point and, at the same time, sending the cut-off to infinity? In QCD, which is defined also at the Gaussian point [4], this procedure results in a cut-off independent, interacting field theory, which describes — as far as we can see — correctly the strong interactions.

For the SU(2) gauge theory with a scalar doublet this procedure gives a cut-off independent theory with three massive vector bosons and a massive Higgs boson. Unfortunately, no interaction survives, the model is free [5]. An interacting model can be defined only if a finite cut-off Λ^{cut} is kept. It is an effective theory in the sense discussed before.

It is a plausible possibility that the full Weinberg–Salam model suffers from the same disease as the SU(2) gauge, doublet Higgs model.

2. The Weinberg–Salam model as an effective theory

Our aim is to find the relation $\Lambda^{\text{cut}} = \Lambda^{\text{cut}}(m_H)$ and the corresponding upper bound on m_H. Even low level perturbative formulas suggest the existence of an upper bound. Really, tree level gives

$$R^2 \equiv m^2_H / m^2_W = 8\lambda / g^2, \tag{2}$$

where we might take λ as the running coupling at the scale m_H : $\lambda = \lambda(m_H)$. Given g^2 (≈ 0.4) and m_W, Eq. (2) relates the Higgs mass to the selfcoupling at the scale

of the Higgs mass. According to perturbation theory, λ is an increasing function of the scale (non-aysmptotic free), and hits the point $\lambda = \infty$ at some finite scale Λ^{cut} (Landau pole). Increasing m_H/m_W increases $\lambda(m_H)$, which results in a smaller ratio Λ^{cut}/m_H.

Unfortunately, for two reasons, this perturbative consideration is not trustworthy:

1. perturbation theory assumes small couplings throughout the whole calculation and this is not the case here (we have to talk even about $\lambda = \infty$).

2. even if the couplings are small, the radiative corrections contain dangerous logarithms of the type $\ln \frac{\Lambda^{\text{cut}}}{m_H}$.

The second problem is treated usually by RG considerations. In our special case, as we shall see, both problems are solved by a RG transformation.

By performing the RG transformation $\Lambda^{\text{cut}} \to e^{-t}\Lambda^{\text{cut}} (t > 0)$, the cut-off is pushed towards the physical scales (without changing the physical content of the theory) and the large logarithms are suppressed. We shall make the assumption that the effect of gauge fields and fermions on the RG flows of the scalar theory is small and negligible [2]. This is our 1st approximation, which is made plausible by the small values of the gauge and Yukawa couplings (assuming the top quark is not too heavy) in the standard model.

3. RG transformation in an O(4) scalar field theory

We have to consider, therefore, the flow of the coupling constants under the transformation $\Lambda^{\text{cut}} \to e^{-t}\Lambda^{\text{cut}}$ in a four component scalar theory. We shall follow these flows with the help of a nonperturbative, approximate RG relation derived and discussed in [6]. In this approximation the action, obtained after the RG transformation, is projected back to the subspace, where no couplings with derivatives occur:

$$S = \int d^d x \left[1/2 \sum_{\alpha=1}^{N} (\partial_\mu \phi_\alpha)^2 + V\left(\sum_{\alpha=1}^{N} \phi_\alpha^2 \right) \right]$$

$$\xrightarrow[\Lambda^{\text{cut}} \to e^{-t}\Lambda^{\text{cut}}]{} S(t) = \int d^d x \left[1/2 \sum_{\alpha=1}^{N} (\partial_\mu \phi_\alpha)^2 + V\left(\sum_{\alpha=1}^{N} \phi_\alpha^2; t \right) \right], \qquad (3)$$

where $V(x,t)\left(x^2 = \sum_{\alpha=1}^{N} \phi_\alpha^2\right)$ satisfies a non-linear partial differential equation. Written for

$$N^{1/2} f\left(x/N^{1/2}, t\right) \equiv \frac{\partial}{\partial x} V(x,t) \qquad (4)$$

it reads [6]:

$$\dot{f} = \frac{A_d}{2}\left[\frac{1}{N}\frac{f''}{1+f'} + \frac{N-1}{N}\frac{xf'-f}{x^2+xf}\right] + \left(1-\frac{d}{2}\right)xf' + \left(1+\frac{d}{2}\right)f, \qquad (5)$$

where "$\cdot$" $\equiv \partial/\partial t$, "$'$" $\equiv \partial/\partial x$, d is the space-time dimension (=4), N is the number of scalar fields (=4), A_d is a number ($= \frac{1}{8\pi^2}$ in $d = 4$). The RG equation describes qualitatively correctly all the fixed points, which are expected to occur in $2 \leq d \leq 4$, and gives reasonable results for the leading and subleading critical exponents for the $d = 3$ ferromagnetic fixed point. It is clearly not exact however, and this is our 2nd approximation.

The function f and the variable x are dimensionless. In Eq.(5) and in the following, the dimensions are carried by the (actual) cut-off.

The starting function $f(x, t = 0)$ corresponds to the original potential in the W–S model

$$f(x, t = 0) = rx + 4\lambda x^3. \tag{6}$$

First, we have to find the region in the (r, λ) plane, where the O(4) symmetry is spontaneously broken. Next, at every point of the broken phase we have to calculate Λ^{cut}/m_H and $R = \frac{m_H}{m_W}$. We expect a structure sketched in Fig. 1. Our aim is to find the envelope giving the maximum value of R for any given Λ^{cut}/m_H. We shall now briefly discuss this procedure, the details are published elsewhere [7].

4. The phase diagram and the masses

Given Eq. (5), it is relatively easy to find $\lambda_c = \lambda_c(r)$ separating the spontaneously broken phase from the symmetric phase [8]. The phase diagram, as obtained by solving Eq. (5) numerically, is given in Fig. 2.

Even if at $t = 0$ the function f has the simple form Eq. (6), it will enter an infinite dimensional coupling constant space under the RG transformation. In this infinite dimensional space the symmetric and broken phases are separated by the

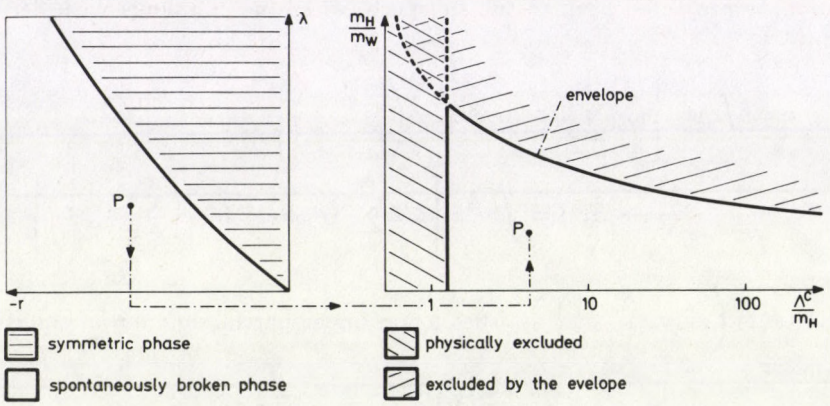


Fig. 1. Sketch of the procedure to be followed: Λ^{cut}/m_H and $R = m_H/m_W$ are calculated for any point in the broken phase of the (r, λ)-plane. These points are expected to be bounded from above by an envelope, determining the functional relation $\Lambda^{\text{cut}} = \Lambda^{\text{cut}}(m_H)$. Excluding the area $\Lambda^{\text{cut}}/m_H \lesssim O(1)$, we get an upper bound for m_H

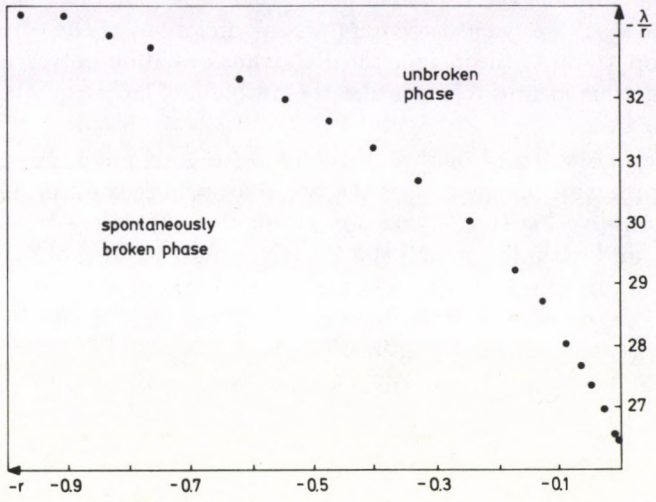


Fig. 2. The phase diagram, as obtained by solving Eq. (5) numerically

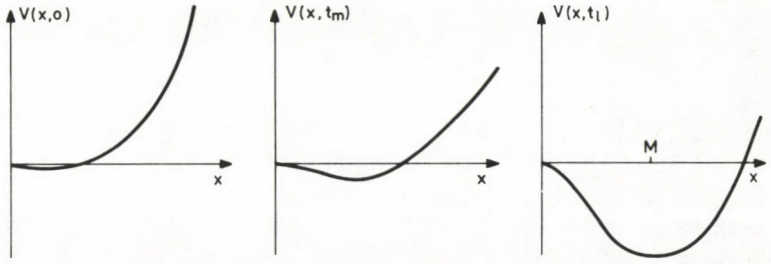


Fig. 3. The change of the potential (in the spontaneously broken phase) under RG transformations.
$t = 0$: the original potential, where perturbation theory is not applicable.
$t = t_m$: the potential at an intermediate t, where perturbation theory is not yet applicable.
$t = t_l$: the potential for large t. We apply now perturbation theory around $x = M$ and calculate radiative corrections.

critical hypersurface. The critical surface contains one fixed point, the Gaussian point (origin: $r = 0$, $\lambda = 0$, all other couplings = 0). This scalar theory is not expected to have any other fixed points ("triviality of ϕ^4") [9], and in our numerical study of the approximate RG equation Eq. (5) we did not find any.

The Gaussian fixed point has no relevant directions along the critical surface in $d = 4$. Therefore, starting anywhere in the vicinity of the critical surface, the RG flow will first run towards the Gaussian point, then diverge away from the critical surface along the single relevant direction ("mass" direction) of the Gaussian point. Due to this drop towards the origin the RG transformation solves not only the problem of dangerous logarithms, but also the problem of large couplings (problem 1. in Section 2).

The strategy is sketched in Fig. 3. At $t = 0$ (original potential) perturbation theory is not applicable in general for the two reasons discussed in Section 2. By applying the nonperturbative RG transformation Eq. (5) the system is becoming weakly coupled, and when the cut-off and the Higgs mass become of the same order, there will be no dangerous logarithms either. At this stage even tree level formulas should work. The validity of perturbation theory can be checked by calculating radiative corrections using the potential $V(x,t)$.

On the tree level one obtains (remember, the dimensions were carried by the actual cut-off)

$$\frac{(\Lambda^{\mathrm{cut}})^2}{m_H^2} = e^{2t}\frac{1}{V''(x = M(t),t)}, \qquad \frac{m_H^2}{m_W^2} = \frac{V''(x = M(t),t)}{\frac{1}{4}g^2M^2(t)}, \tag{7}$$

where $M(t)$ is the position of the minimum of the potential $V(x,t)$.

5. The results

Every point in the spontaneously broken part of the (r,λ) plane gives a pair of mass ratios $(\frac{\Lambda^{\mathrm{cut}}}{m_H}, \frac{m_H}{m_W})$ (Fig. 1). These pairs are plotted in Fig. 4. The indicated errors were estimated by comparing the tree level predictions with the 1-loop corrected results.

For a given $\Lambda^{\mathrm{cut}}/m_H$ the ratio $\frac{m_H}{m_W}$ is increasing with increasing λ. For small λ values this increase is strong, for larger λ values R seems to converge rapidly forming the expected envelope. Unfortunately, we could not investigate the coupling constant region beyond $\lambda \sim 30$. The reason is that Eq.(5) has a singularity when, for some $x_0, f(x_0)$ approaches -1 (the dimensions are carried by the cut-off), and this singularity was hit already at small t values for $\lambda > 30$).*We were not able to treat this singularity neither analytically nor numerically.

For this reason the upmost curves in Fig. 4 are only lower bounds for the envelope. The rapid convergence in λ suggests, however, that the curves for $\lambda \sim 30$ are very close to the envelope.

*We believe, this singularity is due to the specific (sharp cut-off) RG transformation and not to the approximation. Therefore, it would also be present in the exact integro-differential equation of Wegner and Houghton [10].

6. Discussion of the results

When the ratio $\frac{m_H}{m_W}$ drops below 6 or so, $\frac{\Lambda^{\text{cut}}}{m_H}$ becomes so large that no practically interesting constraint on the scale of new physics comes from this consideration. For large $\frac{m_H}{m_W}$ the cut-off comes dangerously close to the Higgs mass itself, which gives an upper bound $\frac{m_H}{m_W} \lesssim 10$.

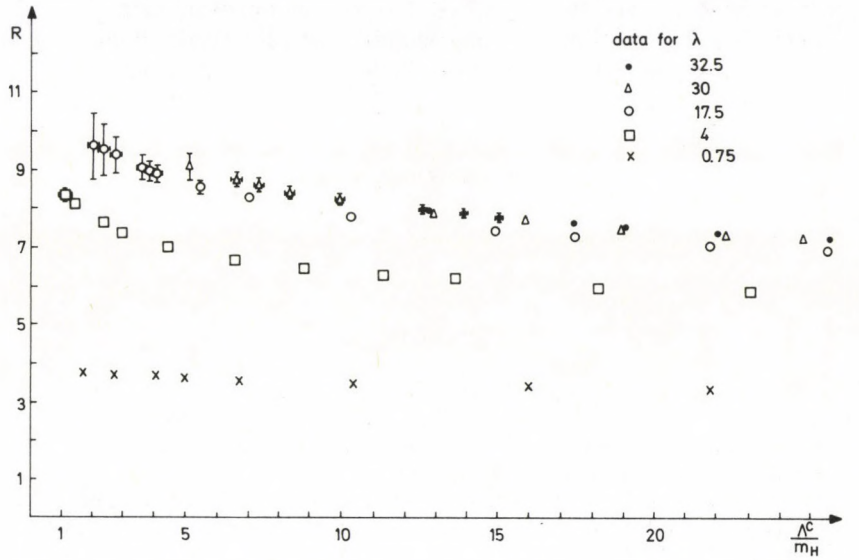


Fig. 4. Results for $R = m_H/m_W$ and Λ^{cut}/m_H for couplings λ up to 32.5

At the end let us collect again the approximations, which entered this calculation.

1^{st} *approximation*

The effect of gauge and fermion fields on the RG flow of the scalars was neglected. This is the approximation, which would be most difficult to remove in our scheme.

2^{nd} *approximation*

For the O(4) scalar theory an approximate (but nonperturbative) RG equation was used. This approximation can be replaced or improved rather easily. One might turn to Wilson's approximate recursion relation [11], use large N, or strong coupling expansions, even an MCRG analysis might be feasible.

3^{rd} *approximation*

After the RG transformation the resulting weakly coupled model with small correlation length was studied perturbatively. This seems to be safe. The radiative corrections were checked.

4^{th} *approximation*

No stronger coupling than $\lambda \sim 30$ was considered. This is a specific technical problem related to the RG equation analysed here. However, $\lambda \sim 30$ is a strong coupling already, and the curves in Fig. 4 show rapid convergence.

There are ways to improve the reliability of the results. Since the problem addressed here is physically relevant, we believe it is worth the effort.

Acknowledgement

The authors are indebted to Anna Hasenfratz and Herbert Neuberger for discussions.

References

1. D.J.E. Callaway, Nucl. Phys., *B223*, 189, 1984;
   M.A. Beg, C. Panagiotakopoulos and A. Sirlin, Phys. Rev. Lett., *52*, 883, 1984.
2. R. Dashen and H. Neuberger, Phys. Rev. Lett., *50*, 1897, 1983.
3. J. Jersak, Aachen prepr. PITHA 85/25, 1985;
   J. Jersak, C.B. Lang, T. Neuhaus and G. Vones, Phys. Rev., *D32*, 2761, 1985;
   V.P. Gerdt, A.S. Ilchev, V.K. Mitrjushkin, I.K. Sobolev and A. M. Zadorozhny, Nucl. Phys., *B26* [FS 15], 145, 1986;
   V.P. Gert, A.S. Ilchev, V.K. Mitrjushkin and A.M. Zadorozny, Z. Phys., *C29*, 363, 1985;
   H.G. Evertz, J. Jersak, D.P. Landau and T. Neuhaus, PITHA 85/23, 1985;
   K. Decker, I. Montvay and P. Weisz, Nucl. Phys., *B268*, 362, 1986;
   J. Jersak, in Advances in Lattice Gauge Theory, p. 241, Eds. D.W. Duke, J.F. Owens, World Sci., 1985;
   D.J.E. Callaway and R. Petronzio, Nucl. Phys., *B267* [FS 12], 253, 1986;
   I. Montvay, DESY, 85-0 05, 1985;
   W. Langguth and I. Montvay, Phys. Lett., *165B*, 135, 1985;
   I. Montvay, in Advances in Lattice Gauge Theory, p. 266, Eds. D.W. Duke, J.F. Owens, Wordl Sci., 1985;
   I. Montvay, W. Langguth and P. Weisz, DESY 85-138, 1985.
4. For a discussion on this question, see e.g. P. Hasenfratz, talk given at the Berkeley Conference (1986), Bern preprint BUTP-86/21.
5. A. Hasenfratz and P. Hasenfratz, Phys. Rev.,*D34*, 3160, 1986.
6. A. Hasefratz and P. Hasenfratz, Nucl. Phys., *B270* [FS 16], 685, 1986.
7. P. Hasenfratz and J. Nager, Z. Phys.,*C37*, 477, 1988.
8. The details of this procedure for $N = 1$ are discussed in ref. [6].
9. K.G. Wilson, Phys. Rev., *D6*, 419, 1971;
   K.G. Wilson and J. Kogut, Phys Rep., *12C*, 76, 1974;
   M. Aizenman, Phys. Rev. Lett., *47*, 1, 1981; Commun. Math. Phys., *86*, 1, 1982;
   G.A. Baker and J. M. Kincaid, J. Stat. Phys., *24*, 469, 1981;
   D.C. Brydges, J. Fröhlich and T. Spencer, Commun. Math. Phys., *83*, 123, 1982;

J. Fröhlich, Nucl. Phys., *B200*, 281, 1982;
A.D. Sokal, Ann. Inst. H. Poincaré, *A37*, 317, 1982;
M. Aizenman and R. Graham, Nucl. Phys., *B225* [FS 9], 261, 1983;
C. Aragao de Carvalho, S. Caracciolo and J. Fröhlich, Nucl. Phys., *B215*, 209, 1983;
K. Gawedzki and A. Kupiainen, Phys. Rev. Lett., *54B*, 92, 1985;
B. Freedmann, P. Smolensky and D. Weingarten, Phys. Lett., *113B*, 481, 1982;
D.J.E. Callaway and R. Petronzio, Nucl. Phys., *B240* [FS 12], 577, 1984;
C.B. Lang, Phys. Lett., *155B*, 399, 1985; Nucl. Phys., *B265* [FS 15], 630, 1986.
10. F.J. Wegner and A. Houghton, Phys. Rev., *A8*, 401, 1972.
11. See, e.g. the first two references in [9].

Acta Physica Hungarica 64 (1-3), pp. 157-162 (1988)

HELICITY METHOD FOR HEAVY FERMIONS*

Z. KUNSZT**

Theoretische Physik, ETH-Hönggerberg
CH-8093 Zürich, Switzerland

(Received 8 January 1987)

We show how to generalize the helicity method of calculating complex Feynman diagrams with massless kinematics for massive fermions. The method is illustrated by a simple example.

Hadron collider physics requires the evaluation of the cross-sections of complex higher order QCD reactions [1]. Both the UA1 and UA2 collaborations have data e.g. on gauge boson + two jet production, four jet production, bottom pair production with an additional jet etc. [2]. The theoretical description of these reactions is given by several hundreds of Feynman diagrams. Obviously, to calculate these cross-sections improved computational techniques are required.

Straightforward numerical methods have been attempted [3]. It has been found that they give too slow computer programs for practical use. Brute force algebraic methods also fail since they generate millions of terms. The key to resolve these computational problems has been provided by the development of the CALKUL collaboration [5] and also by the application of supersymmetry as a technical trick [6].

The CALKUL method makes advantages of the simplification given by massless kinematics, by the freedom given by gauge invariance and by the powerful use of the helicity amplitudes. In this paper we shall show how to reduce the case of massive fermions to the massless formalism.

Let us first consider a simple illustration of the technique given by the example of the Compton amplitudes with massless electrons

$$\gamma(-k_1) + \gamma(-k_2) \longrightarrow e(p_1) + e(p_2). \tag{1}$$

The amplitude is given in terms of two Feynman diagrams as

$$M^{\lambda_1\lambda_2}_{fi,s_1s_2} = e^2 \left(\frac{N^{\lambda_1\lambda_2}_{s_1s_2}(k_1,k_2)}{2k_1p_1} + \frac{N^{\lambda_2\lambda_1}_{s_1s_2}(k_2,k_1)}{2k_2p_1} \right), \tag{2}$$

*Dedicated to Prof. G.Marx on his 60th birthday.

**On leave of absence from Central Research Institute for Physics of the Hungarian Academy of Sciences.

where

$$N^{\lambda_1\lambda_2}_{s_1s_2} = \bar{u}(p_1, s_1)\hat{e}(k_1, \lambda_1)(\hat{k}_1 + \hat{p}_1)\hat{e}(k_2, \lambda_2)v(p_2, s_2). \tag{3}$$

Due to parity, C-parity invariance and helicity conservation, only two helicity amplitudes have to be calculated M^{++}_{--} and M^{+-}_{++}.

The improved CALKUL method [7] defines the photon polarization vectors in terms of massless spinors with one lightlike reference momentum l

$$e^{+-}_{\mu}(k, l) = \pm\frac{< k \mp |\gamma_\mu| l \pm >}{\sqrt{2} < k \pm | l \mp >}, \tag{4}$$

where we used the bra-ket notation for the massless spinors

$$< p - | q + > = \bar{u}(p, +)u(q, -) \tag{5}$$

and

$$u(p, +) = v(p, -) = \frac{1}{\sqrt{2}}\begin{pmatrix} \sqrt{p}_+ \\ \sqrt{p}_-\exp(i\phi_p) \\ \sqrt{p}_+ \\ \sqrt{p}_-\exp(i\phi_p) \end{pmatrix} \tag{6}$$

and

$$u(p, -) = v(p, +) = \frac{1}{\sqrt{2}}\begin{pmatrix} \sqrt{p}_-\exp(-i\phi_p) \\ -\sqrt{p}_+ \\ -\sqrt{p}_-\exp(-i\phi_p) \\ \sqrt{p}_+ \end{pmatrix}. \tag{7}$$

Their phases are fixed by the condition

$$C(\bar{u}(\pm))^T = v(\mp),$$

where

$$C = \begin{pmatrix} 0 & i\sigma^2 \\ -i\sigma^2 & 0 \end{pmatrix}. \tag{8}$$

The momentum components are defined with respect to any convenient axis.

We note some simple properties of the inner products. First products as $< p - | q - >$, $< p + | q + >$ and $< p - | p + >$ are vanishing. Furthermore it is easy to see from Eqs (6–7) that for $p_0 > 0$ and $q_0 > 0$

$$< p - | q + > \equiv < pq > = \sqrt{p - q} + e^{i\varphi_p} - \sqrt{p + q} - e^{i\varphi_q},$$

where

$$\begin{aligned} p\pm &= p_0 \pm p_3, \\ \cos(\varphi_p) &= \frac{p_1}{\sqrt{p_1^2 + p_2^2}}. \end{aligned} \tag{9}$$

We can also define inner product for negative energy states as

$$< -pq >= i < pq >, \quad < p-q >= i < pq >, \quad < -p-q >= - < pq > . \tag{10}$$

We also have conjugate inner products

$$< q+|p- > \equiv < p-|q+ >^{\dagger} . \tag{11}$$

It can be easily seen that for $p_0, q_0 > 0$

$$\begin{aligned} < pq >^{\dagger} &= < pq >^{*}, \\ < -pq >^{\dagger} &= i < pq >^{*}, \\ < -p-q >^{\dagger} &= - < pq >^{\dagger}, \end{aligned} \tag{12}$$

where $*$ denotes complex conjugation. Note also that the inner products are anti-symmetric

$$< pq >= - < qp > \tag{13}$$

and

$$| < pq > |^2 = 2p_{\mu}q^{\mu} = 2(pq). \tag{14}$$

In Eq. (4) the light like four momentum is an arbitrary reference momentum with the only constraint that it cannot be parallel with k_{μ}. If we choose the reference momentum $l = p_1$ for both the photon polarization vectors $e_1(k_1, \lambda_1)$ and $e_2(k_2, \lambda_2)$ we can see that $M_{--}^{++} = 0$ and

$$M_{++}^{+-} = 2e^2 \frac{< p_1|k_2 >}{< p_1|k_1 >} * \text{phase factor.} \tag{15}$$

Taking the absolute squared and summing up for all helicities we obtain the known result

$$\sum_{\text{spins}} |M|^2 = 8e^4 \left(\frac{t}{u} + \frac{u}{t} \right) . \tag{16}$$

It is remarkable that we got this result with few line calculation and without any Dirac algebra.

The question arises whether this impressive simplification can partially survive for massless fermions or not. We shall give first a representation of the massive spinors in terms of massless spinors then we shall calculate the cross-section of the reaction (1) but with massive electrons. We shall see that although many simplifications given by massless kinematics survive, the advantage of the method is not as striking.

Let us first give a decomposition of a massive four vector in terms of two light like vectors as follows

$$p = l + k = l + \frac{m^2}{2kl} k. \tag{17}$$

The second form suggests that presumably we may obtain a representation for the massive spinors $U(p,\pm), V(p,\pm)$ as follows

$$\begin{aligned} U(p,\pm) &= |l\pm> + \frac{m}{<l\pm|k\mp>}|k\mp>, \\ \bar{U}(p,\pm) &= <l\pm| + \frac{m}{<k\mp|l\pm>}<k\mp| \end{aligned} \tag{18a}$$

and

$$\begin{aligned} -iV(p,\pm) &= -|l\pm> - \frac{m}{<l\pm|k\mp>}|k\mp>, \\ -i\bar{V}(p,\pm) &= -<l,pm| - \frac{m}{<k\mp|l\pm>}<k\pm|. \end{aligned} \tag{18b}$$

We can verify with an explicit calculation that indeed these spinors defined by Eqs (18a–b) satisfy the Dirac equation

$$(\hat{p}-m)U(p)=0, \qquad (\hat{p}+m)V(p)=0 \tag{19}$$

and satisfy the completeness relations

$$\begin{aligned} \sum_s U(p,s)\bar{U}(p,s) &= (\hat{p}+m), \\ \sum_s V(p,s)\bar{V}(p,s) &= (-\hat{p}+m). \end{aligned} \tag{20}$$

The amplitude of the process can be given as

$$M_{s_1s_2}^{\lambda_1\lambda_2} = \bar{U}(p_1,s_1)\left(\frac{\hat{N}_{1s_1s_2}^{\lambda_1\lambda_2}}{d_1} + \frac{\hat{N}_{2s_1s_2}^{\lambda_1\lambda_2}}{d_2}\right)V(p_2), \tag{21}$$

where d_1 and d_2 denote the denominators

$$d_i = (p_1+k_i)^2 - m^2 = 2p_1k_i = 2k_il_1, \tag{22}$$

where i =1,2. With the use of the Dirac equation we can write for $\hat{N}_1$ and $\hat{N}_2$

$$\begin{aligned} \hat{N}_1 &= \hat{e}_1\hat{k}_1\hat{e}_2 + 2(e_1p_1)\hat{e}_2, \\ \hat{N}_2 &= \hat{e}_2\hat{k}_2\hat{e}_1 + 2(e_2p_1)\hat{e}_1. \end{aligned} \tag{23}$$

With these simple forms of the numerator matrices and with the choice for the photon polarization vectors

$$\hat{e}_{1,2}^{\pm} = \pm\frac{\sqrt{2}}{<l_{1,2}\mp|k_{1,2}\pm>}\{|k_{1,2}\mp><l_{1,2}\mp| + |l_{1,2}\pm><k_{1,2}\pm|\} \tag{24}$$

we can easily obtain the simple forms for the numerator functions of the helicity amplitudes as follows

$$\begin{aligned} N_{1++}^{++} &= -4d_1 < k_1 + |k_2- >< l_1 - |l_2+ >, \\ N_{1+-}^{++} &= 4d_1 m < k_1 + |k_2- > \frac{< l_1 - |k_1+ >}{< l_2|k_1+ >}, \\ N_{1--}^{++} &= 0, \\ N_{1-+}^{++} &= 0 \end{aligned} \tag{25}$$

and

$$\begin{aligned} N_{2+-}^{++} &= -\frac{d_2}{d_1} N^{++}1 + -, \\ N_{2++}^{++} &= 4d_2 < k_1 + |k_2- >< l_1 + |l_2- > (1 + \frac{m^2}{d_1}), \\ N_{2--}^{++} &= 4m^2 < k_1 + |k_2- >^2 \frac{< l_1 - |k_2+ >< l_1 - |k_1+ >}{< k_1 + |l_1- >< l_2 - |k_1+ >}, \\ N_{2-+}^{++} &= -4\frac{m}{< k_1 + |l_1- >} < k_1 + |k_2- >^2 < l_1 - |l_2+ >< l_1 - +k_2+ >, \end{aligned} \tag{26}$$

with similar expressions also for $N_{s_1 s_2}^{+-}$.

The absolute squared of the helicity amplitudes now can be easily obtained using Eqs (10–14) and (26–27)

$$\begin{aligned} |M_{++}^{++}|^2 &= \frac{4m^4(d_1 + d_2)}{d_1^2 d_2}, & |M_{--}^{++}|^2 &= \frac{d_1^2}{d_2^2}|M_{++}^{++}|^2, \\ |M_{+-}^{++}|^2 &= 0, & |M_{-+}^{++}|^2 &= -4xm^2\frac{(d_1 + d_2)^2}{d_1 d_2^2} \end{aligned} \tag{27}$$

and

$$\begin{aligned} |M_{++}^{+-}|^2 &= 4x^2\frac{d_1}{d_2}, & |M_{--}^{+-}|^2 &= 4x^2\frac{d_2}{d_1}, \\ |M_{-+}^{+-}|^2 &= 0, & |M_{+-}^{+-}|^2 &= -4m^2 x\frac{(d_1 + d_2)^3}{d_1^2 d_2^2}, \end{aligned} \tag{28}$$

where

$$x = 1 + \frac{m^2(d_1 + d_2)}{d_1 d_2}$$

the remaining amplitudes are obtained by parity. Summing up the contributions of all the helicity amplitudes we obtain the known Klein–Nishina formula

$$\sum -\lambda_1\lambda_2, s_1 s_2 |m_{s_1 s_2}^{\lambda_1\lambda_2}|^2 = 8\left[\frac{t}{u} + \frac{u}{t} + \frac{4m^2}{tu}(1 - \frac{m^2 s}{tu})\right], \tag{29}$$

where $t = d_1$ and $u = d_2$.

Although we have obtained the full spin information and the cross section formula without any Dirac algebra, the calculation is more involved then in the massless case. The method appears also to be very convenient for the calculation of the $O(m^2/s)$ corrections to the massless results [4]. Finally we note that in Eq. (3) the reference momentum k has definite physical meaning: it defines the direction of the spin in the rest frame of the massive fermions. Therefore there is no need to perform any integration over the twobody phase space volume. Of course it is also a possibility to get the spin averaged answer. (See Kleiss and Stirling for a discussion [6]). We also tested the method in the case of the more complicated QCD subprocess [7]

$$g + g \longrightarrow g + Q + \bar{Q}, \tag{30}$$

which is described by sixteen Feynman diagrams. The advantage of the method, in comparison with the standard covariant technique, is rather marginal since we are not interested in polarisation effects in this case.

In conclusion we have demonstrated how we can use the massless helicity formalism also in case of massive fermions. As an illustration of the method we have rederived the Klein–Nishina formula without any trace algebra. The method appears powerful in calculating only $O(m^2/s)$ effects (see also [3]).

References

1. G. Altarelli, Phys. Rep., *81*, 1, 1982.
2. Z. Kunszt, Talk given at the XVIII International Symposium on Multiparticle Dynamics, Seewinkel, Austria, 16–20 June, 1986.
3. Z. Kunszt, Nucl. Phys., *B247*, 339, 1984.
4. CALKUL collaboration, F.A. Berends et al, Phys. Lett., *105B*, 215, 1981; ibid. *114B*, 203, 1982; Nucl. Phys., *B206*, 53, 1982; ibid. *B239*, 382, 1982.
5. S. Parke and T. Taylor, Fermilab-Pub-85/118-5, 1985; Fermilab-Pub-85/162-T, 1985; Z. Kunszt, Nucl. Phys., *B271*, 333, 1986.
6. Z. Xu, D.H. Zhang and L. Chang, Tsinghua preprints TUTP 84/3-6, 1984; J.F.Gunion and Z. Kunszt, Phys. Lett., *161B*, 333, 1985; R. Kleiss and W.J.Stirling, Nucl. Phys., *B262*, 235, 1985.
7. J.F. Gunion and Z. Kunszt, Phys. Lett., *B178*, 296, 1986.

Acta Physica Hungarica 64 (1-3), pp. 163-171 (1988)

PARTICLE PHYSICS WITH CELESTIAL ACCELERATORS*

G.Domokos and S.Kovesi-Domokos

Department of Physics and Astronomy, Johns Hopkins University
Baltimore, Maryland 21218, USA

(Received 8 January 1987)

X-ray binary systems emit high energy photons and neutrinos. Experimental data indicate an unexpected excess of muons in showers correlated with the binaries. We suggest that the muon excess can be understood in terms of a substructure of leptons and quarks.

1. Introduction

It has been known for some time that X-ray binary systems constitute the most powerful particle accelerators. In all probability, they emit both charged and neutral particles. However, due to interstellar (and intergalactic) magnetic fields, the stable charged particles arrive to the top of the Earth's atmosphere, in essence, isotropically. Only stable, neutral particles, typically photons and neutrinos can be identified with the source. From the point of view of the particle physicist, an X-ray binary emits a pulsed photon (and, in all probability, neutrino) beam of energy $E_L \leq 10^6$ TeV. When this beam falls on the Earth's atmosphere, it interacts with the "average air nucleus" which has a baryon number $A \simeq 14.4$ and electric charge, $Z \simeq 7.2$. Roughly speaking, the incident photon (and/or neutrino) meets a swarm of light quarks, with u and d quarks in equal numbers. Primary energies up to about 10^5 TeV have been observed; this corresponds to about 10 TeV in the center of mass. (By comparison, the total CMS energy at the Superconducting Supercollider — now in the planning stage — will be 40 GeV). The luminosities, however, are not very high by accelerator standards; for instance, the photon luminosity of CYG X3, one of the best studied X-ray binaries, is given by the approximate formula, $EdN/dE \simeq 4 \times 10^{-4} E^{-1} (\mathrm{m}^{-2}\mathrm{s}^{-2}\mathrm{GeV}^{-2})$, so that the frequency of photons with energy $E > 10^4$ TeV is about $10^{-4}\mathrm{m}^{-2}$ year $^{-1}$. (This is not quite true; CYG X3 behaves in a very erratic way: there are brief and intense bursts, followed by periods of quiescence.)

Given these facts, the astrophysics of such binaries is quite intriguing but it would be hardly of relevance for particle physics. However, extensive air showers (EAS) associated with the binaries are "*anomalous*": they appear to contain too many muons [1] . Moreover, a number of groups, operating proton-decay detectors underground, have reported the observation of muon bursts correlated with the phases of the binaries.

*Dedicated to Prof. G. Marx on his 60th birthday.

The data (as it is the case with most cosmic ray observations) are not quite unambiguous: the total number of events is not very high, there are technical difficulties with some of the measuring devices, etc. The evidence and the controversy surrounding it is reviewed in a recent article by D'Ettore Piazzoli [2]. One will have a clearer picture of the experimental situation a few years from now. There is, however, a growing consensus in the physics community that at least some of the data reflect a real effect.

If this is the case, *the data cannot be understood in terms of the physics we are familiar with today* [3].

In the following Sections we briefly review the facts and point out those aspects of the data which prevent one from giving a satisfactory explanation in terms of the standard model of particle interactions. We then suggest a possible explanation in terms of a substructure of leptons and quarks and briefly discuss some further qualitative consequences of the proposed explanation.

2. The muons nobody wanted

This Section gives a very brief qualitative summary of those aspects of the data which are relevant from the point of view of the subsequent discussion. For more details, cf. [2,3].

i) The particles emitted by the X-ray binaries are *light*: upper limits on the mass are obtained from the fact that a pulse of the particles can be correlated with the phase of the binary as measured e.g. in X-rays. The best upper limit was reported by the Minnesota group [4]; they claim $m \leq 2\text{MeV}$.

ii) The particles must be *stable*, in essence, because they have to survive the trip from the binary to the Earth without decay. Using elementary kinematics, one estimates $\tau \geq dm/E$, where d is the distance of the binary. Taking again data from CYG X3, ($d \approx 4 \times 10^{22}$ cm) one knows that particles of $E \geq 1\text{TeV}$ are observed from it. On using the rest mass estimate $m \approx 2$ MeV, we get $\tau \geq 2 \times 10^8$ s.

The only likely candidates for such particles are photons and neutrinos, see e.g. Berezinskii et al for a discussion [5].

This conclusion appears to be clearly incompatible with the following observations:

a) A number of groups operating extensive air showers (EAS) detectors reported a substantial number of muons in EAS correlated with X-ray binaries.

b) A number of groups operating proton decay detectors underground reported muon bursts correlated with X-ray binaries.

c) All groups reporting positive results agree that the events exhibit a zenith effect: more events are seen when the binary is overhead than when it is near the horizon, indicating a *substantial interaction cross section in the atmosphere.* (The interaction cross section must be $\sigma \geq 10^{-30}\text{m}^2/\text{nucleon}$). (=0.1mb/nucleon).

According to conventional wisdom, c) alone is sufficient to exclude neutrinos as primaries. Moreover, according to the same conventional wisdom, a) and b)

exclude photons as primaries. In essence, the argument runs as follows. One knows that, upon interacting with matter, photons produce very few muons. The dominant process is electron pair production. In air, this has a cross section, $\sigma_e \simeq 5 \times 10^{-27}\text{m}^2$ (=500 mb). Direct pair production of muons is smaller by a factor $(m_e/m_\mu)^2 \approx 10^{-4}$, hence, it is negligible. The main source of muons in photon–air interactions is the decay of charged pions into $\mu + \nu_\mu$. As a rough estimate (born out by more detailed calculations) we take the total photoproduction cross section to be $\sigma_\gamma \simeq \sigma_h \alpha$, where σ_h is a hadronic total cross section, α is the fine structure constant. At CMS energies of relevance here, $\sigma_h \simeq A \times 10^{-27}\text{m}^2 \simeq 1.44 \times 10^{-26}\text{m}^2$, so that $\sigma_\gamma \simeq 1.1 \times 10^{-28}\text{m}^2$. (In units used in particle physics, these formulae read: $\sigma_h \simeq A \times 100\text{mb} \simeq 1440\text{mb}$, and $\sigma_\gamma \simeq 11\text{mb}$, respectively.) (It is to be remembered that the "average air nucleus" has $A \simeq 14.4$, $Z \simeq 7.2$). The multiplicity of "leading", high energy pions is 0(1) even at such high energies, so that $N_\mu/N_e \simeq (\sigma_\gamma/\sigma_e) f \simeq 0.02 f$, where f is the fraction of muons surviving until sea level. An elementary estimate gives $f \simeq 0.1 \div 0.2$ at the relevant energies. Thus, at sea level or in an underground detector $N_\mu/N_e \simeq 10^{-3}$. (This number agrees, within a factor of two or so, with the result of a careful computation by Stanev et al [6]. Instead, the EAS data indicate a muon to electron ratio about a factor 7 bigger at high energies; this is roughly consistent with the underground data.

With all the *caveats* alluded to in the Introduction, it seems very difficult, if not impossible, to dismiss all the evidence for a muon excess as spurious. Moreover, it appears now (as predicted in [7]) that the phenomenon is universal rather than being associated with one particular source. (For instance, the SOUDAN group reported muon bursts associated with CYG X3, HER X1 and 1E 2259+586).

With the present experimental constraints, it seems that the primaries cannot be anything but photons or neutrinos. (In particular, a light neutral hadron with an interaction cross section of the order of 10^{-30}m^2 (=0.1 mb) or larger would be produced copiously enough in present accelerator experiments to show up as a spectacular missing energy signal).

Thus, either all data regarding a muon excess must be classified as spurious, or some new phenomena begin to take place between the presently explored highest energies ($\simeq$ 0.5 TeV CMS) and $\simeq$ 10 TeV CMS. This clearly contradicts *all* theories which at present are regarded as well establised and/or internally consistent: "conventional" grand unified theories and superstring (or superstring inspired) theories alike.

There is a class of theoretical ideas according to which the occurrence of "something new", and, in particular, an increase in photon and neutrino cross sections at energies $\simeq$ 1 TeV CMS is a "natural" phenomenon: these are "composite models" of quarks and leptons [8]. *None* of the specific models proposed so far is particularly attractive from the points of view of internal consistency, beauty of their structure or even their agreement with presently existing data. The general idea of the existence of structures beyond the presently known "elementary" particles is, however, an attractive one and we argue that the cosmic ray data we referred to provide the first hint at such a "substructure" [9].

3. Strongly interacting photons and neutrinos?

In view of the uncertainties presently surrounding composite models ("preon models"), we adopt the following strategy.

i) *We admit ignorance concerning details of a preon model.* Almost certainly the dynamics of preon models is quite different from what we are familiar with today: quarks and leptons look more and more pointlike as experiments achieve increasingly higher accuracies. Yet, one of the main motivations of composite models is an attempt to understand the electroweak symmetry breaking mechanism. This means that the characteristic energy scale of a preon model cannot be much higher than 1 TeV. Therefore, if any "size" of quarks or leptons deduced from the analysis of data (e.g. Bhabha scattering) becomes much smaller than 1 TeV^{-1}, one will have a very strong hint at some new type of dynamics at work.

ii) We try to *abstract* simple and generic features of various composite models, which, one hopes, will persist when a "good" model is found. Most preon models are constructed in such a manner that

a) leptons and quarks have at least some constituents in common: typically the common constituents are non-trivial multiplets of $SU(3)_{color}$.

b) Gauge fields, or at least gauge fields corresponding to exact symmetries are elementary; hence, preons can be classified, inter alia, according to quantum numbers of $SU(3)_{color} \times U(1)_{e.m.}$

c) Overall color neutrality of the leptons is assured by an appropriate choice of the quantum numbers. However, all models considered are asymptotically free and thus, large momentum transfer processes should reveal the presence of "colored" preons inside leptons.

Keeping these features in mind, we can now imagine the following scenario.

A photon or a lepton falls upon a target mainly consisting of quarks. In a virtual process, the projectile dissociates into two groups of preons of large transverse momenta. One group of colored preons interacts with the target quark, the other one fragments into a high transverse momentum jet which, presumably, contains a comparable number of hadrons and leptons.

Remarks

1) In a completely analogous manner, one may consider the fragmentation of the target as well: however, in the laboratory system, where the quarks are approximately at rest, the target fragmentation typically leads to slow secondaries.

2) Leptons (electrons) in the target play a negligible role: the amplitude of the corresponding process is proportional to some dissociation amplitude squared, whereas the lepton–quark interaction is linear in the dissociation amplitude.

3) It is necessary to observe the products of the projectile fragmentation, otherwise we cannot be assured that the process we are considering tests the preon structure. By contrast, the jet arising from the preon–quark interaction does not

appear to carry any characteristic signature of the substructure and its components will be summed over.

We have now separated the description of the large $-P_T$ lepton–quark interaction into four distinct parts.

I. We need a model of the dissociation amplitude of leptons or photons into preons.

II. We need a model of the fragmentation functions, (preon)→ (leptons, hadrons).

III. We need a model of the preon–quark interaction.

IV. We have to join steps I thru III in some approximation.

Part IV is the easiest one: it involves hardly more than kinematics and the observation that the exchange of light preons (rest mass $\simeq 0$) dominates; we also consider the exchange of spin 1/2 preons, although this does not seem to be very important.

Parts I and II contain very difficult, unsolved problems of composite models: in fact, no detailed prediction can be made about either the dissociation amplitudes or of the fragmentation functions without having a good composite model. It is at this point that we are trying to abstract some features of the teory, based on lessons learned from QCD. Before doing that, however, let us proceed as if we had a theory at hand.

Let us denote the momenta of the incident neutrino or photon by k, that of the incident quark by p; the quark carries a fraction x of the total momentum of a nucleon and its structure function is denoted by $F(x)$. The momentum of the outgoing, observed, fragmentation product (typically, a hard muon) is denoted by q. All rest masses are neglected.

We are interested in the contribution of the projectile fragmentation to the inclusive cross section.

$$\left.\begin{matrix}\nu\\ \gamma\end{matrix}\right\} + N \to \nu + X. \tag{3.1}$$

We first compute the cross section on quarks. An elementary calculation leads to an expression of the form,

$$\frac{g^0 d\sigma}{d^3 q} = \int d^4\ell |\varphi(q,\ell)|^2 |d(\ell,k)|^2 |g(\ell)|^2$$

$$\times Tr((\gamma\cdot\ell - \gamma\cdot k)^{-1} A(p, k-\ell)(\gamma\cdot\ell - \gamma\cdot k)^{-1}). \tag{3.2}$$

In this expression, $A(p_1, p_2)$ is the absorptive part of a forward scattering amplitude describing the preon–quark interaction, $g(\ell)$ is a preon propagator,

$$g(\ell) = \begin{cases} (\gamma\cdot\ell)^{-1} & \text{for primary photons,} \\ \ell^{-2} & \text{for primary neutrinos,} \end{cases}$$

whereas d and φ are the corresponding dissociation and fragmentation amplitudes. We shall presently argue that the amplitude, A, can be estimated fairly reliably.

Very little is known about the functions φ and d (cf. items I and II above). Consider, however, the limiting case in which the four-momentum q is almost equal to ℓ : in intuitive terms, a hard preon is dressed up by means of a soft process into the observed fragment. It is intuitively obvious (and it can be justified by some plausible reasoning) that in this case $|\varphi g|^2$ can be, to a good approximation, replaced by δ-functions, viz.

$$|\varphi(q,\ell)g(\ell)|^2 \propto \delta^{(4)}(q-\ell)\delta(g^2). \tag{3.3}$$

This limiting case is of utmost importance: one needs hard leptons to survive the penetration of the atmosphere and the overburden ($2 \times 10^5 \mathrm{g/cm^2}$ or more) above an underground detector. Clearly, the approximation (3.3) simplifies matters considerably. However, the amplitude d is still undetermined.

Fortunately, it appears that the exact form of $|d|^2$ does not have a great influence on the final result. We can reason as follows. Adopting the approximation (3.3), $|d|^2$ depends, in essence, on one invariant, say, $(q-k)^2$. Given the fact that an incoming photon or neutrino is color-neutral, $|d|^2$ must be negligibly small until the resolution reaches the compositeness scale, say, at $(q-k)^2 \approx \Lambda^2$. Thereafter, $|d|^2$ is 0(1) until the dissociation amplitude is cut off either by the preon dynamics or by some natural cutoff in primary energy, e. g. by the Greisen–Zatsepin cutoff [10].

We are unable to report a firm result stating the "irrelevance of the cutoff" or on the robustness of our calculation with respect to changes in $|d|^2$. However, we *did* experiment with extreme shapes, such as, e. g.

$$\begin{aligned} |d|^2 &\simeq \Theta(|(q-k)^2| - \Lambda^2), \\ |d|^2 &\simeq \delta(|(q-k)^2| - \Lambda^2) \end{aligned} \tag{3.4}$$

and many other shapes (Gaussian, Lorentzian,...) in-between, without finding a significant difference in the qualitative properties of the end result.

This evidence encourages us to put in *some* convenient functional form for $|d|^2$ (we ended up using a step function) and hope for the robustness of the result.

Once this is done, the expression of the inclusive cross section is drastically simplified.

Formally, one uses a *local* vertex to describe the photon/neutrino $\rightarrow$ (preon-pair) dissociation, one identifies the observed fragmentation product (typically, a muon) with the preon fragment and also (since no polarizations are measured), one expresses A, *via* a *formal* optical theorem, in terms of a preon–quark total cross section, $\hat{\sigma}$. (It is to be noted, of course, that most of the steps involved are purely formal: for instance, one cannot construct an accelerator to produce preon beams.... Nevertheless, such formal steps provide a crutch for the intuition in exploring unkown realms). The end result is very simple. By integrating over the momentum fraction, we give the cross section on nucleons.

The result reads as follows:

$$\frac{q_0 d\sigma}{d^3q} = \frac{1}{2}\int_0^1 dx F(x)\Theta(q_T - \Lambda)\frac{K}{t\hat{s}}\hat{\sigma}(\mu^2,t)\Theta(\mu^2), \tag{3.5}$$

where $\hat{\sigma}$ is the preon–quark total cross section. The kinematic invariants are defined in the standard fashion, viz.

$$\begin{aligned} \hat{s} &= (p+k)^2 = xs, \\ t &= (q-k)^2, \\ \hat{u} &= (q-p)^2 = xu, \\ \mu^2 &= \hat{s}+\hat{u}+t, \end{aligned} \tag{3.6}$$

$$K = \begin{cases} \hat{u} & \text{(neutrino primary)}, \\ \hat{s}+\hat{u} & \text{(photon primary)}. \end{cases}$$

It is convenient to introduce scaling variables by the definition,

$$x_T^2 = 4tu/s^2, \qquad x = -u/s,$$

so that x_T and z are seen to be the transverse momentum fraction and LAB energy fraction of the observed particle. In terms of these,

$$\frac{d^2\sigma}{dx_T dz} = \frac{\pi}{2}\left\{ \begin{matrix} z \\ l-z \end{matrix} \right\} \Theta\left(x_T^2 - \frac{4\Lambda^2}{s}\right) \int_0^1 dx F(x) \Theta\left(x(l-z) - \frac{x_T^2}{4z}\right) \hat{\sigma}.$$

We now argue that the off-shell preon–quark cross section is determined by "low-energy" physics. Indeed, we can write $\hat{\sigma} = \mu^{-2} f$, where f is dimensionless and it satisfies a renormalization group equation,

$$\left(-2\mu^2 \frac{\partial}{\partial \mu^2} - 2t\frac{\partial}{\partial t} + \beta_1(g)\frac{\partial}{\partial g} + \beta_2(G)\frac{\partial}{\partial G}\right) f = 0, \tag{3.7}$$

where g and G stand for the coupling constant of QCD and of some hypercolor force responsible for binding preons. Eq. (3.7) determines the off-shell ($t \neq 0$) extrapolation of the cross section, in terms of its value at some $t = t_0 << \Lambda$. (We cannot set $t_0 = 0$, because of infrared problems.) Given an initial value of f, $f(\mu^2, t_0, g, G) = \varphi(\mu^2, g, G)$, the solution of (3.7) is:

$$f = \varphi\left(\frac{t_0\mu^2}{t}, g(t), G(t)\right),$$

with

$$\begin{aligned} 2t\frac{dg(t)}{dt} &= \beta_1(g(t)) \\ 2t\frac{dG(t)}{dt} &= \beta_2(G(t)), \end{aligned} \tag{3.8}$$

i.e. $g(t)$ and $G(t)$ are the usual running couplings. The essential point is that (apart from the edge of the accessible phase space) $|t_0\mu^2/t| << \Lambda$; hence the role of $G(t)$ is hidden in the structure of the target quark. The cross section, $\mu^{-2}\varphi$, is basically determined by ordinary gluon exchange. Thus, apart from some numerical factors, $\hat{\sigma} \simeq \Lambda^2_{QCD}$. Thus, once the energy is large enough (we need $s \geq 4\Lambda^2$ in order to produce large transverse momenta), the neutrino cross section should rise spectacularly. (No *spectacular* rise is expected in the case of photon and charged lepton, due to the large electromagnetic cross sections). One can derive a simple asymptotic formula for the cross section, which is valid if s is not too close to $4\Lambda^2$. The structure function of quarks inside a nucleon is expected to behave as $F(x) \sim x^{-1}(x << 1)$. This gives the asymptotic formula,

$$\sigma_{\nu N} \sim \frac{\hat{\sigma}}{4}(ln\, s/4\Lambda^2)^2. \tag{3.9}$$

4. Discussion

Having come so far, it is worth asking: can the mechanism proposed here explain the muon excess observed?

To this end, we note that the integral of the inclusive cross section, strictly speaking, gives the total cross section times the average multiplicity. Therefore, if one primary of energy E penetrates an absorber of thickness t (measured in nucleons/cm^2) [11] the average number of muons produced is: $< N >_E \simeq t \int \frac{d\sigma}{d^3p} d^3p/(\sigma_T)$. The total number of muons is then given by integrating $< N >_E$ over the primary flux. In particular, we assumed that the flux and spectrum of neutrinos is the same as that of photons [12]. An approximate integration gives then the following formula for CYG X3:

$$N_{\text{tot}} \simeq 2.4.10^{-3} \left[\frac{\hat{\sigma}}{mb}\right] < N > (\Lambda/\text{TeV})^{-2}(t/1000\text{g/cm}^{-2})\text{m}^{-2}\text{yr}^{-1}, \tag{4.1}$$

where $< N >$ is the multiplicity averaged over energies.

M. L. Marshak et al reported the observation of muon flux at SOUDAN, [13], $N_\mu \simeq 21\text{m}^{-2}\text{yr}^{-1}$, under an absorber of 20000 g/cm^2. One does not understand these data in quantitative terms yet; in particular, Eq. (4.1) does not take multiple interactions in the absorber into account. It is known, however, [13] that the muon multiplicity in each individual event is not very high. Therefore, substantial cross sections of the order of 10^{-29}m^2, i.e. a few millibarns, are needed in order to explain the data. More will be known as better data become available.

Thus, it appears that the celestial accelerators open a new chapter in particle physics. What we wanted to stress in this paper is that a natural explanation can be found in terms of a substructure of quarks and leptons. We still do not know whether this is the *correct* explanation, but probably, the answer will be known before the end of this decade as new data become available. "Grau, teurer Freund, ist alle

Theorie, und grün des Lebens goldner Baum", says Mephistopheles in Goethe's Faust. In physics, the ultimate judge of theories should be the experiment.

We thank F. Halzen, M. L. Marshak. Leon Madansky and S. Nussinov for enlightening discussions, and to A. K. Harding for pointing out an error in the original version of this paper. This research was supported in part by the U. S. Department of Energy, under Grant NO. DE-FG012-85 ER 40211.

References

1. For a summary, see A.A. Watson, Rapporteur paper, 19th International Cosmic Ray Conference, La Jolla, NASA Scientific and Technical Information Branch, Springfield, VA (1985).
2. B. D'Ettore Piazzoli, I1 Nuovo Saggiatore, *2*, 32, 1986.
3. A. De Rujula, in Proc. Europhysics Conference, Bari, 1985. Eds: L. Nitti and G. Preparata. European Physical Society, Petit Lancy, 1985.
4. M.L. Marshak, invited talk, XXIII International Conference on High Energy Physics, Berkeley, 1986, to be published.
5. V.S. Berezinskii, John Ellis and B.L. Ioffe, Phys. Lett., *B172*, 423, 1986.
6. T. Stanev, T.K. Gaisser and F. Halzen, Phys. Rev., *32*, 1244, 1985.
7. G. Domokos and S. Kovesi-Domokos, Johns Hopkins University preprint JHU-HET 8605, 1986; to be published.
8. For a recent summary of composite models, cf. H. Harari, in "Fundemental Forces". Eds: D. Frame and K.J. Peach. Scottish Universities Summer School in Physics, Edinburgh, 1985.
9. Cf. also, G. Domokos, and S. Kovesi-Domokos, Johns Hopkins University preprint, JHU-HET 8603, 1986 and G. Domokos and S. Nussinov, Phys. Lett., *B187*, 372, 1987.
10. K. Greisen, Phys. Rev. Lett., *16*, 648, 1966;
G.T. Zatsepin and V.A. Kuzmin, Zh. Eksp. Teor. Fiz., *4*, 78, 1966.
11. Measuring absorbers in nucleons/cm^2 is the most convenient for very high energy reactions; the conversion to conventional units is given by the formula: 1g/cm^2 $\simeq 0.6 \times 10^{24}$ nucleons/cm^2.
12. This may not be quite true: due to the different lifetimes of π^0 and $\pi^{\pm}$, the atmosphere of the companion of a compact object may distort the production spectrum differently. (A.K. Harding, private communication.)
13. M.L Marshak et al, Phys. Rev.Lett., *55*, 1965, 1985.

Acta Physica Hungarica 64 (3-4), pp. 173-176 (1988)

RENORMALIZATION OF THE DIPOLE FIELD IN A YUKAWA INTERACTION*

K. L. NAGY

Institute for Theoretical Physics, Roland Eötvös University
1088 Budapest, Hungary

(Received 8 January 1987)

An indefinite metric Abelian model is discussed where both confinement and asymptotic freedom may be presemt. Physical unitarity in the naive form is violated.

In the early days of the confinement problem higher order or multipole ghost theories where proposed [1,2,3] as a possible source of a resolution. It turned out that the correct description of covariant Lagrangian formalism of non-Abelian gauge fields (gravitation included) cannot be carried out without multipole fields [4,5]. In this case appropriate auxiliary conditions guarantee the physical unitarity. In this paper, after performing the renormalization in the dipole case, we make some comments.

The free Froissart (dipole) field [6] is described by the Lagrangian

$$L = \partial^\mu A \partial_\mu B - m^2 AB + \frac{\lambda}{2} A^2.$$

The quantum theory of this field can be completely solved (c.f.e.g. [7]), leading to Green's functions proportional also to p^{-4} providing some possibility of a confinement [1,2,3].

Power counting shows us that a theory with multipole fields in a Yukawa interaction is renormalizable. Here we point out some peculiarities. First, we turn to an interaction picture. Dyson's renormalization requires that it has to be done in such a way that the term λA^2 has to be considered as an interaction. Therefore we take

$$L_1 = g\bar{\psi}\gamma\psi\phi + \frac{\lambda}{2} A^2 + \text{counter terms},$$

where

$$\varphi = \alpha_1 A + \alpha_2 B,$$

γ is some matrix, Lorentz or inner indices are omitted. Let the first basic field be

$$a = k^{-1}\varphi, \qquad k = \sqrt{|2\alpha_1\alpha_2|} \neq 0,$$

*Dedicated to Prof. G. Marx on his 60th birthday.

the second one

$$b = k^{-1}(\alpha_1 A - \alpha_2 B).$$

Then the boson part of L_0 is

$$L_0^B = \frac{\varrho}{2}(-\partial^\mu a \partial_\mu a - m^2 a^2 + \partial^\mu b \partial_\mu b + m^2 b^2),$$
$$\varrho = \text{sign } \alpha_1 \alpha_2 = \pm 1,$$

i. e. one of them is quantized with indefinite metric: ($\varrho = +1$ means the field $a(x)$ is "normal")

$$[a, a] = -[b, b] = i\varrho\Delta, \qquad [a, b] = 0.$$

With these fields the interaction Lagrangian becomes

$$L_1 = g_0 \bar{\psi}\gamma\psi a + \frac{c_1}{2}a^2 + \frac{c_2}{2}b^2 + c_{12}ab.$$

$$g_0 = gk_1, \quad c_1 = c + \delta m_1^2, \quad c_2 = c + \delta m_2^2, \quad c_{12} = c + \delta_k,$$
$$c = \frac{\lambda}{2}\left|\frac{\alpha_2}{\alpha_1}\right|;$$

other counter terms, necessary to compensate singularities originating from the Yukawa interaction (e. g. $\sim \varphi^4$) or from ψ are omitted. The L_0 contains the real masses. It can be seen that for S_F' and Γ no peculiar new problem arises, therefore one can carry out the renormalization in the usual way, thus we have $S_F' = Z_2 S_F^R$, $\Gamma = Z_1^{-1}\Gamma^R$. One may notice that terms in Σ and Γ containing powers of λ are overall finite. Therefore we discuss the causal functions Δ_F. Let us denote the Fourier transforms of these by

$$\Delta_F(a) = \hat{a a} = \varrho(m^2 - p^2 - i\varepsilon)^{-1} = -\Delta_F(b) = -\hat{bb}$$

as Δ_a, Δ_b and analogously for the primed and renormalized quantities. For this interaction the Schwinger–Dyson equations are

$$\Delta_a' = \Delta_a + \Delta_a(\pi + c_1)\Delta_a' + c_{12}^2 \Delta_a \Delta_b^T \Delta_a',$$
$$\Delta_b' = \Delta_b + \Delta_b c_2 \Delta_b' + c_{12}^2 \Delta_b \Delta_a^T \Delta_b',$$

where

$$\Delta_b^T = \Delta_b + \Delta_b c_2 \Delta_b^T,$$
$$\Delta_a^T = \Delta_a + \Delta_a(\pi + c_1)\Delta_a^T.$$

Here

$$\pi = \pi^2 + \pi^1(m^2 - p^2) + \pi_c \quad \sim g_0^2 \int S_F' S_F' \Gamma dp$$

is the sum of all compact a-boson self energy contributions. Then with

$$\Delta_a^T = Z_3\Delta_a^{TR}, \quad \Delta_a^{'} = Z_3\Delta_a^R, \quad \Delta_b^T = \Delta_b^{TR}, \quad \Delta_b^{'} = \Delta_b^R,$$

step by step, requiring $c \to c^R$, $(\lambda \to \lambda^R)$ choosing (necessary to compensate the divergences)

$$\pi^0 + \delta m_1^2 = 0, \qquad cZ_3 = c_R, \qquad Z_3^{-1} = 1 - \varrho\pi^1(g_0),$$
$$g_R = Z_1^{-1}Z_2Z_3^{1/2}g_0, \qquad \pi_c^R = Z_3\pi_c, \qquad c_R = \frac{\lambda_R}{2}\left|\frac{\alpha_2}{\alpha_1}\right|,$$
$$c_2 = c + \delta m_2^2 = c_R = cZ_3, \qquad \text{i. e.} \qquad \delta m_2^2 = c(Z_3 - 1),$$
$$c_{12} = (c + \delta\kappa)^2 Z_3 = c^2 Z_3^2, \qquad \text{i. e.} \qquad \delta\kappa = c(Z_3^{1/2} - 1),$$

we have the equations for the renormalized causal functions as

$$\Delta_a^R = \Delta_a + \Delta_a(\pi_c^R + c_R)\Delta_a^R + c_R^2\Delta_a\Delta_b^{TR}\Delta_a^R,$$
$$\Delta_b^R = \Delta_b + \Delta_b c_R\Delta_b^R + c_R^2\Delta_b\Delta_a^{TR}\Delta_b^R,$$

where

$$\Delta_b^{TR} = \Delta_b + c_R\Delta_b\Delta_b^{TR},$$
$$\Delta_a^{TR} = \Delta_a + \Delta_a(\pi_c^R + c_R)\Delta_a^{TR}.$$

Do notice:

i. According to the graph rules, in the remaining real vertex a—b we must substitute

$$\Delta_a^{'}(c + \delta\kappa)\Delta_b^{'} \to Z_3^{1/2}\Delta_a^R(c + \delta\kappa)\Delta_b^R = \Delta_a^R c_R\Delta_b^R,$$

which gives for $\delta\kappa$ the same value as above.

ii. Contrary to earlier notions [7] for free fields, the cases $m = 0$ or $m \neq 0$ can be treated on the same footing.

iii. The renormalization multiplier Z_3 contains the sign factor ϱ explicitly. The renormalized equations give

$$\Delta_a^{R-1} = (\Delta_a + \Delta_a\Delta_a c_R)^{-1} - \pi_c^R,$$
$$\Delta_b^{R-1} = \Delta_b^{-1}\frac{\Delta_a^{-1} - \pi_c^R - c_R\Delta_a\pi_c^R}{\Delta_a^{-1} - \pi_c^R - c_R}.$$

Having performed the renormalization one can calculate the Gell-Man–Low function β

$$\mu\frac{dg_R}{d\mu} = \beta(g_R)$$

in the usual way. We have in lowest order in g_R e. g. for the vector coupling

$$\beta = \varrho \frac{g_R^3}{12\pi^2},$$

and for the scalar coupling

$$\beta = \varrho \frac{5g_R^3}{16\pi^2}.$$

Therefore for $\varrho=-1$, as it is usually argued, for high energies we have asymptotic freedom. At the same time the term $\sim c_R p^{-4}$ in Δ_a^R provides the confinement in large distances.

Since such a model in not physically unitary, some artificial unitarization can always be applied. In the spirit of [4] our best hope is that considering $a(x)$ $b(x)$ as gauge fields of certain type, by means of a suitable subsidiary condition the unitarity can be restored.

References

1. J. E. Kiskis, Phys. Rev. *D12*, 3583, 1975.
2. K. L. Nagy, Acta Phys. Hung., *42*, 377, 1977.
3. H. Narnhofer and W. Thirring, Phys. Letters, *76B*, 428, 1978.
4. T. Kugo and J. Ojima, Progr. Theor. Phys. Suppl., *66*, 1, 1979.
5. N. Nakanishi, Publ. Math. Sci., Kyoto University, *Vol. 19*, No. 3, 1983.
6. M. Froisart, Nuovo Cim. Suppl., *14*, 197, 1959.
7. N. Nakanishi, Progr. Theor. Phys. Suppl., *No. 51*, 1, 1972.

Acta Physica Hungarica 64 (1–3), pp. 177–197 (1988)

RECENT RESULTS OF THE EUROPEAN MUON COLLABORATION* **

E. NAGY

Central Research Institute for Physics
1525 Budapest, Hungary

(Received 8 January 1987)

Strange particle production is in agreement with the basic expectations from the Quark Parton Model (QPM). The energy distribution of *inclusively produced vector mesons* becomes gradually harder as the mass of the vector meson increases. *In exclusive vector meson production* the process changes from a soft diffractive type interaction to a hard pointlike electromagnetic interaction with increasing mass of the virtual photon. The *u-valence quark distribution* derived from the inclusive distribution of charged pions and protons provides additional confirmation of the QPM.

Introduction

In a long series of experiments starting in 1977 and ending in 1985 the European Muon Collaboration (EMC)[1] had unique possibilities to study the deep inelastic scattering of muons off nucleons and nuclei and thereby to reveal the structure of these latter up to distances as small as 10^{-16} cm. The 400 GeV Super Proton Synchrotron at CERN enabled to produce a high energy (100–280 GeV) and record intensity ($\sim 4.10^7$ muon/burst) muon beam of extremely high purity [2].

In the first phase (NA-2) of the experiment the EMC detector [3], depicted in Fig. 1, measured the scattered muons and the forward going charged particles. The target was interchangeable enabling to measure the quark structure functions on different nuclei [4] which lead among other important results to the discovery of the EMC Effect [5]. Also a detailed analysis of the forward going particles has been carried out [6, 7] including the production of charmed mesons in a multimuonic and hadronic final state [8, 9].

In a later stage, called NA-9, the detector was extended by a large angle spectrometer [10] which made possible to detect the charged particles in the full 4π angular region (Fig. 2). The 1m H_2 and D_2 target was surrounded by a streamer chamber in a superconducting magnet and by additional time of flight hodoscopes and Čerenkov counters which assured high identification power for the charged particles. The obtained streamer chamber pictures were analysed among other

*Dedicated to Prof. G. Marx on his 60th birthday.

**Lectures given by the author in the University of California San Diego and in the NEVIS Laboratory of the Columbia University (New York) in July 11 and August 4, 1986.

laboratories in the Central Research Institute for Physics, Budapest, by highly automatized devices [11] (Fig. 3).

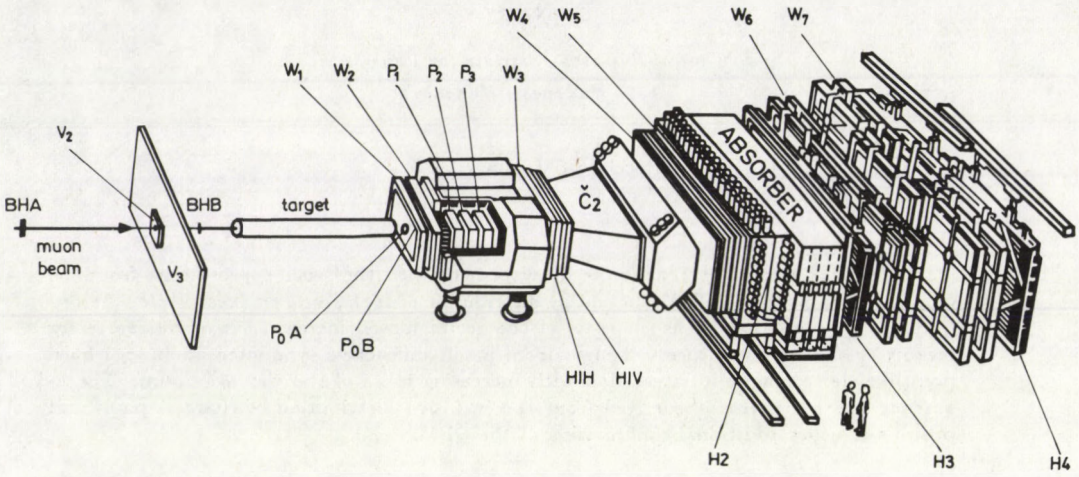


Fig. 1. The Detector of the NA-2 Spectrometer. Halo muons are vetoed by scintillation counters V. The trajectories of the charged particles are measured by proportional (P) and drift chambers (W). The momentum of the charged particles is determined by measuring the bending caused by the spectrometer magnet. The Čerenkov counter provides identification of the forward going charged particles. The event is triggered by a series of scintillation hodoscopes (BH and H). The absorber is used to identify the muons in the final state, the H_2 calorimeter measures the deposited electromagnetic and hadronic energies of the forward particles.

Table I
The number of the observed strange particles

$K^{\pm}$	K^0	Λ	$\bar{\Lambda}$
4300	1903	704	302

Finally, in the third stage of the experiment (NA-2') the large angle spectrometer has been replaced by a target containing polarized protons and by a series of targets of heavy nuclei aiming at a measurement of the structure function on polarised protons and at the systematic study of the EMC Effect on different nuclei.

In this lecture I report on some new results of the EMC concerning the study of the hadronic final state in the NA-9 and NA-2' phase, namely on the i) inclusive production of strange particles, ii) inclusive and exclusive production of vector

mesons and iii) determination of the valence quark distribution in the proton from the inclusive hadron distribution.

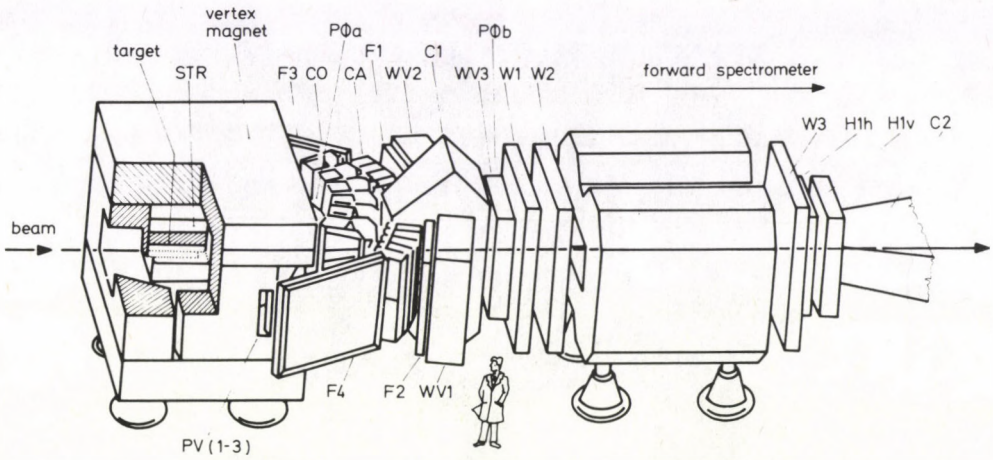


Fig. 2. The Large Angle Spectrometer of the NA-9 Experiment. The target is inside a streamer chamber which in turn is embedded in a superconducting magnet. The magnet is surrounded downstream by a series of Čerenkov counters (C), time-of-flight hodoscopes (F) which provides high identification power for charged particles. Additional proportional (PV) and drift tubes (WV) to the forward spectrometers assure detection of charged particles in almost 4 π solid angle

Inclusive production of strange particles

In the Quark Parton Model (QPM) [12, 13] the deep inelastic leptoproduction can be written down as it is depicted in Figs 4 and 5. The lepton (muon) emits a virtual photon which hits one of the quarks of the nucleon. The struck quark follows the direction of the incident virtual photon in the Centre of Mass System (CMS). The quark and the remaining diquark stretches a colour field between them which materializes into a few quark–antiquark (diquark–antiquark) pairs as is shown in Fig. 5. The final hadrons are then the result of the recombination of the quarks with the diquarks and antiquarks. An important parameter of the above fragmentation process is γ_s/γ_u the relative probability that a strange (s) or a non-strange (u) quark–antiquark pair is formed in the stretched colour field. This parameter can be determined by studying the inclusive distributions of the final state strange particles provided that one has a reliable model wich describes the resonance production and their decays. For this purpose the LUND model [13] has been used.

The analysis includes about 60 thousand deep inelastic events after all kinematic cuts on hydrogen and deuterium targets. The number of the observed strange particles are shown in Table I.

Fig. 3. The monitor and switchboard of the RIMA automatic measuring device for the analysis of the streamer chamber pictures in the Central Research Institute for Physics, Budapest

Fig. 6 shows the average multiplicity of the strange mesons as a function of W, the total CMS energy of the incoming virtual photon and the nucleon. A logarithmic rise is clearly visible. For comparison the average multiplicity of all charged particles is also shown. One can see that the slopes of the charged kaons are similar to that of all charged particles whereas the neutral kaons exhibit a somewhat smaller slope.

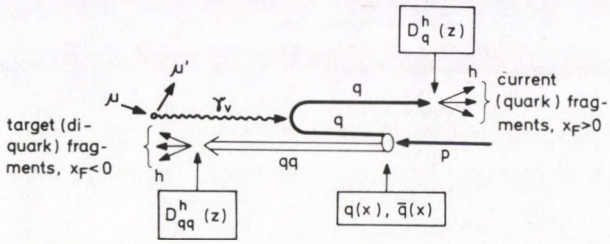


Fig. 4. Deep Inelastic Scattering of Muons according to the QPM. $q(x)$, $\bar{q}(x)$ are the quark and antiquark momentum distributions inside the nucleon, $D_q^h(z)$ is proportional to the probability that the quark q fragments into a hadron h. z is the energy of the hadron in units of the energy of the virtual photon

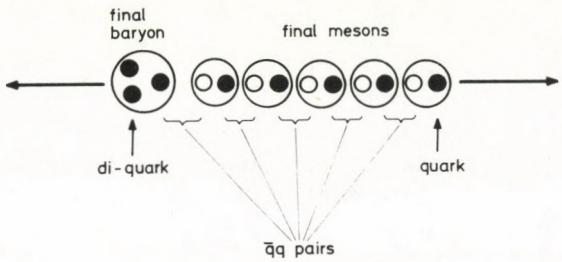


Fig. 5. Production of hadrons in a simplified fragmentation chain model. The original quark and diquark stretches out a colour field in which quark–antiquark pairs are materialized. These latter recombine with the original quark and diquark to form the final hadrons

The logarithmic rise of the multiplicity of the strange mesons manifests itself already at lower CMS energies as can be seen in the Figure, where by open symbols we have plotted the neutral kaon multiplicity observed in ν–Ne interacion [14]. It is a simple consequence of the increase of the available phase space with increasing W, namely the logarithmic rise of the the rapidity range, where the CMS rapidity is defined as

$$y^* = \frac{1}{2}\ln\frac{E^* + p_{\|}^*}{E^* - p_{\|}^*},$$

E^* and $p_{\|}^*$ being the CMS energy and longitudinal momentum of the hadron. Indeed, the EMC has shown [15] that the y^* distribution of the charged pions exhibits

a flat plateau, the width of which increases logarithmically with W, whereas its heignt remains constant.

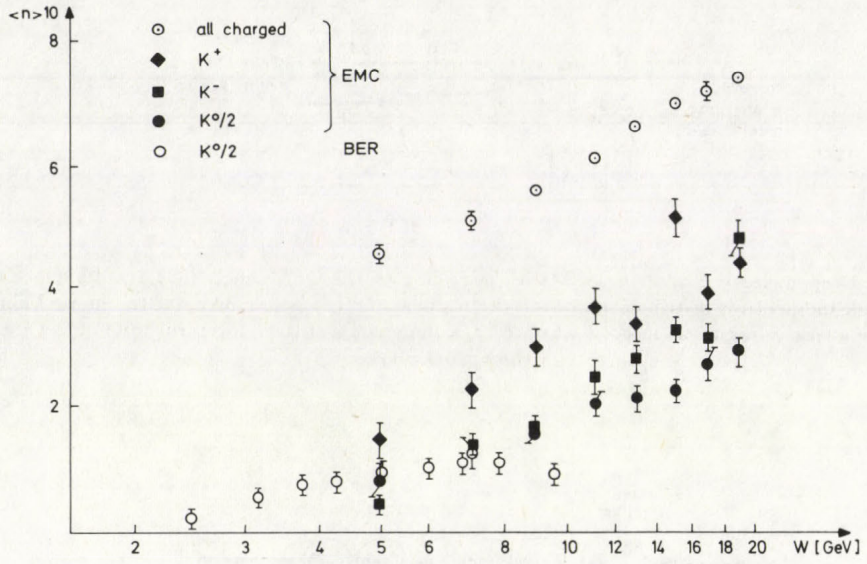


Fig. 6. Average multiplicities of the strange mesons as a function of W. Full symbols represent measurements from the EMC. Open symbols are used for the data from the Brookhaven–Columbia–Rutgers [14] Collaboration

The strange baryon multiplicities are shown in Fig. 7. Here the rise becomes clear only at higher energies ($W > 10$GeV), well above the threshold of the $\Lambda\bar{\Lambda}$ pair production. From the ν-Ne experiment [14] one can conclude that the average Λ multiplicity is constant in the $2 \lesssim W \lesssim 10$ GeV region indicating that most of the Λ particles originate from the recombination of the target diquark with an s quark in the fragmentation chain.

This observation is supported by the x_F distributions, where x_F is the longitudinal momentum of the particle in the CMS in units of the maximum available energy, $W/2$. The x_F distribution of the $\bar{\Lambda}$ is more or less symmetric in the forward ($x_F > 0$) and backward ($x_F < 0$) hemisphere (Fig. 8b) in a striking contrast with the Λ particles (Fig. 8a) which are mainly produced in the target fragmentation region ($x_F < 0$). One can see also that the production is practically independent whether the target is hydrogen or deuterium as expected from the QPM.

The parameter γ_s/γ_u can be determined the best from the strange to non-strange multiplicity ratio, which is directly proportional to it. In Fig. 9 this ratio is plotted against W. The solid lines correspond to the LUND model predictions with different values of the parameter. One can be see that γ_s/γ_u is between 0.25

and 0.35, however there is an indication that it depends on W.

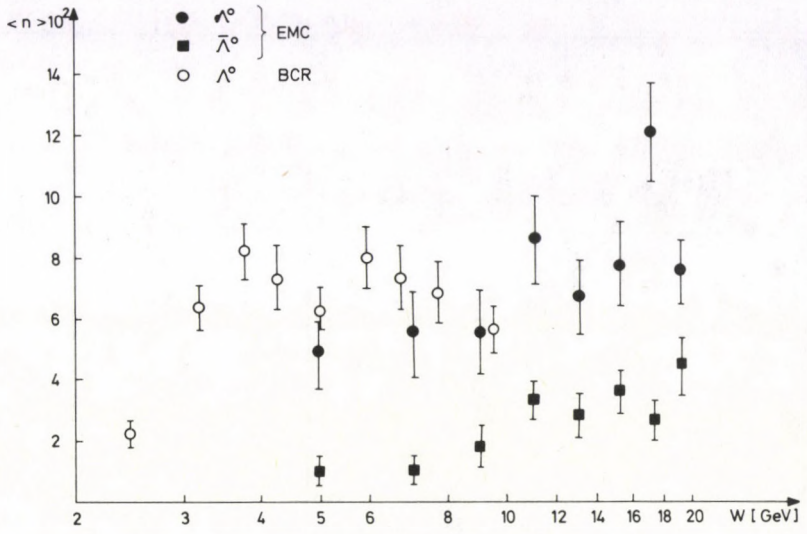


Fig. 7. Average multiplicity of Λ and $\bar{\Lambda}$. Symbols are used as indicated in Fig. 6

Strange particle production enables a further test on the QPM, the mechanism depicted in Fig. 4. If we plot the CMS rapidity distribution of the total observed strangeness normalised to the number of events (Fig. 10) one can observe a clean separation of the forward (quark) and backward (diquark) fragmentation region. In the forward region ($y^* > 0$) the net strangeness is positive due to the fact that the struck quark is mainly a u quark which recombines with an s quark forming a K^+. On the other hand, in the backward ($y^* < 0$) region the remaining diquark picks up an s quark to form a strange baryon which has a negative strangeness. This is, however, only true if x, the fraction of the momentum carried by the quark, is high (greater than 0.1), otherwise the virtual photon strikes a quark (or antiquark) in the strange-symmetric sea which does not result in a separation of the forward and backward region as can be seen in Fig. 11.

The hadronisation process depicted in Fig. 5 can be tested in more detail by looking at the pair correlation of strange mesons. Although smeared somewhat by the decay of unstable particles it is generally expected from the Figure that meson pairs with opposite strangeness are close in the CMS rapidity, whereas pairs with the same strangeness show no such correlation. This is clearly demonstrated in Fig. 12 where the average rapidity of one of the particles from the pair is plotted against the rapidity of the other particle.

In conclusion, we can say that all of the above observations in strange particle production are in a good agreement with the QPM of the hadronisation process.

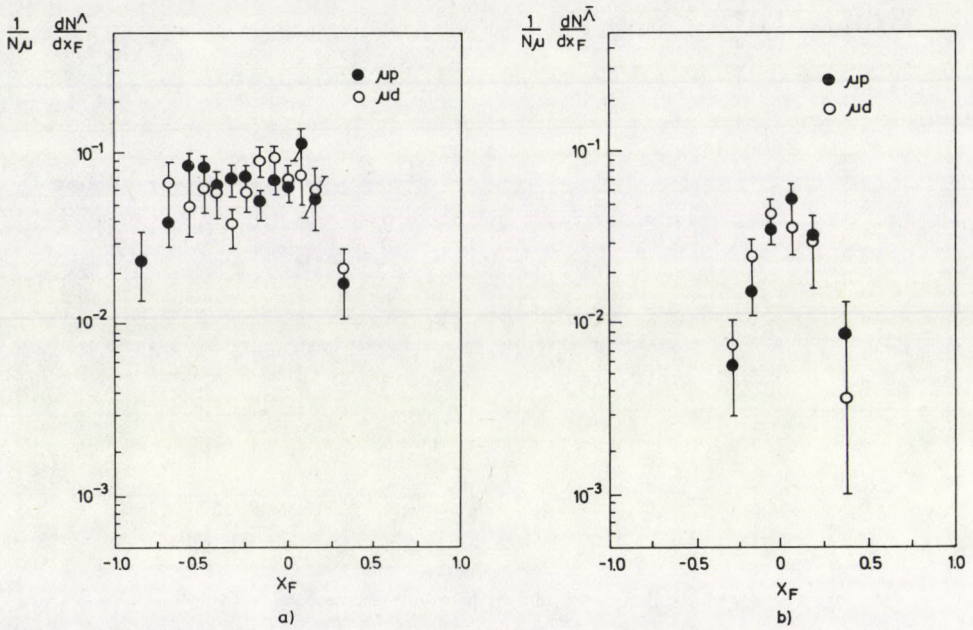


Fig. 8. x_F distribution of a) Λ and b) $\bar{\Lambda}$ particles. The full symbols correspond to data on hydrogen, the open symbols to those on deuterium target

Vector meson production

Inclusive production

Vector Mesons (VM) have the same quantum numbers as the virtual photon and therefore one may expect particularities in their production in electromagnetic interactions. Indeed, if we compare the scaled energy distribution of a light pseudoscalar (PS) meson, e. g. the π^0 [16] with that of a heavy VM (e. g. the J/ψ[8]) in Fig. 13a, where $z = E/\nu$, E being the energy of the particle and ν that of the virtual photon in the laboratory frame, respectively, one observes a striking difference. The mechanisms responsible for the production of these two particles are clearly different. The π^0 is thought to be produced by the hadronisation process depicted in Fig. 5 whereas J/ψ production can be described by the Photon–Gluon-Fusion (PGF) mechanism [17] depicted in Fig. 14. Whether the difference in the production mechanism is due to the spin, isospin or to the quark composition of the particles is of considerable importance therefore the EMC has carried out a systematic study of the inclusive production of light and medium heavy VM's like the ϱ, ω and ϕ.

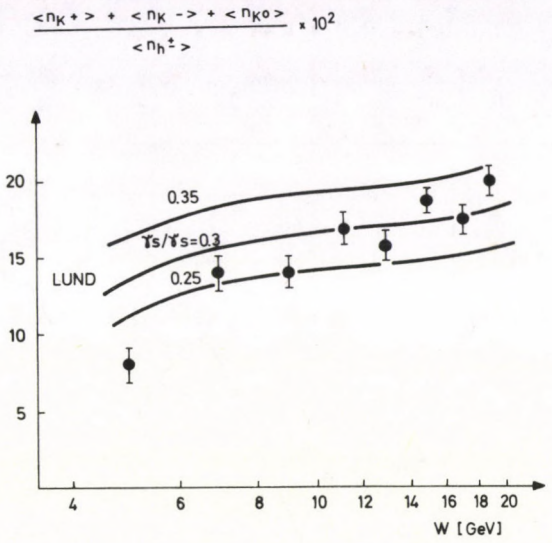


Fig. 9. The ratio of the average $< K^+ > + < K^- > + < K^0 >$ multiplicity to the average charged multiplicity as a function of W. Solid lines are the LUND model predictions with $\gamma_s/\gamma_u = 0.25, 0.30$ and 0.35, respectively

The number of the observed VM's after subtraction of the background is given in Table II. The z distributions are shown in Fig. 13b. The following remarks can be made:

Table II
The number of the observed vector mesons

ρ^0	ω	ϕ
7820 ±600	469±82	68.6±16.5

At low and medium z values the inclusive distributions drop very fast with increasing mass of the VM. Similarly, with increasing mass the disagreement with the LUND model becomes more and more significant. However, as z increases the different curves approach each other. In particular the distribution of the π^0 agrees well with that of the ϱ and ω for $z \geq 0.8$. One would expect three times as many ϱ and ω as π^0 because of the statistical population of their spin states. Thus the observed similarity in the production rate for the π^0 and ϱ (ω) may be due to a mass suppression for the latter. Interestingly, ϱ and ω production is practically the

same in the full z range. Since these two particles have mass values very close to each other one can conclude that the different number of the isospin states (3 for the ϱ and 1 for the ω) does not influence the hadronisation process.

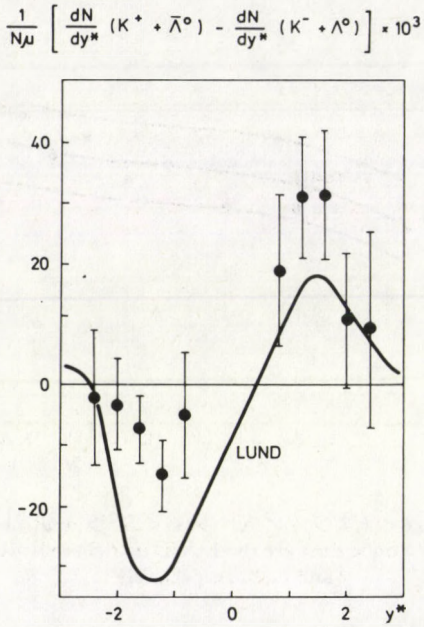


Fig. 10. y^* distribution of the observed total strangeness: $K^+ + \bar{\Lambda} - K^- - \Lambda$. The solid line is the LUND model prediction with $\gamma_s/\gamma_u = 0.3$

There is another difference in the inclusive π^0 and J/ψ production. While the former one is practically independent of Q^2 the latter exhibits a characteristic propagator effect [8]. In inclusive ϱ-production no Q^2-dependence was observed, concerning the ϕ meson no conclusion can be made on this point owing to the insufficient statistics.

Exclusive production

At high z ($z > 0.95$) one can study the process depicted in Fig. 14 where the VM is produced in a quasi-elastic scattering of the virtual photon on the nucleon. It has been found [18, 19] that at low Q^2 the process can be described by the Vector-Meson-Dominance Model (VDM)[20] in which the virtual photon first transforms into a virtual VM which in turn scatters elastically on the incident nucleon (Fig. 15). If however Q^2 is increasing the lifetime of the virtual photon is becoming

shorter and shorter leaving no time for the transformation into a virtual hadron. Therefore in this case one would expect a hard (electromagnetic) scattering in the nucleon vertex in Fig. 14.

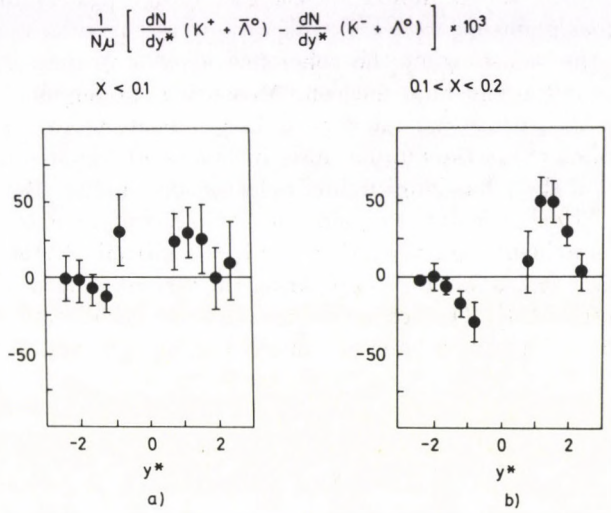


Fig. 11. y^* distribution of the observed total strangeness in two different x regions: a) $x < 0.1$ and b) $0.1 \leq x \leq 0.2$

In the following we shall demonstrate that this is indeed the case. The process has been studied by the EMC in hydrogen and ammonia target (the main component of the polarised target). Typical t' distributions on ammonia target are shown in Fig. 16 where $t' = t - t_{\min}$, t being the four-momentum transfer squared between the initial and the final state nucleon and $t_{\min}$ is its minimum value. One can observe an exponential fall with some excess of events on the top of it around $t' = 0$. This excess is the result of the coherent scattering of the virtrual photon on several nucleons in the ammonia nucleus, whereas the exponentially falling points extending to high t' values are due to the incoherent scattering on individual nucleons, as only this component is present in a measurement on hydrogen target. One can clearly see from the Figure that independently of the energy as Q^2 increases the t' distributions become more and more flat. This is illustrated in Fig. 17, where the slope b of the incoherent scattering is plotted together with the real photoproduction experiment ($Q^2 = 0$)[18]. In the range of Q^2 from 0 to 14 GeV^2 the slope parameter drops from a value typical for hadronic diffraction to a value characteristic for point-like scattering.

At the same time one can observe in Fig. 16 that the ratio of coherent to incoherent contribution also decreases with increasing Q^2. Defining

$$\sigma_{\text{inc}} = a \int_0^\infty e^{-b|t'|} d|t'|$$

and

$$\sigma_{coh} = \text{number of events} - \sigma_{inc}$$

we plot their ratio in Fig. 18. It is evident that as Q^2 increases the coherent scattering on several nucleons of the ammonia nucleus dies away—again indicating that the scattering becomes point-like.

Furthermore, the vanishing of the coherence is accompained by a helicity transfer between the initial and final nucleon. Measuring the angular distribution of one of the pions in the ϱ decay one can deduce the polarisation of the ϱ. The value of the parameter, called r^{00} of the angular distribution is 0 if the ϱ is transversally polarised, and it is 1 if the ϱ has longitudinal polarisation. In Fig. 19 r^{00} is shown as a function of Q^2. The change of the ϱ polarisation from transverse to longitudinal as Q^2 is increased is evident. On the other hand, the virtual photon has always transverse polarisation as was found by comparing the virtual photon cross sections at two different energies [21]. One can thus conclude that while at small Q^2 the helicity is conserved, at high Q^2 a helicity transfer can be observed in the nucleon vertex in Fig. 14.

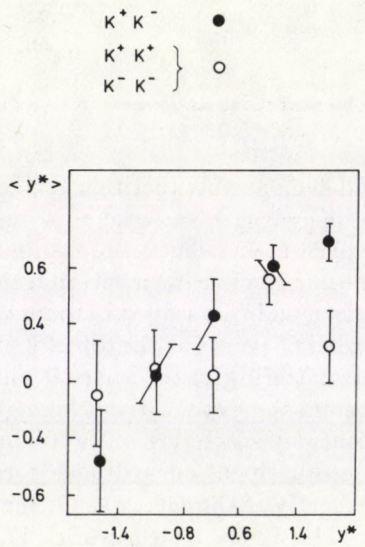


Fig. 12. Average rapidity $< y^* >$ of a strange meson versus the rapidity of another strange meson from the same event: full symbol is used for total strangeness zero and open symbol is used when the strangeness of the pair is ± 2

In conclusion, in elastic ϱ production a change of regime is clearly seen: as Q^2 increases the process changes from a soft hadronic interaction to a hard point-like electromagnetic interaction. It is therefore somewhat surprising that nevertheless

the total production rate of the process shows a universal feature, namely the rate for different types of VM can be described by a universal formula valid for all the three types of VM we have studied: ϱ, ϕ and J/ψ and in the *full* Q^2 range:

$$\text{rate} = c \cdot \frac{g_{SU3}}{(Q^2 + M_{VM}^2)^2}. \tag{1}$$

Here g_{SU3} is the SU3 coupling constant of the photon to the VM, M_{VM} is the mass and c is a universal constant chosen in such a way that for every Q^2 the rate of the ϱ meson equals 9.

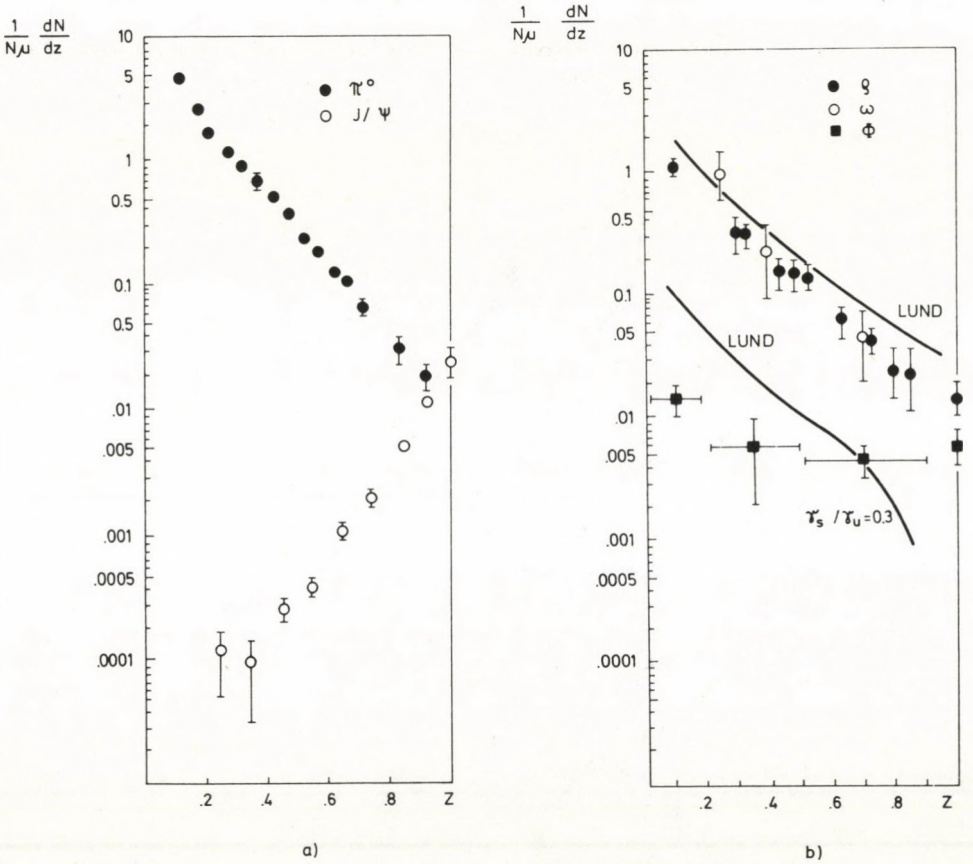


Fig. 13. z distributions normalized to the number of events a) for π^0 (full) and J/ψ (open symbol) b) for ϱ (full circle), ω (open circle) and ϕ (full square). The solid lines are the LUND model predictions for ϱ and ϕ

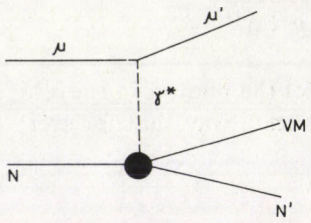


Fig. 14. Feyman graph for elastic Vector Meson production

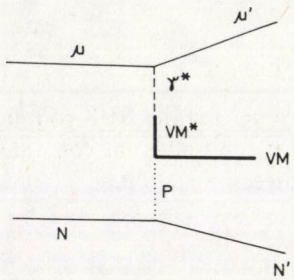


Fig. 15. Feynman graph for elastic VM production according to the Vectormeson – Dominance Model

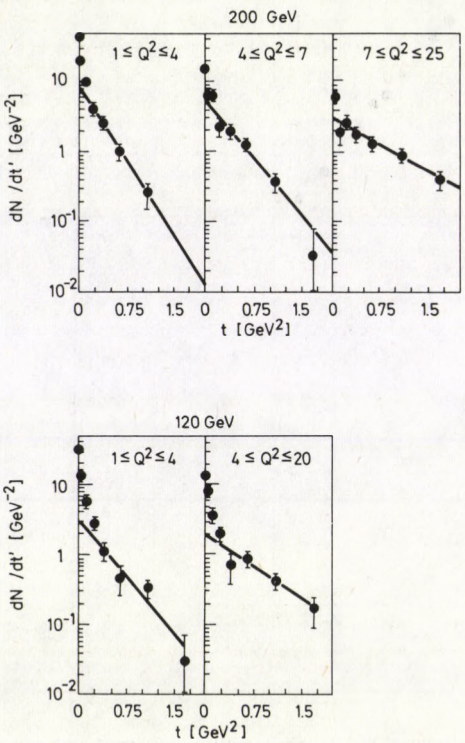


Fig. 16. t' - distributions of exclusively produced ϱ mesons in μ NH_3 scattering at different beam energies and Q^2

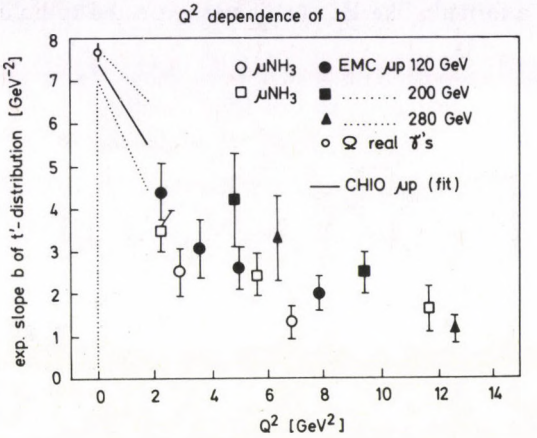


Fig. 17. The slope of the t'-distributions as a function of Q^2. Full symbols: μp scattering, open symbols: μNH_3 incoherent scattering, measured by the EMC. Low Q^2 measurements taken from [18] and [19] are also shown, the latter data are represented by lines

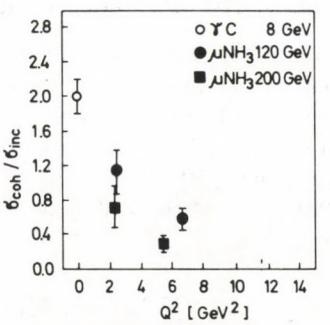


Fig. 18. $\sigma_{coh}/\sigma_{inc}$ (see text) vs Q^2 for two different beam energies observed by the EMC in μNH_3 scattering. Also shown is the value obtained in a real photoproduction experiment on carbon target

The experimentally observed rates are confronted with Eq. (1) in Table III. One can see excellent agreement. It would be very interesting to see if the validity of Eq. (1) could be extended also to PS particles like the π^0 or η^0. According to the

above picture elastic PS particle production is not expected at low Q^2 (where VDM is valid), on the other hand at high Q^2 PS particles can be produced elastically in the electromagnetic scattering of the virtual photon with one of the quarks of the nucleon. Therefore a formula like Eq. (1) is not expected to hold for the whole Q^2 range.

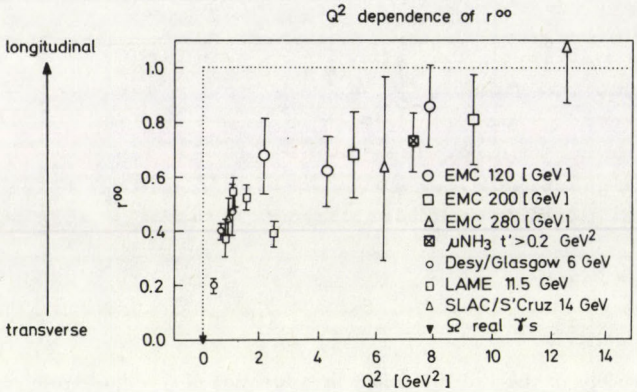


Fig. 19. The ϱ polarisation, r^{00} vs Q^2. Full symbols are the results from the EMC, open symbols represent data from low Q^2 measurements

Determination of the u valence quark distribution

If the simple picture of deep inelastic leptoproduction, as shown in Figs 4 and 5, is valid one can in principle determine the quark momentum distribution inside the nucleon from the observed hadronic final state. The comparison of this result with the same distribution obtained from the measurement of the scattered lepton provides then a strong test of the QPM.

According to the QPM (Fig. 4) the scaled energy distribution of the hadrons in the forward direction is given by

$$\frac{1}{N_\mu}\frac{dN^h}{dz} = \frac{\sum_q e_q^2 q(x) D_q^h(z)}{\sum_q e_q^2 q(x)}, \tag{2}$$

where N^h and N_μ are the number of hadrons and the number of events, respectively, $q(x)$ and e_q are the momentum distribution and the charge of the quark q, finally $D_q^h(z)$ is proportional to the probability that the quark q fragments into a hadron h. The denominator is the F_2 structure function. The validity of Eq. (2) has been extensively tested by the EMC, among others an interesting prediction, that Eq. (2) when applied to all unidentified hadrons does not depend

on x and thus on the type of the target [22]. Fig. 20 demonstrates that indeed $1/N_\mu \cdot dN^{\pm}/dz$ is the same for both H_2 and D_2 targets.

Table III
Vector meson production ratios

ρ	ϕ	J/ψ	Type/Remarks	
9	2	8	g_{SU3}	
9	0.68±0.1	3. ±0.1 10^{-2}	Data	$Q^2=0$
9	0.65	3. 10^{-2}	Eq. (1)	
9	0.63±0.1*	0.17±0.03	Data	$Q^2 = 1\text{GeV}^2$
	1.2 ±0.2**			
9	1.2	0.18	Eq. (1)	
9	1.6 ±0.4	2.0 ±0.4	Data	$Q^2=7.5\text{GeV}^2$
9	1.8	1.8	Eq. (1)	

*E_μ=150GeV **E_μ=100GeV

Working out Eq. (2) for the *difference* of positive and negative pions in case of μp scattering using the relations following from charge conjugation and isospin symmetry:

$$D_u^{\pi^+} = D_{\bar{u}}^{\pi^-}, \qquad D_u^{\pi^+} = D_d^{\pi^-},$$
$$D_d^{\pi^+} = D_{\bar{d}}^{\pi^-}, \qquad D_d^{\pi^+} = D_u^{\pi^-},$$

one obtains

$$h^\pi(x) \equiv \frac{1}{N_\mu}\int\left(\frac{dN^{\pi^+}}{dz} - \frac{dN^{\pi^-}}{dz}\right)dz = \frac{x}{9F_2^h(x)}[4u(x) - d(x)]\int(D_u^{\pi^+} - D_u^{\pi^-})dz, \tag{3}$$

where $u(x)$ and $d(x)$ are the distributions of the u and d valence quarks inside the proton. A similar expression for the difference of protons and antiprotons can be obtained:

$$h^p(x) \equiv \frac{1}{N_\mu}\int\left(\frac{dN^p}{dz} - \frac{dN^{\bar{p}}}{dz}\right)dz = \frac{x}{9F_2^h(x)}[4u(x) + d(x)]\int(D_u^p - D_u^{\bar{p}})dz. \tag{4}$$

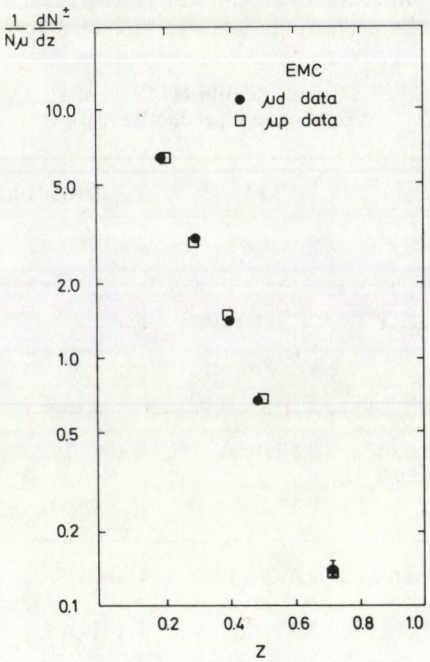


Fig. 20. $1/N_\mu \cdot dN^{\pm}/dz$ for all charged hadrons on hydrogen (full) and deuterium (open symbols) targets

Combining Eqs (3) and (4) one can extract $u(x)$ and $d(x)$ provided

$$c^{\pi} \equiv \int (D_u^{\pi^+}(z) - D_u^{\pi^-}(z))dz$$

and

$$c^{p} \equiv \int (D_u^{p}\ (z) - D_u^{\bar{p}}\ (z))dz$$

are known. These latter can be determined from the same combinations of the final state hadrons in μd scattering:

$$d^{\pi}(x) \equiv \frac{1}{N_\mu} \int \left(\frac{dN^{\pi^+}}{dz} - \frac{dN^{\pi^-}}{dz} \right) dz = \frac{3x}{18F_2^d(x)} [u(x) + d(x)] \int (D_u^{\pi^+} - D_u^{\pi^-})dz, \tag{5}$$

$$d^{p}(x) \equiv \frac{1}{N_\mu} \int \left(\frac{dN^{p}}{dz} - \frac{dN^{\bar{p}}}{dz} \right) dz = \frac{5x}{18F_2^d(x)} [u(x) + d(x)] \int (D_u^{p} - D_u^{\bar{p}})dz. \tag{6}$$

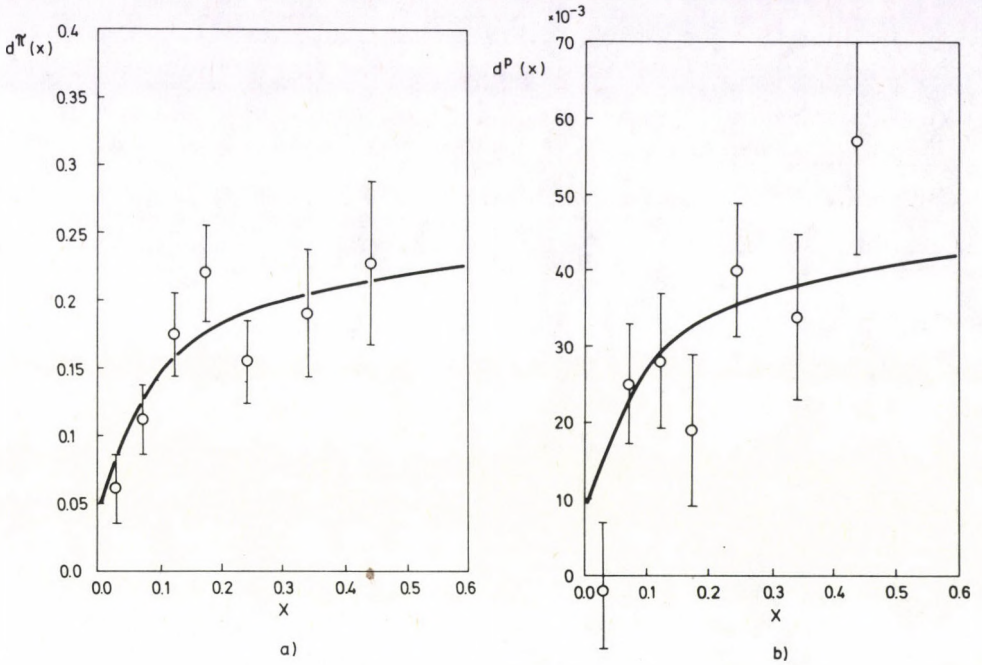


Fig. 21. a) $d^{\pi}(x)$ and b) $d^{p}(x)$ (see Eqs (5) and (6) in the text). The solid lines are the fits to the Ansatz given by Eqs (7) and (8) in the text

At high x $F_2^d(x)$ is proportional to $(u(x) + d(x))$ therefore the x-dependence cancels in Eqs (5) and (6) and $d^{\pi}(x)$ as well as $d^{p}(x)$ approaches to constant values proportional to c^{π} and c^{p}. This behaviour is taken into account by the parametrisation

$$d^{\pi}(x) = \frac{3}{5} \frac{x^{\alpha}}{1 + \gamma(1-x)^{\beta}} \cdot c^{\pi}, \tag{7}$$

$$d^{p}(x) = \frac{x^{\alpha}}{1 + \gamma(1-x)^{\beta}} \cdot c^{p}. \tag{8}$$

The fit of Eqs (7) and (8) to the experimental data is shown in Figs 21 a and b. Using the obtained values of c^{π} and c^{p} from the fit and the F_2^h structure function on hydrogen measured by the EMC [4] one can deduce $u(x)$:

$$xu(x) = \frac{9F_2^h(x)}{8} \left[\frac{3}{5} \cdot \frac{h^{\pi}(x)}{c^{\pi}} + \frac{h^{p}(x)}{c^{p}} \right]. \tag{9}$$

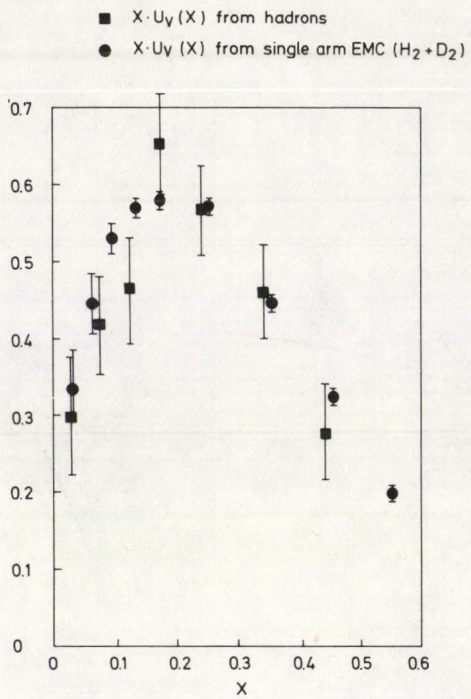


Fig. 22. The $x.u(x)$ valence quark distribution as determined from the inclusive distribution of identified hadrons using Eq. (9) in the text (full squares). Also shown is the same quantity derived from the single arm scattering data on H_2 and D_2 target [23] (full circles)

The result is plotted in Fig. 22. In the same Figure we also show the same quantity obtained from the measurement of the scattered muon on H_2 and D_2 target [23].

It is clear that the mesurement of $u(x)$ from the identified hadrons — mainly because of the low rate of forward produced protons — cannot compete with that from the scattered lepton. However, the good agreement between the two kinds of determination shows once again that QPM is able to describe correctly even an as complex process as the production of hadrons in deep inelastic scattering of leptons.

Acknowledgements

The results presented here have been obtained in a collective effort by the members of the EMC. I thank all of them, especially Drs Terry Sloan, Roland Windmolders, Francois Montanet, Klaus Hamacher, Klaus Mönig and Shafik Benchouk, for their contributions.

I am indebted to Prof. V. N. Gribov for useful discussions on the subject of exclusive VM production.

The kind hospitaly of Prof. Julius Kuti in the Physics Department of the UCSD and that of Prof. Charles Baltay in NEVIS Lab of the Columbia University is gratefully acknowledged.

References

1. The European Muon Collaboration consists of the following laboratories:
   III. Physikalisches Inst. A, Physikzentrum, Aachen, Germany
   CERN, Geneva, Switzerland
   DESY, Hamburg, Germany
   Fakultät für Physik, Universität Freiburg, Germany
   II. Inst f. Experimentalphysik, Univ. Hamburg, Germany
   II. Inst. f. Kernphysik, Universität Kiel, Germany
   Lab. de l'Accélérateur Linéaire, Orsay, France
   Dept. of Physics, University of Lancaster, England
   LAPP IN2P3, Annecy-le-Vieux, France
   Dept. of Physics, University of Liverpool, England
   CPPM, Faculté des Sciences de Luminy, Marseille, France
   Faculté des Sciences, Université de l'Etat à Mons, Belgium
   Max-Planck-Institut f. Physik u. Astrophysik, München, Germany
   Nucl. Phys. Laboratory, University of Oxford, England
   Rutherford and Appleton Lab., Chilton, Didcot, England
   Dept. of Physics, University of Sheffield, England
   Istituto di Fisica, Università di Torino, Italy
   Gustav Werners Institut, University of Uppsala, Sweden
   Physics Institute, University of Warsaw, Poland
   Institute for Nuclear Studies, Warsaw, Poland
   Fachbereich Physik, Universität Wuppertal, Germany.
   The Budapest team from the Central Research Institute for Physics: G. Eszes, G. Jancsó, E. Nagy, P. Ribarics, J. Tóth and L. Urbán participates in the EMC through the LAPP.
2. EMC, Preprint CERN/SPSC/74-78.
3. EMC, Nucl. Instr. Methods, *179*, 445, 1981.
4. EMC, Nucl. Physics, *B259*, 189, 1985; EMC, Physics Letters, *123B*, 123, 1983; EMC, Physics Letters, *105B*, 322, 1981.
5. EMC, Physics Letters, *123B*, 275, 1983.
6. EMC, Physics Letters, *100B*, 433, 1981
7. EMC, Physics Letters, *114B*, 373, 1982.
8. EMC, Nucl. Physics, *B213*, 1, 1983; EMC, Nucl. Physics, *B213*, 31, 1983.
9. EMC, Physics Letters, *167B*, 127, 1986.
10. EMC, Nucl. Instr. Methods, *212*, 111, 1983.
11. L. Diósi, G. Farkas, F. Ferenczy, T. Gémesy, K. Hajdu, A. Holba, J. Koch, T. Mózes, Zs. Oláh, Gy. Pintér, J. Spitzer, L. Szabó, Gy. Thaler, A. Zarándy and J. Urbán. Proc. Xth Int. Symp. on Nucl. Electronics, Dubna–ZFK Rossendorf–TU Dresden, Vol II, p. 50, Dresden, 1980.
12. R. P. Feynman, Photon–Hadron Interactions, Benjamin Press, New York, 1972; R. D. Field, R. P. Feynman, Nucl. Physics, *B136*, 1, 1978.
13. B. Andersson, G. Gustafson, G. Ingelman and T. Sjöstrand, Z. f. Phys., *C9*, 233, 1982.
14. N. J. Baker et al., Preprint RU-86-23.
15. EMC, Zeit. f. Phys., *C31*, 1, 1986.
16. EMC, Zeit. f. Phys., *C18*, 189, 1983.
17. L. M. Jones and H. W. Wyld, Phys. Rev., *D17*, 759, 2332, 1978. Further references can be found in ref. [8].
18. D. Aston et al., Nucl. Physics, *B209*, 56, 1982.
19. D. Schambroom et al., Phys. Rev., *D26*, 1, 1982.
20. T. H. Bauer, R. D. Spital and D. R. Yennie, Rev. Mod. Phys., *50*, 261, 1978.
21. EMC, Physics Letters, *161B*, 203, 1985.
22. E. Nagy. EMC Internal Report B5/12.
23. Valery Gibson, Thesis, Nucl. Phys. Laboratory, Univ. of Oxford, to be published.

Acta Physica Hungarica 64 (1-3), pp. 199-208 (1988)

COMPUTER SIMULATION OF THE NEUTRINO DETECTOR SYSTEM AT LAKE BAIKAL*

P. KAKAS and D. KISS

Central Research Institute for Physics
1525 Budapest, Hungary

(Received 8 January 1987)

A deep underwater muon and neutrino detector system is due to be completed at Lake Baikal, USSR. The primary purpose of this system is to measure cosmic ray neutrino flux at extremely high energies. The experimental set-up and the computer simulation of the detector acceptance are presented. The detector system strongly depends on the arrangement of its constituent parts, the photomultipliers.

1. Introduction

The building of the neutrino detector in Lake Baikal was started in 1983. For a better understanding of its operation it is useful to review the history of neutrino measurements.

In 1930 W. Pauli foretold the discovery of neutrino. His assumption of the parameters of the neutrino: the 1/2 spin, the zero or nearly zero mass, and the extremely low interactional cross section were justified by later measurements. It was in 1953 that F. Reines and G.A. Cowan directly demonstrated the existence of the neutrino at a nuclear reactor (nuclear reactors can be considered as high intensity sources of neutrinos due to the beta activity of the fission products) with the help of induced β^+ decay.

In the experiment it was found that the international cross section of the neutrino is of the order of magnitude of 10^{-44}cm^2 (the actual value strongly depends on the energy of the neutrino). Since that time, the problem concerning the detection of the neutrino has become well known. The device for cosmic ray neutrino flux measurements should be located at great depth underground or under water to avoid the noise due to the high background radiation and it needs to be large because of the low flux of the neutrino. In the seventies, the idea of using water as the medium of the neutrino detector arose. A neutrino interacting with nucleons can create muons. According to the Bethe–Bloch formula, a relativistic muon moving in water loses about 2 MeV energy per cm. Since the muon's energy range lies in the TeV order of magnitude (10^{12} eV), this particle may travel several thousand meters in water before being absorbed. Within this interval of time and distance the direction of its movement can be well studied — with the help of the emitted

*Dedicated to Prof. G. Marx on his 60th birthday.

Cerenkov radiation — the direction of which radiation is in good correlation with that of neutrinos.

Measurements of this kind are due to be carried out by two independent teams of scientists: one team plans to build a neutrino detector in the Pacific Ocean near the Hawaii Islands at a depth of 5 km; the other team (of Soviet scientists) has chosen Lake Baikal for the location of their detector [1]. Two favourable conditions speak in favour of the latter version, viz.: the building of the detector is made essentially easier by the freezing over of the Lake during winter on the one hand and, on the other, due to chemical or biological processes the photon background in the water of Lake Baikal is lower than that of the Pacific Ocean near Hawaii at a depth of 5 km.

To sum up, the task is the following: muons produced by the interactions of neutrinos have to be detected on the basis of their Cerenkov radiation in water of great volume.

2. Construction

It is planned that the completed neutrino detector should consist of many photomultipliers located near each other.

The individual detector units (two multiplier pairs) are located at a distance of about 15 m from each other. 96 multipliers are coupled to one cable.

Nineteen cables form a detector plane with the cables also being at a distance of 15 m from each other. The whole detector system comprises 6 detector planes.

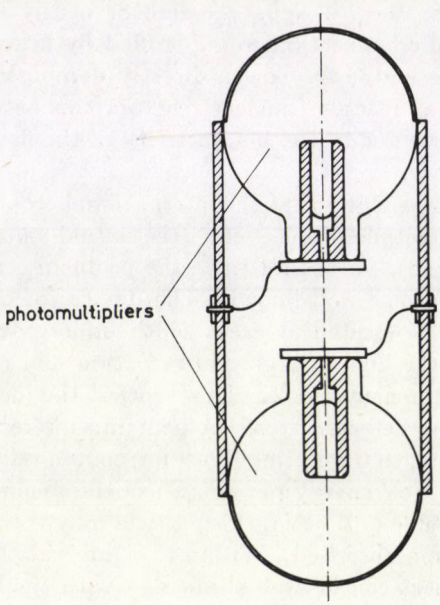


Fig. 1. Two photomultipliers forming a pair

In this experiment we are only interested in the muons originating from the interactions of neutrinos. The detector system should, accordingly, operate solely with horizontally directed muons since it is only in this direction that the secondary muons are able to produce an effect at the given detector depth. Muons from any direction other than horizontal are excluded from detection by increasing the distance between the detector planes. The distance between the detector planes is 40 m. The dimensions of the whole detector system are, accordingly: $270 \times 345 \times 200$ m. The uppermost photomultipliers of the arrangement are at a depth of 1000 m. From the above arrangement it follows for the number of photomultipliers that the 19 cables of each individual plane carry 1824 photomultipliers while the complete system of 6 planes comprises 10 944 photomultipliers.

3. Computer simulation

i. Simulation for a single photomultiplier

For simulation purposes, the first step is to calculate the signal to be obtained for one photomultiplier at a high-energy muon.

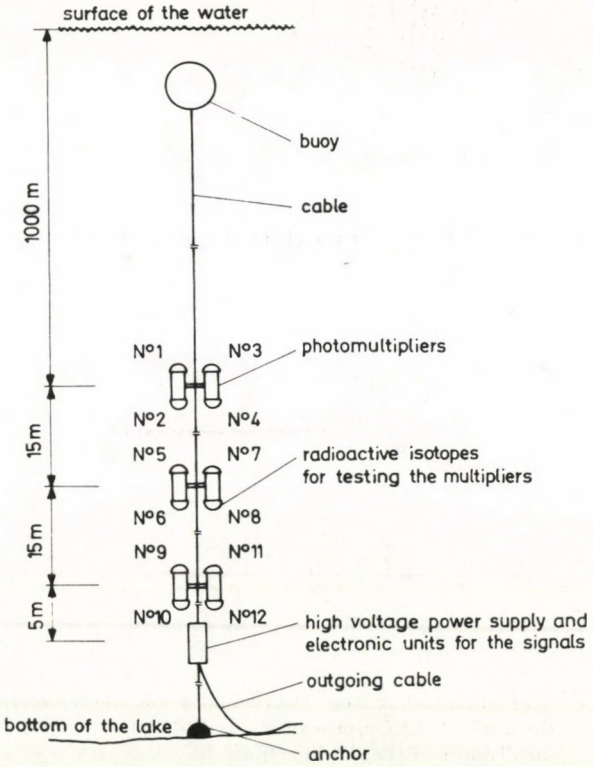


Fig. 2. Construction of a cable with photomultipliers

In a path section of length l the Cerenkov light of the muon yields photons between the wavelength $\lambda + d\lambda$:

$$dN = 2\pi\alpha \sin^2 \Theta_C \cdot l \cdot d\lambda/\lambda^2. \tag{1}$$

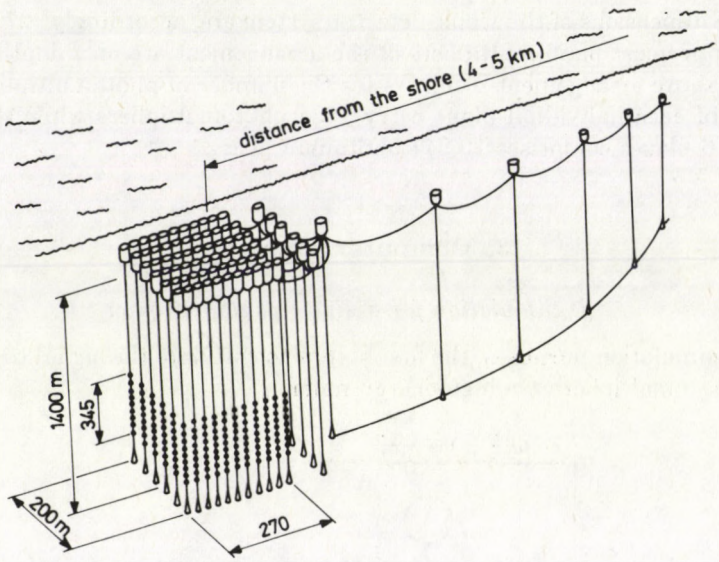


Fig. 3. The neutrino detector system

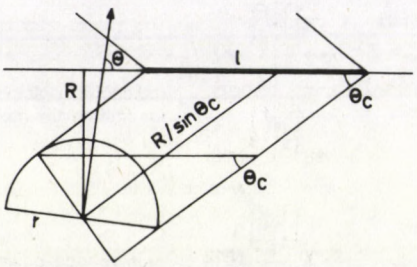


Fig. 4. Schematic diagram for calculating detection. A semicircle represents the photomultiplier
Θ_C - the angle of the Cerenkov radiation (41°)
R - the distance of the detector from the muon path
Θ - the angle formed by the axis of the detector and the muon path
r - the radius of the detector

Here α is a fine structure constant (1/137). Since the light passes through water before reaching the photomultiplier, its intensity decreases due to dispersion, absorption, etc. as a function of the path length and the transparency of the water.

$$dN = 2\pi\alpha \cdot \sin^2 \Theta_C \cdot l \cdot e^{-\beta(\lambda)R/\sin\Theta_C} d\lambda. \tag{2}$$

Here the last term represents the effect of water. The function $\beta(\lambda)$ was obtained in a former measurement [2] and is illustrated in Fig. 5.

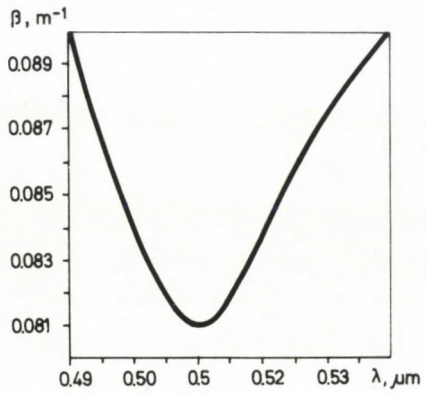


Fig. 5. Absorption coefficient β as a function of the wavelength of light

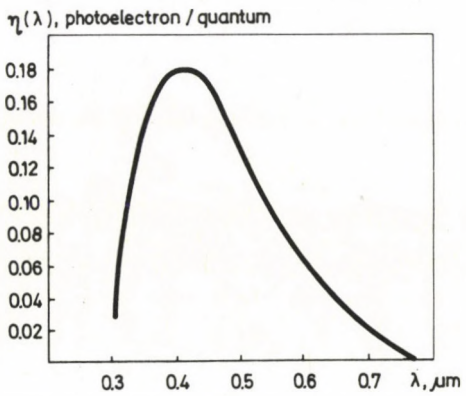


Fig. 6. Photon–electron conversion efficiency as a function of the wavelength

The function curve cannot be described by a mathematical function therefore a polynomial was fitted to the measurement data for the subsequent numerical integration. It has to be taken into account that the direction of the Cerenkov light forms an angle with the detector axes; that is, one has to calculate the projection of the detector surface, S_d, to that part of the spherical surface cut by the conical solid angle of the Cerenkov light, S_c, and the ratio of the two surfaces:

$$Y = S_d/S_c. \tag{3}$$

The final form of this expression is

$$\gamma = \frac{r(1+\cos(\Theta-\Theta_C))}{4(2R + r\cos\Theta_C(\cos(\Theta-\Theta_C)-1))}. \tag{4}$$

Hence formula (2) takes the form

$$dN = \frac{\pi\alpha}{2} \frac{r^2 \sin\Theta_C (1+\cos(\Theta-\Theta_C))^2}{2R + r\cos\Theta_C(\cos(\Theta-\Theta_C)-1)} \cdot e^{-\beta(\lambda)\cdot R/\sin\Theta_C}. \tag{5}$$

During the operation of the detector, the photocathode emits electrons upon the influence of the photons. The efficiency of this process depends only on the wavelength and can be measured. The shape of the function is shown in Fig. 6 [3]. Our results can be summed up by the expression written in integral from:

$$N(R,\Theta) = \frac{\pi\alpha}{2} \frac{r^2 \cdot \sin\Theta_C (1+\cos(\Theta-\Theta_C))^2}{2R + r\cos\Theta_C(\cos(\Theta-\Theta_C)-1)} \int_{\lambda_{low}}^{\lambda_{upp}} \frac{\eta(\lambda)}{\lambda^2} \cdot e^{-\frac{\beta(\lambda)\cdot R}{\sin\Theta_C}} \cdot d\lambda. \qquad (6)$$

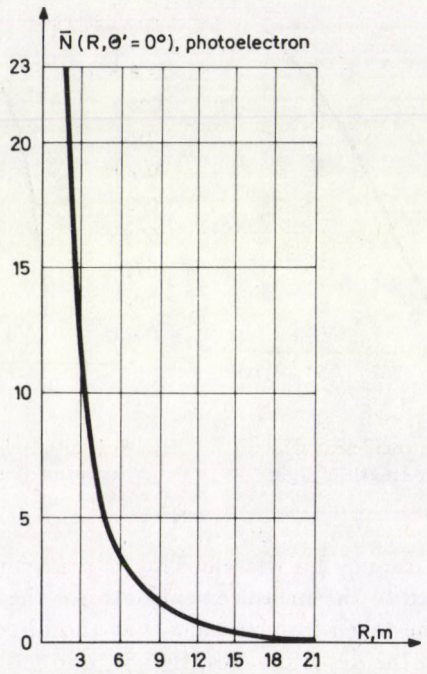


Fig. 7. Number of photoelectrons as a function of distance from the detector

The limits of the integration are determined by the sensitivity range of the photomultiplier and their values are: $\lambda_{lower} = 300$ nm; $\lambda_{upper} = 700$ nm. The result is illustrated in Fig. 7 as a function of R with $\Theta = 41°$.

ii. The detector system

So far as its operation is concerned the system can be divided into cells (see Fig. 8) which operate independently of each other and can be considered as separate

units. For the further calculations let us place the cell into the coordinate system shown in Fig. 9. Within the cell we are only interested in the measurement results for the finite number of points shown in Fig. 10. Hereafter, it is to be considered that the photoelectrons emitted from the photocathode obey a probabilistic law:

$$P(k, \bar{N}) = \frac{(\bar{N})^k}{k!} \cdot e^{-\bar{N}}, \tag{7}$$

where $\bar{N}$ is the mean value of the number of photoelectrons detected (in our case $\bar{N} = N$) and k is the number of photoelectrons obtained.

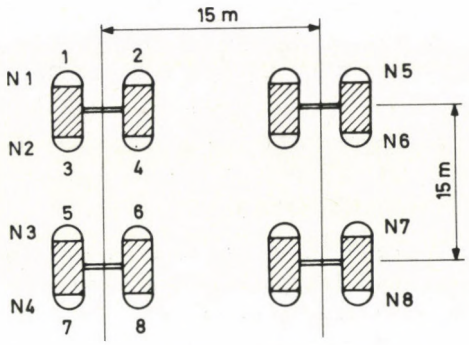


Fig. 8. One of the cells of the detector system

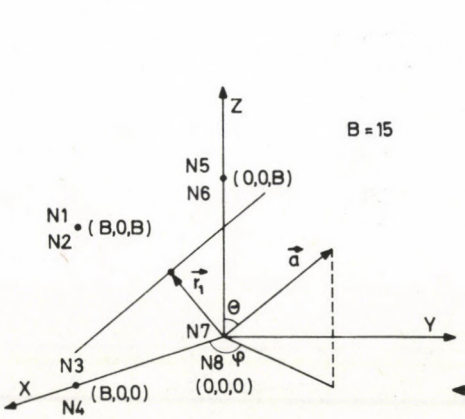


Fig. 9. The cell shown in Fig. 8 placed in a coordinate system

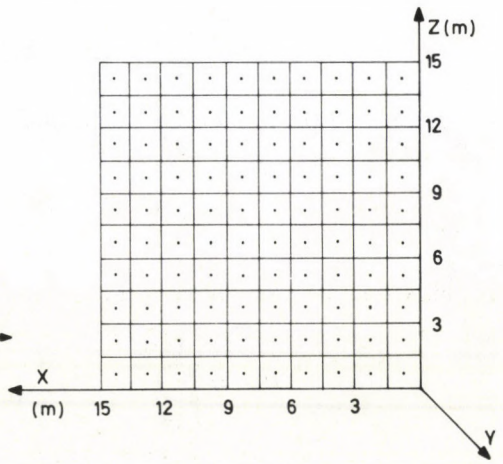


Fig. 10. Point system of a cell for which calculation was performed

The probability of the photomultiplier not detecting any event is given by

$$P(0, \bar{N}) = e^{-\bar{N}}. \tag{8}$$

For us, it is important only if the detector records at least one event, that is

$$g = 1 - e^{-\bar{N}}. \tag{9}$$

As can be seen from Fig. 8, the detectors having the same direction are near each other. Independent pairs of such detectors yield the following detection probability:

$$P = (1 - e^{-\bar{N}})^2 = g^2. \tag{10}$$

We now have to calculate the probability of obtaining an event by the detectors located in the four angles for all the points of the given lattice. With such a system it is a reasonable requirement that the efficiency of the detection should be higher than 90 per cent. Accordingly, we defined three selection conditions which are partly determined by the behaviour of the muons to be detected and partly required in order to improve the detection efficiency.

a. First condition of selection. Here we consider the cases when the detectors operate in pairs and at least two pairs give a response at the same time. The probability of detecting more than m events is

$$P_{\geq m} = \sum_{j=m}^{N} (-1)^{j-m} \binom{j-1}{m-1} S_j, \tag{11}$$

$$S_1 = \sum_i P\{E_i\}; \qquad S_2 = \sum_i \sum_k P\{E_i \cap E_k\}. \tag{12}$$

In our case all the events are independent

$$P(E_1 \cap \cdots \cap E_n) = \prod_{i=1}^{n} P(E_i) \tag{12a}$$

In this case, the probability will have the form

$$P_{\geq 2} = \sum_{j=2}^{4} (-1)^{j-2} \binom{j-1}{1} S_j = S_2 - 2S_3 + 3S_4, \tag{13}$$

$$S_1 = P_1P_2 + P_1P_3 + P_1P_4 + P_2P_4 + P_3P_4,$$

$$S_2 = P_1P_2P_3 + P_1P_2P_4 + P_1P_3P_4 + P_2P_3P_4,$$

$$S_3 = P_1P_2P_3P_4,$$

$$P^1_{TR1} = P_{\geq 2}. \tag{14}$$

This information already enables the first condition of selection to be calculated for the detectors on one of the cables. Performing the operation for the second cable in the same way, we can summarize the results as follows:

$$P_{\square}^{1} = 1 - (1 - P_{CA1}^{1})(1 - P_{CA1}^{2}) = P_{CA1}^{1} + P_{CA1}^{2} - P_{CA1}^{1} P_{CA1}^{2}. \tag{15}$$

Hence we can calculate the probabilities determined by the first condition of selection for all points. By taking the arithmetic mean of these, we obtain a value characteristic of the whole plane. This value is 0.51 which is far below the expected efficiency of 90 per cent. To improve this value, a second condition of selection was chosen.

b. Second condition of selection. With this condition we consider a detector group of four (i. e. two pairs) and calculate the probability of a minimum of three of them giving a response:

$$P_{\geq 3} = \sum_{j=3}^{4} (-1)^{j-1} \binom{j-1}{2} S_j = S_3 - 3S_4, \tag{16}$$

$$\begin{aligned} S_3 &= g_1 g_2 g_3 + g_1 g_2 g_4 + g_1 g_3 g_4 + g_2 g_3 g_4, \\ S_4 &= g_1 g_2 g_3 g_4, \\ P_{CA1}^{1} &= P_{\geq 3}. \end{aligned} \tag{17}$$

The probabilities obtained for two detector groups of four can be summarized as

$$P_{1-}^{2} = P_{CA2}^{1} + P_{CA2}^{2} - P_{CA2}^{1} P_{CA2}^{2}. \tag{18}$$

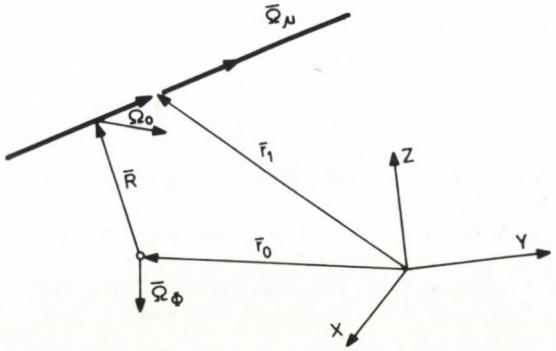


Fig. 11. Schematic diagram for calculating the detector system acceptance.
$\bar{\Omega}_\mu$ - unit vector characteristic of the muon path
$\bar{r}_0$ - coordinate vector of the detector
$\bar{\Omega}_\Phi$ - normal vector of the detector surface
$\bar{r}_1$ - coordinate vector of the point through which the muon passes
$\bar{\Omega}_0$ - vector of direction of the Cerenkov light

Now we perform the same calculation for the second cable and summarize the results obtained for the two cables:

$$P_{\square}^2 = P_{1-}^2 + P_{-1}^2 - P_{1-}^2 P_{-1}^2. \tag{19}$$

The value resulting from the combination of the first and the second condition of selection is 0.84. This figure is still considerably lower than the required 90 per cent efficiency so we employ a further condition of selection.

c. Third condition of selection. This condition permits the following events: From the upper detector group of four on a given cable the lower two respond, the upper two do not; and from the lower group of four, one of the upper detectors operates and the other does not. By taking this arrangement in every possible variation we obtain a third trigger. By a combined calculation of all the three conditions of selection and by taking their arithmetic mean for the lattice, we arrive at an efficiency that is already higher than 90 per cent.

Figure 11 shows how we calculate the above probability values; one needs to know the distance of the given muon path from the individual detectors and the angle between the muon path and the axis of the detector. Knowing R and Θ, we can obtain the probabilities from the data already calculated. For a detailed analysis the detector system can be divided into sub-planes of 15×15 m.

4. Summary

At the present stage of the experiment only two cables have been suspended. The main aim of the basic experiment is to verify the calculations described above. Another such experiment will be to register the Cerenkov light of the proton decays caused by the catalytic effect of magnetic monopoles.

References

1. L. B. Bezrukov, Proc of 11th Int. Conf. of neutrino physics and astrophysics at Nordkirchen near Dortmund, June 11-16, 1984.
2. V.P. Zelov, Izucsenie Vavilova-Cerenkova i ego primenenie v fizike viszokih energij, Atomizdat, Moscow, 1968.
3. E.V. Bugaev, J.A.M. Djilkibaev and N.D. Galperin, Nucl. Instr. and Meth., *A248*, 219, 1968.

Acta Physica Hungarica 64 (1–3), pp. 209–218 (1988)

ANGULAR DISTRIBUTION OF FISSION FRAGMENTS FROM ^{235}U, ^{238}U AND ^{237}Np NEAR THE $(n,2nf)$ THRESHOLD* **

S. OUICHAOUI***, S. JUHÁSZ, M. VÁRNAGY and J. CSIKAI

Institute of Experimental Physics, Kossuth University
4001 Debrecen, Hungary

(Received 8 January 1987)

The angular distribution of fission fragments from the fast neutron induced fission of ^{235}U, ^{238}U and ^{237}Np were measured near the $(n,2nf)$ threshold by polycarbonate nuclear track detector. A function of the form $W(\Theta) = a_0 + a_2 \cos^2 \Theta$ has been fitted to the data measured. The anisotropy parameters determined around 14 MeV show a definite energy and Z^2/A dependence. The differential fission cross sections were also deduced.

Introduction

It is well known [1] that the angular distribution of the fission fragments can provide valuable information about the quantum numbers K belonging to the saddle point deformation of the fissioning nucleus: K being the projection of the total angular momentum J along the nuclear symmetry axis.

Moreover, prior knowledge of the fragment angular distribution is required to determine the fission cross sections, even when a 4π geometry is used for the measurements.

The fission of nuclei such as ^{235}U, ^{238}U and ^{237}Np in the neutron energy range above 12 MeV involves the (n,f), $(n,n'f)$ and $(n,2nf)$ reactions, therefore, the fission fragment angular distributions are superpositions of the first, second and third-chance fission contributions.

It is expected from the theory [2–6] that the anisotropy parameter $W(0°)/W(90°)$ increases abruptly with the neutron energy whenever a multi-chance fission threshold is reached. The experimental results described in the previous literature [7–12] support such an increase in $R(E_n)$ at the (n,f) and $(n,n'f)$ thresholds, $E_f^0 \simeq 1.6$–2 MeV and $E_f^1 \simeq 6$–7 MeV. On the other hand, no noticeable increase in the anisotropy was observed near the $(n,2nf)$ threshold, $E_f^2 \simeq 10.5$–13 MeV, although a great deal of work [7, 10, 13–21] has been carried out in this energy region.

*Dedicated to Prof. G. Marx on his 60th birthday.

**This work was supported in part by the International Atomic Energy Agency, Vienna and the Hungarian Academy of Sciences

***IAEA Fellow. Permanent address: Institut de Physique Nucléaire, Commissariat aux Energies Navelles, B. P. 1017. Alger-gare, Algerie

Experimental procedure

The irradiation of fissile samples was performed by fast neutrons produced in the ^{3}II$(d,n)^4$IIe reaction using a 180 kV neutron generator. The intensity of the analysed D^+ beam was about 200 μA in all experiments. A collimator of 5 mm in diameter was used to determine the beam spot.

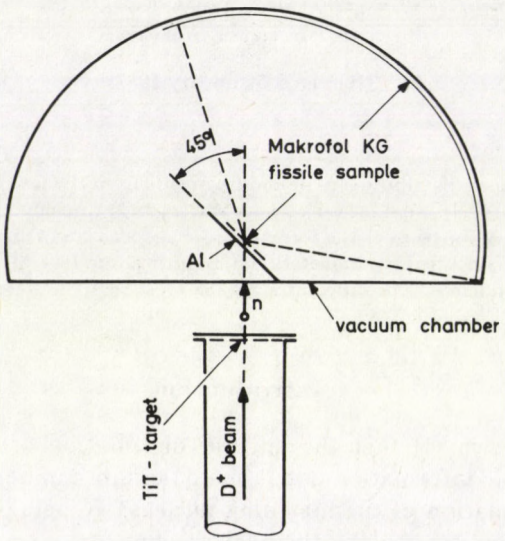


Fig. 1. Experimental set-up for irradiation

A schematic drawing of the experimental arrangement is shown in Fig. 1. Samples of 19 mm diameter deposited onto aluminium backing plates of 0.2 mm thickness and 40 mm diameter were placed in a vacuum chamber together with the detector foils. Two Makrofol KG detector foils of 12μm thickness and 26 mm width were placed at 10 cm from the sample, symmetrically, with respect to the plane of the fission process, to assure the simultaneous recording of fragments at 24 laboratory angles through the same solid angle of about 10^{-3}sr.

The angle of the fissile samples, placed at 5 cm from the neutron source was 45° relative to the neutron beam. The energy of neutrons was changed by the emission angle by placing the samples in different directions to the deuteron beam.

The isotopic composition and a real density of the samples are summarized in Table I. Since the fission cross sections are relatively high for these nuclei, irradiation times of about two hours were enough to obtain suitably high track densities (~1000–1500 tracks/cm^2) in the detector foils.

After irradiation, the detector foils were etched in a 28% KOH solution for one hour at 60°C; then washed in distilled water and dried before counting. The tracks

due to the passage of fission fragments through the detector foils, were evaluated by a jumping spark counter, constructed [22] according to the idea of Cross and Tommasino [23] .

Table I
Isotopic composition and areal density of the fissile samples used in the present experiment

Sample	Isotopic composition			Areal density [$\mu g/cm^2$]
^{235}U	^{234}U	^{235}U	^{236}U	170
	0.0010	99.9955	0.0035	
^{238}U	is depleted by a factor of 230			190
^{237}Np	^{237}Np	^{239}Pu	^{241}Am	148
	99.9933	0.0062	0.0005	

Results and discussion

The absolute experimental differential cross sections and the corresponding fitted curves plotted against the laboratory angles at different neutron energies are shown in Figs. 2–4. The relative angular distributions for all fragments were fitted with a series of Legendre polynomials of the form

$$W(\Theta) = \sum_{n=0}^{N} a_{2n} \cos^{2n} \Theta. \tag{1}$$

The expansion coefficients a_{2n} were obtained by a least-squares fitting procedure. The anisotropy parameters, derived from the fitted angular distributions are given in Table II together with the Z^2/A data of the compound nucleus. The absolute differential cross sections were deduced from the fitted angular distributions by normalization to the integrated cross sections of ^{235}U, ^{238}U, and ^{237}Np:

$$\sigma_f(E_n) = 2\pi \int_0^{\pi} W(E_n, \Theta) \sin \Theta d\Theta. \tag{2}$$

The errors indicated in Figs 2–4 and in Table II include both the uncertainty in the measured data and the goodness of fitting. The $\sigma_f(E_n)$ values were taken from the literature [24] .

Table II
The parameters of the fitted angular distribution functions

Target	E_n[MeV]	a_0	a_2	R	$(Z^2/A)\sigma_f \cdot (10^{-28}\mathrm{m}^2)$ [24]	
^{235}U	14.12	143.99±2.339	63.48±8.062	1.441±0.056	35.86	2.0754
	14.45					2.0932
	14.80	150.52±0.930	50.13±3.118	1.333±0.022		2.1014
^{238}U	14.12	76.63±1.192	40.48±4.254	1.528±0.055	35.41	1.2470
	14.45	79.58±0.900	39.77±3.183	1.499±0.040		1.2388
	14.80	83.63±0.905	35.20±3.120	1.421±0.037		1.2300
^{237}Np	14.12	166.71±1.250	62.50±4.244	1.375±0.025	36.34	2.3556
	14.45	175.20±1.396	48.40±4.589	1.276±0.026		2.4044
	14.80	181.00±1.126	41.91±3.647	1.232±0.020		2.4500

The sources of errors are as follows:

a) The self-absorption of the fission fragments can affect the angular distributions by less than 1%;

b) The uncertainty in the solid angle used can contribute to the anisotropy values less than 1%;

c) The main correction which is applied for the thermal neutron induced fission of ^{235}U. The thermal neutron background was checked by measuring the fission cross section ratio for ^{238}U, ^{235}U as a function of the distance from the neutron source. No significant variation in this ratio has been observed indicating that the contribution of the thermal neutron induced fission to the total one can be neglected. In order to observe the change in the shape and anisotropy of the angular distribution caused by low energy neutrons the vacuum chamber was surrounded by a paraffin shielding. An isotropic angular distribution, which can be attributed to the almost completely thermalized fast neutron beam, was observed as can be seen in Fig. 4. The statistical error of the track density is also indicated in the last part of Fig. 4.

The data in Table II show the expected decrease in the anisotropy by increasing the Z^2/A parameter value. The results obtained are summarized in Table III and compared with the literature data (see Refs. [4, 7, 10, 13, 19, 21, 25–29]). The present results show a systematic decrease in the anisotropy parameter with the increase in neutron energy beyond the $(n, 2nf)$ threshold. No definite trend could be observed in the $R(E_n)$ function in the evaluation of the previous data.

In the vicinity of the $(n, 2nf)$ threshold the angular distribution of fission fragments is interpreted as a superposition of the first, second and third-chance

fission processes (see [18]) which can be expressed as:

$$W(\Theta) = \sum_{i=0}^{i=2} \gamma_i W_i(\Theta) \tag{3}$$

and

$$\gamma_i = \frac{\sigma_{fi}}{\sum_{i=0}^{i=2} \sigma_{fi}}, \tag{4}$$

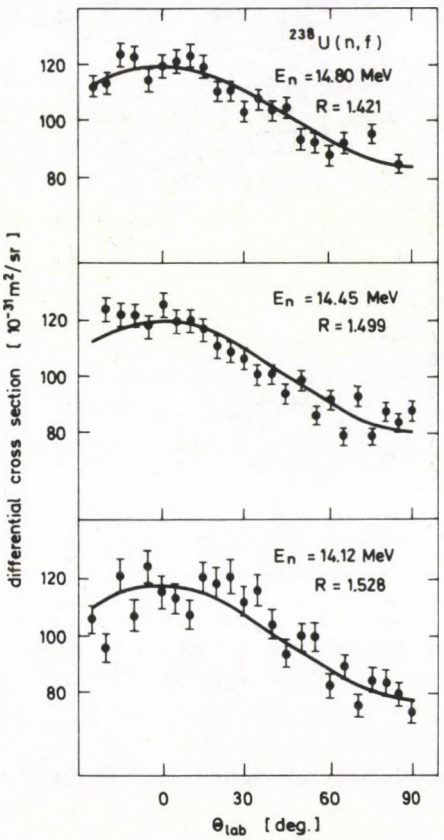


Fig. 2. Measured differential fission cross sections for ^{238}U at around 14 MeV neutron energy

where γ_i is the probability of the contribution to the fission process by the i-th transition nucleus with an angular distribution of $W_i(\Theta)$ and σ_{fi} the corresponding i-th chance fission cross section. Accepting the expression

$$W_i(\Theta) = 1 + A_i \cos^2 \Theta \tag{5}$$

for the approximation of the angular distribution, the total anisotropy can be derived in the form:

$$R(E_n) = 1 + \gamma_0 A_0 + \gamma_1 A_1 + \gamma_2 A_2. \tag{6}$$

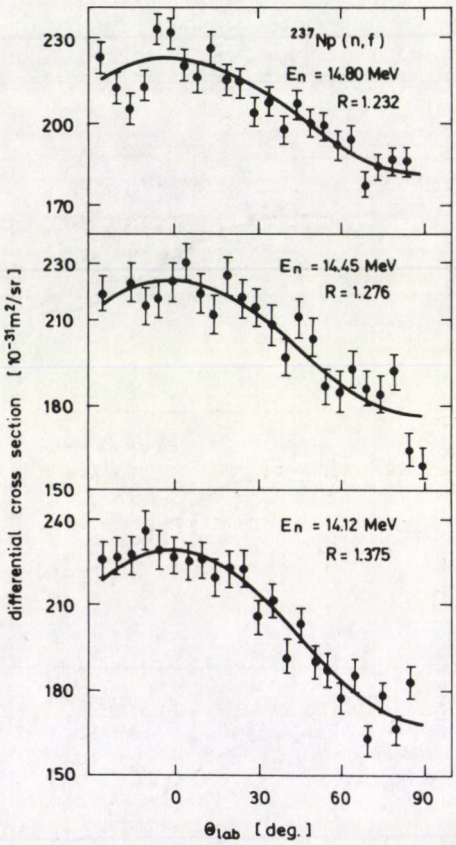


Fig. 3. Measured differential fission cross sections for ^{237}Np at around 14 MeV neutron energy

The calculated values of the anisotropy parameters were obtained by deriving the γ_0, γ_1 and γ_2 probabilities from the avaible fission cross section data [30,31] and accepting the A_0 and A_1 values given in [21] . The $A_2(E_n)$ anisotropy parameter values, which refer to the contribution of the $(n, 2nf)$ reaction, were tentatively determined from the expression [32] :

$$A_2(E_n) = \frac{L_m(L_m+2)}{8K_0^2}, \tag{7}$$

where L_m is the maximum value of the orbital angular momentum of the incident neutron depending only on its energy in the form

$$L_m = (4.2E_n)^{1/2} \tag{8}$$

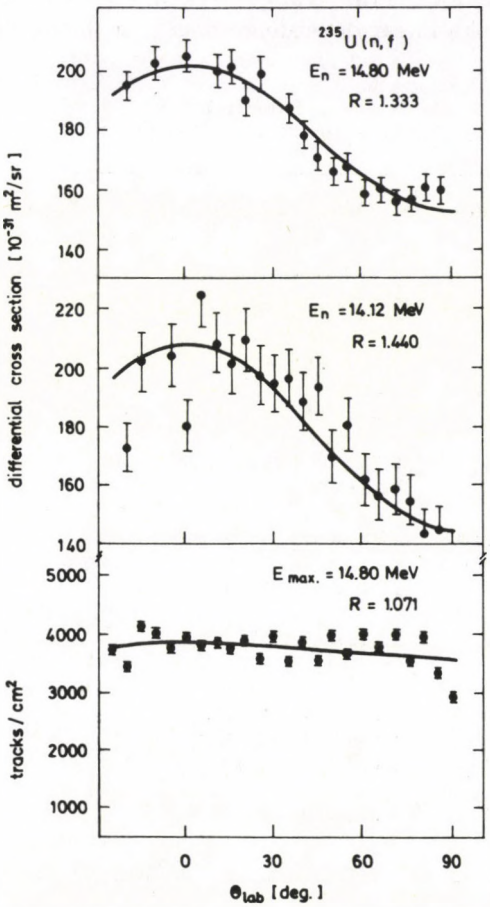


Fig. 4. Measured differential fission cross sections for ^{235}U at around 14 MeV. The track density as a function of emission angle from thermal neutron induced fission on ^{235}U is also indicated

and K_0^2 represents the square of the standard deviation of the Gaussian distribution function that is assumed for the quantum number K.

The calculated values of the anisotropy were obtained in two ways:

a) by estimating the parameter K_0^2 from the experimental anisotropies, using the approximate formula of Griffin [32] :

$$R(E_n) = 1 + \frac{L_m(L_m+2)}{8K_0^2}; \tag{9}$$

b) by deriving K_0^2 from a calculation, using the fitted coefficients given by Simmons and Henkel [8] for the different even–even, even–odd and odd–odd nuclei, under the assumption of a linear dependence of K_0^2 on the excitation energy.

Table III
Experimental anisotropy values compared to the previous data

	E_n[MeV]	R	Reference	Present work E_n[MeV]	R
	14.0	1.27±0.17	12, 25		
	14.0	1.27±0.08	13, 14, 26		
^{235}U	14.0	1.23±0.08	4, 15		
	14.1±0.1	1.27±0.10	1, 10	14.12±0.08	1.441±0.06
	14.8±0.1	1.28±0.07	19, 29	14.80±0.17	1.333±0.02
	14.0	1.31±0.05	13, 14, 26		
	14.0	1.37±0.13	15		
^{238}U	14.1	1.30±0.03	7, 19	14.12±0.08	1.528±0.55
	14.1±0.1	1.31±0.02	4, 10		
	14.5±0.5	1.40±0.14	7, 16	14.45±0.12	1.499±0.040
	14.7	1.43±0.05	7, 19		
				14.80±0.17	1.421±0.037
	14.9	1.25±0.02	17, 27		
	14.0	1.15±0.04	13, 14, 26		
	14.0	1.14±0.04	4, 15		
^{237}Np	14.1±0.1	1.12±0.05	4, 10	14.12	1.375±0.025
				14.45	1.276±0.026
	14.7	1.16±0.02	18, 28		
				14.80	1.232±0.020

In the latter case, the mean excitation energy E_x above the $(n, 2nf)$ threshold was estimated as:

$$E_x = E_n - E_f^2 - W_n, \tag{10}$$

where E_n is the incident neutron energy and W_n is the mean kinetic energy carried off by the evaporated neutrons before fission estimated as $W_n = 2T$, where T is the nuclear temperature.

The two calculated sets of the anisotropy values in comparison with the experimental data are given in Table IV. There is a good agreement between the

Table IV
Comparison of the experimental and calculeted anisotropy values

Nuclide	E[MeV]	Experimental anisotropy	Calculated anisotropy a)	Calculated anisotropy b)	σ_0 (10^{-28}m^2)	σ_1	σ_2 c)	A_0	A_1
^{235}U	14.12±0.08	1.441±0.056	1.296±0.01	1.669±1.06	1.15	0.67	0.55	0.16	0.4
	14.80±0.17	1.333±0.021	1.271±0.02	1.493±0.42					
^{238}U	14.12±0.08	1.528±0.055	1.299±0.015	1.628±1.12	0.56	0.44	0.20	0.20	0.3
	14.45±0.12	1.499±0.040	1.294±0.010	1.550±0.74					
	14.80±0.17	1.421±0.037	1.281±0.007	1.496±0.52					
^{237}Np	14.12±0.08	1.375±0.025	1.271±0.004	1.554±0.85	1.43	0.94	0.26	0.16	0.41
	14.45±0.12	1.276±0.026	1.261±0.002	1.434±1.6					
	14.80±0.17	1.232±0.020	1.256±0.001	1.379±4.1					

a) Calculated R values using the relation [8] to obtain K_0^2.
b) Calculation performed by assuming a linear dependence of K_0^2 on the excitation energy and using the fitted coefficients from reference [8]
c) Nuclear data collection from previous publications

measured and the calculated anisotropies, within the involved uncertainties and approximations. The same trend can also be observed for the calculated and the experimental anisotropies as a function of neutron energy, i. e. the $R(E_n)$ function increases with the neutron energy going down to the $(n, 2nf)$ threshold.

References

1. R. Vandenbosh and J.R. Huizenga, Nuclear Fission, Academic Press, New York, 1973.
2. J.J. Griffin, Phys. Rev., *116*, 107, 1959.
3. A. Bohr, Proc. Int. Conf. Peaceful Uses of Atomic Energy, Geneva, Paper, P/911, 1955.
4. I. Halpern and V.M. Strutinski, Proc. 2nd Int. Conf. Peaceful Uses of Atomic Energy, Geneva, Vol. 15. p. 408, 1958.
5. J.A. Wheeler, in Fast Neutron Physics, Part, II. eds. J.B. Marion and J.F. Fowler, Interscience, New York, 1963, p. 2051.
6. J.R. Huizenga, in Nuclear Structure and Electromagnetic Interactions, ed. N. Mac. Donald, Oliver and Boyd, Edinburgh, 1965, p. 319.
7. R.L. Henkel and J.E. Brolley, Phys. Rev., *103*, 1262, 1956.
8. J.E. Simmons and R.L. Henkel, Phys. Rev., *120*, 198, 1960.
9. V. Emma, S. Lo Nigro, C. Milone and R. Ricamo, Nulc. Phys., *63*, 461, 1965.
10. R.B. Leachman and L. Blumberg, Phys. Rev., *137*, B 814, 1965.
11. Kh.D. Androsenko and G.N. Smirenki, JETP Lett., *8*, 108, 1968.
12. S. Lo Nigro and C. Milone, Nucl. Phys., *A151*, 182, 1970.
13. J.E. Brolley, and W.C. Dickinson, Phys. Rev., *94*, 640, 1954.
14. A.A. Varfolomeev, A.S. Romanceva and V.M. Kutukova, Dokl. Akad. Nauk SSSR, *105*, 1955.
15. A. Katase, Mem. Frac. Eng., Kyushu Univ., *21*, 81, 1961.
16. J. Csikai and S. Nagy, J. Nucl. Energy, *21*, 375, 1967.
17. F. Gonnenwein and E. Pfeiffer, Z. Phys., *207*, 209, 1967.
18. V. Emma, S.Lo Nigro and C. Milone, Nucl. Phys., *A157*, 449, 1970.
19. E. Barutcugil, S. Juhász, M. Várnagy, S. Nagy and J. Csikai, Nucl. Phys., *A173*, 571, 1971.
20. R. Abbeg and R. Wagner, Phys. Rev., *15C*, 1171, 1977.
21. N.K. Chaudhuri, V. Natarajan, R. Sampathkumar, M.L. Sagu and R.H. Iyer, Nucl. Tracks., Vol. *3*, 69, 1979.
22. M. Várnagy, L. Vasváry, E. Gyarmati, S. Juhász, T. Scharbert and T. Sztaricskai, Nucl. Instr. Meth., *141*, 489, 1977.
23. W.G. Cross and L. Tommassino, Rad. Eff., *5*, 85, 1970.
24. ENDF/B-V. File for fission cross sections of ^{235}U, ^{238}U, ^{237}Np, IAEA, Vienna, 1980.
25. J.E. Brolley and W.C. Dickinson, Phys. Rev.,*90*, 388. 1953.
26. J.E. Brolley and W.C. Dickinson, Phys. Rev.,*99*, 159, 1955.
27. A.N. Protopopov, V.P. Eismont and I.A. Baranov, JETP, *9*, 1143 1959.
28. A.N. Protopopov, I.A. Baranov and V.P. Eismont, JETP, *9*, 650. 1959.
29. A.N. Protopopov and V.P. Eismont, JETP, *7*, 173, 1958.
30. V.M. Pankratov, J. Nucl. Energy, *18*, 215, 1964.
31. D.J. Hugues and R.B. Schwartz, Neutron cross sections by BNL-325, 2nd ed. 1958.
32. J.J. Griffin, Phys. Rev.,*127*, 1248, 1962.

Acta Physica Hungarica 64 (1–3), pp. 219–226 (1988)

ENERGETIC GAMMA RAYS IN COINCIDENCE WITH LIGHT FRAGMENTS FROM COLLISIONS OF 35MeV/u ^{14}N IONS*

F.DEÁK, A.KISS

Depeartment of Atomic Physics, Roland Eötvös University
1088 Budapest, Hungary

Z. SERES

Central Research Institute for Physics
1525 Budapest, Hungary

and

G. CASKEY,**A. GALONSKY, B. REMINGTON***

National Superconducting Cyclotron Laboratory and Department of Physics and Astronomy, Michigan State University
East Lansing, Michigan 48824-1321, U.S.A.

(Received 8 January 1987)

Yields of gamma rays with energies above 1.3 MeV have been studied in coincidence with light fragment isotopes from collisions of carbon, nickel and holmium targets with 35 MeV/u ^{14}N ions. The angular distribution of the gamma rays is isotropic. The total number of the coincident gamma rays falls exponentially with fragment angle, but the number of gamma rays per fragment (the multiplicity) is almost independent of fragment angle. For a given fragment isotope the multiplicities with the nickel and holmium targets are equal, but the multiplicity obtained with the carbon target is lower by a factor of from two to six. The hard gamma-ray multiplicities for the nickel and holmium targets tend to be lower when the mass of the coincident light fragment is higher. The results are qualitatively consistent with the assumption that the spectra are dominated by incoherent bremsstrahlung from nucleon–nucleon collision in the nuclear matter heated in the collisions.

Introduction

The investigation of the spectra of energetic (over several MeV) gamma rays from intermediate energy (20 MeV $\leq E/A \leq$ 100 MeV) heavy-ion collisions has attracted much recent interest [1–8]. For these gamma-ray energies the contributions

*Dedicated to Prof.G. Marx on his 60th birthday
**Present address: Donnelly Corp., Holland, Michigan
***Present address: Lawrence Livermore National Laboratory

from electromagnetic transitions ($E \leq 1$ MeV [9]) in the residual nuclei are small. The spectra of energetic gamma rays should be dominated by bremsstrahlung photons.

The majority of such gamma rays could be created during the initial slowing down of the colliding nuclei or from nucleon–nucleon collissions at a subsequent stage [5–7] of the reaction. The former component is coherent, mostly of quadrupole nature, and is expected to have a characteristic angular distribution. It is produced cooperatively by the participant projectile and target nucleons. The second type of contribution is incoherent, therefore isotropic, and it reflects the nucleon–nucleon interactions which take place in the participant zone and in the heated spectators formed in the heavy ion reactions.

As all these gamma rays are emitted prior to, or in competition with particle emissions, and as they represent weakly interacting probes, the investigation of the nuclear electromagnetic bremsstrahlung has been suggested to be a valuable new tool for probing the dynamics of intermediate energy heavy ion collisions [6]. In particular, the study of the gamma-ray yields in coincidence experiments could be of special value. With various triggers we may study heavy-ion reactions of different dynamics. The bremsstrahlung photons could trace the corresponding processes back to early stages of the collisions.

Unfortunately, in-beam experimental work on the emission of energetic gamma rays from violent heavy-ion collisions faces inherent difficulties of the available experimental methods. The complicated and energy dependent response functions of the gamma-ray detectors at elevated energies [10], the determination of efficiencies, and the intense particle backgrounds are some of the problems in the field. Up to now only a few experimental investigations have been published on this topic [1,3,4,7].

The aim of the present paper is to report on experimental results on the yields of energetic gamma rays emitted from collisions of 35 MeV/u ^{14}N ions with carbon, nickel and holmium targets. Our main interest has been devoted to the angular distributions and to the target mass dependence of the relative gamma-ray yields and multiplicities in coincidence with projectile-like fragment (PLF) isotopes. First we briefly describe the experimental method and data acquisition, then the results will be presented. The question of whether the experimental results are consistent with an interpretation that the spectra are dominated by contributions from bremsstrahlung will be discussed.

The experimental method

The experiment was performed at the National Superconducting Cyclotron Laboratory using the 35MeV/u $^{14}N^{5+}$ beam. The targets were natural carbon (3.1 mg/cm^2), nickel (4.6 mg/cm^2) and holmium (8.9 mg/cm^2) foils.

The gamma rays were observed at angles $\pm 10°$, $\pm 30°$, $\pm 70°$, $\pm 110°$ and $\pm 160°$ relative to the beam. The detectors were NE213 scintillator in cylinders of 12.7 cm diameter and 7.62 cm height at distances of 1.6 m to 2.4 m from the target. The

combination of pulse shape discrimination and time-of-flight between the PLf-s and the gamma rays (typical time resolution was $\sim$ 1 ns) proved to be important to simultaneously get rid of the backgrounds of neutrons and delayed gamma rays.

The coincident light fragments were detected by solid state detector telescopes which allowed isotope resolutions for the PLF-s. The telescopes were positioned at $+7°$, $-10°$, $+15°$, $-18°$, $+23°$ in the plane of the gamma detectors and at 15° out of the plane.

The data were recorded event by event and analysed off-line. The details of the experimental set-up have been described in an earlier paper [11].

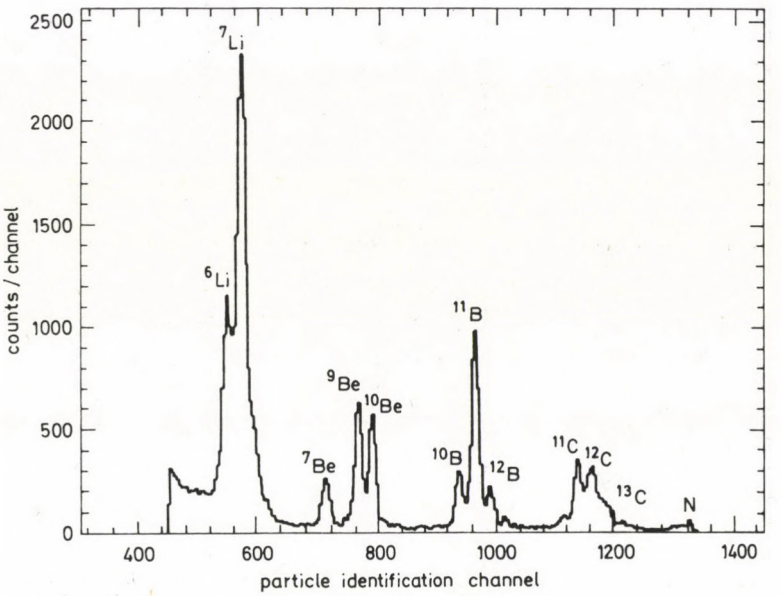


Fig.1. Mass distribution of the PLF-s in coincidence with energetic gammas. Both the PLF telescope and the gamma-ray detector were at 10°

Data acquisition and experimental results

In the data acquisition a pulse-height threshold at the ^{60}Co edge (1.3 MeV) was used. For this level the gamma-ray efficiencies of the detectors were estimated on the basis of the absorption coefficients of the detector material [10]. The efficiencies slowly grow from the threshold to $\sim$ 3 MeV gamma energy and from there they remained practically constant up to $\sim$ 50 MeV or more. The threshold energy for detection of the coincident PLF was 7 MeV/u.

Figure 1 shows a typical mass distribution of the PLF-s at 10° in coincidence with gamma rays also at 10°, for the holmium target. The spectra for each of

the six PLF telescope and ten gamma-ray detector combinations for each of the three targets were divided into lithium, beryllium, boron and carbon PLF element regions. Each region was fitted by Gaussian distributions superimposed on a linear background for each isotope which was seen with significant intensity. For Li the significant isotopes had A=6 and 7, for Be A=7, 9 and 10, for B A=10, 11 and 12 and for C A=11, 12 and 13.

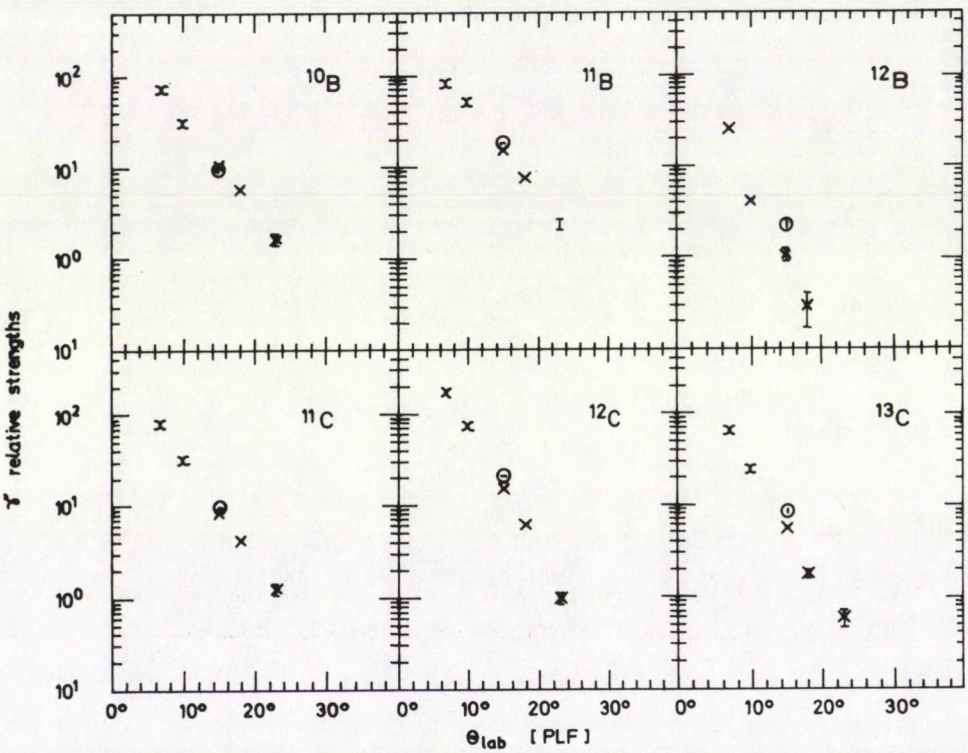


Fig. 2. Relative strength of the hard gamma rays in coincidence with boron and carbon PLF isotopes when the target is carbon. The circles are the data for the out-of-plane PLF telescope

Although our data do not include a determination of energies of the gamma rays, the energy-integrated yields exhibit some interesting properties. Most prominent is the observation that the gamma-ray yields are very different for the various PLF angles and isotopes, but for each particular case the angular distribution of the gamma rays was found to be isotropic up to a precision consistent with the standard deviations ($\sim$ 15%).

The relative yields of the hard gamma rays in coincidence with the PLF isotopes at a given telescope angle were determined from the data summed-up

for all of the gamma detectors The target thickness, the PLF solid angles and the collected charge were taken into account in determining the relative strengths. Figure 2 displays typical results of this procedure for the boron and carbon PLF isotopes when the target is carbon. The relative strengths show nearly exponential fall-offs with PLF telescope angle. The in-plane and out-of-plane 15° strengths were very close to each other for all PLF isotopes.

The multiplicities of the hard gamma rays have been determined from the relative strengths using a calculated average value of 0.07 for the efficiency of the gamma-rays and the cross sections of the PLF singles.

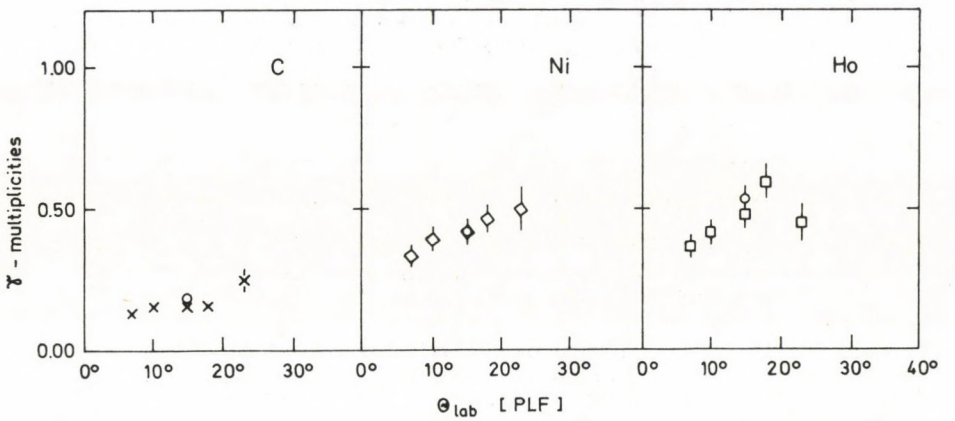


Fig. 3. Multiplicities of the hard gamma rays in coincidence with the ^{11}B PLF-s for the carbon, nickel and holmium targets. The circles are the data for the out-of-plane PLF telescope

Figure 3 displays results for all three targets when the coincident PLF is ^{11}B. These angular distributions are representative of most of the other data. Accordingly, the multiplicities for the different PLF angles for a given target and for the same coincident isotope do not scatter more than $\sim$ 15% around the average value. There is a slight trend that the multiplicities are bigger for bigger PLF angles.

Table I contains the estimated absolute multiplicities for the carbon, nickel and holmium targets in coincidence with the significant PLF isotopes. The data in Table I were calculated by averaging the results for the different PLF angles and assuming isotropic gamma-ray angular distributions. These multiplicities may be somewhat unreliable due to the unknown energy spectra at the lower energies. Assuming that the gamma-ray spectra in the present experiment do not deviate significantly from those in [1], [4], [5]and [7], we estimate that the absolute gamma-ray multiplicities are correct up to a factor of two. Their relative values may have a deviation which corresponds to the average scatter of the different data points in the angular distributions (15%).

The multiplicities in Table I do not show any systematic dependence on PLF isotope in the case of the carbon target. Their average value is about 0.15. The multiplicities for the nickel and holmium targets are considerably (by factors between 2 and 6) greater than for the carbon target, but their values for the same PLF isotope are similar to each other. Furthermore, the multiplicities for the two heavier targets seem to be lower for higher-mass PLF isotopes.

Table I

Multiplicities of energetic gamma rays in coincidence with PLF isotopes for cabon, nickel and holmium targets ($E_{\text{gamma}} > 1.3$ MeV; $E_{\text{PLF}} > 7$ MeV/u)

PLF	Target		
isotope	carbon	nickel	holmium
^{6}Li	0.112±0.030	0.504±0.068	0.665±0.100
^{7}Li	0.097±0.018	0.581±0.029	0.623±0.045
^{7}Be	0.116±0.056	0.389±0.094	0.679±0.047
^{9}Be	0.102±0.014	0.493±0.043	0.579±0.056
^{10}Be	0.276±0.051	0.648±0.078	0.670±0.063
^{10}B	0.123±0.006	0.424±0.030	0.475±0.035
^{11}B	0.169±0.021	0.423±0.028	0.473±0.041
^{12}B	0.204±0.074	0.418±0.076	0.366±0.044
^{11}C	0.191±0.027	0.472±0.043	0.434±0.021
^{12}C	0.140±0.009	0.389±0.062	0.396±0.034
^{13}C	0.140±0.029	0.392±0.089	0.366±0.041

Discussion of the results

The requirement of the detection of a PLF coincident with each hard photon probably means a selective trigger on the reaction mechanism. The shapes of the PLF energy distributions change considerably with PLF detection angle [11, 12]. At small angles (say up to $\sim 15°$) the quasi-elastic, peripheral mechanism dominates, while at larger angles the strongly-damped processes become important.

The fact that the hard gamma rays do not show any angular anisotropy for any of the targets, for any of the PLF isotopes, or for any of the PLF angles is the most striking result of the present experiment. It has been found to be true even

when the PLF-s were detected at 15° out of the plane of the gamma-ray detectors. This, together with the observation that the corresponding data are similar for the 15° in-plane and out-of-plane cases, makes it probable that the yields are isotropic for energetic gamma rays. This may be strong evidence that the origin of the bulk of the hard photons is not the statistical transitions between highly excited states of the rapidly rotating nuclei developed after particle emission in a peripheral collision. It also means that the coherent part of the bremsstrahlung cannot be dominant in the production of the detected gamma-rays for any of the targets. This is consistent with the results of Ko et al [6] who suggested that even for higher-Z projectile and/or target nuclei there should be a large background from the incoherent dipole component of the bremsstrahlung. (We note that the isotropy — with the precision quoted earlier — is consistent with isotropic emission from recoiling source at our gamma-ray threshold and possible source velocities [7].)

The discussed independence of the multiplicities can be interpreted as additional evidence that the hard gamma-rays were emitted by bremsstrahlung processes in an early stage of the heavy-ion collision. This is because the PLF isotopes detected at different angles are produced in reactions of different types [11,12]. It is difficult to imagine that the gamma-ray multiplicities would not depend on them if they were emitted later. Consequently, for those processes in which a PLF is formed, it seems probable to us that most of the energetic photons are results of bremsstrahlung from the hot participant zone and the heated spectators.

This interpretation is consistent with the result that for each PLF isotope the multiplicities for the nickel and holmium targets are similar to each other and considerably higher than for the carbon target (Table I). These observations again could not be understood if the sources of the hard protons were the excited nuclei formed after the particle emissions. For then the highest multiplicities would be expected for the holmium target for which the level densities are the highest. If, however, a hot participant zone is formed by approximately the same number of particle and target nucleons, its size and temperature would be nearly independent of the target mass [11, 12]. In the cases of the nickel and holmium targets the colliding systems contain many nucleons and the average number of the nucleon collisions in both the hot zone and the heated spectators could be about the same. Furthermore, the bigger the mass of the detected PLF isotope, the smaller is the size of the original participant zone [12] and one expects fewer bremsstrahlung gamma rays.

For the carbon target, however, there are only a few nucleons present, and so the number of nucleon – nucleon collisions should be considerably smaller. In addition, here the colliding heavy-ion systems are very nearly symmetric, and the target-like and projectile-like fragments cannot be differentiated experimentally. This could lead to an approximate doubling of the probability of emission of PLF-s with a corresponding reduction of the gamma-ray multiplicities.

The support of both the Hungarian Academy of Sciences and the U.S. National Science Foundation, the latter under Grants INT-80-15333, PHY-83-12245, and PHY-84-16025, is gratefully acknowledged.

References

1. J.J Gaardhøje, O. Andersen, R.M. Diamond, C. Ellegaard, L. Grodzins, B. Herskind, Z. Sujkowski and P.M. Walker, Phys. Lett., *139B*, 273, 1984.
2. E.M. Nyman, Phys. Lett., *136B*, 143, 1984.
3. A. Charvet, R. Béraud, R. Duffait, T. Ollivier, M. Meyer, R. Kossakowski, S. André, D. Barnéoud, C. Foin, J. Genevey, A. Gizon, and J. Tréherne, Z. Phys. A - Atoms and Nuclei, *320*, 605, 1985.
4. Th. Arctaedius, Chr. Bargholtz, R.E. Ekström, K. Fransson, B. Ritzén and P.-E. Tengér, Phys. Lett., *158B*, 205, 1985.
5. H. Nifenecker and J.P. Bondorf, Nucl. Phys., *A442*, 478, 1985.
6. Che Ming Ko, G. Bertsch and J. Aichelin, Phys. Rev., *C 31*, 2324, 1985.
7. J. Stevenson, K.B. Beard, W. Benenson, J. Clayton, E. Kashy, A. Lampis, D.J. Morrissey, M. Samuel, R.J. Smith, C.L. Tam and J.S. Winfield, Phys. Rev. Lett., *57*, 555, 1986.
8. R. Shyam and J. Knoll, Nucl. Phys., *A448*, 322, 1986.
9. M.A. Deleplanque, R.M. Diamond, F.S. Stephens, A.O. Macchiavelli, Th. Døssing, J.E. Draper and E.L. Dines, Nucl. Phys., *A448*, 495, 1986.
10. P.Marmier and E. Sheldon, Physics of Nuclei and Particles, Academic Press, New York and London, 1969, p. 151.
11. B.A. Remington, G. Caskey, A. Galonsky, C.K. Gelbke, L. Heilbronn, J, Heltsley, M.B. Tsang, F. Deák, A. Kiss, Z. Seres, J. Kasagi, J.J. Kolata, Phys. Rev., *C34*, 1685, 1986.
12. F. Deák, A. Kiss, Z. Seres, G. Caskey, A. Galonsky, C.K. Gelbke, B. Remington, M.B. Tsang and J.J. Kolata, Nucl. Phys., *A464*, 133, 1987.

Acta Physica Hungarica 64 (1-3), pp. 227-235 (1988)

SURFACE EFFECTS IN NUCLEAR MONOPOLE DEFORMATIONS*

Cs. SÜKÖSD

Department of Atomic Physics, Roland Eötvös University
1088 Budapest, Hungary

(Received 8 January 1987)

A semiclassical two-parameter collective model is outlined to describe the general features of the nuclear giant monopole resonance. In variance with the generally used scaling model, this approach includes also surface effects. Using energy density functionals and a modified Skyrme interaction for describing the nuclear matter properties, a deformation energy map is generated. The properties of this map indicate that the scaling model is not always appropriate for describing the giant monopole oscillation. This model shows how the excitation energy, the oscillator strength and the shape of the transition density are related through the nuclear matter properties. These quantities were treated independently so far.

Introduction

The existence of the giant monopole resonance is experimentally well established, mostly from inelastic hadron scattering experiments on heavy nuclei [1,2,3,4]. Recently, nearly 100% of the EO EWSR has been found also in some light nuclei in fragmented small resonances [5].

The experimental data were generally analysed by the DWBA method. In view of the high bombarding energies and the small scattering angles involved, the DWBA seems to be appropriate. In the DWBA the characteristics of a specific nuclear excitation is taken into account by the inelastic scattering form factor. Within the framework of the double-folding model this form factor is related to the ρ_{Tr} transition density of the target nucleus. Because of the collective nature of the nuclear giant resonances, usually collective transition densities are used, although some microscopic transition densities are also available from RPA calculations [6].

For describing the monopole excitation the most generally used collective transition density is the Tassie density [7] or Version I. of Satchler [8]. This particular density is closely related to the scaling model [9] of the giant monopole oscillation.

$$\rho_{Tr} = 3\rho_0(r) + r\frac{\partial \rho_0}{\partial r}. \tag{1}$$

There is, however, experimental evidence [4] that this particular model does not give a consistent description of the measured cross-sections, when results of

*Dedicated to Prof. G.Marx on his 60th birthday

inelastic (α, α') and (d, d') scattering experiments are compared. In [4] a consistent description has been achieved only if the transition density had stronger surface components than the Tassie density.

In this paper we discuss different monopole transition densities deduced from the shape of the monopole deformation energy map generated by a modified Skyrme-type effective interaction. We use the collective model and we allow for the independent variation of the radius-parameter and the surface diffuseness of the nucleus as well.

1. Description of monopole deformations

1.1 Collective macroscopic transition densities

In the early days the sharp cut-off model was used, where the nuclear density was taken constant within the radius R of the nucleus, and exactly zero outside. Thus, for describing the oscillation the nuclear radius was the only relevant quantity. This is a crude approximation, since the nuclear surface is known to be diffuse, and the alteration of the surface diffuseness may represent another degree of freedom for the oscillation, which cannot be a priori excluded.

In our model we restrict ourselves to spherical deformation of spherical nuclei, but we allow also for the variation of the surface diffuseness independently from the variation of the nuclear radius. We do not take Coulomb effects into account, and do not distinguish between neutrons and protons (isoscalar excitation).

We suppose that the nuclear density can be written in the general form

$$\rho(r) = n \cdot f(x), \tag{2}$$

where

$$x = (r - R)/a. \tag{3}$$

"R" being the radius parameter, "a" being the surface diffuseness and "n" is related to the central density. At this point we do not need to specify the form of $f(x)$ yet. The three parameters (n, R, a) are relevant for describing the nuclear density. For a given nucleus these are not independent, since the integral of the total density gives the total number of particles (denoted by "A" in the following).

Up to the first order terms any variation $\delta\rho$ of the density can be written as

$$\delta\rho = \frac{\partial\rho}{\partial n} \cdot \delta n + \frac{\partial\rho}{\partial R} \cdot \delta R + \frac{\partial\rho}{\partial a} \cdot \delta a, \tag{4}$$

using (2) and (3) it follows that

$$\frac{\partial\rho}{\partial n} = \frac{\rho}{n}; \qquad \frac{\partial\rho}{\partial R} = -\frac{\partial\rho}{\partial r}; \qquad \frac{\partial\rho}{\partial a} = -\frac{r - R}{a} \cdot \frac{\partial\rho}{\partial r}. \tag{5}$$

Thus the density variation around the equilibrium can be written as

$$\delta\rho = \rho \cdot \frac{\delta n}{n_0} - R_0 \cdot \frac{\partial \rho}{\partial r} \cdot \frac{\delta R}{R_0} - (r - R_0)\frac{\partial \rho}{\partial r} \cdot \frac{\partial a}{a_0}. \tag{6}$$

The mass should be conserved during the deformation:

$$4\pi \int r^2 \delta\rho(r) = 0. \tag{7}$$

Substituting (6) into (7), and taking into account the normalization we get

$$\frac{\delta n}{n_0} \cdot A + 2\left(\frac{\delta R}{R_0} - \frac{\delta a}{a_0}\right) \cdot 4\pi \int_0^\infty \frac{R_0}{t} \cdot \rho \cdot r^2 dr + 3A \cdot \frac{\delta a}{a_0} = 0. \tag{8}$$

Introducing the notation

$$< \frac{R_0}{r} >= \frac{1}{A} \cdot 4\pi \int_0^\infty \frac{R_0}{r} \rho r^2 dr, \tag{9}$$

we write

$$\left(\frac{\delta n}{n_0} + 3\frac{\delta a}{a_0}\right) + 2 < \frac{R_0}{r} > \left(\frac{\delta R}{R_0} - \frac{\delta a}{a_0}\right) = 0. \tag{10}$$

Expressing $\frac{\delta n}{n_0}$, and substituting into (6) we get for the density perturbation

$$-\delta\rho = \rho\left(3\frac{\delta a}{a_0} + 2 < \frac{R_0}{r} > \left(\frac{\delta R}{R_0} - \frac{\delta a}{a_0}\right)\right) + \frac{\partial \rho}{\partial r}\left(R_0 \frac{\delta R}{R_0} + (r - R_0)\frac{\delta a}{a_0}\right). \tag{11}$$

Here ρ and $\frac{\partial \rho}{\partial r}$ both refer to the equilibrium density distribution.

This expression contains all of the collective monopole transition densities used in the literature as special cases.

We get the scaling mode, or Tassie density with $\frac{\delta R}{R_0} = \frac{\delta a}{a_0}$, and the version II of Satchler with $\delta a = 0$ (diffuseness conserving oscillation).

1.2 Exhaustion of the monopole energy weighted sum rule (EWSR)

If an excited state $< f|$ of E_x excitation energy exhausts 100% of the energy weighted sum rule, then

$$E_x \cdot | < f|r^2|0 > |^2 = S(EWSR), \tag{12}$$

where r^2 is the isoscalar monopole operator. This equation can be expressed also using the collective ρ_{Tr} transition density:

$$|4\pi \cdot \int_0^\infty \rho_{Tr}(r) \cdot r^4 dr|^2 = \frac{S}{E_x}. \tag{13}$$

This means that the expectation value of the r^2 operator calculated with the transition density is closely related to the strength of a state relative to the EWSR. The amplitudes in the above introduced transition densities (e.g. $\delta R/R_0$ in Eq. (11)) determine the percentage of the exhaustion of the EWSR of the corresponding state. Or, inversely, if an excited state exhausts 90% or 100% of the EWSR, the amplitude of the corresponding transition density can be calculated using Eq. (13).

2. The monopole deformation energy

In this Section we calculate the $V(R, a)$ potential energy of the monopole oscillation of a spherical nucleus. We restrict ourselves to Fermi-type nuclear densities: $f(x) = 1/(\exp(x) + 1)$. These functions are used most commonly, because they are analytically simple and both the radius and surface diffuseness are taken into account. On the other hand, microscopic self-consistent calulations show that the general behaviour of the density function can be well approximated by Fermi functions, choosing appropriate radius and diffuseness parameters.

For calculating the monopole deformation energy we assume that the actual shape of the nuclear density can be described by a Fermi function, changing only the two free parameters R and "a".

This assumption restricts the basis of the calculation to a relatively small class of functions. However, this basis is much larger than that of the sharp cut-off model, and it is large enough to demonstrate the importance of surface effects in nuclear monopole oscillations. We calculate for quasistatical nuclear deformations and suppose that the nuclear energy is a functional of the matter-density:

$$V = V[\rho]. \tag{14}$$

Since the density depends on two parameters "R" and "a", so does the energy too. This way we get an "energy-map" $V(R, a)$. The minimum of this energy map determines the ground state radius and diffuseness parameters. But this map can be used to determine not only the ground state parameters, but also the energies needed for different monopole deformations characterised by R and "a" (Fig.1).

The main idea behind this procedure is the following. There are energy-functionals published in the literature that describe reasonably well the ground state properties of the nuclei throughout the periodic table. This proves that they contain a piece of general information about the way how energy and nuclear matter density are related. We use this information to calculate the energy needed for a slight deformation of a particular nucleus too. Thus no additional free parameter is introduced.

An additional remark should be made: the energy functionals mentioned above describe reasonably well the observed ground state properties of nuclei when using them in self-consistent calculations [10]. This means that not only the parameters, but also the shapes of the ground state nuclear densities are calculated. The shapes may differ from a Fermi function, so it may seem to be inconsistent to use Fermi functions with these functionals. We show in the Discussion, however, that the ground state parameters determined by the minimum of our energy map are very close to the self consistent (and also to the measured) parameters, indicating that the restriction of the function space to Fermi functions does not cause important effects.

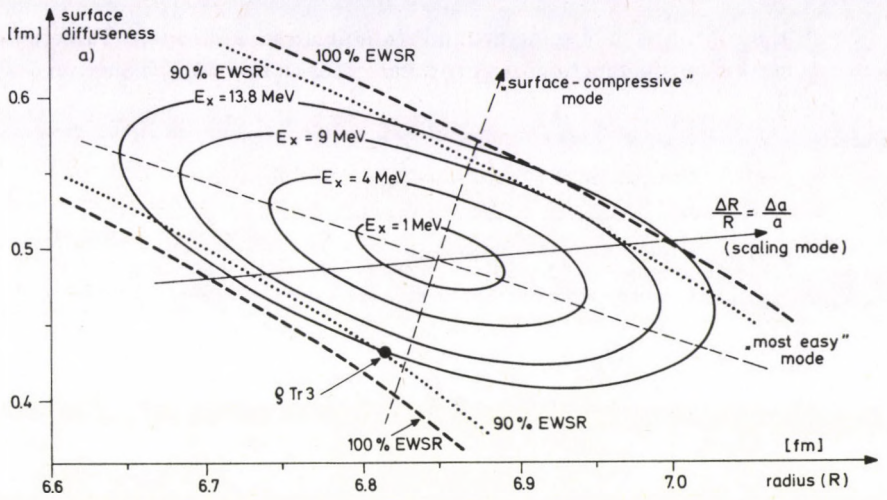


Fig. 1. Deformation energy map of quasistatic monopole deformations of Pb-208, as a funktion of the nuclear radius (R) and diffuseness (a) parameters. The dashed and dotted lines represent transition densities, where the corresponding monopole excitation would exhaust the energy weighted sum rule (EWSR) in the percentage indicated. The point denoted by ρ_{Tr3} corresponds to the empirically determined monopole transition density of [4].

Since our previous experimental results were achieved for the Pb-208 nucleus, these calculations were made also for it. We tried several energy-functionals available in the literature. The main features of the obtained energy-map were similar for all, differing only slightly in numerical details. We present here only one of them where the energy functional was taken from [11].

$$V[\rho] = \int t[\rho] d^3r + \int v[\rho] d^3r \tag{15}$$

the kinetic energy being

$$t[\rho] = \frac{\hbar^2}{2m} \cdot \left(\frac{3}{5} \cdot \left(\frac{3}{2}\pi^2 \right)^{2/3} \cdot \rho^{5/3} + \frac{1}{4}\eta \cdot \frac{(\nabla\rho)^2}{\rho} \right), \tag{16}$$

and a Skyrme-type modified effective interaction was used in form

$$v[\rho] = a_2 \cdot \rho^2 = a_3 \cdot \rho^3 + a_s \cdot \nabla\rho)^2. \tag{17}$$

The values for the different parameters were identical to Set II in [11]:

$$n = 4/9, \qquad a_2 = -408.4(\mathrm{MeVfm}^3), \qquad a_3 = 1079.4(\mathrm{MeVfm}^6), \tag{18}$$

$$a_s = 67.7(\mathrm{MeVfm}^5).$$

3. Discussion of the results

First, the comparison of some ground state properties shown in Table I yields evidence that the Fermi functions approximate reasonably well the ground state density profiles.

Table I
Comparison of ground state parameters for $A=208$

	This work	Ref. [11]
E/A	-11.096 MeV	-11.2 MeV
RMS-radius	5.613 fm	5.6 fm
diffuseness	0.494 fm	0.5 fm
central density	0.147 fm^{-3}	0.155 fm^{-3}

On the energy-map in Fig.1 it is clearly seen that the potential energy is of harmonic oscillator type for not too large deformations. It is also evident that the interplay between the radial and surface deformations causes diferent effective incompressibilities. The direction, where the smallest effective incompressibility occurs is called "most easy mode", and that of the largest incompressibility is called "surface compressive mode". These directions represent also the main axis of the energy surface. It is obvious that these axes do not coincide either with the scaling-mode direction, or with Satchler's Version II, which would be represented as a horizontal line.

In Fig. 2 we show the outer part of the deformed nuclear densities and the corresponding transition densities for these three special deformations.

In Fig. 1 those (R, a) points are also plotted with dotted and dashed lines, where the amplitude of the transition density is so large, that the corresponding excited state would exhaust the monopole EWSR in 90% and in 100%, respectively.

It should be mentioned that this energy map plays the role of a potential energy for the monopole vibration, inspite of the kinetic energy functionals used in the calculation (Eq. (16)). This kinetic energy describes the quantum nuclear motion within a static nuclear matter, and it should be distinguished from the kinetic energy of a collective motion..

In the scaling model of monopole oscillation there is only one dynamic variable for the oscillation: the RMS radius of the nucleus. In our two-parameter model there are two generalised coordinates (and momenta) in the Hamiltonian. Inspecting the potential surface one notices that the most appropriate generalised coordinates would be those combinations of "R" and "a", which would represent the two main axes of the potential surface. Expressing also the kinetic energy operator with their canonical momenta the two parameter oscillation Hamiltonian can be separated and solved. These calculations are in progress and the results will be presented elsewhere.

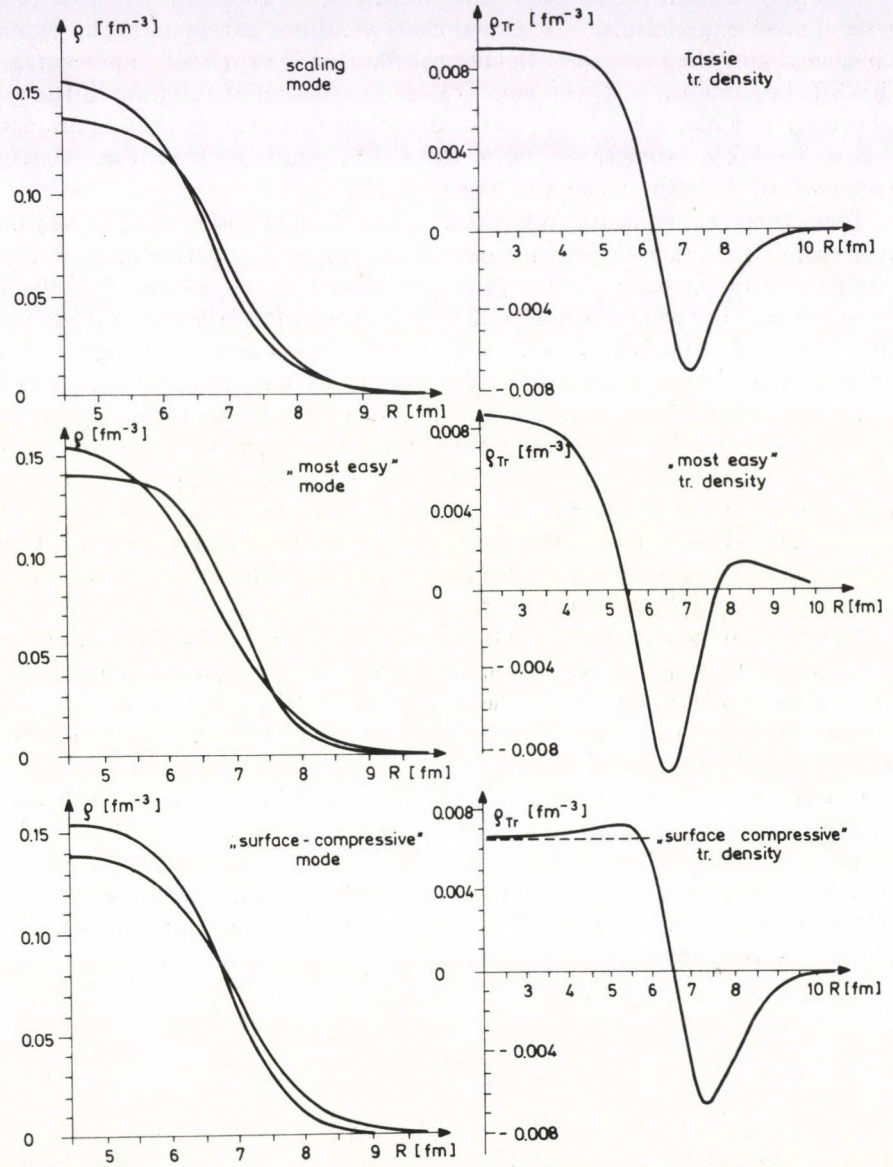


Fig. 2. The outer part of the deformed nuclear densities (left side) and the corresponding transition densities (right side) of three special monopole deformations are shown. The names refer to the deformation modes indicated in Fig.1

Now we would like to discuss an interesting aspect of the comparison of experimental parameters of the giant monopole resonance with the deformation energy surface.

Generally a fixed transition density form (e.g. Tassie form) is used in the analysis of most experiments. The experiments yield two parameters: the E_x excitation energy, and the strength of the resonance, usually expressed in percentage of the EWSR. In some cases [4] the empirical determination of the form of the transition density was also possible. For Pb-208 nucleus these experimental parameters are: E_x=13.8 MeV, strength= 90% of the EWSR, and the transition density is characterised by $\delta R/R$= −0.005 and $\delta a/a$= −0.12 [4].

These three experimental informations are thought to be more or less independent from each other. Now, the point of the empirically determined transition density plotted in the potential energy map (denoted by $\rho_{T_{r3}}$ in Fig. 1.) falls precisely on the crossing point of the E_x=13.8 MeV curve and of the 90% EWSR curve. Additionally, this point is very near to the smaller main axis. Although we think that this very precise agreement is a matter of chance and using other parameters slight deviations might occur, however, this coincidence seems to be very remarkable. That is because for constructing the potential energy map no information has been used about the monopole excitation itself, only the energy functionals were used, which are fitted to describe the systematics of the ground state properties of many nuclei. This coincidence may indicate that the excitation energy, transition strength and transition density are deeply related with each other through the general properties of the nuclear matter.

This fact might also explain, why the monopole excitations are concentrated in heavy nuclei to form an experimentally well detectable giant resonance bump exhausting an essential amount of the monopole EWSR, and why they are more fragmented in light nuclei, where only small resonances can be observed exhausting 10–13% of the EWSR at most [5].

Our results indicate also that the generally used Tassie transition density is not always adequate for the analysis of the experimental data. A better choice would be a transition density closer to the surface compressive mode. If these types of densities would be used also for the investigation of other nuclei, the deduced EWSR strengths would change sligthly and the calculated incompressibility values should also be revised.

References

1. D. H. Youngblood, C. M. Rozsa, J.M. Moss, D. R. Brown and J.D. Bronson, Phys. Rev. Lett., *39*, 1188, 1977;
   D. H. Youngblood, P. Bogucki, J. D. Bronson, U. Garg, Y. M. Lui and C. M. Rozsa, Phys. Rev., *C23*, 1997, 1981;
   J. P. Blaizot, Phys. Rep., *64*, 171, 1980.
2. M. Buenerd, C. Bonhomme, D. Lebrun, P. Martin, J. Chauvin, G. Duhamel, G. Perrin and P. de Saintignon, Phys. Lett., *84B*, 305, 1979;
   M. Buenerd, D. Lebrun, Ph. Martin, P. de Saintignon and C.Perrin, Phys. Rev. Lett., *45*, 1667, 1980;
   M. Buenerd, Proc. of the Int. Symp. on Highly Excited States and Nuclear Structure, Orsay, France, 1983; J. Phys. (Paris) Colloq. 45 (1984) C4-115 and references therein.

3. N. Marty, A. Willis, M. Morlet, R. Frascaria, V. Comparat and P. Kitching, Proc. of the EPS Topical Conference on Large Amplitude Nuclear Collective Motion, ed. A. Kiss, J. Németh, J. Zimányi, MTA-KFKI, Budapest 1979;
F. Zwarts, A. G. Drentje, M. N. Harakeh and A. Van der Woude, Phys. Lett., *125B*, 123, 1983;
F. Zwarts, A. G. Drentje, M. N. Harakeh and A. Van der Woude, Nucl. Phys., *A439*, 117, 1985.
4. H. P. Morsch, C. Sükösd, M. Rogge, P. Turek, H. Machner and C. Mayer-Böricke, Phys. Rev., *C22*, 489, 1980.
5. S. Brandenburg, R. de Leo, A. G. Drentje, M. N. Harakeh, H. Sakai and A. Van der Woude, Phys. Lett., *130B*, 9, 1983;
Y. W. Lui, J. D. Bronson, D. H. Youngblood and Y. Toba, Phys. Rev., *C31*, 1643, 1985;
H. J. Lu, S. Brandenburg, R. de Leo, M. N. Harakeh, T. D. Poelhecken and A. Van der Woude, Phys. Rev., *C33*, 1116, 1986.
6. J. Wambach, V. A. Madsen, G. A. Rinker and J. Speth, Phys. Rev. Lett., *39*, 1443, 1977;
P. Ring and J. Speth, Nucl. Phys., *A235*, 315, 1975;
J. Speth and A. Van der Woude, Rep. Prog. Phys., *44*, 719, 1981.
7. L. J. Tassie, Aust. J. Phys., *9*, 407, 1956.
8. R. Satchler, Particles and Nuclei, *5*, 105, 1973.
9. H. Überall, Electron Scattering form Complex Nuclei, Part B, Academic Press, New York, 1971.
10. H. Krivine and J. Treiner, Phys. Lett., *88B*, 212,1979.
11. G. Eckhart, G. Holzwarth and J. P. da Providencia, Nucl. Phys., *A364*, 1, 1981.

Acta Physica Hungarica 64 (1-3), pp. 237-245 (1988)

VISCOSITY FROM A THREE-COMPONENT FLUID MODEL OF HEAVY ION REACTIONS*

J. NÉMETH, T. CSÖRGŐ

Theoretical Physics Department, Roland Eötvös University
1088 Budapest, Hungary

and

C. NGO

CEN Saclay DPII NMF
91191 Gif sur Yvette Cedex France

(Received 8 January 1987)

A three-component fluid model is introduced for heavy ion reactions to determine the viscosity term and the entropy production of the process. The entropy production turns out to be connected with the modified two-nucleon scattering cross section, while the viscosity coefficient is related only to the densities and to the velocity gradients, as in ordinary hydrodynamics.

1. Introduction

In different energy domains the collision of two heavy nuclei has to be described with different models. At low energies the evolution of the system can be characterized with the time-dependent Hartree–Fock equations, while at high energies among others different cascade calculations can be applied. At the intermediate energy domain neither of these approaches are satisfactory. It seems to us that the best possibility is either to apply a hydrodynamical model, or to determine the phase-space distribution function from some transport equation. In the latter case the calculations are very tedious, and since none of the transport equations are satisfactory enough, it is doubtful whether it is worth while to carry out the long numerical calculations. On the other hand in the hydrodynamical model one needs some further assumptions, which cannot be sufficiently justified. Among those the most doubtful points are the local thermal equilibrium or the expression used for the viscosity and the dissipation term. An additional difficulty is caused by the fact that the nuclear density is varying from zero the central density value, contrary to the ordinary hydrodynamics, where the fluid-density is almost zero.

In order to get some reasonable expression for the nuclear viscosity, a three-component fluid dynamical model is introduced for the description of heavy ion

*Dedicated to Prof. G. Marx on his 60th birthday.

collisions. In this model every single component is assumed to be in thermal equilibrium separately, which seems to be a better justified approximation. In this way the viscosity of the total, non-equilibrium system, as well as the entropy production can be determined without any further approximations.

In Chapter 2 a three-component fluid model in introduced on the basis of a modified Boltzmann-type of transport equation. In Chapter 3 the viscosity term and the entropy production is analysed, while in Chapter 4 the results and some extensions of the model are discussed.

2. The three-component fluid model

Hydrodynamical equations can be deduced starting from a Boltzmann-Vlasov type of transport equation [1]. The phase space distribution function can be determined by

$$Df(\underline{r},\underline{p},t) = \frac{\partial f}{\partial t} + \frac{p_i}{m}\frac{\partial f}{\partial x_i} - \frac{\partial U}{\partial x_i}\frac{\partial f}{\partial p_i} = C^{(+)} - C^{(-)}, \tag{2.1}$$

where $U(\rho)$ is the mean field single particle potential and $C^{(+)}$ and $C^{(-)}$ are the gain and the loss terms. In the simplest case for example the loss term is

$$C^{-}(\underline{p}) = \frac{1}{m}\int d^3p_1 f(\underline{p}) f(\underline{p}_1)|\underline{p} - \underline{p}_1|\sigma(|\underline{p} - \underline{p}_1|). \tag{2.2}$$

$C^{(-)}$ describes the changes of f due to a $\underline{p}+\underline{p}_1 \to \underline{p}' + \underline{p}'_1$ type of two-nucleon collision inside the nucleus. In Eq. (2.2) σ is the cross section of this collision. (It has to be emphasised that it is not the free two-nucleon cross section; it includes the modifying effect of the nuclear background, such as the Pauli principle and the many-body correlations.)

The hydrodynamical equations can by derived from (2.1) by multiplying it with 1, p_i, $\frac{p^2}{2m}$ respectively, and integrate it over d^3p. On the basis of general principles it can be proved that the gain and loss term integrals cancel each other and the three resulting equations express the mass, the momentum and the energy conservation. The first two of these equations take the following forms:

$$\frac{\partial \rho}{\partial t} + \text{div}\,(\rho \underline{u}) = 0, \tag{2.3}$$

$$\frac{\partial(\rho u_i)}{\partial t} + \partial_k(\rho u_1 u_k) + \frac{\rho}{m}\partial_i U + \partial_j P_{ij} = 0, \tag{2.4}$$

where ρ is the density, u_i the velocity and P_{ij} the pressure of the fluid

$$\rho = K\int f(\underline{p})d^3p, \tag{2.5a}$$

$$m\rho u_i = K\int p_i f(p)d^3p, \tag{2.5b}$$

$$P_{ij} = K\int p_i p_j f(p)d^3p - m^2 u_i u_j \rho \tag{2.5c}$$

and

$$K = \frac{4}{(2\pi\hbar)^3}.$$

If f is in local equilibrium, the momentum distribution is symmetric around the u fluid velocity. If slight deviations from the equilibrium are allowed, the pressure is

$$P_{ij} = \frac{2}{3} m\varepsilon\delta_{ij} + m^2 G_{ij}, \tag{2.6}$$

where

$$\varepsilon = K \int p^2 f d^3p - \frac{1}{2} mu^2 \rho \tag{2.5d}$$

is the kinetic energy of the system and G_{ij} is the viscosity term.

If two nuclei collide with high velocity, in the overlapping region a hot, dense nuclear matter drop is formed, whereas a big part of the colliding nuclei is still cold. To assume about this total system that it is either in thermal, or in momentum equilibrium is an oversimplification. Since in the hydrodynamical modes the entropy production (temperature increase) and the viscosity are closely connected, in order to get experimentally observed temperature values, big viscosity terms have to be introduced. In addition a simple hydrodynamical model cannot give answers to some basic questions. For example: in head-on collisions will the original nuclei be completely absorbed in the hot drop or will part of them emerge after the collision as big cold fragments with almost unchanged velocities? It is also an interesting question, how big the maximal central density becomes and what the compressional–heat energy relations are during the collision.

To try to get at least some vague answers to some of these questions, we applied a modified version of three-component fluid model of Csernai et al [2] to describe the nuclear collision [3]. The f distribution function will be divided into three parts:

$$f = f_1 + f_2 + f_3, \tag{2.7}$$

where components 1 and 2 denote the bombarding and target nuclei and component 3 denotes the hot fluid created in the collisions. Any nucleon which collided at least once, will belong to component 3. This means that component 1 and 2 will decrease all the time, and component 3 will increase. The transport equation (2.1) can be divided into three parts

$$Df_1 = -C^{(-)}(f_1, f_2) - C^{(-)}(f_1, f_3), \tag{2.8a}$$

$$Df_2 = -C^{(-)}(f_2, f_1) - C^{(-)}(f_2, f_3), \tag{2.8b}$$

$$\begin{aligned} Df_3 = {} & C^{(+)}(f_1, f_2) + C^{(+)}(f_2, f_1) + C^{(+)}(f_1, f_3) + \\ & + C^{(+)}(f_2, f_3) + C^{(+)}(f_3, f) - C^{(-)}(f_3, f), \end{aligned} \tag{2.8c}$$

where a term like $C^{(-)}(f_1, f_3)$ means that the component 1 is decreasing because two nucleons from component 1 and 3 collide with each other.

In the following we assume that components 1 and 2 are cold nuclei, moving with $\underline{u}_{1,2}$ velocities, while component 3 is a hot nucleus with temperature T and velocity $\underline{u}_3$. Since we know from our earlier calculations [4], that the temperature is very nearly a global quantity in the one-fluid model, we shall assume that the temperature of fluid 3 is space-independent. Since according to the model each of the single fluids are in local equilibrium, viscosity terms are neglected.

The single components can be in equilibrium only if they do not overlap in the phase-space: that means, that the model is applicable only if the fluid veloicities $\underline{u}_1$ and $\underline{u}_2$ are bigger than the Fermi velocities. On the other hand, a non-relativistic approximation can be applied only for not too high velocoties. This means that the model can be used only in the intermediate energy domain.

Taking into account the above considerations the (2.3)–(2.4) hydrodynamical equations for the three components can be written as follows:

$$\frac{\partial \rho_\alpha}{\partial t} + \partial_i(\rho_\alpha u_\alpha^i) = -I_\alpha, \quad \alpha = 1,2 \tag{2.9a}$$

$$\frac{\partial \rho_3}{\partial t} + \partial_i(\rho_3 u_3^i) = I_1 + I_2, \tag{2.9b}$$

$$\frac{\partial(\rho_\alpha u_\alpha^i)}{\partial t} + \partial_k(\rho_\alpha u_\alpha^i u_\alpha^k) + \frac{2}{3m}\partial_i \varepsilon_\alpha + \frac{\rho_\alpha}{m}\partial_i U = -\Gamma_\alpha^i, \tag{2.10a}$$

$$\frac{\partial(\rho_3 u_3^i)}{\partial_t} + \partial_k(\rho_3 u_3^1 u_3^k) + \frac{2}{3m}\partial_i \varepsilon_3 + \frac{\rho_3}{m}\partial_i U = F_1^i + \Gamma_2^i, \tag{2.10b}$$

where I and the Γ are the collision integrals

$$I_1 = I_{12}^{(-)} + I_{13}^{(-)} = \rho_1 \left[\rho_2 < \sigma v_{12} > + \rho_3 < \sigma v_{13} >\right], \tag{2.11a}$$

$$\Gamma_1^i = \Gamma_{12}^i + \Gamma_{13}^i = \rho_1 \left[\rho_2 < \sigma v_{12} u_1^i > + \rho_3 < \sigma v_{13} u_2^i >\right]. \tag{2.12a}$$

Here $v_{12} = |\underline{u}_1 - \underline{u}_2|$ is the relative velocity of two components. In the intermediate energy domain

$$\Gamma_\alpha^1 \sim u_\alpha^i I_\alpha, \quad \alpha = 1,2. \tag{2.13}$$

The Euler type of equations (2.10), using (2.9) and (2.13), take a simpler form

$$\frac{\partial u_\alpha^i}{\partial t} + (u_\alpha^k \partial_k) u_\alpha^i + \frac{1}{m}\partial_i U + \frac{2}{3m}\frac{1}{\rho_\alpha}\partial_i \varepsilon_\alpha = 0, \tag{2.14a}$$

$$\frac{\partial u_3^i}{\partial t} + (u_3^k \partial_k) u_3^i + \frac{1}{m}\partial_i U + \frac{2}{3m}\frac{1}{\rho_3}\partial_i \varepsilon_3 = K_i, \tag{2.14b}$$

where K_i is as follows

$$K_i = \frac{1}{\rho_3}\left[(\underline{u}_1^i - \underline{u}_3^i) I_1 + (\underline{u}_2^i - \underline{u}_3^i) I_2\right]. \tag{2.15}$$

In the case of fluid 1 and 2, applying the

$$\psi_\alpha = \sqrt{\rho_\alpha} e^{iS_\alpha} \tag{2.16}$$

Madelung transformation, Eqs (2.9a)–(2.11a) can be included into a single equation

$$i\hbar \frac{\partial \psi_\alpha}{\partial t} = -\frac{\hbar^2}{2m} \Delta \psi_\alpha + V_\alpha \psi_\alpha + i W_\alpha \psi_\alpha, \tag{2.17}$$

where

$$V_\alpha = U + \frac{\hbar^2}{2m} \left(\frac{3\pi^2}{2} \right)^{2/3} \rho_\alpha^{2/3} + \frac{\hbar^2}{2m} \frac{\Delta \rho_\alpha^{1/2}}{\rho_\alpha^{1/2}}, \qquad \alpha = 1, 2, \tag{2.18a}$$

$$W_\alpha = -\frac{\hbar}{2\rho_\alpha} I_\alpha, \qquad \alpha = 1, 2. \tag{2.18b}$$

In the case of the third component the Madelung transformation can be applied only if rot $\underline{K} = 0$. In this case there exists a $\underline{H}$ function, for which

$$\nabla H = \underline{K}. \tag{2.19}$$

(2.19) is generally not satisfied. For axial symmetric collision however we substitute K in (2.14b) with ∇H, where

$$\frac{\partial H}{\partial z} = K_z. \tag{2.19a}$$

This way we use an approximation; but at least we assure that the total momentum is conserved. Later calculations showed that the comitted error is insignificant.

Substituting (2.19) into (2.14b) the Madelung transformation can be carried out even for the third component. The potentials occurring in Eq. (2.17) are

$$V_3 = U + \frac{\hbar^2}{2m} \frac{\delta \rho_3^{1/2}}{\rho_3^{1/2}} - H + \eta T, \tag{2.20a}$$

$$W_3 = \frac{\hbar}{2\rho_3} (I_1 + I_2), \tag{2.20b}$$

where η in the chemical potential of the hot system, and T is its temperature:

$$\begin{aligned} \rho_3 &= cT^{3/2} F_{1/2}(\eta) \\ \varepsilon_3 &= cT^{5/2} F_{3/2}(\eta). \end{aligned} \tag{2.21}$$

In equations (2.21) c is a constant and $F_{\nu/2}$ is the Fermi integral

$$F_{\nu/2}(\eta) = \int \frac{x^{\nu/2}}{1 + e^{x-\eta}} dx.$$

With the potentials (2.18), (2.20) one has to solve for each component a (2.17) Schrödinger-type of equation with complex potential.

The temperature of the third fluid can be determined from the third fluid-dynamical equation. Since we use, however, a space-independent temperature, it is simpler to determine T directly from the total energy conservation:

$$\frac{dE}{dt} = 0. \tag{2.22}$$

From (2.22) we get a time-dependent differential equation for the temperature

$$A\frac{dT}{dt} + BT - C = 0, \tag{2.23}$$

where the coefficients of Eq. (2.23) are as follows

$$A = \int \rho \left[\frac{5}{2}\frac{f_{3/2}(\eta)}{F_{1/2}(\eta)} - \frac{9}{2}\frac{F_{1/2}(\eta)}{F_{-1/2}(\eta)}\right] dV, \tag{2.24a}$$

$$B = \int \nabla(u_3\rho_3)\left(\eta - 3\frac{F_{1/2}(\eta)}{F_{-1/2}(\eta)}\right) dV, \tag{2.24b}$$

$$C = \int I_1 \left[\frac{1}{2}mv_{13}^2 + \left(\frac{3\pi^2}{2}\rho_1\right)^{2/3}\frac{\hbar^2}{2m}\right] dV + \\ + \int I_2 \left[\frac{1}{2}mv_{23}^2 + \left(\frac{3\pi^2}{2}\rho_2\right)^{2/3}\frac{\hbar^2}{2m}\right] dV. \tag{2.24c}$$

Solving the three (2.17) equations together with (2.23) we can determine all the relevant quantities of the nuclear collision.

3. Entropy production and viscosity

Comparing the results of the three-fluid description with a one-fluid model, we can determine the viscosity term and the dissipation. Substituting (2.6) into Eq. (2.3), the Navier-Stokes equation is

$$\frac{\partial(\rho u_1)}{\partial t} + \partial_k(\rho u_i u_k) + \frac{\rho}{m}\partial_i U + \frac{2}{3m}\partial_i \varepsilon + \partial_j G_{ij} = 0, \tag{3.1}$$

where G_{ij} is the viscosity term. Adding the three (2.10) equations we get a one-fluid Navier-Stokes equation from the three-component model:

$$\frac{\partial(\rho u_i)}{\partial t} + \sum_{\alpha=1}^{3}\partial_k(\rho_\alpha u_i^\alpha u_k^\alpha) + \frac{\rho}{m}\partial_i U + \frac{2}{3m}\sum_{\alpha=1}^{3}\partial_i \varepsilon_\alpha = 0. \tag{3.2}$$

Comparing Eqs (3.1) and (3.2), the G_{ij} viscosity tensor turns out to be

$$G_{ij} = \left(\sum \rho_\alpha u_i^\alpha u_j^\alpha - \rho u_i u_j\right) + \frac{2}{3m}\left(\sum \varepsilon_\alpha - \varepsilon\right). \tag{3.3}$$

The total fluid density, velocity and energy can be expressed in terms of the three-component model densities, velocities and energies

$$\rho = \rho_1 + \rho_2 + \rho_3, \tag{3.4}$$

$$\rho \underline{u} = \rho_1 \underline{u}_1 + \rho_2 \underline{u}_2 + \rho_3 \underline{u}_3, \tag{3.5}$$

$$E = \int \left(\frac{1}{2} m\rho u^2 + c\tau^{5/2} F_{3/2}(\zeta) + W(\rho)\right) dV =$$

$$= \int \left(W(\rho) + \sum_\alpha \left(\varepsilon_\alpha + \frac{1}{2} m\rho_\alpha \rho_\alpha^2\right)\right) dV. \tag{3.6}$$

In Eq. (3.6) τ is the temperature of the total fluid, ζ is its chemical potential and W is the total mean field potential energy density. τ and ζ can be determined from the density and the energy:

$$\rho = c\tau^{3/2} F_{1/2}(\zeta), \tag{3.7a}$$

$$\varepsilon = c\tau^{5/2} F_{3/2}(\zeta), \tag{3.7b}$$

where ε is the kinetic energy of the total system. It can be expressed from Eq. (3.6) as follows

$$\varepsilon = \frac{m}{2}\left[\sum_\alpha \rho_\alpha u_\alpha^2 - \rho u^2\right] + \sum_\alpha \varepsilon_\alpha. \tag{3.8}$$

Substituting (3.8) into (3.3), the viscosity term turns out to be

$$G_{ij} = \sigma_{ij} - \frac{1}{3}\delta_{ij}\sigma_{ll}, \tag{3.9}$$

where σ_{ij} is

$$\sigma_{ij} = \frac{1}{\rho}\left[\rho_1\rho_2 u_{12}^i v_{12}^k + \rho_1\rho_3 v_{13}^i v_{13}^k + \rho_2\rho_3 v_{23}^i v_{23}^k\right]. \tag{3.10}$$

As expected, $\mathrm{Tr}G_{ij} = 0$, that is G_{ij} is a traceless tensor.

In the three-component fluid model the entropy increase is due only to the third component, because the other two nuclei are cold ones. The total entropy and its time derivative can be written as

$$S = \int \left[\frac{5}{3} cF_{3/2}(\eta) T^{3/2} - \eta\rho_3\right] dV, \tag{3.11}$$

$$\frac{\partial S}{\partial t} = \frac{1}{T}\int I_1 \left(\frac{1}{2} m v_{13}^2 + \frac{h^2}{2m}\left(\frac{3\pi^2}{2}\rho_1\right)^{2/3} - \eta T\right) dV +$$

$$+ \frac{1}{T}\int I_2 \left(\frac{1}{2} m v_{23}^2 + \frac{h^2}{2m}\left(\frac{3\pi^2}{2}\rho_2\right)^{2/3} - \eta T\right) dV, \tag{3.12}$$

where we have used Eq. (2.21) for the time derivative of the temperature.

It is interesting to observe that while the viscosity term is only a function of the velocities and densities, the entropy increase is directly proportional to the collision integrals. We do not find the close connection between viscosity and dissipation, as in ordinary hydrodynamics.

In the above calculations we have assumed that the temperature of the third fluid is a constant all over component 3. From equations (3.4)–(3.7) it is easy to see, that the temperature of the total fluid cannot be constant throughout the whole system, that is grad $\tau \neq 0$. This means in ordinary hydrodynamics that the entropy production is mainly caused by the heat current term.

4. Results and discussion

To solve the three-component fluid dynamical problem, an expression is needed for the σ cross section. In the usual Boltzmann equation the system is a weakly interacting gas, there is no mean field term, so in the collision integrals one can use the free two particle collision cross section for the two particle interaction. Nuclei are, on the other hand, strongly interacting systems. This means partly that a mean field interaction has to be introduced as a self-consistent potential (Vlasov equation), and partly that in the collision integral the two-nucleon interaction within the nuclear matter has to be taken into account. From relativistic calculation we know that even in the nonrelativistic limit the nucleon potential itself changes due to the nuclear surrounding. If we take into account the many body effects, this change is even more significant. There are relativistic Bruckner Hartree–Fock calculations, which determine the correct expression to be used in the collision integral [5]. They show that the free-nucleon cross section has to be decreased by a factor of two or even more, depending on the energy. This makes our calculations a bit more difficult, because all our results strongly depend on the collision cross section. One simplification arises, however: the integrals of the loss terms can be expressed with the imaginary part of the optical potential determined in many body calculations by different authors [6]. So in the first calculation we introduced the simplest possible expressions for $< \sigma v >$ (constant or proportional to v) and varied its numerical coefficient. Later we are going to use an expression taken from the many body calculations. (In the one-fluid hydrodynamical calculations the problem does not arise, because the integrals of the gain and loss term cancel each other. In the calculations however, where the distribution function is determined directly by the transport equation, the difficulty of the double counting has to be considered too. In fact the situation is even worse, because in the non-integrated form the complex optical potential cannot be used for the gain terms).

In our calculations we considered $Mg^{24} - Ca^{40}$ and $Mg^{24} - Pb^{208}$ head on collisions, with 200–800 MeV/particle bombarding energy. In this energy domain the applicability conditions of the models are fulfilled. The details of the numerical calculations will be published elsewhere, here we simply collect the important features of the results.

1. The probability of finding a big, fast, relatively cold fragment depends first of all on the numerical value of the cross section, and much less on the bombarding energy and the velocity dependence of σ. If the results of the Brueckner Hartree-Fock calculations are correct, such fragments can be found, with the free two-nucleon cross section without Pauli principle the colliding nuclei melt into the hot component.

2. The velocity of the cold nuclei decreases very little during the collision, so $\underline{u}_1$, and $\underline{u}_2$ in the CM system are strongly forward and backward peaked, while the velocity of the component 3 is at the beginning almost completely perpendicular to the collision direction. As the time evolves $u_{3\|}$ and $u_{3\perp}$ will have roughly the same values, the hot spot is expanding in every direction.

3. The temperature of the third component increases fast up to 20–50 MeV (depending on σ and the collision energy) and after a time it starts to cool down due to the expansion. In our calculation the central nuclear density in the overlap region is only slightly higher than the nuclear matter density, so we did not succeed to produce very dense matter. Further calculations concerning this question are in progress. (It has to be investigated, how the central density depends on the compressibility of the mean field.)

4. In the case of very asymmetric collisions the hot spot is smaller compared to the total mass of the nuclear system, and the incoming nucleus is completely absorbed in it. That means that we cannot expect that the hydrodynamical side-flow angle is the same for symmetric and asymmetric collisions.

5. The high temperatures indicate that the thermal pion creation has to be taken into account.

The above results can be considered only as preliminary ones, calculations including pion creation and more exact two-nucleon collision cross section inside nuclear matter are under way.

References

1. H. Stöcker and W. Greiner, Phys. Rep., *137*, 277, 1986. and see references there.
2. L. Csernai, I. Lovas, J. A. Maruhn, A. Rosenhauer, J. Zimányi and W. Greuer, Phys. Rev., *C 26*, 149, 1982.
3. J. Németh, M. Barranco, C. Ngo and E. Tommasi, Proc. of Hirschegg Workshop (Technische Hochschule Darmstadt), *p. 243*, 1986;
   J. Németh, Acta Phys. Hung., *62*, 365, 1987.
4. J. Németh, M. Barranco, C. Ngo and E. Tommasi, Zeitschr. Phys., *A. 323*, 419, 1986.
5. R. Malfliet, B. ter Haar and W. Botermans, Topical Meeting of Trieste, 1985, World Scientific, Singapore,, *p. 47*, 1985.
6. A. Lejeune, P. Grauge, M. Mertzloff and J. Cugnon, *NPA, 453*, 189, 1986.

Acta Physica Hungarica 64 (1-3), pp. 247-253 (1988)

STRINGENT TESTS OF THE POTENTIAL SEPARABLE EXPANSION METHOD*

B. GYARMATI, A.T. KRUPPA and Z. PAPP

Institute of Nuclear Research of the Hungarian Academy of Sciences
4001 Debrecen, Hungary

(Received 8 January 1987)

It is shown that the approximation to the resonant solution of a Schrödinger equation, based on separable expansions of the asymptotically negligible part of the potential gives not only excellent resonance energies but also well-behaved wave functions even in large distances from the centre of the potential and even if relatively small bases are used.

1. Introduction

Some of the most important quantum mechanical approximation methods are htose using sets of known functions to approximate the unknown solutions of the Schrödinger equation.

There are two versions of these basis set methods that are especially often used. One is based on the representation of the wave function by a linear combination of a number of basis functions and the determination of the coefficients (and the possibile non-linear parameter(s)) from a variational principle. The other is the expansion of the wave function in a (truncated) complete set of basis functions and the diagonalization of the Hamiltonian in the truncated Hilbert space. If the Hamiltonian operator is Hermitean and the basis functions are square integrable, the variational method exhibits an extremely advantageous property, viz. by enlarging the basis the approximate energies approach the exact one monotonically. When the basis size is fixed the wave function expansion (WFE) method can be recognized as a special case of the variational method containing only linear variational parameters. The drawback of both methods is that, if the asymptotics of the basis functions are different from that of the exact one, the farther we need the exact wave function the larger the basis size we have to apply. If e. g. very weakly bound or resonant solutions are calculated it is hardly hopeful to obtain the correct asymptotics with a tolerable number of square integrable functions. Considering that to a number of important transition matrix elements (electromagnetic transitions, nucleon or nuclear cluster transfer, etc.) considerable contributions may come from the outer part of the wave function, a poor representation of that part may lead to serious errors. That is why there is really a high demand for methods yielding asymptotically correct wave functions.

*Dedicated to Prof. G. Marx on his 60th birthday.

In the variational method very good results have been obtained in this respect by using the ad hoc (nonetheless intuitively natural) suggestion that the trial function should contain at least one term possessing the correct asymptotics (Siegert type trial function).

A more systematic way of finding an asymptotically correct solution to the equation

$$H|\psi>=E|\psi> \tag{1}$$

is to handle correctly, from the outset, that part of the equation which influences the asymptotic behaviour. The potential separable expansion method [1]which will be briefly reviewed in Section 2 is such a method. In Section 3 and 4 two new examples will be presented to convince the reader about the exceptionally good characteristics of this technique.

2. The potential separable expansion (PSE) method

Looking for an asymptotically correct approximate solution to Eq. (1) we split $H = H_0 + H_s$ so that H_0 may comprise all terms that cannot be neglected at infinity and we approximate only $H_s \equiv V$. We seek for the solution to (1) as the $N \to \infty$ limit of the solutions to the intermediate problems

$$(H_0 + V_N)|\psi_N>=E_N|\psi_N> \qquad N=0,1,\ldots, \tag{2}$$

where V_N is a projection of V to a $N+1$ dimensional subspace of the Hilbert space. It was proven in [1]that if V is locally square integrable and bounded, as well as H_0 is self-adjoint, and the projector is

$$P_N = \sum_{i=0}^{N} |i><i|, \tag{3}$$

where $|i>$ is an element of an orthonormalized complete set, the procedure is convergent. It turned out that, unfortunately, the limit may be reached after unpleasant oscillations: the procedure is thus convergent but it is not variational. To remedy the oscillatory behaviour an arbitrary smoothing factor σ_i^N was introduced in P_N so as to keep the proof of the convergence valid. We chose σ_i^N as

$$\sigma_i^N = \frac{1 - \exp\left[-(\alpha\frac{i-N-1}{N+1})^2\right]}{1 - \exp\left[-\alpha^2\right]} \qquad i=0,1,\ldots N, \tag{4}$$

where α is a non-linear parameter of the procedure. Then using

$$V_N = \sum_{i,j=0}^{N} |i> V_{ij}^{\sigma} <j|,$$
$$V_{ij}^{\sigma} = \sigma_i^N <i|V|j> \sigma_j^N, \tag{5}$$

Eq.(2) with purely out-going (bound or resonant) boundary conditions can be arranged into the form of a homogeneous Lippmann–Schwinger equation,

$$|\psi_N >= (E - H_0)^{-1} V_N |\psi_N > .$$

Denoting the Green's function $(z-H_0)^{-1}$ belonging to H_0 by $G_0(z)$ and the numbers $< j|\psi_N >$ by c_j^N we have

$$|\psi_N >= \sum_{i,j=0}^{N} c_j V_{ij}^{\sigma} G_0(E_N)|i > . \tag{6}$$

For the coefficients c_j^N, a system of homogeneous algebraic equations derives by multiplying Eq. (6) by $< k|$

$$\sum_{j=0}^{N} \left\{ \delta_{kj} - \sum_{i=0}^{N} < k|G_0(E_N)|i > V_{ij}^{\sigma} \right\} c_j^N = 0, \qquad k = 0, 1, \ldots N. \tag{7}$$

The roots E_N of the transcendent equation

$$D(E) \equiv \det \left\{ \delta_{kj} - \sum_{i=0}^{N} < k|G_0(E)|i > V_{ij}^{\sigma} \right\} = 0 \tag{8}$$

are the approximate eigenvalues, which, together with the solutions c_j^N of Eq. (7) at E_N, determine the approximate eigenvectors through Eq. (6).

It is worth quoting here a result in [2], namely that Eq. (6), which is a composition of $N+1$ non-orthogonal functions $G_0(E_N)|i >$ with correct asymptotic behaviour, can identically be rewritten in a sum of the combination of N square integrable and orthonormal functions $|i >$ and 1 function $G_0(E_N)|0 >$, which bears the correct asymptotics. So the systematic PSE procedure helps to understand the success of the Siegert type trial functions in the framework of the variational method.

In what follows we will be concerned with complex E_n eigenvalues, more precisely with those defined by the relations

$$\sqrt{2mE_N/\hbar^2} \equiv k_N = \kappa_N - i\gamma_N, \qquad \kappa_N > \gamma_N > 0.$$

The related states are called Gamow or Siegert states. These states do not belong to the Hilbert space. In coordinate representation they are complex and not square integrable, and their real and imaginary parts oscillate between exponentially diverging envelopes. So an inner product containing Gamow state(s) does not exist in the orthodox sense. Nevertheless there are various, essentially equivalent generalizations of the concept of the inner product that make sense if one or both of

the factors are Gamow states and boil down to the usual definition if the factors are bound states. We chose to work with the rules elaborated by T. Berggren and reviewed in [3]. All inner products in Eqs (6)–(8) are to be regarded as generalized ones. The Gamow states are of great conceptual and practical importance in the theory of resonances, and, because of their violent asymptotic behaviour, they provide an ideal testing ground of a method that aims at describing the wave function correctly also in the asymptotic region.

Table I

The energy E of the $S_{1/2}$ resonance in the potential $7.5\ r^2e^{-r}$ as a function of the basis size *

12	3.42834-i 0.0126291
13	3.42741-i 0.0131535
14	3.42701-i 0.0126980
15	3.42674-i 0.0127830
16	3.42674-i 0.0128572
17	3.42667-i 0.0127485
18	3.42663-i 0.0128073
19	3.42652-i 0.0127838
20	3.42652-i 0.0127681
21	3.42648-i 0.0127948
22	3.42645-i 0.0127711
23	3.42642-i 0.0127786
24	3.42642-i 0.0127794
25	3.42642-i 0.0127721
26	3.42641-i 0.0127799
27	3.42640-i 0.0127737
28	3.42640-i 0.0127753
29	3.42640-i 0.0127761

*$b = 1.3$, $\alpha = 6$

3. The $S_{1/2}$ resonance in the $7.5\ r^2e^{-r}$ potential

In this paragraph all quantities are given in atomic units. The potential in the section heading is a famous standard in atomic physics. The best variational value of its $S_{1/2}$ resonance is 3.42639-i 0.0127745 as it was published in 1974 [4]. The "exact" values, i. e. the results of the direct numerical integration given all over this paper were generated by the code GAMOW [5].

As this potential is short-ranged the kinetic energy T plays the role of H_0. It is thus convenient to use for the basis $\{|i>\}$ the harmonic oscillator wave functions (HOWF), since the matrix element of the free Green operator between them can

be computed exactly with a recursion relation. The algorithms and the program of the HOWF-based PSE method called PSEUDO have been published in [6]. The numerical results are based on this program, except that the subroutine FC01A computing $\exp[-z^2]$ Erfc (iz) has been substituted by the extremely precise routine elaborated by A.D. Isaacson. The convergence of the method is displayed in Table I. The b stands for the size parameter of the HOWF basis, and α is the parameter of the smoothing factor (4). It has been checked that the results are nearly independent of their values over a reasonably broad range. Fig. 1 shows the complete agreement between the exact and PSE radial wave functions.

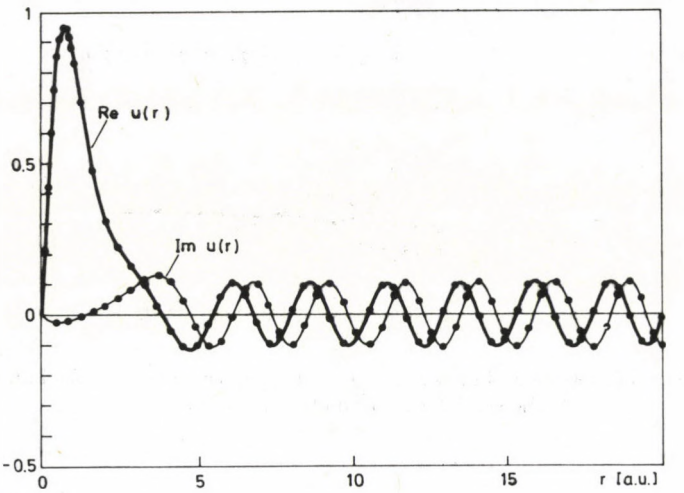


Fig. 1. Real (imaginary) parts of the radial wave function. The continuous lines represent the results of the direct numerical integration, the dots are the PSE values

4. A broad resonance in a Saxon–Woods plus hard-sphere Coulomb potential

In our second example we intend to show a most conspicuous example of a broad $E=13.63 - i\ 3.28$ (MeV) proton resonance in a potential that schematically represents the ^{208}Pb nucleus

$$V = V_0 f(r) + V_{SO}\left(\frac{\hbar^2}{m_\pi c}\right)^2 \frac{1}{r}\frac{df(r)}{dr} \mathbf{l}\sigma + V_c, \tag{9}$$

$$f(r) = -\{1 - \exp[(r - R_0)/d]\}^{-1}, \tag{10}$$

$$V_c = \begin{cases} Z_1 Z_2 (e^2/2R_c)(3 - r^2/R_c^2), & r < R_c \\ Z_1 Z_2 e^2/r, & r > R_c, \end{cases} \tag{11}$$

with the parameter values $V_0 = 50.9$ MeV, $V_{SO} = 5.8$ MeV, $R_0 = R_c = 7.06$ fm, $d = 0.75$ fm, $Z_1 = 82$, $Z_2 = 1$ ($\mathbf{l}, \sigma$ are the orbital and spin angular momentum, m_π is the pion mass, c is the velocity of light, e is the electron charge, $\hbar$ is the Planck constant divided by 2π).

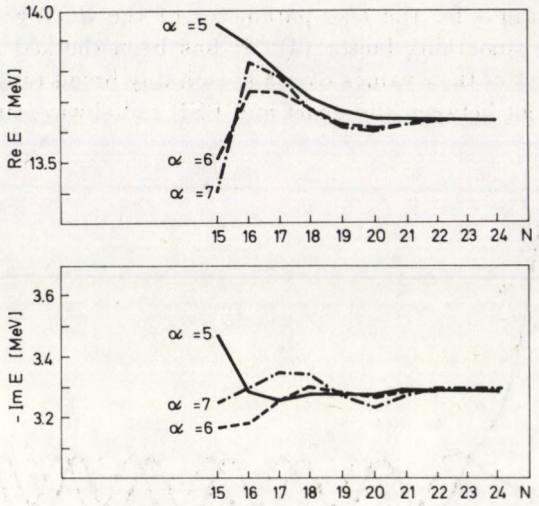


Fig. 2. Covergence of the real and imaginary parts of the eigenenergies at the indicated values of the smoothing parameter α, $b = 2.7$ fm

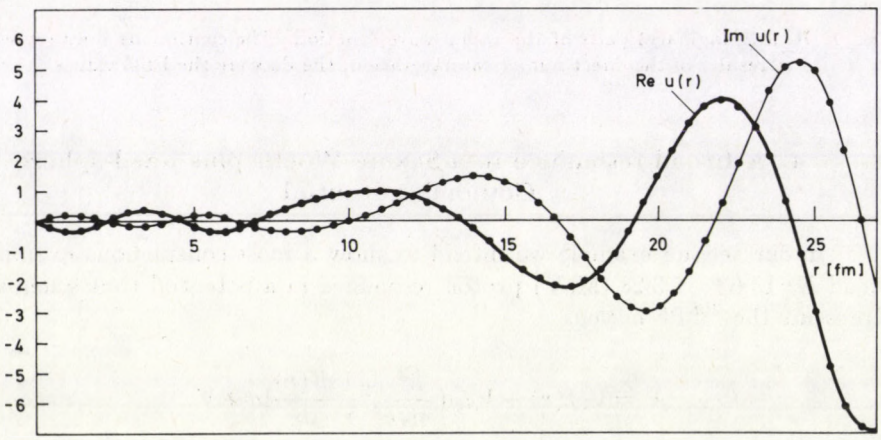


Fig. 3. Real and imaginary parts of the resonance wave function, $\alpha = 6$, b=2.7 fm. N=24. Notation as in Fig 1

In this case $H_0 \equiv H_c = T + Z_1 Z_2 e^2/r$, and the matrix elements of the Coulomb–Green operator G_c belonging to H_c are to be correctly computed on the basis used for the expansion of the rest of the Hamiltonian. In [7] a PSE algorithm is presented based on the Coulomb–Sturmian functions

$$< r|nl >= \left(\frac{n!}{(n+2l+1)!}\right)^{1/2} (2br)^{l+1} e^{-br} L_n^{2l+1}(2br), \qquad (12)$$

here L_n^{2l+1} are the associated Laguerre polynomials, and b is a scaling factor. In Fig. 2 the convergence of the eigenenergies is shown at the indicated values of the smoothing parameter α, the value of the scale parameter of the basis is b=2.7 fm but around this value the results are almost independent of b. In Fig 3 the spectacular performance of the method in reproducing the wave function can be observed.

Acknowledgement

The authors are indebted to J. Németh, K. Ladányi and P. Hraskó whose questions inspired the above work. Thanks are due to A.D. Isaacson for having supplied us with his excellent computer code yielding the complex error function.

References

1. J. Revai J.I.N.R. E4-9429, Dubna, 1975;
   B. Gyarmati, A.T. Kruppa and J. Révai, Nucl. Phys., *A326*, 119, 1979.
2. K.F. Pál, J. Phys., *A18*, 1665, 1985.
3. B. Gyarmati, A.T. Kruppa and Z. Papp, Phys. Rev., *C31*, 2317, 1985.
4. R.A. Bain, J.N. Bardsley, B.R. Junker and C.V. Sukumar, J. Phys., *B7*, 2189, 1974.
5. T. Vertese. K.F. Pál and Z. Balogh, Comput. Phys. Commun., *27*, 309, 1982.
6. A.T. Kruppa and Z. Papp, Comput. Phys. Commun., *36*, 59, 1985.
7. Z. Papp J. Phys., *A20*, 153, 1987.

Acta Physica Hungarica 64 (1–3), pp. 255–265 (1988)

SHAPE ANALYSIS OF THE "CUSP" IN THE SPECTRUM OF ELECTRONS EJECTED INTO FORWARD DIRECTION FROM ION–ATOM COLLISIONS*

D. BERÉNYI, L. GULYÁS, Á. KÖVÉR and GY. SZABÓ

Institute of Nuclear Research of the Hungarian Academy of Sciences
4001 Debrecen, Hungary

(Received 8 January 1987)

The energy dependence of the parameters in the series expansion of the cusp-shape was studied by using the experimental data of the present authors and all the other published data. The recent theoretical interpretations are also discussed.

1. Introduction

It was in 1970 that Crocks and Rudd observed for the first time a very intensive sharp peak (a so called "cusp") in the energy spectrum of electrons ejected in the forward direction from H^+–He collisions [1]. In the same year a similar cusp was observed in the case of solid target (carbon and gold foils) [2]. The position of the cusp in the energy spectrum of forward ejected electrons corresponds to a value of the velocity of the ejected electron which is equal to that of the emergent ion. The origin of the electrons in the cusp can be either the transfer (capture) of a target electron to a low-lying projectile continuum state (ECC, Electron Capture to the Continuum) or the excitation of a projectile electron to a low-lying continuum state (ELC, Electron Loss to the Continuum) if the projectile carries electron(s).

During the last more than one decade a number of experimental and theoretical studies were carried out on this phenomenon (see e.g. the surveys in [3–6] and in 1984 a whole symposium was devoted to this topic in Aarhus [7]). The production cross section, the position, the width, the shape of the cusp were studied as a function of impact energy, the type and charge state of the projectile as well as the emission angle of the electron in detail, both experimentally and theoretically.

Contrary to the very many investigations, there are problems and unelucidated items on practically all the features of the cusp, namely cross section, width and shape. In connection with the shape it has become clear during the studies in the last years that the cusp is nearly symmetric only in the ELC processes (for projectiles carrying electron(s)) but it is skewed towards lower velocities for bare projectiles (ECC process) for gaseous targets (see e.g. in [8]). In cases of solid targets (foils) the symmetry and the shape of the cusp is rather similar to those for

*Dedicated to Prof. G. Marx on his 60th birthday.

ELC cusp (see e.g. [9][10]). The asymmetric character of the ECC cusp, however, cannot be completely explained theoretically up till now (see e.g. in [8][11–12]).

Several years ago, a series expansion and fitting procedure was introduced for the better characterization of the experimental cusp-shape [13][14]. This multipole expansion method is independent of any specific theory and of the actual apparatus. Recently a number of such studies were carried out at light [8][14–16] and heavy ion [17–21] impact even if those with foil targets [13][22][23] are not taken into consideration. In this way there is now a good opportunity to compare them with each other and there is also a possibility for looking at the energy dependence of the coefficients in the series expansion for gaseous targets (where single-collision conditions are maintained) by using all the published values. That is our goal in the present paper.

In this analysis and consideration of the data, however, mainly the cases [8][14][16–18][20–21] (ECC process) are to be studied in which the projectile was bare ion.

2. The series expansion procedure and experimental conditions

The cross section of forward ejected electrons can be described both for ELC and ECC in the following form:

$$\frac{d\sigma}{d\mathbf{v}_e} = \frac{1}{|\mathbf{v}_e - \mathbf{v}_p|} F(v_e, v_p, \cos\Theta'), \tag{1}$$

where $\mathbf{v}_e, \mathbf{v}_p$ is the velocity of the electron and the ion in the laboratory frame, respectively, while Θ is the angle between $\mathbf{v}_e$ and $\mathbf{v}_p$ in the laboratory frame and Θ' is the same angle in the projectile frame (see in Fig. 1).

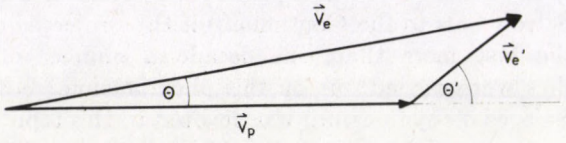


Fig. 1. Velocity vector diagram of the collision process

The observed asymmetry should be associated with the finite $F(v_e, v_p, \cos\Theta')$ which can be expanded by the help of the Legendre polynoms:

$$F(v_e, v_p \cos\Theta') = \sum_{n,l=0} a_l(v_e') P_l(\cos\Theta'), \tag{2}$$

where $\mathbf{v}_e' = \mathbf{v}_e - \mathbf{v}_p . a_l(v_e')$ can be expanded in Taylor series while $\mathbf{v}_e'$ is always small around the cusp:

$$\frac{d\sigma}{d\mathbf{v}_e} = \frac{1}{v_e'} \sum_{n,l=0}^{\infty} B_{nl}(v_p) v_e'^n P_l(\cos\Theta'), \tag{3}$$

where B_{nl} are the Taylor coefficients of the series a_l.

The measured distribution $Q(v_e, \Omega_e)$ of ejected electrons is a convolution of the double differential cross section $(d\sigma/d\mathbf{v}_e)$ in laboratory frame and the electron spectrometer transition function $(S(v_e, \Omega_e))$ where Ω_e is the solid angle of the spectrometer:

$$Q(v_e, \Omega_e) = \int_{v_e} \int_{\Omega_e} \frac{d^2\sigma}{dv_e d\Omega_e} S(v_e, \Omega_e) dv_e d\Omega_e. \tag{4}$$

Substituting Eq. (3) to Eq. (4) we get:

$$Q(v_e, \Omega_e) = \sum_{n,l=0}^{\infty} B_{nl}(v_p) U_{nl} = B_{00} \sum_{n,l=0}^{\infty} \frac{B_{nl}}{B_{00}} U_{nl}, \tag{5}$$

where the form of U_{nl} is in velocity representation:

$$U_{nl} = \int_{v_e} \int_{\Omega_e} v_e^2 (v_e')^{n-1} P_l(\cos\Theta) S(v_e, \Omega_e) dv_e d\Omega_e. \tag{6}$$

The experimental investigation of electrons ejected into the direction of the projectile beam is rather difficult. The projectile beam must be well collimated and cleaned from secondary electrons. It should pass through the electron spectrometer without touching any solid part of it avoiding additional secondary electron production. The cusp electrons are sharply peaked into the beam direction and therefore narrow acceptance solid angle of the electron analyzer is required. The main parameters of the experimental setup used by the different authors can be seen in Table I. Mainly electrostatic analyzers are used for recording the electron spectra.

As it can be seen above, the $S(v_e, \Omega_e)$ expresses the spectrometer transmission function. Meckbach et al [14] showed that for high velocity resolution and small acceptance angle the transmission function is separable:

$$S(v_e, \Omega_e) = R(v_e) G(\Omega_e).$$

The $R(v_e)$ can be well approximated by a triangular or trapezoidal line shape. $G(\Omega_e)$ is calculated from the geometrical conditions of the analyzer [20].

The $Q(v_e, \Omega_e)$ electron distribution should be corrected for the efficiency of the analyzer (mainly for the efficiency of the channel electron detector) before the fitting procedure. For the peak shape analysis only the relative change of the efficiency as a function of the impact electron energy is interesting. In general, the detector

Table I
Comparison of the experimental conditions used for shape analysis

	Projectile beam				Electron spectrometer			
Authors	Type	Velocity int. [a.u.]	Beam divergency	Target	Type	Relative energy resolution	Acceptance half angle(Θ_0)	Correction for efficiency
Dahl [24] (Rødbro & Andersen [25])	H^+	1-3.2		He	electrostatic parallel plate	1%	0.38°	+200V
Meckbach et al [14]	H^+, H°	1.25-3.1	< 0.1°	He	electrostatic cylindrical mirror	0.1-0.5%	0.17°-2.5°	+100 V
Kövér et al [26] [8]	H^+, H^{2+}	2.5-4	0.5°	He	electrostatic cylindrical mirror	0.3%	1.0°	measured efficiency curve
Andersen et al [16][18] Knudsen et al [21]	H^+,H^{2+} C^{6+} O^{8+},$C1^{11+}$ Au^{11+}	2-10		He	electrostatic parallel plate 30°	0.6%	3.4°	constant impact energy on CEM 300 eV
Berry et al [17][20]	C^{6+},O^{8+}	6.3-10	< 0.1°	H,H_2,He	electrostatic spherical mirror 160°	1.1-1.3%	1.8°-2.0°	no
	O^{8+},Ne^{10+} Ar^{18+}	15.0-18.1	< 0.1°		magnetic sector 90°	0.7-1.7%	1.4°-1.5°	no

efficiency increases from 0 eV to 200–300 eV where it reaches its maximum and than slowly decreases. In order to avoid the efficiency correction some authors, when measuring low energy electrons, accelerate those with a fixed potential field [14][24]. Andersen et al [16] and Knudsen et al [18] accelerate or deaccelerate the electrons in front of the channel detector to assure constant impact energy (300 eV). We determined the ejected electron yield from H^+–He collision and compared with the absolute double differential cross section found in the literature for the determination of the efficiency curve [26]. The advantage of our procedure is that we determined not only the efficiency function of the detector but the efficiency of the whole analyzer including the detector, too. Berry et al [17][20] did not use any correction but they measured high energy electrons ($E_e > 500$ eV) where the change of the efficiency of the system in the measuring range is negligible.

Vager et al [27] suggested that the significant fraction of the cusp yield is ejected from the analyzer field-ionized Rydberg states. However, Berry et al [20] indicated that it is insignificant in the commonly used experimental arrangements.

3. Discussion of the experimental results

The leading parameter of the series expansion is the B_{00} value which is related to the total ECC cross section (see Eq. 5). The other parameters are connected with the asymmetry of the cusp and show the dynamics (the transfer velocity distribution in the projectile frame) of the transfer process. The contribution of the different parameters can be seen in Fig. 2.

Recently Andersen et al [16] and Knudsen et al [21] analyzed the energy dependence of the B_{00} parameter. For light projectiles and high impact energy they found that B_{00} is proportional to $Z_p^{2.9\pm0.1} V_p^{-11.3\pm0.2}$ which agrees rather well with the perturbative second order Born theories [16]. For heavy projectiles, however, substantial deviations were found from this Z_p^3 prediction where the projectile velocity is comparable to the orbital velocity [16]. A similar discrepancy was observed for the velocity dependence, too [16].

These results are in contradiction with the Z dependence of the ECC cross section measured by Breinig et al [3] and Kövér et al [26], who found for heavy and light projectiles that the exponent of Z_p in the cross section function is 2.3 ± 0.3 and 2.5 ± 0.3, respectively. This discrepancy among the experimental values may be caused by the fact that they determined the total yield of the cusp and not the B_{00} parameter.

The next parameter is the B_{10} which describes a continuous "background" contribution to the measured cusp, therefore its effect to the overall shape is minor, but helps to find the best fit. Andersen et al [16] neglected this parameter in their fitting procedure.

As it can be seen in Fig. 2 the next parameters are connected with the asymmetry of the cusp. The B_{01} parameter is responsible for the main contribution to the

asymmetry. Sometimes it is called dipol parameter. Fig. 3 shows the impact energy and projectile charge dependence of the B_{01}/B_{00} measured by different authors.

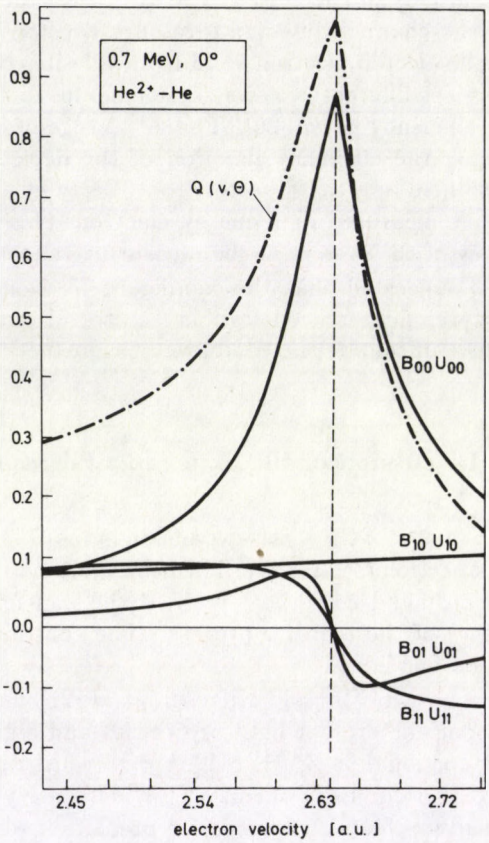


Fig. 2. Contribution of the different parameters to the shape of the cusp [35]

Disregarding the data of Andersen et al [16] and Dahl [24] measured for H^+ projectile, most of the measured values agree fairly well with each other in a certain interval indicating a slight increase towards the higher impact velocities. However, the spread of the data is rather high, therefore it is difficult to find any tendency of the values as a function of the impact energy and charge. The sign of B_{01} is negative, showing that the cusp skewed towards the lower electron velocities.

For light projectiles the data of Andersen et al [16] deviate from the data of Meckbach et al [14] and Gulyás et al [8] which agree rather well with each other and with the data measured for heavy ions. This discrepancy may be caused by the fact that Andersen et al [16] used only two parameters (B_{00} and B_{01}) in the fitting procedure and they found that these parameters are enough for the best fit

assuming that in the small interval around the top of the cusp where $v_e' << 1$ the leading term is the B_{01}. The data of Dahl (showed in Fig. 3) were published by Andersen et al [16]. In the original work of Dahl [24] only the backward-to-forward ratio of emitted electrons in the projectile frame can be found. From these data the value of B_{01} can be determined when neglecting all other parameters.

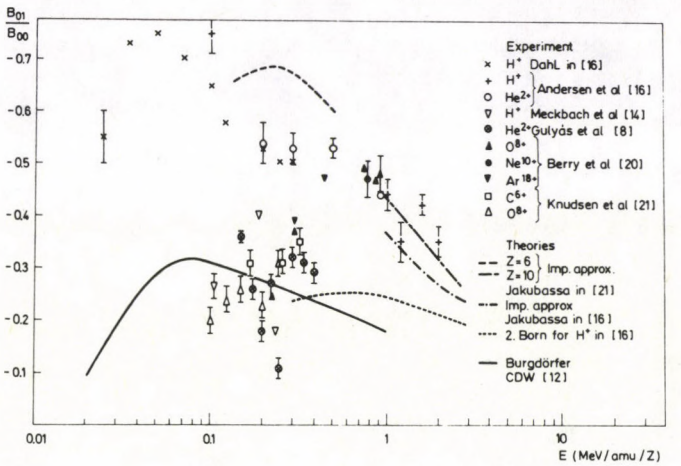


Fig. 3. Comparison of the experimental and theoretical B_{01}/B_{00} parameters determined by different authors

Gulyás et al [8] examined the effect of the length of the interval around the top of the peak where the fitting procedure was performed and they found that the values of the parameters were practically constant only in the interval larger than $(1 \pm 0.05)v_p$. For a smaller region where the assumption of Andersen et al [16] is valid (i.e. two parameters are enough for the fitting) some changes could be observed. A similar conclusion has been drawn by Berry et al [20]. They found that higher order terms are important in the wings of the cusp where $v_e' << 1$ is less valid.

Gulyás et al [8] also investigated the value of the reduced χ^2 by using different parameters for fitting. It was found that four parameters are necessary for the best fit at light projectiles. By using two parameters (B_{00} and B_{01}), the value of B_{01} was higher than it was in the case of four parameters fitting of their data.

As it is seen in Fig. 2 the effect of the B_{11} parameter is similar to the B_{01} but its relative importance is increasing at the wings of the cusp. It is interesting that the value of this parameter shows a sharp decrease towards higher impact energies as it is seen in Fig. 4. Unfortunately, Knudsen et al [21] did not publish these values. For heavy ion–atom collision the B_{02} (quadrupole term) was investigated, too [20–21]. It was found that this value has a negative character and its relative importance is increasing with decreasing impact velocity [21].

Shape analyzes for the ELC processes are rather few, mainly qualitative results can be found. Recently, three dimensional cusp was measured as a function of the electron velocity and its emission angle and the so called contour lines are compared with the theory [15][19]. Among the studies for ELC mechanism, however, multiply ionized Au projectiles having still very many electrons were used and here the coincidence technique was also applied in the measurement concerned [18].

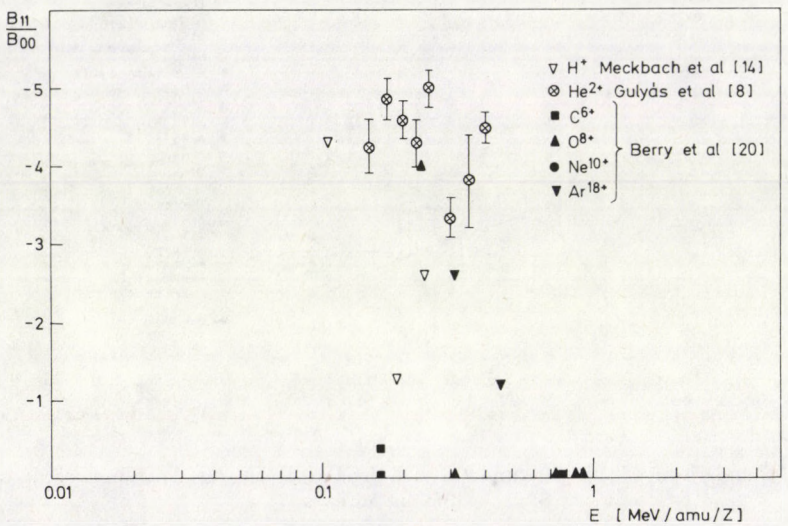


Fig. 4. Comparison of the experimental B_{11}/B_{00} parameters determined by different authors

Gulyás et al [8] studied the shape of the cusp for light projectile (He^+) carrying electron and found that the value of the B_{01} parameter is practically zero but the value of B_{11} is rather high showing a slight asymmetry.

4. Theoretical description

The early first order Born approximations [28–30] give symmetric shape for the ECC cusp which is in contrast with the experimental observations. This contradiction was one of the reasons of the theoretical investigations of higher order approximations. Shakeshaft and Spruch [31] published a second order Born calculation for H-like target and asymptotically high velocity, which gives an asymmetric shape for the ECC cusp. Fig. 3 shows the values of the B_{01} parameter calculated according to this theoretical approximation by Jakubassa-Amundsen [11] for H^+–He collision. As it was noted this approximation is valid only for higher than 0.4 MeV/amu impact energies. A similar calculation was done by Kövér et al [26]

for He^{2+}–He collision in the lower energy region (less than 0.4 MeV/amu). These data are substantially smaller than the experimental values.

The calculation made by Miraglia and Ponce [32] is based on the same assumptions as that of Shakeshaft and Spruch, but differs in the approximations. The value of the series expansion parameters determined by them is rather similar to those which was calculated according to Shakeshaft and Spruch [31].

Another theoretical description was done by Barrachina and Garibotti [33] which is based in the multiple scattering expansion of the ionization amplitudes. The calculated value of the relative B_{01} agreed rather well with the experimental result of Meckbach et al [14] and Gulyás et al [8] but for higher order parameters the agreement is rather poor.

Jakubassa-Amundsen [11][34] interpreted the ECC process with the semiclassical impulse approximation for asymmetric collision systems where the projectile is heavier than the target. This description is valid for H-like target atoms, but it can be expanded for more complicated systems, too. These results for H^+–He and for C^{6+}, Ne^{10+}, Ar^{18+}–He collision systems [11] are displayed in Fig. 3.

Recently, Burgdörfer united the description of the ELC and ECC processes in terms of the density matrix of the low lying continuum states [12]. In this approximation the asymmetry parameters can be determined in terms of the expected values of the Runge–Lenz operators. Fig. 3 shows the results of his calculation in continuum distorted wave approximation for H^+–He system.

The scaling used for the comparison of the different experimental values by Knudsen et al [21] and by us in Fig. 3 is valid theoretically only in impulse approximation [34] for heavy projectiles. However, these theoretical curves are in bad agreement with most of the experimental values except the data of Andersen et al [16] in which case this theory is not valid (the projectile is lighter than the target).

5. Conclusions

The detailed analysis of the shape of the cusp by series expansion seems to be a fruitful method for examining the electron-transfer process. It is independent from the theoretical model and the experimental arrangements, therefore stimulates further theoretical and experimental investigations. Unfortunately, for the most important parameter (B_{01}) the experimental data show a rather large spread. Recently, the different theoretical approximations differ not only from the experimental data but from each other, too. A standard evaluation procedure and interpretation for the series expansion coefficients would be necessary not only for the experimental but for the theoretical data.

Recently, Meckbach et al [15] and Elston et al [19] have measured the ELC angular distribution around the projectile beam determining the transverse electron transmission. It is a promising extension method to study more complex collision problems.

References

1. G.B. Crooks and M.E. Rudd, Phys. Rev. Lett., *25*, 1599, 1970.
2. K.G. Harrison and M.W. Lucas, Phys. Lett., *33A*, 142, 1970.
3. M. Breinig, S.B. Elston, S. Huldt, L. Liljeby, C.R. Vane, S.D. Berry, G.A. Glass, M. Schauer, I.A. Sellin, G.A. Alton, S. Datz, S. Overbury, R. Laubert and M. Suter, Phys. Rev., *A25*, 3015, 1982.
4. I.A. Sellin, Proc. Int. Conf. X84, Leipzig, Aug. 20–24, 1984, ed. A. Meisel and J. Finster. K. Marx Univ., Leipzig, 1984, p. 219.
5. I.A. Sellin, S.B. Elston and S.D. Berry, Proc. 2nd Workshop on High-Energy Ion–Atom Collision Processes, Debrecen, Aug. 27–28, 1984, ed. D. Berényi and G. Hock, Akadémiai Kiadó, Budapest, 1985, p. 249.
6. D. Berényi, Proc. XV Brasov Int. School on Atomic and Heavy Ion Interaction, Poiana-Brasov, Aug. 28–Sept. 8., 1984. Ed. A.L. Berinde, I.A. Dorobantu, V. Zoran. Central Inst. of Phys., Bucharest, 1986, p. 161.
7. "Forward Electron Ejection in Ion Collision". Proc. Symp. Aarhus, June 29–30, 1984, ed. K.O. Groeneveld, W. Meckbach and I.A. Sellin, Springer Vlg., Berlin, 1984.
8. L. Gulyás, Gy. Szabó, D. Berényi, Á. Kövér, K.O. Groeneveld, D. Hofmann and M. Burkhard, Phys. Rev. *A34*, 2751, 1986.
9. I.A. Sellin, Nucl. Instrum. Meth., *B10/11*, 156, 1984.
10. P. Focke, I.B. Nemirovsky, E. Gonzales-Lepera, W. Meckbach, I.A. Sellin and K.O. Groeneveld, Nucl. Instrum. Meth., *B2*, 235, 1984.
11. D.H. Jakubassa-Amundsen, in Ref. 7. p. 17.
12. J. Burgdörfer, Phys. Rev., *A33*, 1578, 1986.
13. M.W. Lucas, W. Steckelmacher, J. Macek and J.E. Potter, J. Phys. B: At. Mol. Phys., *13*, 4833, 1980.
14. W. Meckbach, I.B. Nemirovsky and C.R. Garibotti, Phys. Rev., *A24*, 1793, 1981.
15. W. Meckbach, R. Vidal, P. Focke, I.B. Nemirovsky and E. Gonzales-Lepera, Phys. Rev. Lett., *52*, 621, 1984.
16. L.H. Andersen, K.E. Jensen and H. Knudsen, J. Phys. B: At. Mol. Phys., *19*, L161, 1986.
17. S. Berry, I.A. Sellin, K.O. Groeneveld, D. Hofmann, L.H. Andersen, M. Breinig, S.B. Elston, M.M. Schauer, N. Stolterfoht, H. Schmidt-Böcking, G. Nolte and G. Schiwietz, IEEE Trans. Nucl. Sci., Ns- *30*, 902, 1983.
18. L.H. Andersen, M. Frost, P. Hvelplund and H. Knudsen, J. Phys. B: At. Mol. Phys., *17*, 4701, 1984.
19. S.B. Elston, S.D. Berry, J. Burgdörfer, I.A. Sellin, M. Breinig, R. DeSerio, C.E. Gonzales-Lepera, L. Liljeby, K.O. Groeneveld, D. Hofmann, P. Koschar and I.B.E. Nemirovsky, Phys. Rev. Lett., *55*, 2281, 1985.
20. S.D. Berry, G.A. Glass, I.A. Sellin, K.O. Groeneveld, D. Hofmann, L.H. Andersen, M. Breinig, S.B. Elston, P. Engar, M.M. Schauer, N. Stolterfoht, H. Schmidt-Böcking, G. Nolte and G. Schiwietz, Phys. Rev., *A31*, 1392, 1985.
21. H. Knudsen, L.H. Andersen and K.E. Jensen, J. Phys. B: At. Mol. Phys., *19*, 3341, 1986.
22. J. Macek, J.E. Potter, M.M. Duncan, M.G. Menendez, M.W. Lucas and W. Steckelmacher, Phys. Rev. Lett., *46*, 1571, 1981.
23. P. Focke, W. Meckbach, C.R. Garibotti and I.B. Nemirovsky, Phys. Rev., *A28*, 706, 1983.
24. P. Dahl, J. Phys. B: At. Mol. Phys., *18*, 1181, 1985.
25. M. Rødbro and F.D. Andersen, J. Phys. B: At. Mol. Phys., *12*, 2883, 1979.
26. Á. Kövér, Gy. Szabó, D. Berényi, L. Gulyás, I. Cserny, K.O. Groeneveld, D. Hofmann, P. Koschar and M. Burkhard, J. Phys. B: At. Mol. Phys., *19*, 1178, 1986.
27. Z. Vager, B.J. Zabransky, D. Schneider, E.P. Kanter, Gu Yuang Zhuang and D.S. Gemmel, Phys. Rev. Lett., *48*, 592, 1982.
28. A. Salin, J. Phys. B: At. Mol. Phys., *2*, 631, 1969.
29. J. Macek, Phys. Rev., *A1*, 235, 1970.
30. K. Dettmann, K.G. Harrison and M.W. Lucas, J. Phys. B: At. Mol. Phys., 7, 269, 1974.

31. R. Shakeshaft and L. Spruch, Phys. Rev. Lett., *41*, 1037, 1978.
32. J.E. Miraglia and V.H. Ponce, J. Phys. B: At. Mol. Phys., *13*, 1195, 1980.
33. R.O. Barrachina and C.R. Garibotti, Phys. Rev., *A28*, 1821, 1983.
34. D.H. Jakubassa-Amundsen, J. Phys. B: At. Mol. Phys., *16*, 1768, 1983.
35. Á. Kövér, Gy. Szabó, D. Berényi, L. Gulyás, K.O. Groeneveld, D. Hofmann and M. Burkhard, Abstract of the contributed papers of the XVII. International Conference on Phenomenon in Ionized Gases, ICPIG-XVII., Budapest, 8–12 July, 1985, ed. J.S. Bakos and Zs. Sörlei, p. 395.

Acta Physica Hungarica 64 (1–3), pp. 267–275 (1988)

THE NUCLEUS OF HALLEY'S COMET*

K. Szegő

Central Research Institute for Physics
1525 Budapest, Hungary

(Received 8 January 1987)

In this paper we summarize the most important result of the VEGA, GIOTTO and SUISEI space missions relevant to understand the physics of the nucleus of Halley's Comet.

1. Introduction

Comets, especially Halley's Comet always set people's imagination into work. What are they, where did they come from, how are the visible features, the coma, the tail etc. created? These are only some of the interesting questions which excited not only astronomers but laymen, too.

Comets as we presently understand [1] form an important link between the solar system and the rest of the Universe. The comets created are stored in a cloud at about 50000 AU from the Sun. This cloud, first proposed by Oort, stores about 10^{12} comets with an average mass of 10^{17} g. To retain one comet in the cloud about 20 have to escape, not necesssarily toward us. Due to gravitational effect of passing stars 5 – 10 comets yearly start their journey towards the solar system. The general dynamics of the cloud is not completely satisfactory but no major contradiction is present.

Comets were created from compactification of interstellar grains resulted of gravitational sedimentation. One theory assumes that they were formed in the region of Uranus and Neptune, at 15–30 AU from the Sun in an environment 50–100 K cold. Some grew to planets, most were swept towards the outer regions. Comparing the mass of the Oort cloud to that of the planets of the solar system one may claim that the planets are just the by-products of comet formation. Other theories assume that at 300-1000 AU there was an extended accretion disk where comets were formed at a 20–40 K temperature. Dynamically it is improbable that comets result from external capture; the too long accretion time contradicts to originating them from distant nebula fragments. Generally, it is assumed that the cometary material is 99% volatiles (gas) and 1% interplanetary dust.

It is an important question how pristine comets are [2], which means that one should not only investigate how comets were formed but also how they were stored and during their return what modifications they suffered. During storage the

*Dedicated to Prof. G. Marx on his 60th birthday.

dust environment, accretion of more gas and the irradation may alter the surface materials. During the return phase both the solar wind interaction and the acquired heat may modify its material. The processes have to be understood, too, if from the investigation of comets we would like to make conclusions about the origin of the solar system. M. Wallis [3] raised the interesting question that radioactive heating could have played an important role in cometary evolution. The decay of 26Al may cause heating enough to obtain a fluid core. If that really happens, comets may not be as pristine as it is generally assumed.

As ground observations could not reveal many phenomena related to these, space missions should have been planned to fill some of the gaps in our knowledge.

Cometary fly-by missions had become technically feasible before the next apparition of Halley's Comet was due in 1985–86. The scientific objectives of the missions were very similar: to image the nucleus and the near nucleus region, study jet phenomena, analyze the chemical processes in the cometary atmosphere, analyze the mass distribution and composition of the dust, investigate plasma phenomena including the interaction with the solar wind. Alterations in the objectives were due to the technical capabilities of the individual missions.

This paper is a summary of the investigations of the results of the cometary nucleus.

2. Imaging of the nucleus aboard the VEGA spacecraft

One of the most important goals of the space missions was to image the nucleus of Halley's Comet. This curiosity was understandable since cometary nuclei cannot be observed from ground as the extended dust coma screens it from ground based telescopes. Though astronomers generally assumed one single solid object as nucleus, more extravagant ideas were also present, from multiple core to non solid nucleus.

To image the nucleus, a very complex television system (TVS) was put on board the VEGA s/c [4] because it also had the task to search and find its target. Halley's Comet revolves retrograde, hence the relative speed at the encounter between the comet and the s/c was high, almost 80 km/s. So all the optical experiments were put on a turnable platform pointing toward the nucleus during the observation. As the exact position of the nucleus was not known TVS had to find its target autonomously.

TVS consisted of two telescopes, a narrow angle (1200 mm focal length) for scientific imaging and a wide angle (150 mm focal length) for navigation. The light in both was split into two paths and reached Soviet-made CCD detectors 512×512 pixel wide. One channel of the WAC was used by an autonomous system for back-up navigation. One channel of the NAC was dedicated to make the scientific images. In this channel there was a filter magazin with 8 different filters in the 400–1100 nm interval. In the other channels there were only fixed red filters (red filters are optimal to diminish dust screening). The TVS set the exposure time automatically in a broad (10 ms – 163 s) interval. The automatic tracking and navigation required

a very advanced and sophisticated software, but its details are out of the scope of this paper.

Before launch the system underwent appropriate calibration, and built-in lamps made possible in-flight calibration, too. During ground calibration the dark current was measured at different temperatures, the homogeneity and linearity of the detectors were registered, different transfer functions of the system were measured: spectral transfer function, line transfer function (better than 80% of the theoretical value), charge transfer efficiency (0.9997/line), etc. The resolution was checked by a special grid pattern and proved to be not very far from the theoretical value. The dynamical range of the system was better than 20, the effect of blooming was also investigated. The characteristics of the in-flight calibration lamps were also registered. The ground calibration data in total filled 200 normal size magnetic tapes, their evaluation took more than a year after launch.

It followed from the general considerations of the VEGA mission that imaging of the nucleus was possible on −2, −1, 1, 2 days for 2 hours from a distance of −14, −7, 7 and 14 million km, and for 3 hours around closest approach. As TVS was in transport position till mid-February 1986, calibration earlier was possible only using the built-in lamps; this took place in every two month after launch. After having opened the platform to operational position, Jupiter and Saturn were observed. This was necessary partly for the optical check up of the system and partly to acquire data for the photometric measurements of the nucleus.

The scientific observation of the nucleus was done according to so-called "photosequences", i.e. according to a software program written in the memory before launch. The sequences were compiled to satisfy different goals. On the basis of the spectrum of comet P/Tuttle different filters were used to try to differentiate between emission lines. A special sequence was composed to observe the nucleus around closest approach. It consisted only of three filters covering the 400–650, 630–760, 740–900 nm range and the so-called moving window images. It was optimized to get enough comparable images with the same filter (allowing a margin for the unknown brightness) but to make it possible to reconstruct the real colour and shape of the nucleus. In-flight modification of the "photosequences" was possible: this option was used during the VEGA-2 fly-by.

3. Discussion of the results

The imaging experiment was very successful [5,6,7,8]. This is true in spite of the fact that during the VEGA-2 encounter the orientation microprocessor on board failed and the system worked in a back-up mode transmitting only full frame images with exposure times not optimized for nucleus observation.

Some further problems made the image processing difficult. In the case of VEGA-1, the linear response of the CCD matrix was shifted from zero by an offset of about 40 digital intensity values. Coherent noise was also present both on VEGA-1 and VEGA-2 images which was removed using Fourier transformation.

As of today we have processed only those images in which the nucleus can be

identified. This includes about 63 VEGA-1 and 11 VEGA-2 images taken around closest approach which were, respectively, on 6 March 1986 at 07.20:00 UT at a distance of 8889 km and on 9 March at 07.20:06 UT at a distance of 8030 km from the nucleus.

Fig.1. Image of the nucleus of Halley's Comet taken by the camera on board the VEGA-2 spacecraft on 9 March 1986. It was taken at 07.19:58 UT, two seconds before closest approach, at a distance of 8030 km from the nucleus, using a near-infrared filter (700–900 nm). The direction of the Sun is towards the lower left hand corner, 113° from the vertical axis of the image. The phase angle (the angle between the spacecraft – comet and the Sun – comet lines) is 28.5°. The resolution is 120 m and 160 m. The comet nucleus appears as an aspherical peanut-shaped object with overall dimensions of 16x8x8 km. The region surrounding the nucleus appears bright due to sunlight reflected by dust particles.

Fig. 1 shows the best image of the nucleus taken by VEGA-2 at a distance of 8030 km from the nucleus with a resolution of 120×160 m^2, in the 740–900 nm range.

Our first goal was to obtain the size of the nucleus. On the images only the illuminated part can be seen, but by iterating images, taking into account the fly-by geometry and the rotation period, we reached the conclusion that the overall dimension of the nucleus is $(16-1)\times(8-1)\times(8-1)$ km. On GIOTTO s/c the

camera could define only two sizes [9], those agree with our result. The nucleus is irregularly shaped, similar to a potato or an avocado; the closest geometrical body is a truncated cone capped at each end by hemispheres of 4 km and 2.25 km radii, respectively. The volume is about 500 km^3. Its density is about 0.6 g/cm^3, very much less than it had been anticipated.

The rotation period was obtained by reconstructing in space the major axis of the nucleus. From this, supposing prograde rotation, we obtained 53.5 +/−1 h period [7]. This agrees well with ground based observation of the jet periodicity [11] and brightness variations [12], and with the Lyman-α breathing of the hydrogen coma observed by the SUSEI s/c [13]. GIOTTO has no data for the rotation. We have to mention, however, that some ground based astronomers observing brightness change of the coma have obtained 7.4 days period.

Using both the ground based and spacecraft observations we reanalyzed the rotation of the nucleus [10]. Since it is an irregular body, its rotation is the rotation of an asymmetric top. Based on the stability of the 7.4 d period we accepted that the motion is free, and we assumed that the nucleus is rigid. We also assumed that the orientation of the angular momentum vector is not far from one of the inertial axes of the nucleus, since in that case the rotation of an asymmetric top can be solved in closed form.

We accepted from the observations the existence of a short, 2.2 d and a long, 7.4 d period. The major dimensions, 16 km for the longest and 7 km for the shortest axis were taken from the VEGA observations. We identified the position of the long axis with the position of the shortest inertial momentum axis, which is reasonable from the existing data. Then using the orientation of the long axis during the different encounters (as given earlier) we concluded that the nucleus of P/Halley rotates as a slightly asymmetric top. The orientation of the rotation vector, i.e. the orientation of the angular momentum vector is $b=-54°$, $l=219°$ in the ecliptic system ($l=39°$ in our cometocentric system). The error zone is a cone about this axis with a half opening angle of 15°.

In the rotation of an asymmetric top the rotation axis is not fixed rigidly to the body, so while the nucleus rotates about the axis with a short period of 2.2 +/− 0.05 d, its long axis "nods" periodically with the long period of 7.4+/−0.05 d. The amplitude of the "nodding" is about 14° +/−3° in both directions relative to a plane perpendicular to the rotation axis. The rotation is a so-called short axis mode.

No differences could be found between images made by different filters; so we conclude that the colour of the nucleus is neutral, grey.

At present we are in the process of defining the contour of the nucleus on individual images. This should be corrected by the illumination to obtain the limb of the dark side too, based on the fly-by geometry. At this stage preliminary limbs are available. Using these, in the Space Research Institute in Moscow an approximate 3-dimensional nucleus has been reconstructed, but this needs further improvement. The same group, using a similar technique as in medical tomography, reconstructed the 3-d light intensity distribution in the near nucleus region. The dust jets shield some part of the limb, so it was important for us to understand the jet geometry.

An important observation was that jets are observable mostly around the sunlit hemisphere, after dawn or before dusk the jet activity was limited. The VEGA-2 images made possible the 3-d reconstruction of most of the jet sources. On those images the jet boundaries and cores are clearly identifiable.

For 3-d reconstruction, we assumed that the jet sources could be approximated either by discrete points or by lines. (The edges of a linear source should appear as boundaries on several images.) By knowing the geometry, most but not all of the sources could be spatially reconstructed.

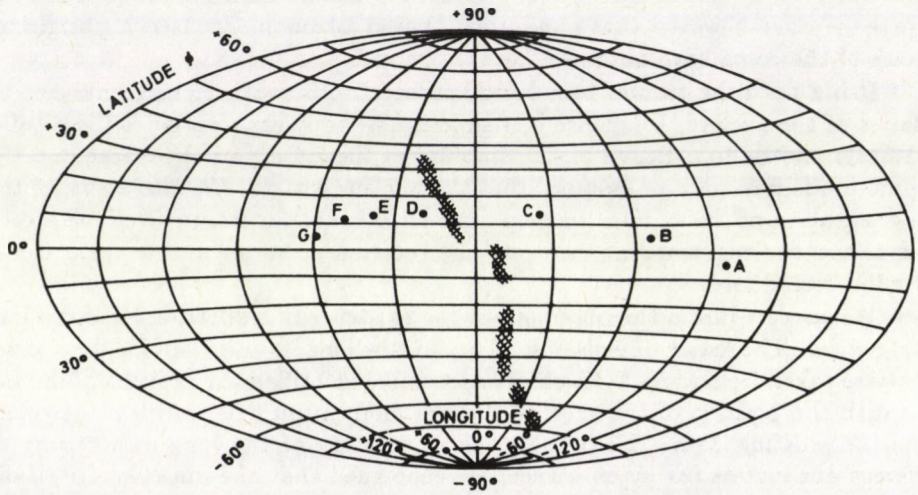


Fig.2. The path of VEGA-2 projected onto a sphere centered on the nucleus. North is ecliptic north on the Figure, the subsolar point defines zero longitude. On the path, dots indicate the recording of the images; e.g. we list for images A–G the recording time relative to encounter (E) and distance from the nucleus. (A: E-370.5 s, 29539 km; B: E-101.7 s, 11197 km; C: R-1.5 s, 8031 km; D: E+98.7 s, 11037 km; E: E+187.3 s, 16462 km; F: E+280.5 s, 22971 km; G: E+558.1 s, 43567 km). The identified line-like jet sources are denoted by the shaded regions

The result is shown in Fig.2. One important discovery is that not only are many of the sources actually pointlike or linear, but they seem to be ordered to form an even larger quasi-linear structure along the surface. One part of this structure crosses the projection of the s/c trajectory onto the nucleus, which means that the s/c crossed this jet. This was actually observed by the SP-1 dust counter [14]. We were able to identify a jet source on the dark hemisphere too, using the shift differentiation technique developed by Larson and Sekanina [15].

The GIOTTO HMC experiment [9] concluded that the very active regions on the surface are less than 15–20 % of the total surface. The Lyman-α imaging on SUISEI [13] also identified localized sources on the surface: two strong (S1, S2) and four weak (W1–W4) sources were identified, resulting in periodic breathing of the

Lyman-α coma. The angular positions of the sources are: S1(0) – W1(20) – W2(75) – S2(150) – W3(204) – W4(326), where the degrees in parentheses are the positions of the sources projected onto an arbitrary circle. The conclusion was reached that VEGA-1 experienced the outburst of S2, whereas VEGA-2 did the same with W4, assuming a 7h time of flight for dust particles from surface to s/c. These conclusions agree well with the general features of VEGA-1 and VEGA-2 images. As GIOTTO met the comet when it was rotated by almost 180 degree in comparison to the VEGA-1 situation, it explains why the GIOTTO images are relatively clear from dust.

It is important to note that on the VEGA images the surface of the nucleus on the non-active parts covered by optically thin dust [8], the optical thickness is less than 0.1. These measurements also revealed that the nucleus is covered by a dark material resulting in a low, 4% albedo.

Work is in progress to identify the observed surface brightness variations on the processed images with topographic features, dust sources, etc. At this stage from the VEGA images no surface feature has been identified unambiuously. The GIOTTO images show some crater-like structures on the surface.

The results of some other experiments are also relevant to understand the nucleus. The IKS experiment on VEGA measured the surface temperature of the nucleus and found it to be 300–400 K, in a region of about 5×3 km^2. There the surface cannot be highly darkened normal ice. Both devices, the PUMA on VEGA [16] and the PIA [18] on GIOTTO measured the dust composition. They confirmed that dust particles fall into different classes: one is quite close to light organic material, another to silicates, the third is close to C1 carbonaceous chondrites but richer in carbon. Ten minutes before closest approach the dust material very frequently had the highest peak in carbon, closer to the nucleus the highest peak was carbon or silicon. Both IKS and TKS on VEGA confirmed thet the activity around the nucleus was very asymmetrical, and confined to the sunward hemisphere. In different spectral ranges jet-like activities were observed [18,19].

What is our present understanding of the nucleus? At the advent of the space missions Whipple's dirty snow ball model was the accepted paradigm. It was developed further e.g. by allowing an outgassed, friable sponge like mantle on top of the ice–dust core [20]. As this model yielded a fairly homogeneous activity, special surface features were considered to explain jet activity.

After the space missions we have to conclude that the nuclear activity is much more inhomogeneous than was previously thought. The most active area is about 20% of the surface, being divided into active spots and crisps, linear features. The inactive or less active areas are covered by dark, refractory material probably rich in carbon. (The IRAS experiment observed three comets all of which were very dark [21]). At some places the solid surface might be almost impervious to the escape of gas. The volatiles are either produced from jet sources or are emitted from the dust in the atmosphere but pure gas outbursts cannot be excluded. The surface material has probably low heat conductivity. If one compares the position of the jet sources measured by VEGA-2 with the active regions measured by SUISEI and as seen by GIOTTO, the activity is highest around noon and varies with the Sun angle. IKS

on VEGA observed about 5–10 degrees of heat lag, corresponding to 1–2 hours of lag. Having so localized jet sources and small active surface, the active areas should almost be covered by ice to account for the observed total gas production rate [22]. However, here we face some difficulty in understanding where the dust comes from. If it is just frozen in material in the ice, dust analysis does not reveal the overall cometary composition. However, this is an unlikely situation, because to account for the total material lost during one apparition, the surface layer has to shrink about 1 m, or, assuming 20% active areas, 5 m deep crisps should cross the surface. But deepening decreases the possibility for the Sun to heat it, so the active areas should change from time to time. It will be interesting to see how the 1910 jet sources coincide with the present observations. If the positions of the crisps are changing, this can be due to heat stresses. In that case the new crisps are covered by "fresh dust" from the surface layer and the dust will give information about the overall composition of the comet. In this picture the icy core is covered by a thicker layer of solid mantle.

There is another option too, that the crisps are not changing position. In that case the nucleus is similar to the icy glue model [23]. This means, however, that the crisps should be deep and wide after many apparitions, slowly leading to the disintegration of the nucleus. Some crisps should be big enough to be observable by the camera having about 200–300 m resolution. This type of analysis is still in progress. We would mention that in this case the missions give information about the composition of the icy glue, but not about the boulders glued together. The difficulty of this model is, however, that it is unclear how large-sized boulders could reach regions outside the big planets.

If we believe that comets were born in the Uranus–Neptunus region (though the low density favours more distant regions), it is conceivable that very big objects were also created with small likelihood. Some of the big ones could have been captured by the giant planets. It is tempting to claim that Umbriel, a satellite of Uranus recently observed by Voyager, is very similar to Halley albeit very much larger.

During the 20th ESLAB Symposium (27–31 October 1986, Heidelberg) Whipple suggested that comets are made of smaller snow balls glued together [24]. The surface probably underwent some changes due to irradiation, etc., while the comets were stored in the Oort cloud, and this yielded the less active or inactive areas. The friable sponge like mantle might be in work only in the active places. Further studies are needed to clarify this situation.

References

1. W. McCrea, Proc. of Comet Nucleus Sample Return Workshop (Canterbury, 1986), ESA SP-249, p. 9.
2. P. Weismann, Proc. of Comet Nucleus Sample Return Workshop (Canterbury, 1986), ESA SP-249, p. 15.
3. M. Wallis, Proc of Comet Nucleus Sample Return Workshop (Canterbury, 1986), ESA SP-249, p. 63.

4. The Venus–Halley Mission, ed. by V. M. Balebanov et al., Imprimerie Louis-Jean, Paris, 1985.
5. R.Z. Sagdeev et al, Nature, *321*, 262, 1986.
6. R.Z. Sagdeev et al, Adv. of Space Res., *5*, 95, 1985.
7. R.Z. Sagdeev et al, Proc. of the 20th ESLAB Symposium, Heidelberg, 26-31 October, 1986, ESA SP-250, II. 289.
8. R.Z. Sagdeev, K. Szegő, KFKI-1987-35/c; to be published in 'Comet Halley 1986', Ellis Horwood Limited, England.
9. H.U. Keller et al., Nature, *321*, 320, 1986.
10. R.Z. Sagdeev et al, KFKI-1987-75/c, submitted to Astron. Journal.
11. Z. Sekanina et al., Proc. of the 20th ESLAB Symposium, Heidelberg, 26-31 October, 1986, in print.
12. M.J.S. Belton, Proc. of the 20th ESLAB Symposium, Heidelberg, 26-31 October, 1986, ESA SP-250, 1, 599.
13. E. Kaneda et al., Nature, *321*, 297, 1986.
14. O.L. Vaisberg et al., Nature, *321*, 274, 1986.
15. Z. Sekanina and S. Larson, Astron. J., *89*, 1408, 1984.
16. J. Kissel et al., Nature, *321*, 280, 1986.
17. J. Kissel et al., Nature *321*, 336, 1986.
18. M, Gogoshev et al., Adv. of Space Res., *5*, 133, 1986.
19. J.M. Lamarre et al., Adv. of Space Res., *5*, 123, 1986.
20. M. Horanyi et al., The Astrophys. J., *278*, 449, 1984.
21. M.S. Hanner, Icarus, *62*, 97, 1985.
22. W.H. Ip., Adv. of Space Res., *5*, 233, 1986.
23. T.I. Gombosi and H. Houpis, preprint, KFKI-1986-23/C.
24. F. Whipple, Proc. of the 20th ESLAB Symposium, Heidelberg, 26-31 October, 1986, ESA SP-250, II. 281.

Acta Physica Hungarica 64 (1–3), pp. 277–283 (1988)

ON THE DETERMINATION OF LARGE SCALE MOTION IN THE UNIVERSE*

A.S. SZALAY and E. REGŐS

Department of Atomic Physics, Roland Eötvös University
1088 Budapest, Hungary

(Received 8 January 1987)

We consider the effect of a nonisotropic selection function on the estimated values of the large scale velocities of the galaxy distribution. We show that the direction of the velocity is correlated with the dipole anisotropy, and it is also effected by quadrupole and higher terms. These anisotropies seem to be unavoidable and their effect should be included when comparing the results of experiments to theoretical predictions.

It is generally accepted that early, close to the Big Bang the Universe was close to homogeneous and isotropic, only small fluctuations of the density and curvature existed. The present structure of the Universe arose from these small primordial fluctuations. Here we attempt to derive observable consequences of these, namely their effect on large scale flows in the galaxy distribution.

The fluctuations are described by a dimensionless density fluctuation field and the peculiar velocity field relative to the Hubble flow. Both are functions of the comoving spatial coordinate $\mathbf{X}$ and time t.

$$\delta(\mathbf{x},t) = \frac{\rho(\mathbf{x},t)}{\rho_0(t)} - 1; \quad \mathbf{v}_{\rm obs}(\mathbf{x},t) = \mathbf{v}(\mathbf{x},t) + H\mathbf{s}.$$

$\mathbf{x} = \mathbf{x}_0 + \mathbf{s}$ and $H = (\dot{a}/a)$, the Hubble constant, a is the expansion factor. Solving the hydrodynamic equations in expanding coordinates a second order differential equation is obtained for the fluctuation amplitude, which contains the second derivative of the pressure. In late epochs ($Z < 200$) the universe is dominated by nonrelativistic matter, thus pressure in negligible. If the fluctuations are small, we can linearize the equations. This approximation can be used to describe the behaviour of large scale density irregularities in the present Universe.

One would like to relate the fluctuation spectrum of the Universe to observational facts. This has been done by calculating *rms* density fluctuations, correlation functions etc., fluctuations in the microwave background radiation. So far measurements of the microwave anisotropy yield only upper limits on the fluctuation amplitude, whereas the presence of galaxies and clusters is a lower limit. Only the galaxy correlation function has been used to provide a normalization for any

*Dedicated to Prof. G. Marx on his 60th birthday.

given spectrum. However, if light does not trace mass, these normalizations can be false. On the other hand peculiar velocities are caused by the perturbations in the mass distribution, so it has been suggested (Peebles and Clutton-Brock [1], Kaiser [2]), that the peculiar motion of relatively large regions be studied instead. These regions — if large enough — are well described by linear theory since nonlinear motions inside the region cancel out, only the collective part remains.

The experimental determination of such large scale motion is indeed possible, and has been carried out recently, although results remain somewhat controversial. Several groups attempted to determine the collective motion of relatively large (about 10–50 Mpc in radius) region of the Universe centered on us (Rubin and Ford [3], Hart and Davies [4], Burnstein et al [5], Collins et al [6]). Using an independent distance determination to subtract the Hubble flow from the observed total radial velocity, the radial peculiar velocity $v_r(\mathbf{s}) = \mathbf{v}(\mathbf{s})\hat{\mathbf{s}}$ can be determined for each object. $\hat{\mathbf{s}} = \mathbf{s}/s$ is a vector of unit length parallel to s. The whole radial velocity is measured from the redshift, and the distances are calculated from secondary distance indicators: absolute luminosity is related to the radial velocity dispersion within an elliptical galaxy (Burnstein et al [5]) or to the dispersion of the 21 cm line in a spiral galaxy (Collins et al [6]).

The various groups determine the net CM velocity of the observed region in different ways. Either they take the vector sum of the radial measurements weighted with a radial selection function that defines the distribution of the observed objects (de Vaucouleurs and Peters [7]), or they use a least square fit or maximum likelihood techniques to the distribution of the radial velocity in various directions. Results depend on the method used and can have large systematic errors. The different samples do not give identical results yet.

Motion on such a large scale is linear, therefore it can be related to the primordial linear fluctuation spectrum in a straightforward way as follows. Using the Fourier expansion:

$$\delta(\mathbf{x}) = \int d^3k e^{\mathbf{ikx}} \delta_{\mathbf{k}}; \quad \mathbf{v}(\mathbf{x}) = \int d^3k e^{\mathbf{ikx}} \mathbf{V}_{\mathbf{k}},$$

where $\delta_{\mathbf{k}}$ has random phases and $P(k) = |\delta_{\mathbf{k}}|^2$ is the power spectrum. From the continuity equation:

$$\mathbf{v}_k = -i(Haf)\frac{\mathbf{k}}{k^2}\delta_{\mathbf{k}} = \hat{\mathbf{k}} v_{\mathbf{k}}$$

with $Hf = (\dot{D}/D)$ and D is the growing solution of the linear equation on $\delta_{\mathbf{k}}$ (Peebles [8]).

Given a selection function describing the observed region one can predict a center-of-mass velocity, based upon the fluctuation spectrum. The magnitude of this prediction is given by the rms (*root mean square*) value, which is compared to the observed velocity. If the two strongly disagree, then our choice or normalization of the fluctuation spectrum can be rejected. In order to make strong statements, our calculation should attempt to simulate the observations as close as possible.

The expected value of the CM velocity at a given point is given by the integral of $\mathbf{v}(\mathbf{x})$ over the observed area:

$$\tilde{\mathbf{V}}(\mathbf{x}_0) = \frac{\int d^3s\Phi(\mathbf{s})\mathbf{v}(\mathbf{s})}{\int d^3s\Phi(\mathbf{s})}.$$

The selection function is $\Phi(\mathbf{s})$, normalized such that the denominator becomes 1. All theoretical calculations of this kind so far have assumed an isotropic selection function. The dispersion $< V^2 >$ can be expressed through the power spectrum and the scalar window function $W(k)$, which is the Fourier transform of the isotropic $\Phi(s)$:

$$< \tilde{\mathbf{V}}^2 >= \int d^3k|v_{\mathbf{k}}|^2W(k)^2.$$

The assumption of isotropy is generally not correct (clumpy distribution of galaxies, galactic exinction, etc). Here we take the anisotropies into account, and show that they have a strong effect. We consider the galaxy distribution in a spherical shell describing the selected region and expand it in multipoles. We derive the dispersion of the CM velocity of the region selected this way, assuming that it is at a random place in the Universe and compare it to the results obtained with an isotropic selection.

We have analytically calculated the case when there is a dipole component along the z axis and there are quadrupole components as well. Doing a least squares fit to determine the $\mathbf{U}$ CM velocity from the $v^i = \mathbf{v}^i\hat{\mathbf{s}}^i$ observed radial velocities:

$$\sum_i (v^i - \mathbf{U}\hat{\mathbf{s}}^i)^2 = \min ,$$

leading to

$$\mathbf{U} = M^{-1}V;$$

where

$$M_{\mu\nu} = \sum_i \hat{s}^{i*}_\nu \hat{s}^i_\mu = \int d^3s\Phi(\mathbf{s})\hat{s}^*_\nu\hat{s}_\mu;$$

$$V(x_0) = \sum_i v^i\hat{s}^i =< v_r(s)\hat{s} >= \int d^3s\Phi(\mathbf{s})\hat{\mathbf{s}}v_r(\mathbf{s});$$

thus

$$< |U_\nu|^2 >= M^{-1}_{\rho\nu}M^{-1}_{\nu\mu} < V^*_\rho V_\mu > .$$

First we determine $\mathbf{V}$, the vector-sum of the radial velocities

$$V_\alpha = \int d^3s\Phi(\mathbf{s})\hat{s}_\alpha\hat{s}_\beta v_\beta(\mathbf{x}) = \int d^3k e^{\mathbf{ikx}_0}v_{\mathbf{k}\beta}W_{\alpha\beta}(\mathbf{k}),$$

where we have a tensor window function, similar to the one used by Grinstein et al [9]

$$W_{\alpha\beta}(\mathbf{k}) = \int d^3s e^{i\mathbf{ks}} \hat{s}_\alpha \hat{s}_\beta \Phi(\mathbf{s}).$$

We use a spherical coordinate system: $V_\nu = U^+_{\nu\alpha} V_\alpha$ where $\mathbf{U}$ is a unitary rotation matrix

$$\mathbf{U} = \begin{pmatrix} -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ i/\sqrt{2} & 0 & i/\sqrt{2} \end{pmatrix},$$

$$\hat{s}_\alpha = \sqrt{\frac{4\pi}{3}} U_{\alpha\mu} Y_{1_\nu}(\Omega_s),$$

$$V_\nu(\mathbf{x}_0) = \int d^3k e^{i\mathbf{kx}_0} V_{\mathbf{k}\nu} = \int d^3k e^{i\mathbf{kx}_0} \left(v_{\mathbf{k}\mu} \int d^3s \Phi(\mathbf{s}) e^{i\mathbf{ks}} \frac{4\pi}{3} Y_{1\nu}(\Omega_s) Y^*_{1\mu}(\Omega_s) \right),$$

$$v_{\mathbf{k}\mu} = v_{\mathbf{k}} \sqrt{\frac{4\pi}{3}} Y_{1\mu}(\Omega_{\mathbf{k}}).$$

We assumed that our position is random location, thus the dispersions in coordinate- and Fourier space are equal, so the expectation value of $|V_\nu(\mathbf{x}_0)|^2$ is:

$$< |V_\nu(\mathbf{x}_0)|^2 >= (2\pi)^3 \int d^3k |V_{\mathbf{k}\nu}|^2.$$

Let us expand $V_{\mathbf{k}\nu}$ in multipoles and determine the coefficients

$$V_{\mathbf{k}\nu} = v_{\mathbf{k}} \sum_{l,m} q_{lm,\nu}(\mathbf{k}) Y_{lm}(\Omega_{\mathbf{k}}),$$

$$< |V_\nu|^2 > = (2\pi)^3 \sum_{l,m} \int d^3k |v_{\mathbf{k}}|^2 q_{lm,\nu}|^2.$$

From these equations:

$$q_{lm,\nu} = (\frac{4\pi}{3})^{3/2} \sum_\mu \int d\Omega_{\mathbf{k}} Y^*_{lm}(\Omega_{\mathbf{k}}) Y_{1\mu}(\Omega_{\mathbf{k}}) \int d^3s \Phi(s) e^{i\mathbf{ks}} Y_{1\nu}(\Omega_s) Y^*_{1\mu}(\Omega_s).$$

Using the Rayleigh expansion:

$$e^{i\mathbf{ks}} = \sum_{l'm'} 4\pi i^{l'} j_{l'}(ks) Y^*_{l'm'}(\Omega_{\mathbf{k}}) Y^*_{l'm'}(\Omega_{\mathbf{s}}).$$

and

$$\int d\Omega Y^*_{lm} Y_{1\mu} Y_{l'm'} = \sqrt{\frac{3}{4\pi}} \sqrt{\frac{2l'+1}{2l+1}} C(l'1l;00) C(l'1l;m-\mu,\mu) \delta_{m',m-\mu},$$

$$Y^*_{1\mu} Y^*_{l',m-\mu} = \sum_L \sqrt{\frac{2l'+1}{2L+1}} \sqrt{\frac{3}{4\pi}} C(l'1L;\mu-m,-\mu)(-1)^{l'+1-L} Y^*_{Lm},$$

where we apply the integral for $\Omega_{\mathbf{k}}$ and the product for $\Omega_{\mathbf{s}}$. Using the properties of the Clebsch–Gordan coefficients we arrive at:

$$q_{lm,\nu} = 4\pi \sum_{L=l\pm 1} \frac{C(l1L;00)C(l1L;-m,\nu)}{\sqrt{(2l+1)(2L+1)}} \int d^3s\Phi(\mathbf{s})Y_{L,\nu-m}(\Omega_{\mathbf{s}})Q_l(ks),$$

where

$$Q_l(z) = i^{l-1}(lj_{l-1}(z) - (l+1)j_{l+1}(z)).$$

Let us expand $\Phi(s)$ in multipoles:

$$\Phi(\mathbf{s}) = F(s) \sum_{l'm'} a_{l'm'} Y^*_{l'm'}(\Omega_s),$$

thus

$$\int d^3s\Phi(\mathbf{s})Y_{L\mu}\frac{Q_l(ks)}{(2l+1)} = a_{L\mu}G_l(k),$$

where

$$G_l(k) = \int_0^\infty dss^2F(s)\frac{Q_l(ks)}{(2l+1)}4\pi.$$

Calculating the C coefficients we get:

$$q_{lm,\pm 1} = \left(-\sqrt{\frac{(l\pm m)(l\pm m-1)}{2(2l-1)(2l+1)}}a_{l-1,\pm 1-m} + \right.$$

$$\left. + \sqrt{\frac{(l\mp m+1)(l\mp m+2)}{2(2l+1)(2l+3)}}a_{l+1,\pm 1-m}\right) G_l,$$

$$q_{lm,0} = \left(\sqrt{\frac{l^2-m^2}{(2l-1)(2l+1)}}a_{l-1,-m} + \sqrt{\frac{(l+1)^2-m^2}{(2l+1)(2l+3)}}a_{l+1,-m}\right) G_l.$$

Now let us calculate M:

$$M_{\mu\nu} = \int_0^\infty dss^2F(s)\sum_{lm} a_{lm}\int d\Omega_s Y_{lm}\frac{4\pi}{3}Y^*_{1\nu}Y_{1\mu} =$$

$$= \frac{3}{\sqrt{4\pi}a_{00}}\sum_{lm} a_{lm}\sqrt{\frac{4\pi}{2l+1}}C(11l;00)C(11l;m-\nu,\nu)\delta_{\mu,m-\nu},$$

hence only a_{00} contribute to the normalization of $\Phi(\mathbf{s})$.

Since only $l=0$ and $l=2$ are non-vanishing:

$$M_{\mu\nu} = \frac{\delta_{\mu\nu}}{3} + \sqrt{\frac{2}{15}}\frac{1}{a_{00}}a_{2,\mu+\nu}.$$

Analyzing the data of Burnstein et al we can consider the case when there is a dipole along the z axis and a rotationally symmetric quadrupole around the equatorial since we cannot see through the plane of the Milky-Way Galaxy. Thus only a_{00}, a_{10} and a_{20} are not zero. In this case we can easily invert M. Hence in the mentioned simplified case for the multipole components $a_{lm} = a_{l0}\delta_{m0}$ we arrive at $q_{lm,\nu} = q_{lm,m}\delta_{m,\nu}$ and

$$< V_\rho^* V_\mu > = \delta_{\mu\rho} < |V_\mu|^2 >,$$
$$< |U_\nu|^2 > = M^{-1}_{\nu\mu} M^{-1}_{\mu\nu} < |V_\mu|^2 > .$$

Thus through the only non-vanishing $q_{lm,m}$:

$$< |V_0|^2 >= (2\pi)^3 \frac{4\pi}{3}(H_0 af)^2 \times$$

$$\times \int dk |\delta_k|^2 \left(a_{00}^2 \frac{G_1^2(k)}{3} a_{10}^2 \left(\frac{1}{3}G_0^2 + \frac{4}{15}G_2^2 \right) + a_{20}^2 \left(\frac{9}{35}G_3^2 + \frac{4}{15}G_1^2 \right) + \frac{4}{3\sqrt{5}} a_{00} a_{20} \right)$$

$$< |V_{\pm 1}|^2 >= (2\pi)^3 \frac{4\pi}{3}(H_0 af)^2 \times$$

$$\times \int dk |\delta_k|^2 \left(a_{00}^2 \frac{1}{3}G_1^2(k) + a_{10}^2 \frac{1}{5}G_2^2 + a_{20}^2 \left(\frac{6}{35}G_3^2 + \frac{1}{15}G_1^2 \right) - \frac{2}{3\sqrt{5}} a_{00} a_{20} \right).$$

The dispersion of **U** comes from the elements of $\mathbf{M}^{-1}$. We present our results in terms of the coefficients K_ν: the correction to be applied to the results of calculations based upon the assumption of isotropy:

$$K_\nu = \frac{< |U_\nu|^2 >}{< |\tilde{U}_\nu|^2 >}.$$

Taking a delta-function as radial selection due to the thin shell K_ν becomes independent of the radius. The bulk of the contribution to the k integrals is coming from the large scale part of the spectrum, thus we use plain Zel'dovich spectrum ($P(k) = k$). Calculating the integrals we get:

$$K_0 = 1.748 \frac{1}{(1 + 0.89\frac{a_{20}}{a_{00}})^2} \left(1 + 2.14 \left(\frac{a_{10}}{a_{00}} \right)^2 + 1.06 \left(\frac{a_{20}}{a_{00}} \right)^2 + 1.79 \frac{a_{20}}{a_{00}} \right),$$

$$K_{\pm 1} = 1.748 \frac{1 + 0.2(\frac{a_{20}}{a_{00}})^2}{(1 - 0.2(\frac{a_{20}}{a_{00}})^2)^2} \left(1 + 0.30 \left(\frac{a_{10}}{a_{00}} \right)^2 + 0.37 \left(\frac{a_{20}}{a_{00}} \right)^2 - 0.89 \frac{a_{20}}{a_{00}} \right).$$

Due to extinction, galaxy groups, the location of the telescope the dipole component in the galaxy distribution can be as large as 1/3. The Burnstein position data give 0.25; analyzing the different shells of de Vaucouleurs and Peters we got values about 0.3. The quadrupole term is amounting to 0.5 considering the

Burnstein data. All these give a 91 per cent systematic error in the longitudinal component and 35 per cent in the transverse one.

In conclusion, we have made several improvements over the usual ways of calculating large scale motions in the Universe. We took into account that observers measure radial velocities, this resulted in a tensor window function. Furthermore, we have calculated the effects of a non-isotropic galaxy distribution previously not considered and we have shown that it has important consequences on the expected motions: the direction of the velocity will be correlated with the dipole moment of the galaxy distribution.

References

1. M.Clutton-Brock and P.J.E. Peebles, Astr. J., *86*, 1115, 1981.
2. N. Kaiser, Astrophys. J. Lett., *273*, L17, 1983.
3. V.C. Rubin, W.K. Ford, N. Thonnard and M.S. Roberts, Astr. J., *81*, 687, 1976.
4. L. Hart and R.D. Davies, Nature, *297*, 191, 1976.
5. D. Burnstein, R.L. Davies, A. Dressler, S.M. Faber, D. Lynden-Bell, R. Terlevich and G. Wegner, Galaxy distances and deviations from universal expansion (NATO ASI Series) ed. B. Tully, Reidel, Dordrecht, 1986, p. 123.
6. C.A. Collins, R.D. Joseph and N.A. Robertson, ibid. p. 131.
7. G. de Vaucouleurs and W.L. Peters, Astrophys. J., *287*, 1, 1984.
8. P.J.E. Peebles, The large scale structure of the Universe, Princeton Series in Physics, Princeton, 1980.
9. B. Grinstein, J. Politzer and R. Wise, Caltech preprint CALT-68-1358, 1986.

Acta Physica Hungarica 64 (1-3), pp. 285-308 (1988)

INFLATION IN THE UNIVERSE, CIRCA 1986*

MICHAEL S. TURNER

Fermilab Astrophysics, Fermi National Accelerator Laboratory
Batavia, IL 60510 USA
and
Department of Physics and Astronomy and Astrophysics
Enrico Fermi Institute, The University of Chicago
Chicago, IL 60637 USA

(Received 8 January 1987)

The hot big bang cosmology, or the standard cosmology as it is appropriately known, is a highly successful model, providing a reliable and tested accounting of the Universe from 0.01 s after the bang until today, some 15 Gyr later. However, very special initial data seem to be required in order to account for the observed smoothness and flatness of our Hubble volume and for the existence of the small primeval density inhomogeneities required for the formation of structure in the Universe. Inflation offers a means of accounting for these special initial data which is based upon physics at sub-Planck energy scales ($\ll m_{pl} \simeq 10^{19}\,\mathrm{GeV}$) and is motivated by contemporary ideas in particle theory. Here I review the status of the "Inflationary Paradigm," Circa 1986. At present essentially all inflationary models involve a very weakly-coupled (quantified by the presence of a dimensionless parameter of order 10^{-12} or so) scalar field which is displaced from the minimum of its potential. Regions of the Universe where the scalar field is initially displaced from its minimum undergo inflation as the scalar field relaxes, resulting in a Universe today which resembles ours in regions much larger than our present Hubble volume ($\simeq 10^{28}\,\mathrm{cm}$), but which on very large scales ($\gg 10^{28}\,\mathrm{cm}$) may be highly irregular. At present, the most conspicuous blemish on the paradigm is the lack of a compelling particle physics model to implement it. I also briefly review some other unresolved issues facing inflation, including the confrontation between the predictions of inflation and observational data.

Preface

George Marx is one of the physicists who pioneered research at the interface of particle physics and cosmology. In 1972 he and Sándor A. Szalay wrote a paper speculating about the role of massive neutrinos in cosmology and placing a cosmological bound on the mass of the neutrino (Marx and Szalay [1]). The paper has become a classic reference in the field, and it could be argued that this paper marked the birth of the field of cosmology and particle physics. Physics at the interface of cosmology and elementary particle physics is now flourishing. The idea of the neutrino-dominated Universe has spawned other-ino (or WIMP) dominated Universe models. One of the other flowers that has blossomed in this field is the

*Dedicated to Prof. G. Marx on his 60th birthday

inflationary Universe scenario. From the very beginning George has taken a keen interest in the inflationary Universe scenario and so I have chosen as the topic of my contribution to this volume celebrating George's 60th birthday, the present status of the inflationary Universe. Because the inflationary scenario predicts $\Omega = 1.0$ and primordial nucleosynthesis precludes $\Omega_{\rm baryon} \gtrsim 0.15$, an inflationary Universe must be dominated by something other than baryons, the most promising candidate being relic WIMPs. One of the toughest challenges facing the inflationary scenario is reconciling the prediction of $\Omega = 1.0$ with the observational evidence that on scales $\lesssim 30$Mpc $\Omega_{\lesssim 30} \simeq 0.2 \pm 0.1$. The neutrino-dominated Universe offers a very attractive resolution to this dilemma, as neutrinos, by virtue of their very large damping length, would probably be smoothly distributed on these scales and thereby would not affect the determinations of Ω on these scales.

Successes of the standard cosmology

The standard cosmology is a remarkable achievement. Based upon the Friedmann–Robertson–Walker (FRW) homogeneous and isotropic cosmological model, it provides us with an accurate description of the evolution of the Universe from about 10^{-2}s after the bang (when the temperature of the Universe was about 10 MeV) until the present some 3×10^{17}s later (and temperature 2.75 K). Support for the standard cosmology is based upon a triad of observations. First, the isotropic Hubble flow and homogeneous distribution of galaxies; light from the most distant galaxies and QSO's (redshifts of order 3–4) left these objects only a few billion years after the bang and therefore test the model to within a few billion years of the bang. Second, there is the cosmic microwave background radiation (CMBR) whose spectrum is consistent with that of a black body at a temperature of 2.75 ± 0.05K (Smoot et al [2]; Peterson et al [3] and which is spatially uniform to about a part in 10^4 on angular scales from a few arcminutes to 180 degrees (Wilkinson [4]). [The only anisotropy unambiguously detected thus far is the dipole component whose magnitude is of order 10^{-3} and whose simplest interpretation is being due to our motion with respect to the cosmic rest frame.] The surface of last scattering for the CMBR is the Universe at $t \simeq 10^{13}$ s and $T \simeq 1/3$ eV, and so it provides a probe of the standard model to within a few 100,000 yrs of the bang. Finally, there is the concordance of the cosmic abundances of D,^{3}He,^{4}He, and ^{7}Li with the predictions of big bang nucleosynthesis, providing that the present baryon-to-photon ratio $\eta \simeq (4 - 7) \times 10^{10}$ (equivalently, $0.014 \lesssim \Omega_B h^2 \lesssim 0.035$, where as usual $H_0 = 100h\text{kms}^{-1}\text{Mpc}^{-1}$) (Yang et al [5]; Boesgaard and Steigman [6]). According to the standard cosmology there was an epoch of primordial nucleosynthesis from about $t \simeq 0.01$s–300s ($T \simeq 10$MeV–0.1MeV), and so the cosmic abundances of these light elements serve to test the model at times to within a fraction of a second after the bang. In addition, the standard cosmology provides a general framework for understanding how structure in the Universe evolved (see, e.g., Efstathiou and Silk [7]): once the Universe became matter-dominated ($t \simeq 10^{10}$s and $T \simeq 10$eV),

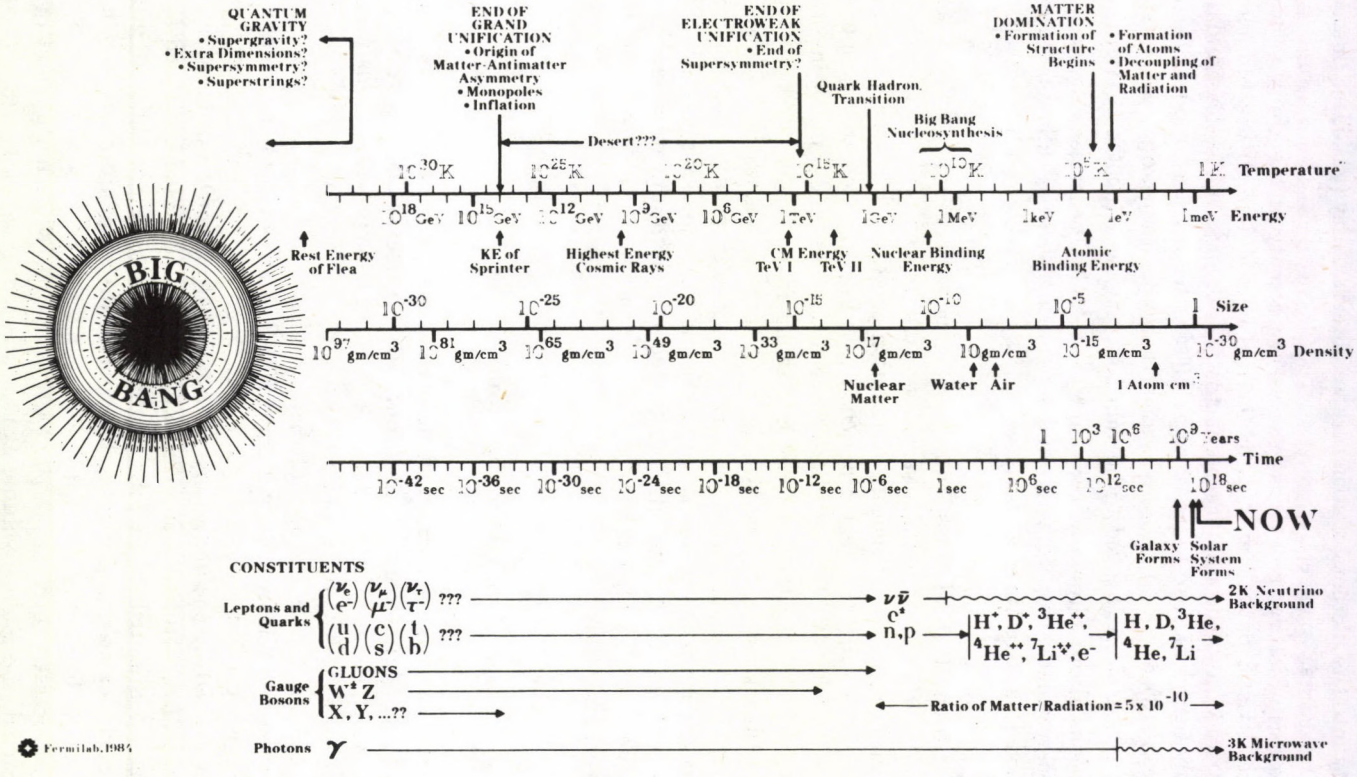


Fig. 1. The Complete History of the Universe, according to the Standard Cosmology and currently fashionable ideas in elementary particle physics

small ($\simeq 10^{-4}$) primeval density inhomogeneities grew via the Jeans (or gravitational) instability into the plethora of structure we observe today (galaxies, clusters of galaxies, superclusters, voids, etc.).

During its early history ($t \leq 10^{10}$s) the energy density of the Universe was dominated by relativistic particle species in thermal equilibrium with a temperature $T \simeq (t/\text{s})^{1/2}$MeV, and the cosmic scale factor $R(t) \propto t^{1/2}$. While the standard cosmology is only tested back to times of order 10^{-2}s, the standard model of particle physics,the $SU(3)_c \otimes SU(2)_L \otimes U(1)_Y$ gauge theory of the strong, weak, and electromagnetic interactions (belived to be valid at energies $\lesssim$ 1000GeV), and theoretical speculations about physics at very high ($\gtrsim 10^{14}$GeV) energies (e.g., grand unification, supersymmetry/supergravity, and superstring theories) allow us to extrapolate the model back to times as early as 10^{-43}s and perhaps even earlier (see Fig.1).

[At times earlier than 10^{-43}s (corresponding to temperatures $\gtrsim 10^{19}$GeV) quantum gravitational effects should become very important and extrapolation to times this early necessarily requires a quantum description of gravity.] The speculations have proven very interesting, from extra dimension, to baryogenesis, to monopoles, to cosmic strings, to relic WIMPs, to phase transitions, and finally inflations, the subject of this paper. Of course, all of these interesting speculations could prove to be nothing more than that; however, nothing in our present knowledge of physics would tell us that such speculations are *a priori* wrong. Compare this to the situation some 20 years ago when it was thought that hadrons were fundamental: at times earlier than about 10^{-5}s after the bang inter-particle distances should have been less than the size of a typical hadron, thus precluding sensible speculations about times earlier than this.

Shortcomings of the standard model

As successful as it is, the standard model has its shortcomings. They involve a number of very fundamental facts about the Universe we observe within our Hubble volume, which it can accommodate, but by no means provides fundamental explanations for (in contrast, the standard cosmology provides a fundamental explanation for the abundance of the light elements). These *cosmological conundrums* are by this time very well known; they include: (1) The smoothness (isotropy and homogeneity) of our present Hubble volume (radius $\simeq H^{-1} \simeq 10^{28}$cm) on scales $\gg$ 10Mpc, as evidenced by the uniformity of the CMBR and of the distribution of galaxies. The size of our Hubble volume is conveniently quantified by the entropy within it, which is dominated by the relic photon and neutrino seas and is of order 10^{88}. Because of the existence of particle horizons in the standard cosmology it is essentially impossible to account for a smooth volume this large as having evolved due to physical processes operating in the early Universe: when matter and radiation last interacted, the Hubble volume at the time contained an entropy of only about 10^{82}, so that particle interactions could not account for such a large smooth

volume. (2) The origin and nature of the primeval fluctuations required to explain the rich array of structure in the Universe today; curvature fluctuations of order 10^{-4} on mass scales $10^{8}M_{\odot} - 10^{15}M_{\odot}$ are required. In the standard cosmology, curvature fluctuations cannot arise spontaneously (again because of the existence of particle horizons; see, e.g., Bardeen [8] and must be put in *ab initio*. [It is possible that the requisite fluctuations are isocurvature fluctuations and were created during the early history of the Universe, perhaps during a phase transition; a promising example which has attracted a great deal of attention lately is cosmic strings (see e.g., Vilenkin [9]; Albrecht and Turok[10]; Turok [11]).] (3) The apparent flatness of our Hubble volume; the radius of curvature ($R_{\rm curv} \equiv R(t)|k|^{-1/2} = H^{-1}/|\Omega - 1|^{1/2}$) in our vicinity must be at least comparable to the radius of our Hubble volume.Had the radius of curvature been of order the Planck length ($\simeq 10^{-33}$cm) at the Planck time ($\simeq 10^{-43}$s), it would only be of order 0.1cm when the Universe reached a temperature of 3K. [Put another way, in order that Ω still not be too different from unity, at the Planck time it must have been equal to 1 to within a part in 10^{60}.] (This dilemma and the naturalness of the flat, Einstein–deSitter model has been emphasized by Dicke and Peebles [12]. (4) The net baryon number within our Hubble volume, quantified as the baryon-to-entropy ratio, $n_B/s \simeq \eta/7 \simeq (7-10) \times 10^{-11}$. Of course one of the great successes of the InnerSpace/OuterSpace connection is baryogenesis, the modern theory of the origin of the baryon asymmetry, and it appears that now we at least have a framework for understanding the origin of this very fundamental quantity (for a review see Kolb and Turner [13]). (5) The dearth (thank goodness) of monopoles and other topological beasts which would have been produced in great excess during the earliest moments of the Universe ($t \leq 10^{-34}$s) had the standard model been valid and if the interactions of nature are unified by a semi-simple gauge group, such as SU_5, SO_{10}, or E_6 (for a review see Preskill [14]). (6) The smallness of the present cosmological term. With the possible exception of supersymmetry, no symmetry forbids such a term in the Einstein–Hilbert action, and so on dimensional grounds one would expect such a term to be of order m_{pl}^2, corresponding to a vacuum energy of order m_{pl}^4. In any case, contributions to the vacuum energy of order M^4 arise due to spontaneous symmetry breaking (SSB) at the energy scale M. The measured expansion rate of the Universe restricts the present vacuum energy contribution to be $\lesssim 10^{-46}\text{GeV}^4$. Even the contribution from chiral symmetry breaking ($M \simeq few100\text{MeV}$)—a phenomenon that particle physicists think they know something about, violates this bound by some 42 orders-of-magnitude!

All of these cosmological facts can be accommodated by the standard model, but seemingly at the expense of highly special initial data (the possible exception being the monopole problem). In 1973, Collins and Hawking [15] pointed out that the set of initial data which evolve to a Universe such as ours is of measure zero providing that the stress energy in the Universe has always satisfied the strong and dominant energy conditions. Over the years there have been a number of attempts to try to understand and/or explain this apparent dilemma of initial data. Inflation is the most recent attempt and I believe shows great promise. Let me begin by briefly mentioning the earlier attempts:

-*Mixmaster paradigm.* Starting with a solution with a singularity which exhibits the features of the most general singular solutions known (the so-called mixmaster model) Misner and his coworkers hoped that they could show that particle viscosity would smooth out the geometry. In part because horizons still effectively exist in the mixmaster solution this program has proven unsuccessful (see, e.g., Misner [16], [17]; Matzer and Misner [18]).

-*Nature of the initial singularity.* Penrose [19] explored the possibility of explaining the observed smoothness of the Universe by restricting the kinds of initial singularities which are permitted in Nature (those with vanishing Weyl curvature). In a sense his approach is to postulate a law of physics governing allowed initial data.

-*Quantum gravity effects.* The first two solutions involve appealing to classical gravitational effects. A number of authors have suggested that quantum gravity effects might be responsible for smoothing out the space-time geometry (deWitt [20]; Parker [21]; Zel'dovich [22]; Starobinskii [23]; Anderson [24]; Hartle and Hu [25]; Fischetti et al. [26]). The basic idea being that anisotropy and/or inhomogeneity would drive gravitational particle creation, which due to back reaction effects would eliminate particle horizons and smooth out the geometry. Recently, Hawking and Hartle [27] have advocated the Quantum Cosmology approach to actually compute the initial state. All of these approaches necessarily involve events at times $\lesssim 10^{-43}$s and energy densities $\lesssim m_{pl}^4$.

-*Anthropic principle.* Some (Carr and Rees [28]; Barrow and Tipler [29]) have suggested (or in some cases even advocated) "explaining" many of the puzzling features of the Universe around us (and in some cases, even the laws of physics!) by arguing that unless they were as they are intelligent life would not have been able to develop and observe them! Hopefully we will not have to resort to such an explanation.

The approach of inflation is somewhat different from previous approaches. Inflation (at least from my point-of-view) is based upon well-defined and reasonably well-understood microphysics (albeit, some of it very speculative). That microphysics is:

- Classical Gravity (general relativity), at least as an effective, low energy theory of gravitation;

-"Modern Particle Physics"—grand unification, supersymmetry /supergravity, field theory limit of superstring theories at energy scales $\lesssim m_{pl}$.

As I will emphasize, in all viable models of inflation the inflationary period (at least the portion of interest to us) takes place well after the Planck epoch, with the energy densities involved being far less than m_{pl}^4 (although semi-classical quantum gravity effects might have to be included as non-renormalizable terms in the effective Lagrangian). Of course, it could be that a resolution to the cosmological puzzles discussed above involves both "modern particle physics" and quantum gravitational effects in their full glory (as in a fully ten dimensional quantum theory of strings).

I will not take the space here to review the historical development of our present view of inflation; I refer the interested reader to the interesting paper on this

subject by Lindley [30]. It suffices to say that Guth's very influential paper of 1981 [31] initiated the inflation revolution, and that Guth's doomed original model (Guth and Weinberg [32]; Hawking et al. [33]) was revived by Linde's [34] and Albrecht and Steinhardt's [35] variant, "new inflation". I will focus all of my attention on the present status of the "slow-rollover" model of Linde [34] and Albrecht and Steinhardt [35].

Basic mechanics of new inflation

Stated in the most general terms, the current view of inflation is that it involves the dynamical evolution of a very weakly-coupled scalar field (hereafter referred to as ϕ) which is, for one reason or another, initially displaced from the minimum of its potential (see Fig. 2).While it is displaced from its minimum, and is slowly-evolving toward that minimum, its potential energy density drives the rapid (exponential) expansion of the Universe, now known as inflation.

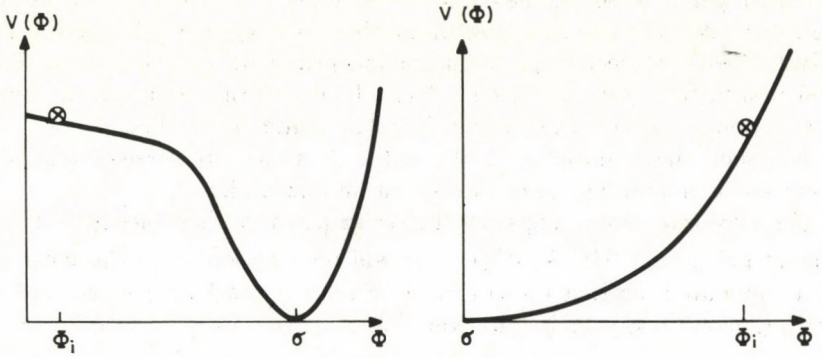


Fig. 2. Stated in the most general terms, inflation involves the dynamical evolution of a scalar field which was initially displaced from the minimum of its potential, be that minimum at $\sigma = 0$ or $\sigma \neq 0$

The usual assumptions which are made (often implicitly) in order to analyze inflation are:

–A FRW spacetime with scale factor $R(t)$ and expansion rate

$$H^2 \equiv (\dot{R}/R)^2 = 8\pi\rho/3m_{pl}^2 - k/R^2, \tag{1}$$

where the energy density is assumed to be dominated by the stress energy associated with the scalar field (in any case, other forms of stress energy rapidly redshift away during inflation and become irrelevant).

–The scalar field ϕ is spatially constant (at least on a scale $\gtrsim H^{-1}$) with initial value $\phi_i \neq \sigma$, where $V(\sigma) = V'(\sigma) = 0$, and stress energy tensor

$$T^{\mu}_{\nu} = \text{diagonal}(-\rho, p, p, p), \tag{2a}$$

$$\rho = V + \dot{\phi}^2/2(+\nabla\phi^2/2R^2), \tag{2b}$$

$$p = -V + \dot{\phi}^2/2(-\nabla\phi^2/6R^2). \tag{2c}$$

(I have indicated the contribution of the spatial gradient terms for future reference.)

– The semi-calassical equation of motion for ϕ provides an accurate description of its evolution; more precisely,

$$\phi(t) = \phi_{cl}(t) + \Delta\phi_{QM}, \tag{3a}$$

$$\ddot{\phi}_{cl} + 3H\dot{\phi}_{cl} + \Gamma\dot{\phi}_{cl} + V'(\phi_{cl}) = 0, \tag{3b}$$

where the quantum fluctuations (characterized by size $\Delta\phi_{QM} \simeq H/2\pi$) are assumed to be a small perturbation to the classical trajectory $\phi_{cl}(t)$, and Γ is the decay width of the ϕ particle. Throughout I use units where $\hbar = k_B = c = 1$; overdot indicates a derivative with respect to proper time and prime with respect to ϕ. From this point forward I will drop the subscript "*cl*". I will return later to these assumptions to discuss how they have been or can be relaxed and/or justified.

The semi-classical evolution of ϕ naturally splits into three phases: slow-roll; coherent scalar field oscillations; and quantum fluctuations.

(1) *Slow-roll.* Assuming that the scalar potential is sufficiently flat (the requirement being that $9H^2 \gtrsim |V''|$) there will be a period when the motion of ϕ is friction-dominated, so that the $\ddot{\phi}$ term in Eq. (3b) can be neglected and $\dot{\phi}^2 \ll V$ can be neglected in Eq. (2b). The equation of motion for ϕ becomes

$$3H\dot{\phi} \simeq -V',$$

$$H^2 \simeq 8\pi V(\phi)/3m_{pl}^2.$$

The growth of the scale factor during the slow-roll is approximately exponential since $H \simeq$ constant, and

$$R_f/R_i = \exp(\int H\,dt) \equiv \exp(N), \tag{4a}$$

$$N \simeq -3\int H^2 d\phi/V' \simeq 8\pi m_{pl}^{-2}\int -V(\phi)d\phi/V'(\phi). \tag{4b}$$

The exponential growth or inflation occurs during the slow-roll phase.

(2) *Coherent scalar field oscillations.* As ϕ approaches σ the potential steepens sufficiently (or $V(\phi)$ becomes sufficiently small) so that the motion of ϕ is no longer friction-dominated (which occurs when $|V''| \gtrsim 9H^2$) and ϕ begins to oscillate about

the minimum of the potential $(\phi - \sigma)$ on a timescale ($\simeq |V''|^{-1/2} = m_\phi^{-1}$) short compared to the Hubble time. During this phase its oscillatory motion can be time-averaged, and the equation for its evolution becomes

$$\dot{\rho}_\phi + 3H\rho_\phi = 0, \tag{5}$$

whose solution is $\rho_\phi \propto \exp(-\Gamma t)/R^3$–precisely that of an unstable, NR particle species. The coherent field oscillations behave like NR matter (as they should, since they correspond to the zero momentum mode of the field), and decay in a time Γ^{-1} ($= \tau_\phi$, the lifetime of the ϕ particle) due to particle creation by the oscillating ϕ field. From the particle point-of-view, the oscillations correspond to a very cold condensate of ϕ particles, which then decay. During this phase the scale factor $R(t) \propto t^{2/3}$. Assuming that the decay products thermalize quickly (or at least are relativistic) and neglecting any relativistic particles present before inflation as they 've been exponentially diluted, the evolution of the energy density in radiation produced by the decay of the coherent field oscillations is governed by

$$\dot{\rho}_R + 4H\rho_R = \Gamma\rho_\phi. \tag{6}$$

The evolution of ρ_ϕ, ρ_R, and the entropy per comoving volume ($\equiv S \propto R^3\rho_R^{3/4}$) are shown in Fig. 3.

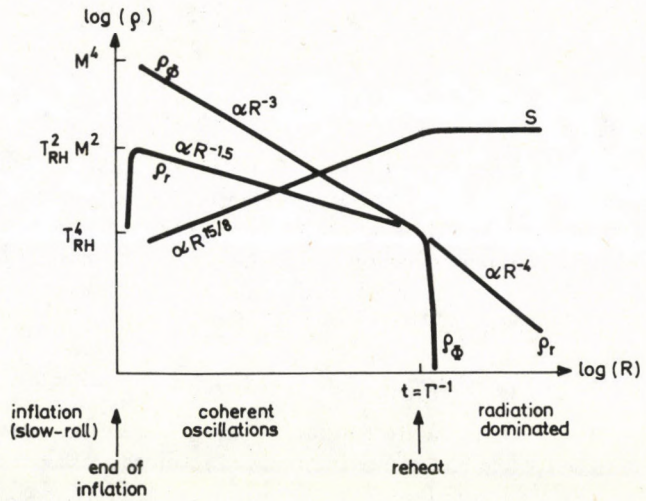


Fig. 3. The evolution of the energy density in the scalar field (ρ_ϕ), in radiation (ρ_R), and of the entropy per comoving volume, $S \propto R^3\rho_R^{3/4}$

At early times, $t \ll \Gamma^{-1}, \rho_r \propto t^{-1}$ and $S \propto t^{5/4}$. The entropy per comoving volume levels off when most of the ϕ particles have decayed, at $t \simeq \Gamma^{-1}$, and thereafter the

energy density in coherent field oscillations decreases exponentially. The reheating process is essentially complete at this time (save for the possible thermalization of the decay products of the ϕ particles) and the temperature is about

$$T_{RH} \simeq (\Gamma m_{pl})^{1/2}. \tag{7}$$

Figure 3 summarizes the evolution of ϕ and the reheating process. For further discussion see Turner [36].

(3) *Quantum fluctuations.* During inflation ϕ is on the flat part of its potential and can be treated as an effectively massless scalar field. The spectrum of de Sitter space quantum fluctuations is given by (Bunch and Davies [37])

$$\Delta\phi^2_{QM} \equiv |\delta\phi_k|^2 k^3/(2\pi)^3 = H^2/16\pi^3, \tag{8}$$

where $\delta\phi_k$ is the kth Fourier component of $\delta\phi$ and $k_{\rm phys} = k/R(t)$ is the physical wavenumber of the mode with comoving wavenumber k. For scales $k_{\rm phys} \gtrsim H^{-1}$ these perturbations are treated quantum mechanically; as $k_{\rm phys}$ becomes $\lesssim H^{-1}$, and a given scale crosses outside the "physics horizon" (or Hubble radius) H^{-1} the fluctuation on that scale is taken to evolve classically thereafter (the quantum fluctuation is *assumed* to "freeze in" as a classical perturbation in the metric). The evolution of the metric perturbations due to the classical fluctuations is straightforward to compute. They give rise to curvature (scalar mode) fluctuations of amplitude (Bardeen et al [38]; Guth and Pi [39]; Starobinskii [40]; Hawking [41])

$$(\delta\rho/\rho_{HOR}) \simeq k^{3/2}|\delta_k|/(2\pi)^{3/2} \simeq \begin{cases} (H^2/\pi^{3/2}\dot{\phi})|_{t_1}, & \text{(9a)} \\ (H^2/10\pi^{3/2}\dot{\phi})|_{t_1}, & \text{(9b)} \end{cases}$$

when they reenter the horizon. Perturbations which reenter when the Universe is radiation-dominated do so as pressure waves (in the baryon–photon fluid) of the amplitude indicated in Eq. (9a); those which reenter when it is matter-dominated do so as growing mode perturbations with the amplitude indicated in Eq. (9b).

Tensor mode metric (gravitational wave) perturbations also arise and are of dimensionless amplitude (Abbott and Wise [42]; Starobinskii [43]; Rubakov et al [44])

$$h_{GW|HOR} \simeq (H/m_{pl})|_{t_1}, \tag{10}$$

when they reenter the horizon after inflation. In both cases the quantities on the rhs are to be evaluated when the comoving scale of interest crossed outside the Hubble radius ($k/R \simeq H$), at $t = t_1$ (see Fig. 4). Normalizing $R_{\rm today} = 1$, it follows that a given scale $k(\equiv 2\pi/\lambda)$ crossed outside the horizon during inflation

$$N(\lambda) \simeq 45 + \ln(\lambda/{\rm Mpc}) + 2\ln(M/10^{14}{\rm GeV})/3 + \ln(T_{RH}/10^{10}{\rm GeV})/3 \tag{11}$$

Hubble times (or e-folds) before the end of inflation (see Fig. 4). Thus the scales of astrophysical interest crossed outside the horizon 40 ± 10 or so Hubble times before the end of inflation. Although H and $\dot{\phi}$ can vary considerably during inflation, over

such a small range of Hubble times both H and $H^2/\dot{\phi}$ are essentially constant (vary by less than a factor of 2 in all the models I have studied) and so a generic prediction of inflationary models is scale invariant curvature and tensor perturbations. [In models where there are other massless fields, quantum fluctuations of order $H/2\pi$ arise in these fields too during inflation and can give rise to isocurvature (often called isothermal) perturbations. A simple example being axion models where the quantum fluctuations in the axion field result in fluctuations in the local axion to entropy density ratio n_a/s (see Steinhardt and Turner [45]; Linde [46]; Turner [47]).] The mechanics of inflation are described in much greater detail in the review by Turner [48].

Successfully implementing inflation

Now that I have discussed the basic mechanics of inflation, how does one build a model which actually leads to a Universe which resembles ours in regions as large as our current Hubble volume? It's as easy as 1,2,3,4,5,6,7,8,...

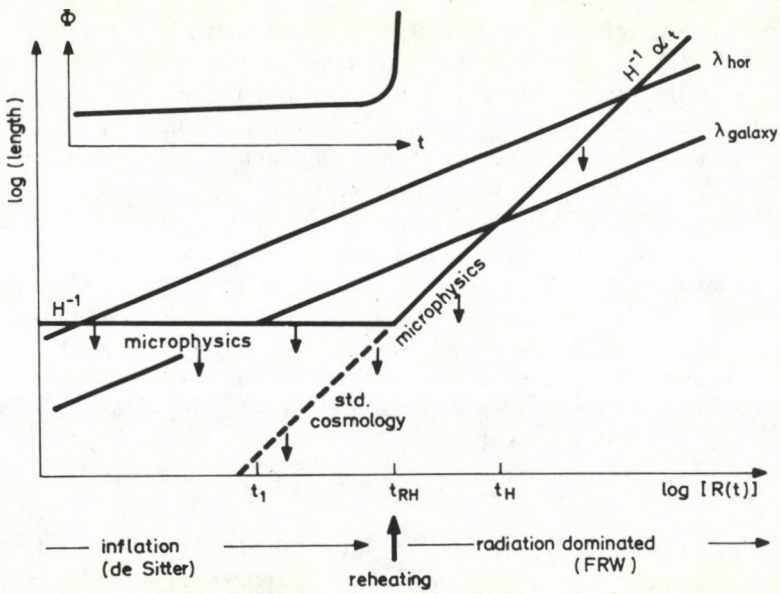


Fig. 4. The evolution of the physical wavelength of a given mode. Early during inflation $\lambda_{phys} \lesssim H^{-1}$ and the scale is inside the physics horizon and fluctuations are treated quantum mechanically. As the mode leaves the physics horizon ($\lambda_{phys} \geq H^{-1}$), the perturbation is assumed to "freeze in" as a classical metric perturbation

(1) *Smoothness/Flatness.*—After inflation there must be smooth/flat regions of the Universe which contain an entropy of at least 10^{88}. Assuming that ϕ was constant in a region of size of order the Hubble radius before inflation, it is straightforward to show that the number of e-folds of inflation required so that the patch contains an entropy of at least 10^{88} after reheating is

$$N \equiv \int H dt \simeq \int -3H^2 d\phi/V' \gtrsim 53 + 2\ln(M/10^{14}\mathrm{GeV})/3 + \ln(T_{RH}/10^{10}\mathrm{GeV})/3. \tag{12}$$

The number of e-folds required to solve the "flatness problem" is equal to this plus $\ln|1-\Omega_i^{-1}|$ where Ω_i is the value of Ω at the beginning of inflation. In general, if a given model inflates at all, it is not difficult to make it inflate enough to solve the smoothness/flatness problems.

(2) *Sufficiently large smooth patch.* In order that the vacuum energy contribution to the energy density of the Universe dominates the gradient terms $((\nabla\phi)^2)$, ϕ must be smooth over a sufficiently large region. [If the gradient term dominates, the stress energy of the scalar field behaves like a fluid with $p = -\rho/3$ and scale factor only grows as $R(t) \propto t$, not exponentially.] The condition that $V \gg (\nabla\phi)^2$ requires ϕ to be smooth over a region of size greater than $\phi_i/V^{1/2} \simeq (\phi_i/m_{pl})H^{-1}$.

(3) *Validity of semi-classical description.* In order that the semi-classical equations of motion be self consistent $\Delta\phi_{QM} \sim H$ must be much less than ϕ_{cl}. I will return to this point later — in general it is not a difficulty.

(4) *Topological beasts.* One must be careful not to produce any of the dangerous topological beasts such as domain walls, monopoles, etc. after inflation. This can be arranged by having the symmetry breaking stages which result in the formation of such objects occur before or early on during inflation.

(5) *Other unwanted garbage.* As is well-known the ratio of the energy density in nonrelativistic (NR) particles relative to that in relativistic (R) particles grows with time: $\rho_{NR}/\rho_r \propto R(t)$. Since our Universe did not become matter-dominated until rather recently, $T \simeq 10$ eV and $t \simeq 10^{10}$s, the ratio of energy density in stable, NR particles (which do not come into thermal equilibrium, or annihilate) to that in R particles after inflation must be small:

$$\rho_{NR}/\rho_R \lesssim 10^{-19}(10^{10}\mathrm{GeV}/T_{RH}) \tag{13}$$

— which is not always an easy thing to do. [Just ask any experimentalist about suppressing something by 19 orders-of-magnitude!] Examples of potentially dangerous forms of NR matter include: gravitinos, weakly-coupled scalar fields which may be put into oscillation after inflation (the so-called Polonyi problem; e.g., see Coughlan et al [49]), etc. There are even stronger limits on unstable NR particles (particularly gravitinos) which follow by considering their effect on primordinal nucleosynthesis (Ellis et al [50]; Khlopov and Linde [51]; Scherrer and Turner [52]).

(6) *Metric perturbations.* While the eventual development of structure in the Universe requires density perturbations of the order of *few* x 10^{-5} or so, the observed isotropy of the CMBR precludes scalar or tensor perturbations of size

greater than about 10^{-4}. Achieving tensor mode perturbations of this size or smaller is not difficult; it only requires that inflation occur at an energy scale $M \simeq V^{1/4} \lesssim 10^{-2} m_{pl}$, putting an upper limit on the reheating temperature: $T_{RH} \lesssim 3 \times 10^{16}$GeV. The scalar perturbations are another matter; thus far, they have posed the most serious obstacle to constructing a successful model of inflation. To achieve curvature fluctuations of amplitude less than about 10^{-4} requires a dimension parameter in the scalar potential of the order of 10^{-12}. For example, for a potential of the form $V = \lambda\phi^4$, λ must be $\lesssim 10^{-13}$; for the potential $V = V_0 + \alpha\phi^2 - \beta\phi^3 + \lambda\phi^4$, $\alpha \lesssim 10^{-15} m_{pl}^2$, $\lambda \lesssim 10^{-11}$, $\beta \lesssim 3 \times 10^{-12} m_{pl}$; for the potential $V = m^2\phi^2$ (i.e., a non-interacting, massive scalar field), $m^2 \lesssim 10^{-9} m_{pl}^2$. Achieving density perturbations of an acceptable amplitude necessitates a very weakly-coupled scalar field; weakly-coupled to all fields in the theory or else radiative corrections would spoil the small coupling put in at tree level. Such a small coupling also implies that inflation takes place at an energy scale much less than the Planck scale; typically, $V \lesssim \lambda\phi^4 \simeq 10^{-12} m_{pl}^4$ or less.

(7) *Sufficient reheating.* In order not to spoil the concordance of the predictions of primordial nucleosynthesis with the observed abundances, the Universe should be radiation–dominated when $t \simeq 0.01$s and $T \simeq 10$MeV at the very latest, i.e., $T_{RH} \gtrsim 10$MeV. This implies that Γ must be $\gtrsim 10^{-23}$GeV—which is not difficult to achieve, even for a very weakly-coupled scalar field. Baryogenesis, however, poses a more formidable challenge. It goes without saying that baryogenesis must follow inflation , as any baryon asymmetry produced before inflation is diluted away by the enormous entropy produced by inflation. If baryogenesis is to proceed in the usual way, T_{RH} must be gerater than about a $m_H/10$, where m_H is the mass of the superheavy boson whose out-of-equilibrium decays produce the baryon asymmetry (see Kolb and Turner [13]). In most unified theories the longevity of the proton requires the masses of superheavy bosons whose interactions violate B-conservation to be greater than about 10^{10}GeV, thereby requiring $T_{RH} \gtrsim 10^9$GeV at the very least. Because the scalar field must be very weakly-coupled (to produce density perturbations of an acceptable magnitude) Γ tends to be very small, and sufficient reheating is often very difficult to achieve. There is, however, an alternative method which does not require such a high reheating temperature: the direct production of the baryon asymmetry by the decays of the scalar field responsible for inflation ($\phi \to q's, l's$ with $\Delta B \neq 0$). In this case the baryon asymmetry produced is

$$n_B/s \simeq \epsilon T_{RH}/m_\phi, \tag{14}$$

where ϵ is the magnitude of the requisite C, CP violation in the decay of the ϕ particle. Note that the asymmetry produced only depends upon the ratio of the reheating temperature to the mass of the ϕ particle, and so it is possible to have a relatively low reheating temperature and still a baryon asymmetry of the required magnitude.

(8) *Part of a unified model which predicts sensible particle physics.* In order to avoid having the tail wag the dog so to speak, the scalar field should be part of a unified theory which predicts sensible particle physics.

Specific models

While the requirements on a successful model of inflation are straightforward they are not simple to satisfy simultaneously, and the path to a successful model is strewn with the remains of many an attractive model that failed for one reason or another. The toughest challenge has been the constraint imposed by the scalar density perturbations. At present there are no successful models which are so elegant as to be compelling (compelling here, meaning attractive to other than the authors of the model!), although there do exist a number of "proof of existence" models. I will describe two particularly simple models here.

(1) *Shafi–Vilenkin–Pi model.* This model (Shafi and Vilenkin [53]; Pi [54]) is based upon an SU_5 nonsupersymmetric GUT (although the gauge group could just as well be SO_{10} or E_6). The scalar field responsible for inflation is a complex Higgs gauge singlet, which in addition to being responsible for inflation, also breaks a Peccei–Quinn (PQ) symmetry and effects GUT symmetry breaking by inducing a negative mass squared for the adjoint Higgs representation which then breaks the GUT down to SU(3) ⊗ SU(2) ⊗ U(1). The part of the Higgs potential which is relevant for inflation is a Coleman–Weinberg type potential

$$V(\phi) = B\sigma^4/2 + B\phi^4[\ln(\phi^2/\sigma^2) - 1/2], \tag{15}$$

where $\sigma \simeq 10^{18}$GeV is the vacuum expectation value of ϕ at low temperatures, which breaks PQ symmetry and induces GUT SSB; $B \simeq 10^{-15}$ arises due to radiative corrections from coupling of ϕ to other scalar fields in the theory. Interestingly enough isocurvature axion fluctuations of similar magnitude to the curvature fluctuations also arise in this model (Seckel and Turner [47]).

(2) *Florida SUGR.* This model (Holman et al [55]) is based on an effective low energy supergravity theory with a superpotential of the form, $I + S + G$, where the three pieces of the superpotential are responsible for inflation, supersymmetry breaking and GUT symmetry breaking respectively. The inflation piece of the superpotential takes a particularly simple form

$$I = (\Delta^2/M)(\phi - M)^2, \tag{16}$$

where Δ is the only adjustable parameter and corresponds to the intermediate scale and $M = m_{pl}/(8\pi)^{1/2} \simeq 2.43 \times 10^{18}$GeV. The resulting scalar potential for ϕ is given by

$$V_I(\phi) = \Delta^4 \exp(\phi^2/M^2)[\phi^6/M^6 - 4\phi^5/M^5 + 7\phi^4/M^4 - 4\phi^3/M^3 - \phi^2/M^2 + 1]$$

$$\simeq \Delta^4[1 - 4\phi^3/M^3 + 6.5\phi^4/M^4 - 8\phi^5/M^5 + \ldots \tag{17}$$

Achieving density perturbations of an acceptable magnitude requires that $(\Delta/M) \simeq 9 \times 10^{-5}$. Note that the coefficient of the ϕ^4 term in V_I is dimensionless and for this value of Δ is about 4×10^{-16}–the small dimensionless parameter which always arises. Of course, in this model it is directly related to the smallness of the intermediate scale relative to the Planck scale, suggesting (or offering hope) that the very small parameter needed for successful inflation is related to fundamental physics. This model reheats to a temperature of about 10^6GeV and baryogenesis is effected directly through the decays of the ϕ field.

Open (or semi-open) questions

"Who is ϕ?"

Inflationary models exist in which the scalar field ϕ: effects SSB of the GUT (Shafi and Vilenkin [53]; Pi [54]), effects SSB of SUSY (Ovrut and Steinhardt [56,57]), induces Newton's constant (in a Landau–Ginzburg model of induced gravity) (Accetta et al [58]; Spokoiny [59]), is $\sim \ln(r_{XEQ})$ (where r_X is the radius of compactified extra dimensions) in theories with extra dimensions which become compactified (Shafi and Wetterich [60,61]), is $\propto$ (scalar curvature)$^{1/2}$ (Starobinskii [23]; Mijic et al [62]), is just some "random" scalar field (Linde [63]), or is merely in the theory to effect inflation (Holman et al [55]; Nanopoulos [64]). Given the number of different kinds of inflationary scenarios which exist, it seems as though inflation is generic to early Universe microphysics, occurring whenever a weakly-coupled scalar field finds itself displaced from the minimum of its potential. Clearly, a key question at this point is just how "the inflation sector" of the theory fits into the Big Picture!

What determines the initial value of ϕ?

One thing is certain, and that is that ϕ must be very weakly-coupled, as quantified by its small dimensionless coupling constant. Because of this fact, it is almost certain that ϕ was not initially in thermal contact with the rest of the Universe and so ϕ_i is unlikely to be determined by thermal considerations (in the earliest models of new inflation, ϕ_i was determined by thermal considerations, however, these models resulted in density perturbations of an unacceptably large amplitude). At present ϕ_i must be taken as initial data. Some have argued that ϕ_i might be determined in an anthropic-like way, as regions of the Universe where ϕ_i is sufficiently far displaced from equilibrium will undergo inflation and eventually occupy most of the physical volume of the Universe. Perhaps the wavefunction of the Universe approach will shed some light on the initial distribution of the scalar field ϕ. Or it could be that due to "as-of-yet unknown dynamics" ϕ was indeed in thermal equilibrium at a very early epoch. It goes without saying that it is crucial that ϕ be initially displaced from its minimum.

Validity of the semi-classical equations of motion for ϕ

While it may seem perfectly plausible that ϕ evolves according to its semi-classical equations of motion, the validity of this assumption has troubled inflationists from the "dawn of new inflation". While a full quantum field theory treatment of inflation is very difficult and has not been effected, a number of specific issues have been addressed. Several authors have studied the role of inhomogeneities in ϕ, and have found that for the very weakly-coupled fields one is dealing with, mode coupling is not important and the individual modes are quickly smoothed by the exponential expansion of their physical wavelengths (Albrecht and Brandenberger [65]; Albrecht et al [66]). I already mentioned the necessity of having ϕ smooth over a sufficiently large region so that the gradient terms in the stress energy do not dominate.

The effect of quantum fluctuations on the evolution of ϕ has been studied in some detail by Guth and Pi [67], Fischler et al [68], Linde [69], and Vilenkin and Ford [70]. The basic conclusion that one draws from the work of these authors is that the use of the semi-classical equations of motion is valid so long as $\phi_{cl} \gg \Delta\phi_{QM} \simeq N^{1/2}H/2\pi$, which is almost always satisfied for the very flat potentials of interest to inflationists (at least for the last 50 or so e-folds which affect our present Hubble volume). [More precisely, the semi-classical change in ϕ in a Hubble time, $\Delta\phi_{\rm Hubble} \simeq -V'/3H^2 \simeq -V' m_{pl}^2/(8\pi V)$, should be much greater than the increase in $< \phi^2 >_{QM}^{1/2}$, which is of order $H/2\pi$, due to the addition of an another quantum mode; see Bardeen et al [38].] At present the validity of the semi-classical equations of motion seems to be reasonably well established.

No hair conjectures

While inflation has been touted from the very beginning as making the present state of the Universe insensitive to the initial spacetime geometry, not much has been done to justify this claim until very recently. As I mentioned earlier, inflation is nearly always analyzed in the context of a flat, FRW cosmological model, making such a claim somewhat dubious. However, it has now been shown that all of the homogeneous models (with the exception of the highly-closed models) undergo inflation, isotropize and remain isotropic to the present epoch providing that the model would have inflated the requisite 60 or so e-folds in the absence of anisotropy (Turner and Widrow [71]; Jensen and Stein-Schabes [72]).

The proof of this result involves three parts. First, Wald [79] demonstrated that all homogeneous models with a positive cosmological term asymptotically approach deSitter (less the aforementioned highly-closed models which recollapse before the cosmological term becomes relevant). Wald's result follows because all forms of "anisotropy energy density" decrease with increasing proper volume element, whereas the cosmological term remains constant, and so eventually triumphs. Of course, inflationary models do not in the strictest sense, have a cosmological term, rather they have a positive vacuum energy as long as the scalar field is displaced from the minimum of its potential. Thus the dynamics of the scalar field comes

into play: does ϕ stay displaced from the minimum of its potential long enough so that the vacuum energy comes to dominate? Due to the presence of anisotropy the expansion rate is greater than if there were only vacuum energy density, and so the friction felt by ϕ as it trys to roll (the $3H\dot{\phi}$ term) is greater and it takes ϕ longer to evolve to its minimum than without anisotropy. For this reason the Universe does become vacuum dominated before the vacuum energy disappears, and in fact the Universe inflates slightly longer in the presence of anisotropy (one or two e-folds)(Steigman and Turner [74]). Finally, is the anisotropy reduced sufficiently so that the Universe today is still nearly isotropic? As it turns out, the requisite 60 or so e-folds needed to solve the other conundrums reduces the growing modes of anisotropy sufficiently to render them small today.

Allowing for inhomogeneous initial spacetimes makes matters much difficult. Jensen and Stein-Schabes [75] and Starobinskii [76] have proven the analogue of Wald's theorem for spacetimes which are negatively-curved. Jensen and Stein-Schabes [75] have gone on to conjecture that spacetimes which have sufficiently large regions of negative curvature will undergo inflation, resulting in a Universe today which although not globally homogeneous, at least contains smooth volumes as large as our current Hubble volume.

Does this improve the situation that Collins and Hawking discussed in 1973? While the work of Jensen and Stein-Schabes [75] seems to indicate that many inhomogeneous spacetimes undergo inflation and even leads one to speculate that the measure of the set of initial spacetimes which eventually inflate is non-zero, it is not possible to draw a definite conclusion without first defining a measure on the space of initial data. In fact, as Penrose [19] pointed out there is at least one way of defining the measure such that this is not the case. Consider the set of all Cauchy data at the present epoch; intuitively it is clear that those spacetime slices which are highly irregular are the rule, and those which are smooth in regions much larger than our current Hubble volume are the exception. Defining the measure today, it seems very reasonable that the smooth spacetime slices are a set of measure zero. Now evolve the spacetimes back to some initial epoch (for example $t = 10^{-43}$s). Using the seemingly very reasonable measure defined today and the mapping back to "initial" spacetimes, one could argue that the set of initial data which inflate is still of measure zero. While I believe that this argument is technically correct, I also believe that it is silly. First, upon close examination of all of those initial spacetimes which led to spacetimes today without smooth regions as large as our present Hubble volume, one would presumably find that the scalar field in most would be very close to the minimum of its potential (in order that they not inflate)—not a very generic initial condition. Secondly, if one adopts the point-of-view of an evolving Universe which has an "initial epoch" (and not everyone does), then there is a preferred epoch at which one would define a measure — the "initial epoch," and at that epoch I believe any reasonably defined measure would lead to the set of initial spacetimes which inflate being of non-zero measure.

Although it is not possible *yet* to claim rigorously that inflation has resolved the problem of the seemingly special initial data required to reproduce the Universe we see today (at least within our Hubble volume), I think that any fairminded per-

son would admit that it has improved the situation dramatically. Extrapolating from the solid results that exist, it seems to me that starting with a general inhomogeneous spacetime, there will exist regions which undergo inflation and which today are much larger than our present Hubble volume, thereby accounting for the smooth region we find ourselves in. From a more global perspective, one might expect that on scales $\gg H^{-1}$ the Universe would be highly irregular.

The present vanishingly small value of the cosmological constant

Inflation has shed no light on this difficult and very fundamental puzzle (nor has anything else for that matter!). In fact, since inflation runs on vacuum energy so to speak, the fate of inflation hinges upon the resolution of this puzzle. For example, suppose there were a grand principle that dictated that the vacuum energy of the Universe is always zero, or that there were an axion-like mechanism which operated and ensured that any cosmological constant rapidly relaxed to zero; either would be a disaster to inflation shorting out its source of power–vacuum energy. [Another possibility which has received a great deal of attention recently is the possibility that deSitter space might be quantum mechanically unstable—of course, if its lifetime were least 60 some e-folds that would not necessarily adversely affect inflation (Starobinskii [23]; Myhrvad [77]; Mottola [78]; Parker [79]; Ford [80]; Anderson [81]; Traschen and Hill [82].]

Inflation confronts observation

No matter how appealing a theory may be, it must meet and pass the test of experimental verification. Experiment and/or observation is the final arbiter. One of the few blemishes on early Universe physics is the lack, thus far, of experimental/observational tests of the many beautiful and exciting predictions. That situation is beginning to change as the field starts to mature. Inflation is one of the early Universe theories which is becoming amenable to verification or falsification. Inflation makes the following very definite predictions (postdictions?):

– $\Omega = 1.0$ (more precisely, $R_{\rm curv} = R(t)|k|^{-1/2} = H^{-1}/|\Omega - 1|^{1/2} \gg H^{-1}$)

– Harrison–Zel'dovich spectrum of constant curvature perturbations (and possibly isocurvature perturbations as well) and tensor mode gravitational wave mode perturbations.

The prediction of $\Omega = 1.0$ together with the primordial nucleosynthesis constraint on the baryonic contribution, $0.014 \lesssim \Omega_B h^2 \lesssim 0.035 \lesssim 0.15$ (Yang et al [5]), suggests that most of the matter in the Universe must be nonbaryonic. The simplest and most plausible possibility is that it exists in the form of relic WIMPs (*W*eakly-*I*nteracting *M*assive *P*articles, e.g., axions, photinos, neutrinos; for a review, see Turner [83]). Going a step further, these two original predictions then lead to testable consequences:

– $H_0 t_0 = 2/3$ (providing that the bulk of the matter in the Universe today is in the form of NR particles)—The observational data both H_0 and t_0 are far from being definitive: $H_0 \simeq 40 - 100 {\rm kms}^{-1}{\rm Mpc}^{-1}$ and $t_0 \simeq 12 - 20$ Gyr, implying only that $H_0 t_0 \simeq 0.5 - 2.0$.

– $\Omega = 1.0$. All of the dynamical observations suggest that the fraction of critical density contributed by matter which is clumped on scales $\lesssim 10 - 30$Mpc is only about: $\Omega_{\lesssim 30} \simeq 0.2 \pm 0.1$($\pm 0.1$ is not meant to be a formal error estimate, but indicates the spread in the observations) (see the recent review by Trimble [84]). If inflation is not be falsified, that leaves but two options: (1) the observations are somehow misleading or wrong; or (2) there exists a component of energy density which is smoothly distributed on scales $\lesssim 10 - 30$ Mpc (and therefore would not be reflected in the dynamical determinations). Candidates for the smooth component include: relic, light neutrinos, which by virtue of the large length scale ($\lambda_\nu \simeq 13 h^{-2}$Mpc) on which neutrino perturbations are damped by freestreaming, would likely still be smooth on these scales; relic relativistic particles produced by the recent decay of an unstable WIMP species (Turner et al [85]; Dicus et al [86]; Olive et al [87]; a relic cosmological term (Turner et al [85]; Peebles [88];) "failed galaxies," referring to a population of galaxies which have the same mix of dark matter to baryons, but are more smoothly distributed and are too faint to observe (at least thus far) (Kaiser [89], [90]; Bardeen et al [91]; a relic population of light strings — either fast moving nonintercommuting strings or a tangled network of non-Abelian strings (Vilenkin [92]). All of these smooth component scenarios have testable consequences (Charlton and Turner [93]) — their predictions for $H_0 t_0$ differ from 2/3; the growth of perturbations is different; the evolution of the cosmic scale factor $R(t)$ is different from the matter-dominated model and various kinematic tests (magnitude-redshift), angular size-redshift, lookback time-redshift, proper volume element-redshift, etc.) can in principle differentiate between them.

– *Microwave fluctuations.* Both the scalar and tensor metric perturbations, cf. Eqs (9.10), lead to fluctuations in the CMBR on large angular scales ($\gg 1°$). On such large scales causal processes (such as reionization) cannot have erased the primordial fluctuations, and so if ever present, they must still be there. The scalar perturbations (if they have anything to do with structure formation) must be of amplitude $\gtrsim$ *few* $\times 10^{-6}$, which is within a factor of 10 or less of the current upper limits on these scales.

– *Two detailed stories of structure formation.* The simplest possibility, namely that the most of the mass density is in relic WIMPs ($\Omega_{WIMP} = 1.0 - \Omega_B \simeq 0.9$) leads to two very detailed scenarios of structure formation: hot dark matter (the case where the dark matter is neutrinos) and cold dark matter (essentially any other WIMP as dark matter). At present, the numerical simulations of these scenarios are sufficiently definite that it is possible to falsify them-and in fact, both of these simplest scenarios have difficulties (see the recent review by White [94]). In the hot dark matter case it is forming galaxies early enough. The large-scale structure which evolves in this case (voids, superclusters, froth) qualitatively agrees with what is observed; however, in order to get agreement with the galaxy–galaxy correlation function, galaxies must form very recently (redshifts $\lesssim 1$) in contradiction to all the galaxies (redshifts as large as 3.2) and QSO's (redshifts as large as 4.0) which are seen at redshifts $\gtrsim 1$.

With cold dark matter the simulations can nicely reproduce galaxy clustering, most of the observed properties of galaxies (masses and densities, rotation curves, etc) (Blumenthal et al [95]; Davis et al [96]). However, the simulations do not seem to be able to produce sufficient large-scale structure. In particular, they fail to account for the amplitude of the cluster–cluster correlation function (by a factor of about 3), large amplitude, largescale peculiar velocities, and voids. [In fairness I should mention that our knowledge of large-scale structure of the Universe is still very fragmentary, with the first moderate sized ($\sim 10^4$), 3-dimensional surveys having just recently been completed.] In order to account for $\Omega = 1.0$, galaxy formation must be biased (i.e. only density-averaged peaks greater than some threshold, typically $2 - 3\sigma$, are assumed to evolve into galaxies which we see today, the more typical 1σ peaks resulting in "failed galaxies" for some reason or another; see Bardeen et al [91]).

[The situation with respect to large scale structure is becoming more interesting every moment. Several groups have now reported large-amplitude (600 – 1000 kmsec^{-1}) peculiar velocities on large scales ($\sim 50\mathrm{h}^{-1}\mathrm{Mpc}$) (Burstein et al [97]; Collins et al [98]). Such large peculiar velocities are very difficult, if not impossible, to reconcile with either hot or cold dark matter (or even smooth component models) and the Zel'dovich spectrum (Vittorio and Turner [99]). If these data hold up they may pose an almost insurmountable obstacle to any scenario with the Zel'dovich spectrum of density perturbations. The frothy structure observed in the galaxy distribution by de Lapparent et al [100], galaxies distributed on the surface on large ($\sim 30\mathrm{h}^{-1}\mathrm{Mpc}$), empty bubbles, although somewhat more qualitative, also seems difficult to reconcile with cold dark matter.]

There are a number of observations/experiments which can and will be done in the next few years and which should really put the inflationary scenario to the test. They include improved sensitivity measurements of the CMBR anisotropy. The microwave background anisotropies predicted in the hot dark matter scenario are very close to the observational upper limits on angular scales of both 5 or so arcminutes and $\gtrsim$ few degrees (Vittorio and Silk [101]; Bond and Efstathiou [102]). With cold dark matter, the predictions are a factor of 3 — 10 away from the observational limits (for the isocurvature spectrum, the quadrupole upper limit may actually rule out this possibility; see, Efstathiou and Bond [103]). An improvement in sensitivity to microwave anisotropies of the order of 3 — 10 could either begin to confirm one of the scenarios or rule them both out, and is definitely within the realm of experimental reality (Wilkinson [4]).

The relic WIMP hypothesis for the dark matter can also be tested. While it was once almost universally believed that all WIMP dark matter candidates were, in spite of their large abundance, essentially impossible to detect because of the feebleness of their interactions, a number of clever ideas have recently been suggested (and are being experimentally implemented) for detecting axions (Sikivie [104]), photinos, sneutrinos, heavy neutrinos, etc (Goodman and Witten [105]). Results and/or limits will be forthcoming soon. With the coming online of the Tevatron at Fermilab, the SLC at SLAC, and hopefully the SSC it is possible that

one of the candidates may be directly produced in the lab. Experiments to detect neutrino masses in the eV mass range also continue.

A geometric measurement of the curvature of the Universe (which uses the dependence of the comoving volume element as a function of redshift) has recently been made by Loh and Spillar [106]). Their preliminary results indicate $\Omega = 0.9^{+0.7}_{-0.5}$(95% confidence) (for a matter-dominated model.) This technique appears to have great cosmological leverage and looks very promising—far more promising than the traditional approach of determining the density of the Universe through the declaration parameter q_0.

Another area with great potential for improvement is 3d surveys of the distribution of galaxies. The largest redshift surveys at present contain only a few 1000 galaxies, yet have been very tantalizing, indicating evidence of voids and froth-like structure to the galaxy distribution (de Lapparent et al [100]). The large, automated surveys which are likely to be done in the next decade could very well lead to a quantum leap in our understanding of the large scale features of the Universe and help to provide hints as to how they evolved.

The peculiar velocity field of the Universe is potentially a very valuable and direct probe of the the density field of the Universe:

$$|\delta v_k| = |\dot{\delta}_k/k| \qquad (= (\lambda H/2\pi)\delta_k \text{ for } \Omega = 1), \tag{18}$$

$$(\delta v/c)_\lambda \simeq (\lambda/10^4 \mathrm{h}^{-1}\mathrm{Mpc})(\delta\rho/\rho)_\lambda, \tag{19}$$

where δ_k and δv_k are the kth Fourier components of $\delta\rho/\rho$ and $\delta v/c$ respectively. The very recent measurements which indicate large amplitude peculiar velocities on scales of $\sim 50^{-1}$Mpc are surprising in that they indicate substantial power on these scales, and are problematic to almost every scenario of structure formation. Should they be confirmed they will provide a very acute test of structure formation in inflationary models.

Of course, theorists are very accommodating and have already started suggesting alternatives to the simplest scenarios for structure formations. As I mentioned earlier, scenarios with a smooth component to the energy density have been put forward to solve the Ω problem. Cosmic strings present a radically different approach to structure formation with their non-gaussian spectrum of density fluctuations.[It is interesting to note that cosmic strings of the right "weight" ($G\mu \simeq 10^{-6}$ or so, where μ is the string tension) seem to be somewhat incompatible with inflation, as they must necessarily be produced after inflation and require reheating to a temperature $\gtrsim \mu^{1/2} \simeq 10^{16}$GeV which seems difficult.] Somewhat immodestly I mention a proposal Silk and I recently made: "double inflation" (Silk and Turner [107]; Turner et al [108]). While the Harrison — Zel'dovich spectrum is a beautiful prediction both because of its geometric simplicity and its definiteness, it may well be in conflict with observation because it does not seem to allow enough power on large scales to account for the recent observations of froth and large amplitude peculiar velocities. In the variant we have proposed there are two (or more) episodes of inflation,

with the final episode lasting only about 40 e-folds or so, so that the amplitudes of perturbations on large scales are set by the first episode and those on small scales by the second episode. This enables one to have very large amplitude perturbations on small scales (of order 10^{-1}) and larger than usual amplitude perturbations on large scales (nearly saturating the large scale microwave limits), thereby providing enough power for the large scale structure which the recent redshift surveys and peculiar velocity measurements indicate. The large amplitude perturbations on small scales allows for very early galaxy formation (and reionization of the Universe, thereby erasing the CMBR fluctuations on small angular scales). If the second episode of inflation proceeds via the nucleation of bubbles, they might directly explain the froth-like structure recently reported by de Lapparent [100].

Epilogue

Despite the absence of a compelling model which successfully implements the inflationary paradigm, inflation remains a very attractive means of accounting for a number of very fundamental cosmological facts by microphysics that we have some understanding of: namely, scalar field dynamics at sub-Planck energies. The lack of a compelling model at present must be viewed in the light of the fact that at present we have no compelling, detailed model for the "*Theory of Everything*" and the fact that despite vigorous scrutiny there has yet to be a *No-Go Theorem* for inflation unearthed. It is my belief that the undoing of inflation (if it should come) will involve observation and not theory. At the very least The Inflationary Paradigm is still worthy of further consideration.

My review of inflation has necessarily been incomplete, for which I apologize. I refer the interested reader to the more complete reviews by Linde [109]; by Abbott and Pi [110]; and by Steinhardt [111]; by Brandenberger [112]; and by myself (Turner [48]). My prescription for successfully implementing inflation borrows heavily from the paper written by Steinhardt and myself (Steinhardt and Turner [113]).

This work was supported in part by the DoE (at Chicago) and by my Alfred P. Sloan Fellowship.

References

1. G. Marx and A.S. Szalay, Proc. Neutrino, *1*, 123, 1972.
2. G.F. Smoot et al, Astrophys. J., *291*, L23, 1985.
3. J. Peterson et al., Phys. Rev. Lett., *55*, 332, 1985.
4. D.T. Wilkinson: In Inner Space/Outer Space, eds. E.W. Kolb, et al. University of Chicago Press, Chicago, 1986.
5. J. Yang, et al., Astrophys. J., *281*, 493, 1984.
6. A. Boesgaard and G. Steigman, Ann Rev. Astron. Astrophys., *23*, 319, 1985.
7. G. Efstathiou and J. Silk, Fund. Cosmic Phys., *9*, 1, 1983.
8. J.M. Bardeen, Phys. Rev., *D22*, 1882, 1980.
9. A. Vilenkin, Phys. Rep., *121*, 263, 1985.
10. A. Albrecht and N. Turok, Phys. Rev. Lett., *54*, 1868, 1985.
11. N. Turok, Phys. Rev. Lett., *55*, 1801, 1985.
12. R.H. Dicke and P.J.E. Peebles, in General Relativity: An Einstein Centenary Survey, eds. S. Hawking and W. Israel, Cambridge University Press, Cambridge, 1979.
13. E.W. Kolb and M.S. Turner, Ann. Rev. Nucl. Part. Sci., *33*, 645, 1983.
14. J. Preskill, Ann. Rev. Nucl. Part. Sci., *34*, 461, 1984.

15. C.B. Collins and S.W. Hawking, Astrophys. J., *180*, 317, 1973.
16. C.W. Misner, Astrophys. J., *151*, 431, 1968.
17. C.W. Misner, in Magic Without Magic, ed. J. Klauder, Freeman, San Francisco, 1972.
18. R. Matzner and C.W. Misner, Astrophys. J., *171*, 415, 1972.
19. R. Penrose, in General Relativity: An Einstein Centenary Survey, eds. S. W. Hawking and W. Israel, Cambridge University Press, Cambridge, 1979.
20. B. DeWitt, Phys. Rev., *90*, 357, 1953.
21. L. Parker, Nature, *261*, 20, 1976.
22. Ya. B. Zel'dovich, JETP Lett., *12*, 307, 1970.
23. A.A. Starobinskii, Phys. Lett., *91B*, 99, 1980.
24. P. Anderson, Phys. Rev.,*D28*, 271; *D29*, 615, 1983.
25. J.M. Hartle and B.-L. Hu, Phys. Rev., *D20*, 1772, 1979.
26. M.V. Fischetti et al., Phys. Rev., *D20*, 1757, 1979.
27. J.M. Hartle and S. W. Hawking, Phys. Rev., *D28*, 2960, 1983.
28. B.J. Carr and M. J. Rees, Nature, *278*, 605, 1979.
29. J.D. Barrow and F. Tipler, The Anthropic Cosmological Principle, Oxford University Press, Oxford, 1986.
30. D. Lindley, Fermilab preprint (unpublished), 1985.
31. A. Guth, Phys. Rev., *D23*, 347, 1981.
32. A. Guth and E. Weinberg, Nucl. Phys., *B212*, 321, 1983.
33. S.W. Hawking et al., Phys. Rev., *D26*, 2681, 1982.
34. A.D. Linde, Phys. Lett., *108B*, 389, 1982.
35. A. Albrecht and P.J. Steinhardt, Phys. Rev. Lett., *48*, 1220, 1982.
36. M.S. Turner, Phys. Rev., *D28*, 1243, 1983.
37. T. Bunch and P.C.W. Davies, Proc. Roy. Soc. London, *A360*, 117, 1978.
38. J.M. Bardeen et al., Phys. Rev., *D28*, 679, 1983.
39. A. Guth and S.-Y. Pi, Phys. Rev. Lett., *49*, 1110, 1982.
40. A.A. Starobinskii, Phys. Lett., *117B*, 175, 1982.
41. S.W. Hawking, Phys. Lett., *115B*, 295, 1982.
42. L. Abbott and M.B. Wise, Nucl. Phys., *B244*, 541, 1984.
43. A.A. Starobinskii, JETP Lett., *30*, 1979.
44. V.A. Rubakov et al., Phys. Lett., *115B*, 189, 1982.
45. P.J. Steinhardt and M.S. Turner, Phys. Lett., *129B*, 51, 1983.
46. A.D. Linde, Phys. Lett., *158B*, 375, 1985.
47. D. Seckel and M.S. Turner, Phys. Rev., *D32*, 3178, 1985.
48. M.S. Turner, in Fundamental Physics and Cosmology, eds. J. Audouze and J. Tran Thanh Van, Editions Frontieres, Gif-sur-Yvette; and in The Architecture of Fundamental Interactions at Short Distances, eds. P. Ramond and R. Stora, North-Holland, Amsterdam, 1986.
49. G. Coughlan et al., Phys. Lett., *131B*, 54, 1983.
50. J.E. Ellis et al., Phys. Lett., *145B*, 181, 1984.
51. M.Yu. Khlopov and A.D. Linde, Phys. Lett., *138B*, 265. 1984.
52. R.J. Scherrer and M.S. Turner, Astropnys. J. in press, 1987.
53. Q. Shafi and A. Vilenkin, Phys. Rev. Lett., *52*, 691, 1984.
54. S.-Y. Pi, Phys. Rev. Lett., *52*, 1725, 1984.
55. R. Holman et al., Phys. Lett., *137B*, 343, 1984.
56. B. Ovrut and P.J. Steinhardt, Phys. Rev. Lett., *53*, 732, 1984.
57. B. Ovrut and P.J. Steinhardt, Phys Lett., *147B*, 263, 1984.
58. F. Accetta et al., Phys. Rev., *D31*, 3046, 1985.
59. B.L. Spokoiny, Phys. Lett., *147B*, 39, 1984.
60. Q. Shafi and C. Wetterich, Phys. Lett., *129B*, 387, 1983.
61. Q. Shafi and C. Wetterich, Phys. Lett., *152B*, 51, 1985.
62. M.B. Mijic et al., Phys. Rev., *D34*, 2934, 1986.
63. A.D. Linde, Phys. Lett., *129B*, 177, 1983.

64. D.V. Nanopoulos, Comments on Astrophys., *X*, 219, 1985.
65. A. Albrecht and R. Brandenberger, Phys. Rev., *D31*, 1225, 1985.
66. A. Albrecht et al., Phys. Rev., *D32*, 1280, 1985.
67. A. Guth and S.-Y. Pi, Phys. Rev., *D32*, 1899, 1985.
68. W. Fischler et al., Nucl. Phys., *B259*, 730, 1985.
69. A.D. Linde, Phys. Lett., *116B*, 335, 1982.
70. A. Vilenkin and L. Ford, Phys. Rev., *D26*, 1231, 1982.
71. M.S. Turner and L. Widrow, Phys. Rev. Lett., *57*, 2237, 1986.
72. L. Jensen and J. Stein-Schabes, Phys. Rev., *D34*, 931, 1986.
73. R.M. Wald, Phys. Rev., *D28*, 2118, 1983.
74. G. Steigman and M. S. Turner, Phys. Lett. *128B*, 295, 1983.
75. L. Jensen and J. Stein-Schabes, Phys. Rev., in press, 1986.
76. A.A. Starobinskii, JETP Lett., *37*, 66, 1983.
77. N. Myhrvad, Phys. Lett., *132B*, 308, 1983.
78. E. Mottola, Phys. Rev., *D31*, 754; *D33*, 2136, 1985.
79. L. Parker, Phys. Rev. Lett., *50*, 1009, 1983.
80. L. Ford, Phys. Rev., *D31*, 710, 1985.
81. P. Anderson, Phys. Rev., *D32*, 1302, 1985.
82. J. Traschen and C.T. Hill, Phys. Rev., *D33*, 3519, 1986.
83. M.S. Turner, in Dark Matter in the Universe, eds. J. Kormendy and J. Knapp, Reidel, Dordrecht, 1986.
84. V. Trimble, Ann. Rev. Astron. Astrophys., *25*, in press, 1987.
85. M.S. Turner et al., Phys. Rev. Lett., *52*, 2090, 1984.
86. D.A. Dicus et al., Phys. Rev. Lett., *39*, 168, 1977.
87. K. Olive et al., Astrophys. J., *292*, 1, 1985.
88. P.J.E. Peebles, Astrophys. J., *284*, 439, 1984.
89. N. Kaiser, Astrophys. J., *273*, L17, 1983.
90. N. Kaiser, in Inner Space/Outer Space, eds. E.W. Kolb et al, University of Chicago Press, Chicago, 1986.
91. J.M. Bardeen et al., Astrophys. J., *304*, 15, 1986.
92. A. Vilenkin, Phys. Rev. Lett., *53*, 1016, 1984.
93. J. Charlton and M.S. Turner, Astrophys. J., in press, 1987.
94. S.D.M. White, in Inner Space/Outer Space, eds. E.W. Kolb et al, University of Chicago Press, Chicago, 1986.
95. G. Blumenthal et al., Nature, *311*, 517, 1986.
96. M. Davies et al., Astrophys. J., *292*, 371, 1985.
97. D. Burstein et al., Astrophys. J., in press, 1986.
98. C.A. Collins et al., Nature, *320*, 506, 1986.
99. N. Vittorio and M.S. Turner, Astrophys. J., in press, 1987.
100. V. de Lapparent et al., Astrophys. J., *302*, L1, 1986.
101. N. Vittorio and J. Silk, Astrophys. J., *285*, L39, 1984.
102. J.R. Bond and G. Efstathiou, Astrophys. J., *285*, L44, 1984.
103. G. Efstathiou and J.R. Bond, Mon. Not. r. Astron. Soc., *218*, 103, 1986.
104. P. Sikivie, Phys. Rev. Lett., *51*, 1415, 1983.
105. M. Goodman and E. Witten, Phys. Rev., *D31*, 3059, 1985.
106. E. Loh and E. Spillar, Astrophys. J., *307*, L1, 1986.
107. J. Silk and M.S. Turner, Phys. Rev., *D35*, in press, 1987.
108. M.S. Turner et al., Astrophys. J., in press, 1987.
109. A.D. Linde, Rep. Prog. Phys., *47*, 925, 1984.
110. L. Abbott and S.-Y. Pi, The Inflationary Universe, World Scientific, Singapore, 1986.
111. P.J. Steinhardt, Comments Nucl. Part. Phys., *12*, 273, 1984.
112. R. Brandenberger, Rev. Mod. Phys., *57*, 1, 1985.
113. P.J. Steinhardt and M.S. Turner, Phys. Rev., *D29*, 2162, 1984.

Acta Physica Hungarica 64 (1-3), pp. 309-312 (1988)

FRACTALS AND THE LOGNORMALITY OF GALAXY COUNTS* **

YAKOV B. ZELDOVICH

Institute of Physical Problems, Moscow, USSR

and

A. S. SZALAY

Department of Atomic Physics, Roland Eötvös University
1088 Budapest, Hungary
Department of Physics and Astronomy, J. Hopkins University
Baltimore, USA

(Received 8 January 1987)

We discuss how a random Gaussian density fluctuation field can dynamically evolve into a fractal through various physical processes creating the luminosity of galaxies. It is shown that in many cases the distribution of light will deviate from Gaussian. It can often have a lognormal distribution in accordance with an observation by Hubble, a fact unexplained so far.

It has been long recognized that the galaxy distribution has a power law correlation function [1] $\xi(r) = (r/r_0)^{\gamma}$. It is well known that such scale invariant structures are called fractals [2]: the galaxy distribution has a nontrivial fractional or fractal dimension [3] D_F. The previous exponent γ and D_F are related:

$$\gamma = 3 - D_F. \tag{1}$$

On the other hand, it is believed that the density fluctuations in the early universe were Gaussian (random phases of the initial Fourier components.) The accepted view has been that nonlinear gravitational clustering can be solely responsible for transforming these small fluctuations into the power law clustering hierarchy that we see today [1]. However, the improving limits of the microwave background fluctuations [4] and the discovery of a very strong cluster–cluster correlation [5,6] pose some interesting questions. Compared to the $\Delta T/T$ limits even the galaxy correlations are too strong. If these are due to nonlinear gravitational clustering, and the cluster–cluster systems are believed to be still linear, why does

*Dedicated to Prof. G. Marx on his 60th birthday

**A note added by Ya. B. Z.: George Marx was the first person whom I have met abroad. There were a lot of parallel lines in our work: from lepton numbers to neutrino masses we had similar ideas. But, the competition has never spoiled our personal friendship!

the cluster correlation function have an identical shape? Besides, it has been noted that the galaxy and cluster correlation functions obey a scaling relation; if we had a picture of each catalog without labels we could not distinguish between them. In mathematical terms this means that both correlation functions are adequately described by the same relation [7]:

$$\xi_L(r) = 0.3 \left(\frac{r}{L}\right)^{-1.8}, \tag{2}$$

where $L = n^{-1/3}$ is the mean distance among the objects in the catalog. This scaling is well satisfied by the well known trend of increasing correlation with richness.

It has been suggested by Kaiser [8], that by the "biasing" of Gaussian fluctuations (taking only the regions which are above a certain threshold) one can enhance the correlations of clusters with respect to galaxies. One can overcome the $\Delta T/T$ limits easily by assuming that the observed galaxies do not follow the distribution of the mass. The general idea behind biasing is to introduce an extra degree of freedom how the mass fluctuations turn into light. Several other papers have carried this idea of biasing further since [9]. Another form of biasing, an exponential relationship between the linear overdensity and the nonlinear number of galaxies has been discussed by Kaiser and Davis [10], a lognormal transformation, that we will consider more in detail below.

In a recent paper by Szalay and Mandelbrot [14] it is suggested that for certain random Gaussian fields biasing can change the correlation properties of the distribution drastically: one can get power law correlations for the biased regions, where the slope of the power law depends on the "threshold". We would like to outline this idea here. Consider the linear density fluctuations $\delta = \delta\rho/\rho$. If $\langle|\delta|^2\rangle = \sigma^2$, one can introduce a normal Gaussian random field $y = \delta/\sigma$, with a linear dimensionless correlation function $w(r) = \langle y_1 y_2\rangle$. Instead of using a sharp threshold, one can use a smooth one by creating a new lognormal random field Y, representing the luminosity density

$$Y = Ae^{\nu y}. \tag{3}$$

Here A is required to normalize the mean of Y to 1. The correlation function of the Y field can be trivially calculated:

$$1 + \xi_\nu(r) = \langle Y_1 Y_2\rangle = e^{\nu^2 \omega(r)}. \tag{4}$$

This is very similar to the approximation obtained by Politzer and Wise [11] for sharp clipping, but this is an exact result. It seems, that the correlation function of the "biased" regions is quite insensitive to the sharpness of the clipping.

The random field y has a natural lower and upper cutoff scale as all physical systems do. For fluctuation spectra describing the distribution of galaxies it is not unreasonable to assume that the small scale cutoff is around the scale of the galaxies λ_o and the large scale cutoff would be determined by the equality of the matter and radiation energy densities λ_x.

Certain fluctuation spectra $|\delta_k|^2 \alpha k^{-3}$ have equal power per octave, so the dispersion integral would diverge logarithmically were there no cutoffs. The dispersion is determined by the logarithm of the two cutoff scales, and the correlation function of the y-field has a logarithmic leading term. Due to this behaviour the correlation function of the Y-field becomes a power law:

$$w(r) = \frac{\ln(r/\lambda_x)}{\ln(\lambda_o/\lambda_x)}; \qquad 1 + \xi_\nu(r) = \left(\frac{r}{\lambda_x}\right)^{-\nu^2/\ln(\lambda_x/\lambda_o)}. \tag{5}$$

For reasonable values of $\lambda_o \approx 1$ Mpc and $\lambda_x \approx 50(\Omega_o h^2)^{-1}$ Mpc ≈ 100 Mpc the observed correlation function slope is given by $\nu \approx 2.5 - 2.7$, a similar value to what was obtained from the cold dark matter N-body simulations in the best case [12]. It was thus shown that power law correlation functions can be generated by the appropriate nonlinear distortion of a Gaussian random field with a spectrum near k^{-3}. The resulting distribution will have the right correlation properties. One cannot do a separate clipping on the scale of clusters, since the filtering and the nonlinear biasing are not interchangeable. There are two separate questions, one of them is the viability of the lognormal process, the other is whether the underlying random field has a k^{-3} spectrum.

The above lognormal transformation has attractive features: it describes such a nonlinear process, which does lead to negative densities. It does not satisfy the continuity equation, however, so it cannot come from nonlinear hydrodynamic motion. Here we address the question, how such a lognormal biasing of the fluctuations can arise physically. We take the view that it is the luminosity which is a nonlinear function of the overdensity, the transformation describes the "relative visibility" of the various regions. In this case the continuity equation does not have to be satisfied.

It turns out that in physics it is quite easy to imagine conditions under which lognormal distributions arise. As discussed in the very enjoyable paper by Montroll and Schlesinger [15], ("The tale of the tails ...") if there is an event with probability P that depends on the occurrence of N independent events with probabilities p_i, $i = 1, ...N$, then

$$\log P = \log p_1 + ... \log p_N. \tag{6}$$

If the p_i are well behaved, i.e. they have second moments, then due to the central limit theorem $\log P$ will have a normal distribution, thus P will be random lognormal. One such case is the "broken-stick" type fragmentation, the random breakup of a stick into smaller parts. The distribution of the fragments will be lognormal, as first pointed out by Kapteyn then applied by Kolmogorov [16] to the size distribution of crushed ore.

In another case we consider the random process y representing fluctuations in the gravitational potential. If the galaxy distribution has a uniform virial temperature T_g, then the density distribution of the galaxies in such a fluctuating potential will be

$$n(\vec{x}) = n_0 \exp\left(-\frac{\Phi(\vec{x})}{kT_g}\right). \tag{7}$$

This functional dependence is precisely of the discussed lognormal type, but instead of a static threshold here the dispersion of the potential Φ and the temperature T_g determine an effective $\nu = \langle|\Phi^2|\rangle^{1/2}/kT_g$. If the density fluctuations are scale-invariant ($n = 1$), the fluctuations of the potential have a k^{-3} spectrum, so we obtain the above mentioned fractal distribution.

One can also turn to an analogy in biology, suggested by one of us [13] for the population growth of bacteria. If for some time there is an exponential growth law for the luminosity of a region

$$\frac{\partial L}{\partial t} = \beta(\vec{x})L; \qquad L(\vec{x},t)\alpha \exp(\beta(\vec{x})t). \tag{8}$$

If the growth coefficient β is a Gaussian random function, one can easily see that the luminosity density will have a lognormal distribution. This fact will not change if we allow for the diffusion-like motion of particles.

The galaxy distribution has indeed such a property, found by Hubble [16]: he made the remark that the distribution of galaxy counts on the different plates is lognormal, log N is very close to a Gaussian. Recently Crane and Saslaw [18] analysed the galaxy counts in equal area cells of the Zwicky catalog. They show, that the distribution is rather well fitted by a lognormal, and they point out that multiple capture on clusters will also result in lognormality.

The galaxy distribution is known to deviate from the Gaussian, having a more extended tail. Biasing is just one way to try to simulate this feature, and lognormality is a smooth and physically plausible version to achieve this. We saw above that it is possible to come up with various models which will generate a power law correlation function and lognormal galaxy counts.

References

1. P.J.E. Peebles, The Large Scale Structure of the Universe, Princeton University Press, Princeton, 1980.
2. B.B. Mandelbrot, The Fractal Geometry of Nature, Freeman, San Francisco, 1982.
3. G. de Vaucouleurs, Astrophys. J., *202*, 319, 1975.
4. J.M. Uson and D.T. Wilkinson, Astrophys. J., *277*, L1, 1984.
5. N.A. Bahcall and R.A. Soneira, Astrophys. J., *270*, 20, 1983.
6. A.A. Klypin and A.I. Kopylov, Pisma A. Zh., *9*, 41, 1983.
7. A.S. Szalay and D.N. Schramm, Nature, *314*, 718, 1986.
8. N. Kaiser, Astrophys. J., *284*, L9, 1984.
9. J.R. Bardeen, J.R. Bond, N. Kaiser and A.S. Szalay, Astrophys J., *304*, 15, 1986.
10. N. Kaiser and M. Davis, Astrophys. J., *297*, 365, 1985.
11. D. Politzer and M.B. Wise, Astrophys. J., *285*, L1, 1984.
12. M. Davis, G. Efstathiou, C.S. Frenk and S.D.M. White, Astrophys J., *292*, 371, 1985.
13. Ya.B. Zeldovich, Doklady AN USSR, *270*, 1369, 1983.
14. A.S. Szalay and B.B. Mandelbrot, in preparation, 1987.
15. E.W. Montroll and M.F. Shlesinger, J. Stat. Phys., *32*, 209, 1983.
16. A.N. Kolmogorov, C.R. Acad. Sci. USSR, *31*, 99, 1941.
17. E. Hubble, Astrophys. J., *79*, 8, 1934.
18. P. Crane and W.C. Saslaw, Astrophys. J., *301*, 1, 1986.

Acta Physica Hungarica 64 (1-3), pp. 313-319 (1988)

TRANSIENT CONDUCTIVITY CHANGE IN PURPLE MEMBRANE SUSPENSION DURING THE PHOTOCYCLE*

G. FRICSOVSZKY, E. PAPP, G. MESZENA and A. EL-LAKKANI

Department of Atomic Physics, Roland Eötvös University
1088 Budapest, Hungary

(Received 8 January 1987)

During the photocycle of bacteriorhodopsin in the purple membrane of Halobacteria halobium protons are pumped through the membrane. In purple membrane suspension the release and uptake of protons cause a transient electric conductivity change. This change is measured in AC electric field oriented purple membrane suspension. The signal is strongly and nonlinearly orientational dependent. The time course of the electric signal is analyzed and it is shown that proton release coincides with the decay of the M-state and the proton uptake with the decay of the O-state.

Introduction

Halobacterium halobium can convert light energy into chemical energy. For this purpose the bacterium develops some special patches on the cell wall, named purple membrane (pm). This membrane contains only one type of protein in a two-dimensional crystal-like ordered array, the bacteriorhodopsin (bR) consisting of a known sequence of amino acids and a photochemically active chromophore linked to the peptide-chain through a Schiff-base. The bacteriorhodopsin protruding through the membrane can pump protons from one side of the membrane to the other. When the chromophore absorbs light the proton pump generates a gradient of H^+-concentration, which can drive the metabolic processes of the bacterium. The chromophore is a retinal like the light-sensing molecule of the eye. The light excitation of retinal is followed by an isomerization process which perturbs the neighbourhood of the Schiff-base, thus initiating the proton pump. After the excitation the process is thermally decaying through some intermediate states of the so-called photocycle recycling the bR again to its ground-state [1].

Though the molecular mechanism of proton pumping is not yet well understood, many aspects are known: the isomerization process of the retinal and the Schiff-base [2], the primary deprotonation step of the protein [3] and proton movement through the membrane [4]. A very important question is the mechanism of proton release and uptake by the membrane. This problem was investigated by different methods: steady state measurements of light-induced pH-change in phospholipid vesicles [5], determining volume changes [6] and using pH-indicator dyes

*Dedicated to Prof. G. Marx on his 60th birthday.

[7]. Marinetti and Mauzerall measured the small electrical conductivity change of pm suspension during the flash-initiated photocycle [8,9].

At pH ~ 7 the pm is negatively charged. The bound (partly surface) charges and the screening ionic cloud play an important role in the behaviour of pm in electric field. In AC fields the pm will be aligned parallel to the field due to the anisotropic polarizability of the ionic cloud around the pm [10]. This alignment can be detected by electric dichroism (see e. g. [11]). We found that the electric conductivity of the pm suspension changes (increases) during the alignment. This observation led us to the idea that perhaps the transiently released charges during the photocycle could be detected much easier on aligned pm suspension. In the following we give some experimental results supporting this idea.

Materials and methods

The measuring circuit is shown schematically in Fig. 1. Two specially designed and matched cuvettes with platinum electrodes contain the pm suspension. Because of the matching the circuit needs only a small phase and amplitude

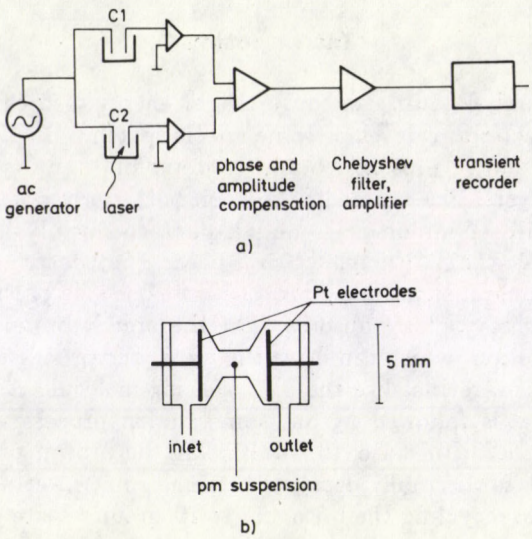


Fig. 1. (a) Experimental arrangement. C1 and C2 are the two measuring cuvettes. (b) Cross section of the cuvette

compensation in order to balance the bridge. The difference signal is amplified and recorded by a transient recorder. The generator supplies high enough AC-voltages (~ 100V, 2kHz) to orient the pm in the cuvettes. At the same time the bridge

allows to detect small conductivity changes in any of the cuvettes. The Chebysev filter strongly attenuates the unwanted higher harmonics of the response signal. The bridge (difference signal) can be compensated almost to 10^{-5}.

For the excitation of the photocycle in one of the cuvettes a nitrogen laser pumped rhodamin 6G dye laser was used with a pulse energy of $\sim 200 \mu J$. The laser light was focused to the central narrow part of the cuvette (Fig. 1b).

The pm suspensions were a gift from the Institute of Biophysics, Biological Research Center, Hungarian Academy of Sciences, Szeged. All measurements were made on pm suspended in distilled water with an absorbance of 1.4 cm^{-1} at 575 nm. The resistance of the suspension in the measuring cuvette was 200 kohm.

Electrodichroic measurements were made with an apparatus described previously [11].

Results and discussion

The electric conductivity change during the alignment of pm in AC electric field is shown in Fig. 2.

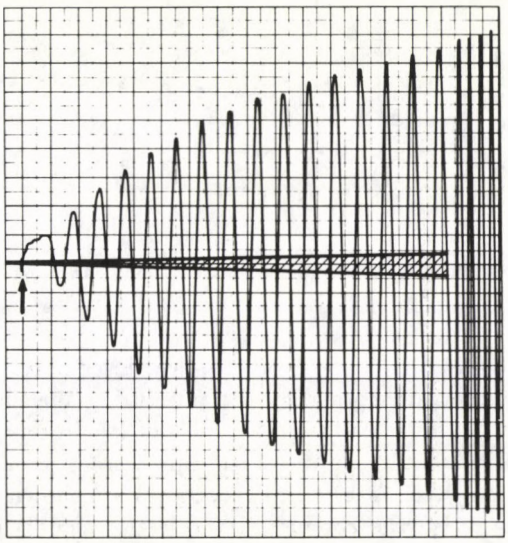


Fig. 2. Orientational increase of the electric conductivity of the pm suspension. AC voltage 100V. One period is 0.5 ms. Relative conductivity change at the end of the recording: $1.6 \cdot 10^{-2}$

At this recording one of the cuvettes in the bridge was replaced by an appropriate resistor. At the beginning, when the voltage is switched on (arrow), the bridge is balanced. As the orientation of the pm in the suspension is going on, we observe an increase in the electric conductivity of the sample. This is much more

pronounced than the heating effect (shown by the shadowed area) which originates in the temperature variation of the cuvette in the bridge. The statement that this conductivity increase is an orientational effect is supported by the observation that the time course of the dichroism and the conductivity increase are similar (Fig. 3). This orientational increase of the conductivity is an interesting problem in itself. We believe that this effect is connected with the anisotropic electrical polarizability of the ionic cloud around the charged membrane. In an AC field this cloud moves collectively, oscillating around the membrane. This oscillation is more pronounced when the pm is parallel to the field, giving maximal contribution to the conductivity of the pm suspension.

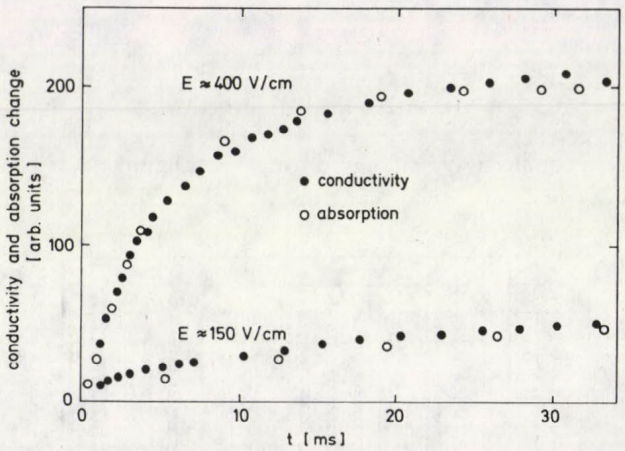
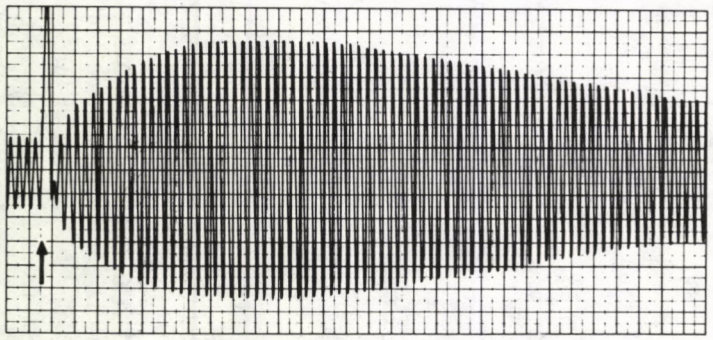


Fig. 3. Conductivity and absorption change after an AC field is switched on the pm suspension

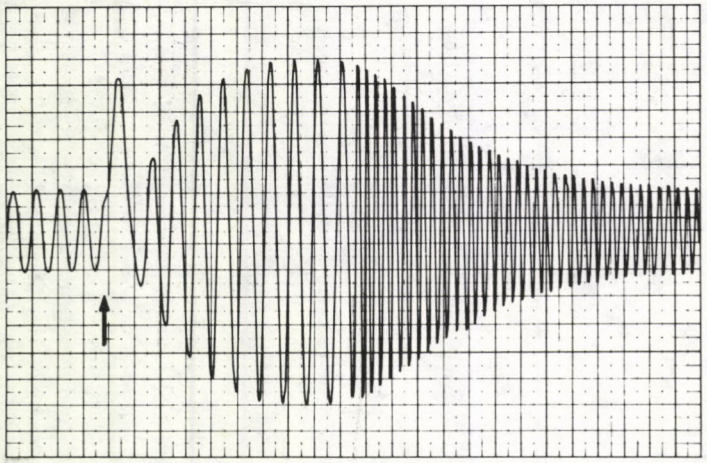
This orientational enhancement of the electric conductivity was applied to detect the charge release and uptake by the pm during the photocycle. After the orientation a laser pulse initiates the photocycle of BR in one of the cuvettes and the charges injected into the medium by the pm cause a transient increase of the conductivity of that cuvette. A few such recordings are shown in Fig. 4.

After the laser pulse the conductivity (electric signal) increases and going through a maximum it returns to the starting level. The signal amplitude at the maximum is strongly and nonlinearly orientational dependent. Fig. 5 shows the signal amplitude dependence on the orienting field strength. For comparison, we give the electric field dependence of the dichroism, too. Without an orientational enhancement the starting linear region, shown in Fig. 5, would be expected for the transient electric signal. Measurements given in [8, 9] were made in this region.

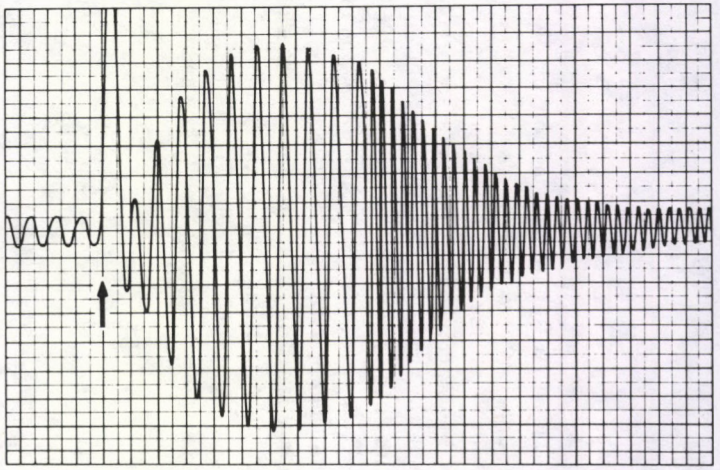
The electric signal was recorded at different temperatures and the kinetics of the charge release and uptake was analyzed. Supposing that the observed conductivity change is proportional to the number of transiently released charges, the simplest



a)



b)



c)

Fig. 4. Transient conductivity increase of the pm suspension during the photocycle at different temperatures. Arrow shows when the laser pulse was given. The sharp change after the laser is an electrical artifact from the laser discharge. (a) t=10, 5C, (b) t=25, 2C, (c) t=28, 2C

kinetics which can reproduce the transient signal is the following:

$$n(t) = A\frac{k_1}{k_1 - k_2}(e^{-k_2 t} - e^{-k_1 t}). \tag{1}$$

This is a first order reaction for the release and the same for the uptake of charges. Here A is proportional to the number of the excited (photocycling) bacteriorhodopsins. The measured signals were fitted to this equation and it was found that the two rate constants are practically equal. The temperature dependence of the rate constants are shown in Fig. 6, where for comparison, the M-state kinetics of the photocycle is shown, too, taken from [12].

From these results it follows that the measured transient conductivity change reflects the appearance of additional free or loosely bounded charge-carriers in the boundary layer of pm. The question whether these charge-carriers are protons or the pumped protons have been exchanged by other ions (originally bounded on the membrane surface [8, 9, 13]) can be answered by the investigation of the ionic strength and pH-dependence of the signal.

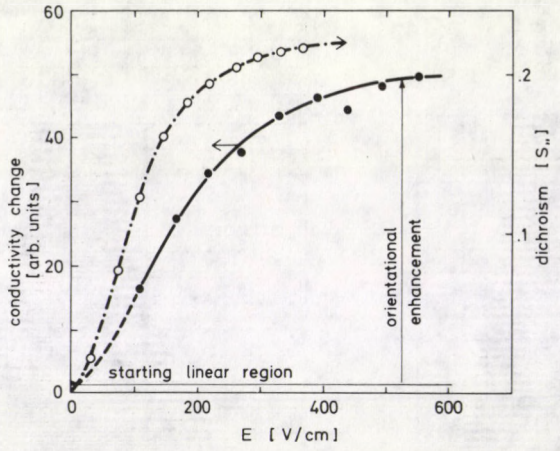


Fig. 5. The electric field dependence of the transient conductivity change caused by the photocycle of pm. For comparison data for electrodichroism are shown, too

The time-dependence of these processes seems to be well described by a simple exponential function, as mentioned above, i. e. they are monophasic in contrast to the biphasic M-state kinetics measured by optical absorption changes during the photocycle [12, 14]. As Fig. 6 shows, the value and the temperature dependence of the rate constant of charge release are well correlated with those of the slow component of the M-state decay. This seems to be in agreement with the result, that the bR pumps protons only in the slow decaying M-state [14] and the fast

M-state is an equilibrium state with the L-form [15], which does not pump protons directly. The rate constant of the decay of the electric signal goes parallel with the decay of the O-state as determined by kinetic absorption measurements (data not shown).

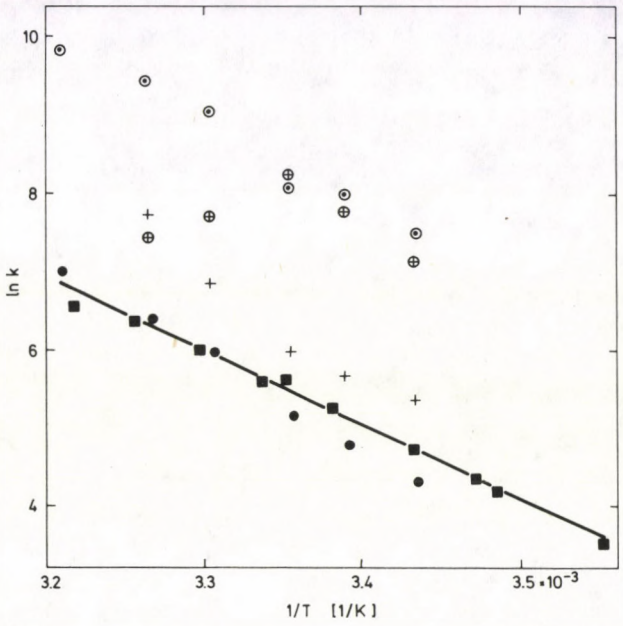


Fig. 6. Temperature dependence of the rate constants: electric signal rise and decay (□), the two M-states rise (⊙ slow and ⊕ fast components) and decay (• slow and + fast components)

References

1. W. Stoeckenius and R.A. Bogomolni, Ann. Rev. Biochem., *57*, 587, 1982.
2. S.O. Smidth, J. Lugtenburg, R.A. Mathies J. Membr. Biol., *85*, 95, 1985.
3. J.M. Fukomoto, J.H. Hanamoto, M.A. El-Sayed, Photochem. Photobiol., *39*, 75, 1984.
4. L. Keszthelyi and P. Ormos. FEBS Lett., *109*, 189, 1980.
5. N.A. Dencher and M.P. Heyn, FEBS Lett.,*108*, 307, 1979.
6. D.R. Ort and W.W. Parson, J. Biol. Chem., *253*, 6158, 1978.
7. R. Govindjee, T.G. Ebrey and A.R. Croffs, Biophys. J., *30*, 231, 1980.
8. T. Marinetti and D. Mauzerall, Proc. Natl. Acad. Sci. USA, *80*, 178, 1983.
9. T. Marinetti and D. Mauzerall, Biophys. J., *50*, 405, 1986.
10. E. Papp, Biophys. Chem., *21*, 243, 1985.
11. E. Papp, G. Fricsovszky and G. Meszena, Biophys. J., *49*, 1089, 1986.
12. G. Fricsovszky, E. Papp, G. Meszena and A. El-Lakkani, Studia Biophysica, *111*, 23, 1986.
13. C.-H. Chang, C.-H. Suh, R. Govindjee and T. Ebrey, Biophys. J., *45a*, 210a, 1984.
14. Q.-Q. Li, R. Govindjee and T. G. Ebrey, Proc. Natl. Acad. Sci. USA, *81*, 7079, 1984.
15. R. Alshuth and M. Stockburger, Photochem. Photobiol., *43*, 55, 1986.

Acta Physica Hungarica 64 (1–3), pp. 321–326 (1988)

STUDY OF THE RADIATION PROTECTION EFFECT OF SELENIUM-METHIONINE BY DETERMINING THE PARAMAGNETIC PROPERTIES OF LIVER TISSUES OF MICE*

VALÉRIA KOVÁCS and A. EL-LAKKANI**

Department of Atomic Physics, Roland Eötvös University
1088 Budapest, Hungary

(Received 8 January 1987)

The paramagnetic properties of the liver tissues were observed after the change which took place during 3, 18 and 24 hours following the administration of s-methionine or alternatively of se-methionine and subsequent irradiation. The signals recorded were those characteristic of P-450 citochrome, free radicals of the semichinon type, Mo-protein and Fe-s-protein. Spectra of these paramagnetic centres generally show the changes in activity of the ferments characteristic for the tissue, which is closely related to the increased or decreased functional activity of the liver tissues [1,2].

We showed that in irradiated animals which contained ordinary methionine the functional activity of the hepocites drops with the increase of doses. In animals treated with se-methionine, however, the functional activity was found to be significantly larger in comparison with the control group of animals which had not been protected. This is believed to be due to the radiation protective effect of se-methionine.

1. Introduction

According to several papers published in the literature, there have been quite a number of attempts to use selenium as antitumorous and a radiation protective agent, based on the well-known antioxidant properties of its compounds. [3, 4, 5, 6, 7].

The present work is one of a series of papers on the mechanisms of the biological activity of selenomethionine. It describes results obtained by investigation of the effects of this compound upon the change of general energetic parameters of cells in normal as well as X-ray irradiated tissues.

Materials and method

Animals investigated were white Balb ♂ mice of 20 gm average weight, the number of animals in each group amounting to ten. Administration of selenium-methionine was carried out in amounts of 1.5 mg/kg by interpertionial injection of

*Dedicated to Prof. G. Marx on his 60th birthday

**On leave from Department of Biophysics, Cairo University, Egypt

0.5 ml of aqueous se-methionine solution. This was performed 10 minutes before irradiation with doses 2, 4.5 and 6 Gy, respectively, of X-rays. Sampling was done 3, 18 and 24 hours after irradiation; the mice were decapitated and dissected, their liver taken out and frozen in liquid nitrogen.

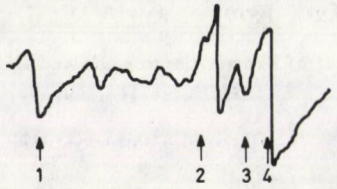


Fig. 1. The EPR spectrum of a frozen intact liver tissues sample of Balb ♂ mice. 1. *p*-450 citochrome (g=2.25) 2. Free radicals of the semichinon type (g=2) 3. Mo-protein (g=1.97) 4. Fe-s-protein (g=1.94)

The paramagnetic resonance method is used to investigate the kinetic nature of the change of the paramagnetic properties of the liver tissues. Figure 1 shows the EPR spectrum of the frozen intact liver tissue samples of the white Balb ♂ mice. The EPR spectrum was determined at low temperature to make possible fixation of the cells in their metabolitically active state. Normal liver tissues are characterized by certain signals of the EPR spectrum differing in their g-factor values and their intensities. In the signals recorded were those characteristic for:

1. P.450 citochrome (g=2.25) which takes part in the processes of detoxication of the disintegration products;
2. Free radicals of the semichinone type (g=2.00) which takes part in the energetic processes;
3. Mo-protein (g=1.97) of the disintegration ferments;
4. Fe-s-protein (g=1.94) electron carriers in the citochrome system.

From the changes in the intensities of the signals for these paramagnetic centres, their biochemical nature as well as the levels of the various metabolic processes in the tissue can be deduced. Figure 2 shows some examples of changes in some different cases.

Results and conclusion

Table I gives the important kinetic changes in relative intensities of the signals of the paramagnetic centres after 3, 18 and 24 hours following administration of s-methionine or se-methionine and irradiation with X-rays in doses 2, 4.5 and 6 Gy.

In determining the concentrations of the paramagnetic centres of the liver tissues of mice, the mean standard error of the results was equal in average to

15%. (Generally this method of paramagnetic resonance yields results whose error is usually in the range of 10% to 30%.) To prove that the tendency observed in the changing of concentrations of the paramagnetic centres is a real one, calculations were made by applying the method of dispersion analysis and linear regression. The results confirmed the tendency derived by us from observations in our study.

Figure 3 shows the change in relative intensities of the signals for the paramagnetic centres as a function of time and doses of irradiation. These changes generally show the changes in activity of the ferments characteristic for the tissue, which is closely related to the change in the functional activity of the liver tissues.

Table I
The relative intensities of the signals for the paramagnetic centres of the liver tissues samples of Balb ♂ mice

Groups of animals			Relative concentrations of the paramagnetic centers			
Type of administration	Time of change [h]	Irradiation dose of X-ray [Gy]	P-450 citoclrome $g=2.25$	Free radial of the semichinon type $g=2$	Mo-protein $g=1.97$	Fe-s-protein $g=1.94$
s-methionine	3	2	1.07±0.20	1.03±0.13	0.94±0.20	0.98±0.20
		4.5	0.92±0.20	1.09±0.20	0.76±0.13	0.93±0.20
		6	0.82±0.16	1.02±0.17	0.81±0.18	0.90±0.17
se-methionine	3	2	0.95±0.17	1.11±0.19	0.84±0.13	1.04±0.21
		4.5	1.04±0.18	1.08±0.20	1.11±0.18	0.99±0.20
		6	0.85±0.15	1.05±0.20	0.90±0.16	0.87±0.18
s-methionine	18	2	0.94±0.21	0.88±0.18	0.83±0.18	0.89±0.15
		4.5	0.82±0.20	0.71±0.16	1.01±0.31	0.79±0.14
		6	0.97±0.12	0.88±0.17	1.09±0.22	0.97±0.16
se-methionine	18	2	1.17±0.21	1.05±0.20	1.05±0.19	1.07±0.20
		4.5	0.94±0.18	1.00±0.20	0.92±0.19	0.94±0.17
		6	1.04±0.13	0.87±0.18	0.96±0.20	0.96±0.18
s-methionine	24	2	1.00±0.16	1.12±0.16	0.93±0.18	0.97±0.19
		4.5	0.74±0.14	0.88±0.13	0.72±0.15	0.73±0.11
		6	0.87±0.16	0.73±0.14	0.87±0.17	0.83±0.14
se-methionine	24	2	1.28±0.31	0.89±0.19	1.16±0.19	1.05±0.22
		4.5	1.15±0.21	1.03±0.17	0.96±0.20	1.00±0.19
		6	1.19±0.15	0.97±0.18	1.23±0.24	1.10±0.18

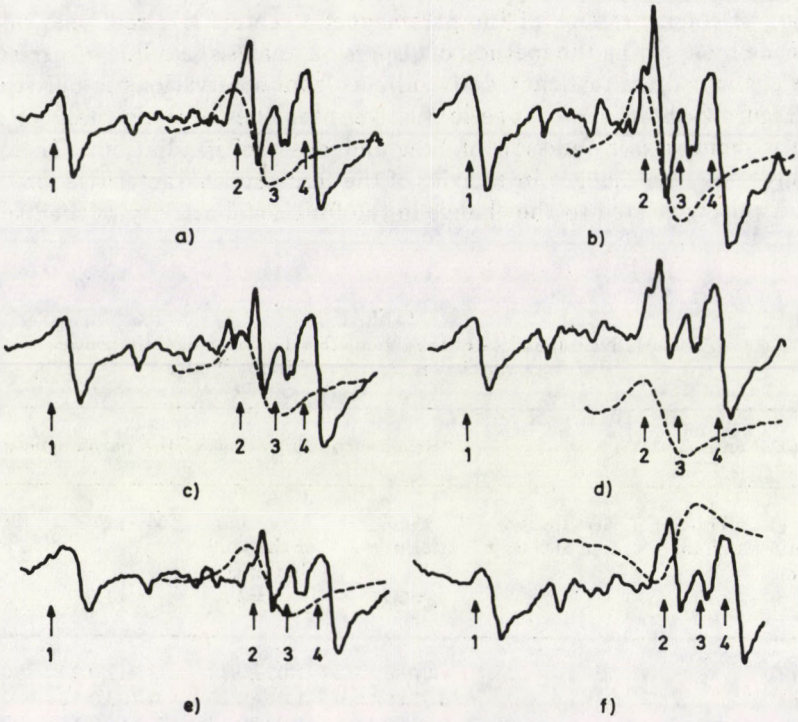


Fig. 2. The change in the EPR spectrum of a frozen liver tissues sample of Balb ♂ mice during the first day after: 1. administering of s-methionine or se-methionine (a, b, respectively) 2. administering of s-methionine or se-methionine (c, d, respectively) and irradiated with a dose of 2 Gy of X-rays 3. administering of s-methionine or se-methionine (e, f, respectively) and irradiated with a dose of 6 Gy of X-rays. The recording is carried out at 77 K, the microwave power is 10 mW and 0.2 mW in case of free radicals and the amplitude of the high frequency modulation is 1 mT (in units of magnetic field)

Taking into account the above mentioned results and supposing, as is well-known, that biologically active substances mobilize the internal potentialities of the organism to keep up its homeostasis against stress impacts, it becomes obvious that se-methionine may exert a certain radiation protective action.

From the results it appears that in animals treated with s-methionine and exposed to a sublethal X-ray dose amounting to 2 Gy the homeostasis is restored during the first 24 hours after irradiation (general adaptive syndrome according to Selye) and the functional activity of the ferment systems becomes normal.

However, in the case of exposure with higher-lethal-doses (6 Gy) a monotonic decrease in functional activity is observed during the first day. As a contrary, for

animals treated with se-methionine the functional activity of the ferment system goes to normal.

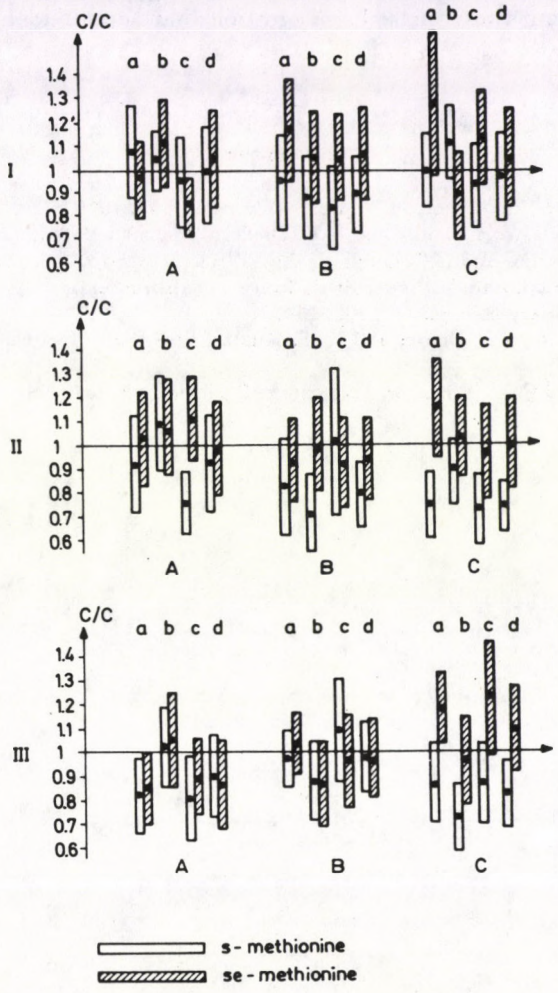


Fig. 3. The change in concentration of the paramagnetic centres (in relative units) as a function of time after administration of s-methionine or se-methionine and irradiation with X-rays in doses 2, 4.5 and 6 Gy (I, II, and III, respectively)

a. p-450 citochrome
b. free radicals of the semichinon type
c. Mo-protein
d. Fe-s-protein
A. during 3 hours
B. during 18 hours
C. during 24 hours after irradiation

Resuming the results, we underline that for irradiated animals which contained ordinary methionine the functional activity of the liver tissues drops with

the increase of doses. In animals treated with se-methionine, however, the functional activity was found to be significantly larger in comparison with the control group of animals which had not been protected.

We deduce from these facts the suggestion that se-methionine has radiation protective effect.

References

1. G.N. Bogdanov and V.M. Smonina, Voprosi Onkologii, *4*, 28, 1975.
2. G.N. Bogdanov, V.H. Varfolomeev and V.V. Diakova, Biofizika, *5*, 881, 1976.
3. G. Perona, R. Cellerino and G.C. Guidi, Scand. J. Haematol., *19*, 116, 1977.
4. G. Perona, G.C. Guidi and A. Piga, British Journal of Haematology, *42*, 567, 1979.
5. D.W. Peter, Veterinary Record, *107*, 193, 1980.
6. Joseph F. Weiss, Roger L. Hoover and K. S. Kumar, Free Rad. Res. Comm.,*3*., N 1-5, 33., 1987.
7. M. A. Beilstein and P. D. Whanger, J. of Inorg. Biochem., *29*, 137, 1987.

Manuscript received by Akadémiai Kiadó: 12 March 1987
Received for TEX typesetting by the Computer Automation Institute
of the Hungarian Academy of Sciences: 18 March 1988
Date of publication: 30 November 1988
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, Budapest

NOTES TO CONTRIBUTORS

I. PAPERS will be considered for publication in Acta Physica Hungarica only if they have not previously been published or submitted for publication elsewhere. They may be written in English, French, German or Russian.

Papers should be submitted to

Prof. I. Kovács, Editor
Department of Atomic Physics, Technical University
1521 Budapest, Budafoki út 8, Hungary

Papers may be either articles with abstracts or short communications. Both should be as concise as possible, articles in general not exceeding 25 typed pages, short communications 8 typed pages.

II. MANUSCRIPTS

1. Papers should be submitted in three copies.
2. The text of papers must be of high stylistic standard, requiring minor corrections only.
3. Manuscripts should be typed in double spacing on good quality paper, with generous margins.
4. The name of the author(s) and of the institutes where the work was carried out should appear on the first page of the manuscript.
5. Particular care should be taken with mathematical expressions. The following should be clearly distinguished, e.g. by underlining in different colours: special founts (italics, script, bold type, Greek, Gothic, etc.); capital and small letters; subscripts and superscripts, e.g. x^2, x_3; small *l* and l; zero and capital O; in expressions written by hand: *e* and *l*, *n* and *u*, *v* and v, etc.

A List of Symbols on a separate sheet should be attached to each paper.

6. References should be numbered serially and listed at the end of the paper in the following form: J. Ise and W. D. Fretter, Phys. Rev., *76,* 933, 1949.

For books, please give the initials and family name of the author(s), title, name of publisher, place and year of publication, e.g.: J. C. Slater, Quantum Theory of Atomic Structures, I. McGraw-Hill Book Company, Inc., New York, 1960.

References should be given in the text in the following forms: Heisenberg [5] or [5].

7. Captions to illustrations should be listed on a separate sheet, not inserted in the text.
8. In papers submitted to Acta Physica all measures should be expressed in SI units.

III. ILLUSTRATIONS AND TABLES

1. Each paper should be accompanied by three sets of illustrations, one of which must be ready for the blockmaker. The other sets attached to the copies of the manuscript may be rough drawings in pencil or photocopies.
2. Illustrations must not be inserted in the text.
3. All illustrations should be identified in blue pencil by the author's name, abbreviated title of the paper and figure number.
4. Tables should be typed on separate pages and have captions describing their content. Clear wording of column heads is advisable. Tables should be numbered in Roman numerals (I, II, III, etc.).

IV. RETURN OF MATERIAL

Owing to high postage costs, the Editorial Office cannot undertake to return *all* material not accepted for any reason for publication. Of papers to be revised (for not being in conformity with the above Notes or other reasons) only *one* copy will be returned. Material rejected for lack of space or on account of the Referees' opinion will not be returned to authors outside Europe.

Periodicals of the Hungarian Academy of Sciences are obtainable at the following addresses:

AUSTRALIA
C.B.D. LIBRARY AND SUBSCRIPTION SERVICE
Box 4886, G.P.O., *Sydney N.S.W. 2001*
COSMOS BOOKSHOP, 145 Ackland Street
St. Kilda (Melbourne), Victoria 3182

AUSTRIA
GLOBUS, Höchstädtplatz 3, *1206 Wien XX*

BELGIUM
OFFICE INTERNATIONAL DES PERIODIQUES
Avenue Louise, 485, *1050 Bruxelles*
E. STORY-SCIENTIA P.V.B.A.
P. van Duyseplein 8, *9000 Gent*

BULGARIA
HEMUS, Bulvar Ruszki 6, *Sofia*

CANADA
PANNONIA BOOKS, P.O. Box 1017
Postal Station "B", *Toronto, Ont. M5T 2T8*

CHINA
CNPICOR, Periodical Department, P.O. Box 50
Peking

CZECHOSLOVAKIA
MAD'ARSKA KULTURA, Národni třida 22
115 66 Praha
PNS DOVOZ TISKU, Vinohradská 46, *Praha 2*
PNS DOVOZ TLAČE, *Bratislava 2*

DENMARK
EJNAR MUNKSGAARD, 35, Nørre Søgade
1370 Copenhagen K

FEDERAL REPUBLIC OF GERMANY
KUNST UND WISSEN ERICH BIEBER
Postfach 46, *7000 Stuttgart 1*

FINLAND
AKATEEMINEN KIRJAKAUPPA, P.O. Box 128
00101 Helsinki 10

FRANCE
DAWSON-FRANCE S.A., B.P. 40, *91121 Palaiseau*
OFFICE INTERNATIONAL DE DOCUMENTATION ET LIBRAIRIE, 48 rue Gay-Lussac
75240 Paris, Cedex 05

GERMAN DEMOCRATIC REPUBLIC
HAUS DER UNGARISCHEN KULTUR
Karl Liebknecht-Straße 9, *DDR-102 Berlin*

GREAT BRITAIN
BLACKWELL'S PERIODICALS DIVISION
Hythe Bridge Street, *Oxford OX1 2ET*
BUMPUS, HALDANE AND MAXWELL LTD.
Cowper Works, *Olney, Bucks MK46 4BN*
COLLET'S HOLDINGS LTD., Denington Estate,
Wellingborough, Northants NN8 2QT
WM DAWSON AND SONS LTD., Cannon House
Folkstone, Kent CT19 5EE
H. K. LEWIS AND CO., 136 Gower Street
London WC1E 6BS

GREECE
KOSTARAKIS BROTHERS INTERNATIONAL BOOKSELLERS, 2 Hippokratous Street, *Athens-143*

HOLLAND
FAXON EUROPE, P.O. Box 167
1000 AD Amsterdam
MARTINUS NIJHOFF B. V.
Lange Voorhout 9–11, *Den Haag*
SWETS SUBSCRIPTION SERVICE
P.O. Box 830, 2160 Sz Lisse

INDIA
ALLIED PUBLISHING PVT. LTD.
750 Mount Road, *Madras 600002*
CENTRAL NEWS AGENCY PVT. LTD.
Connaught Circus, *New Delhi 110001*
INTERNATIONAL BOOK HOUSE PVT. LTD.
Madame Cama Road, *Bombay 400039*

ITALY
D. E. A., Via Lima 28, *00198 Roma*
INTERSCIENTIA, Via Mazzé 28, *10149 Torino*
LIBRERIA COMMISSIONARIA SANSONI
Via Lamarmora 45, *50121 Firenze*
SANTO VANASIA, Via M. Macchi 58
20124 Milano

JAPAN
KINOKUNIYA COMPANY LTD.
Journal Department, P.O. Box 55
Chitose, *Tokyo 156*
MARUZEN COMPANY LTD., Book Department
P.O. Box 5050 Tokyo International, *Tokyo 100-31*
NAUKA LTD., Import Department
2-30-19 Minami Ikebukuro, Toshima-ku, *Tokyo 171*

KOREA
CHULPANMUL, *Phenjan*

NORWAY
TANUM-TIDSKRIFT-SENTRALEN A.S.
Karl Johansgata 43, *1000 Oslo*

POLAND
WĘGIERSKI INSTYTUT KULTURY
Marszalkowska 80, *00-517 Warszawa*
CKP I W, ul. Towarowa 28, *00-958 Warszawa*

ROUMANIA
D. E. P., *Bucuresti*
ILEXIM, Calea Grivitei 64–66, *Bucuresti*

SOVIET UNION
SOYUZPECHAT — IMPORT, *Moscow*
and the post offices in each town
MEZHDUNARODNAYA KNIGA, *Moscow G-200*

SPAIN
DIAZ DE SANTOS Lagasca 95, *Madrid 6*

SWEDEN
ESSELTE TIDSKRIFTSCENTRALEN
Box 62, *101 20 Stockholm*

SWITZERLAND
KARGER LIBRI AG, Petersgraben 31, *4011 Basel*

USA
EBSCO SUBSCRIPTION SERVICES
P.O. Box 1943, *Birmingham, Alabama 35201*
F. W. FAXON COMPANY, INC.
15 Southwest Park, *Westwood Mass. 02090*
MAJOR SCIENTIFIC SUBSCRIPTIONS
1851 Diplomat, P.O. Box 819074,
Pallas, Tx. 75381-9074
READ-MORE PUBLICATIONS, INC.
140 Cedar Street, *New York, N. Y. 10006*

YUGOSLAVIA
JUGOSLOVENSKA KNJIGA, Terazije 27, *Beograd*
FORUM, Vojvode Mišića 1, *21000 Novi Sad*

Acta Physica Hungarica

VOLUME 64 · NUMBER 4, 1988

Akadémiai Kiadó, Budapest

ACTA PHYS. HUNG. APAHAQ 64 (4) 327–416 (1988) HU ISSN 0231-4428

ACTA PHYSICA HUNGARICA

A JOURNAL OF THE HUNGARIAN ACADEMY OF SCIENCES

EDITED BY
I. KOVÁCS

Acta Physica publishes original papers on subjects in physics. Papers are accepted in English, French, German and Russian.

Acta Physica is published in two yearly volumes (4 issues each) by

AKADÉMIAI KIADÓ
Publishing House of the Hungarian Academy of Sciences
H-1054 Budapest, Alkotmány u. 21

Subscription information

Orders should be addressed to

KULTURA Foreign Trading Company
1389 Budapest P.O. Box 149

or to its representatives abroad.

Acta Physica Hungarica is abstracted/indexed in Chemical Abstracts, Current Contents–Physical, Chemical and Earth Sciences, Mathematical Reviews, Science Abstracts, Physics Briefs, Risk Abstracts

CONTENTS

Manuscript received by Akadémiai Kiadó: 22 July 1987
Received for TEX typesetting by the Computer Automation Institute
of the Hungarian Academy of Sciences: 18 March 1988
Date of publication: 12 December 1988
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, Budapest

Acta Physica Hungarica 64 (4), pp. 327-338 (1988)

CONVERSION OF GRAVITATIONAL WAVES INTO ELECTROMAGNETIC WAVES IN A BIANCHI TYPE I UNIVERSE WITH A UNIFORM MAGNETIC FIELD — COSMOLOGICAL IMPLICATIONS

D. BOCCALETTI and W. AGOSTINI

Department of Mathematics, "La Sapienza" University of Rome
Rome, Italy

(Received in revised form 5 January 1985)

The process of conversion of g. waves into e.m. waves is studied in an axisymmetric Bianchi type I universe with a uniform magnetic field. A case is found in which the process is sufficiently effective to maintain a continuous interchange between photons and gravitons. This causes photons and gravitons to have the same frequency.

1. Introduction

The problem of the conversion of gravitational waves into electromagnetic waves and vice versa has been discussed from various points of view, in its own right and in view of detecting gravitational waves or in cosmological and astrophysical implications [1, 2, 3]. It has been shown that in a flat Minkowski background containing a uniform magnetic field gravitational waves are transformed into e.m. waves and v. v. at a rate growing quadratically with distance [2]. In spite of this, one must consider very high strength magnetic fields and paths of cosmological magnitude to get appreciable effects [2].

The case has also been studied in which the magnetic field is embedded in a conducting plasma with anisotropic conductivity; the results depend on the frequency of incoming gravitational waves but are not very different from the case of the empty space [4]. In this paper we study the conversion of g. waves into e.m. waves no longer in a flat background but in an anisotropic cosmological model endowed with a uniform magnetic field.

Our aim is to estimate if this process is working sufficiently to maintain photons and gravitons at the same frequency, taking into account that the inverse process must have the same rate. We proceed in this way: we consider the g. waves as first-order perturbation of the anisotropic metric with magnetic field and try to solve the Maxwell equations in this perturbed metric. As we are interested in the cosmological era preceding the recombination, we insert in the Maxwell equations (in the three cases we consider) asymptotic expressions for the metric coefficients valid near the initial singularity.

This enables us to solve analytically the Maxwell equations which in all the three cases lead to the solution of equations of the Bessel type. If we call

Acta Physica Hungarica 64, 1988
Akadémiai Kiadó, Budapest

η=(produced flux of e.m. waves)/(incident flux of g. waves), then we give an approximate evaluation of η as a function of time.

As we know, only if the primordial magnetic field is created with a strength not exceeding the so-called $B_{\text{crit}} = 4.4 \times 10^{13}$ gauss, can it be treated on a classical ground; furthermore it seems likely that, for fields of infinite strength or of strength $B >> B_{\text{crit}}$, a field organized on a large scale would not emerge as the universe expanded out of the singularity [5].

Therefore we take as fairly unrealistic the second and the third case we consider which, also at times over the Planck time $t_g \sim 10^{43}$s, involve field strengths largely exceeding B_{crit}. In the first case we study, on the contrary, we assume an initial field of the order of B_{crit}.

2. Axisymmetric cosmologies of Bianchi type I with a uniform magnetic field—asymptotic solutions

In the second half of the sixties, after the discovery of the cosmic microwave radiation background, spatially homogeneous anisotropic cosmological models endowed with a uniform magnetic field were extensively studied [6]. The hypothesis of the existence of primordial magnetic fields was largerly favoured [5, 6] because of the necessity of understanding how galactic magnetic fields could have arisen since the big-bang creation of the universe. Also if today we seem to have a satisfactory theory of galactic magnetic fields which do not require magnetic fields frozen into the matter since the origin of the universe, the hypothesis of primordial magnetic fields is not ruled out [7]. We assume the existence of such a primordial field.

We shall consider universe models with metric

$$ds^2 = c^2dt^2 - A^2(t)(dx^2 + dy^2) - W^2(t)dz^2 \tag{2-1}$$

filled with perfect-fluid matter at rest in the coordinate system of equations (2–1) and having an equation of state

$$P = \gamma \cdot \rho_M, \tag{2-2}$$

where $0 \leq \gamma < 1$ and ρ_M is the energy density of the perfect-fluid matter. The stress-energy tensor contains both ρ_M and ρ_B, where ρ_B is the energy density of the magnetic field. Solving the Maxwell equations in the metric (2–1) in the hypothesis of a uniform magnetic field aligned along the z-axis, one obtains for the primeval field

$$\bar{B}_z = B_0(WA^2)^{-1},$$

and for the magnetic energy density

$$\rho_B = B_0^2/8\pi \cdot A^{-4} \tag{2-3[1]}$$

[1]For the definition of the 3-dimensional fields we follow L.D. Landau, E.M. Lifshitz, Classical theory of fields, Pergamon Press, Oxford — New York, 1971, p. 256, from now on referred to as L.L.

(B_0 is a constant to be fixed).

As is known, the Einstein field equations in the case of (2–1), (2–2), (2–3) have been solved analytically in four cases [6]; moreover, there are solutions valid in asymptotic conditions, near the singularity. If we define an adimensional normalized time $\tau =$ at ($a \sim 10^{-17}$ s^{-1}), the asymptotic conditions, near the initial singularity, are characterized by $\tau << 1$ (being $\tau \sim 1$ to day). We shall study three cases:

1) Axisymmetric pancake singularity;
2) Isotropic point singularity;
3) Axisymmetric hard magnetic solution.

The case 1) is defined by

$$A = (1 + \alpha \cdot \tau^{(1-\gamma)}), \quad W = \tau, \quad \rho_B = B_0^2/8\pi \left[1 - 4\alpha \cdot \tau^{(1-\gamma)}\right], \tag{2–4}$$

(α non-negative constant) $0 \leq \gamma < 1$, and is an asymptotic solution. Of the three cases it is the only one which allows for a finite magnetic energy density for $\tau \to 0$; furthermore the solution becomes isotropic for large τ. We fix the constant $B_0 \sim B_{\text{crit}} = 4.4 \times 10^{13}$ gauss. The case 2), also an asymptotic solution, is given by

$$\begin{aligned} A = W = \tau^{\Gamma} (\Gamma = \frac{2}{3}(1+\gamma)), \qquad & 1/3 < \gamma < 1, \\ \rho_B = B_0^2/8\pi \cdot \tau^{-4\Gamma}. & \end{aligned} \tag{2–5}$$

Obviously $\rho_B \to \infty$ when $\tau \to 0$; we fix the constant $B_0 \sim 10^{-8}$ gauss (a possible value for the intergalactic field to day, but leading to a physically inconceivable magnetic field at the early times). The case 3) is at the same time an approximate solution ($\tau << 1$) and an exact solution (but not the general solution) and is given by

$$A = \tau^{1/2}, \qquad W = \tau^{\Lambda} (\Lambda = 1 - \gamma/1 + \gamma), \qquad 1/3 < \gamma < 1, \tag{2–6}$$

$\rho_B = B_0^2/8\pi \cdot \tau^{-2}$, where the constant B_0 is not arbitrary but has the value $B_0 =$ ca $/(G)^{1/2} [(1-\gamma)(3\gamma - 1)]^{1/2} /2(1+\gamma)$. Also this case has $\rho_B \to \infty$ when $\tau \to 0$.

3. The Maxwell equations in the perturbed metric

We have to solve the Maxwell equations:[2]

$$\begin{aligned} \partial F_{ik}/\partial x^l + \partial F_{ei}/\partial x^k + \partial F_{kl}/\partial x^i = 0, \\ (-g)^{-1/2} \partial/\partial x^k \left[(-g)^{1/2} F^{ik}\right] = -4\pi/c j^i, \end{aligned} \tag{3–1}$$

[2]For definitions, notations (except the gravitational constant we denote by G) and all that, see L. L.

in the metric

$$g_{ik} = g^0_{ik} + h_{ik},$$
$$g^{ik} = g^{0ik} - h^{ik},$$

where g^0_{ik} are the coefficients of the metric (2–1) and h_{ik} are first-order perturbations; the solutions of (3–1) are to be considered small of the first order like the h_{ik}: squares are negligible. We choose gravitational waves travelling in the positive x-direction and two independent polarizations denoted as

$$h^3_2(x,t) \text{ and } -h^2_2(x,t) = h^3_3(x,t).$$

Being perturbations, they will be linear and we consider plane waves:

$$\begin{aligned} h^3_2(x,t) &= \varepsilon_a \exp ik(x-ct), \\ -h^2_2(x,t) = h^3_3(x,t) &= \varepsilon_b \exp ik(x-ct), \end{aligned} \tag{3–2}$$

any g. wave being a sum of terms like (3–2), we solve (3–1), separately for each of the two polarizations.

Making use of the variable τ=at, we rewrite (3–2) as

$$\begin{aligned} h^3_2(x,\tau) &= \varepsilon_a \exp [ik(x-c/a\cdot\tau)], \\ -h^2_2(x,\tau) = h^3_3(x,\tau) &= \varepsilon_b \exp [ik(x-c/a\cdot\tau)]. \end{aligned} \tag{3–2'}$$

3a. Polarization h^3_2

In virtue of the assumed plane symmetry (3–1) yield in this case

$$\begin{aligned} \partial/\partial x(WA^2B_x) = 0; \quad \partial/\partial\tau(WA^2B_x) &= 0, \\ a/c\partial/\partial\tau(WA^2B_y) - WA^{-2}\partial/\partial x(W^{-1}A^2E_z) &= 0, \\ a/c\partial/\partial\tau(WA^2B_z) + W^{-1}\partial/\partial x(WE_y) &= 0, \\ W^{-1}A^{-2}\partial/\partial x(WE_x) = 4\pi/cj^0; \quad \partial/\partial\tau(WE_x) &= 0, \\ a/c\partial/\partial\tau(WE_y) + WA^{-2}\partial/\partial x(WA^2B_z) &= 0, \\ a/c\partial/\partial\tau(W^{-1}A^2E_z) - W^{-1}\partial/\partial x(WA^2B_y) &= WA^{-2}B_0\partial h^3_2/\partial x, \end{aligned} \tag{3–1a}$$

having taken into account that $F_{12} = WA^2B_z + B_0$. Rewriting and rearranging the relevant equations, we have:

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_z) + W^{-1}\partial/\partial x(WE_y) = 0, \\ a/c\partial/\partial\tau(WE_y) + WA^{-2}\partial/\partial x(WA^2B_z) = 0, \end{cases} \tag{3–1a'}$$

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_y) - WA^{-2}\partial/\partial x(W^{-1}A^2E_z) = 0, \\ a/c\partial/\partial\tau(W^{-1}A^2E_z) - W^{-1}\partial/\partial x(WA^2B_y) = WA^{-2}B_0\partial h^3_2/\partial x. \end{cases}$$

3b. Polarization $-h_2^2 = h_3^3$

In this case (3–1) yield:

$$\begin{aligned}
&\partial/\partial x(WA^2B_x) = 0; \qquad \partial/\partial\tau(WA^2B_x) = 0,\\
&a/c\partial/\partial\tau(WA^2B_y) - WA^{-2}\partial/\partial x(W^{-1}A^2E_z) = 0,\\
&a/c\partial/\partial\tau(WA^2B_z) + W^{-1}\partial/\partial x(WE_y) = 0,\\
&W^{-1}A^{-2}\partial/\partial x(WE_x) = 4\pi/cj^0; \quad \partial/\partial\tau(WE_x) = 0,\\
&a/c\partial/\partial\tau(WE_y) + WA^{-2}\partial/\partial x(WA^2B_z) = WA^{-2}B_0\partial h_2^2/\partial x,\\
&a/c\partial/\partial\tau(W^{-1}A^2E_z) - W^{-1}\partial/\partial x(WA^2B_y) = 0.
\end{aligned} \tag{3–1b}$$

And, rewriting and rearranging the relevant equations:

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_y) - WA^{-2}\partial/\partial x(W^{-1}A^2E_z) = 0,\\ a/c\partial/\partial\tau(W^{-1}A^2E_z) - W^{-1}\partial/\partial x(WA^2B_y) = 0, \end{cases} \tag{3–1b'}$$

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_z) + W^{-1}\partial/\partial x(WE_y) = 0,\\ a/c\partial/\partial\tau(WE_y) + WA^{-2}\partial/\partial x(WA^2B_z) = WA^{-2}B_0\partial h_2^2/\partial x. \end{cases}$$

In the subsequent sections we shall solve (3–1a') and (3–1b') in the three cases of asymptotic solution for the metric (2–1) we have discussed in Section 2, substituting the coefficients of (3–1a') and (3–1b') by their asymptotic expressions.

4. Axisymmetric pancake singularity

As we have seen in Section 2, this case is characterized by $A(\tau) = 1+\alpha\cdot\tau^{(1-\gamma)}$, $W(\tau) = \tau$ with $0 \leq \gamma < 1$ and α, a non-negative constant. For small τ, (3–1a') asymptotically become:

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_z) + \tau^{-1}\partial/\partial x(WE_y) = 0,\\ a/c\partial/\partial\tau(WE_y) + \tau\partial/\partial x(WA^2B_z) = 0, \end{cases} \tag{4–1a'}$$

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_y) - \tau\partial/\partial x(W^{-1}A^2E_z) = 0,\\ a/c\partial/\partial\tau(W^{-1}A^2E_z) - \tau^{-1}\partial/\partial x(WA^2B_y) = ik\varepsilon_a B_0\cdot\tau\cdot\ \exp\,[ik(x - c/a\cdot\tau)]. \end{cases}$$

The systems (4–1a') have the solutions

$$\begin{aligned}
WA^2B_y &= (A_1e^{-ikx} + B_1e^{ikx})\cdot kc/a\cdot\tau J_1(kc/a\cdot\tau) +\\
&\quad + (-i/5\cdot kc/a\cdot\tau^3 - 1/5\cdot\tau^2)\varepsilon_a B_0\ \exp\,[ik(x - c/a\cdot\tau)],\\
W^{-1}A^2E_z &= -i(A_1e^{-ikx} + B_1e^{ikx})\cdot J_0(kc/a\cdot\tau) +\\
&\quad + (i/5\,kc/a\cdot\tau^2 - 2/5\cdot\tau + 2/5\,ia/kc)\varepsilon_a B_0\exp\,[ik(x - c/a\cdot\tau)],\\
WA^2B_z &= (C_1e^{-ikx} + D_1e^{ikx})J_0(kc/a\cdot\tau),\\
WE_y &= -i(C_1e^{-ikx} + D_1e^{ikx})kc/a\cdot\tau\cdot J_1(kc/a\cdot\tau),
\end{aligned} \tag{4–2a'}$$

where J_0 and J_1 are Bessel functions[3]. Imposing to (4–2a') the obvious initial condition that the left members vanish at $\tau = 0$, the final result is:

$$\begin{aligned} WA^2B_y &= \varepsilon_a \cdot B_0 \{2/5 J_1(kc/a \cdot \tau) \cdot \tau \cdot e^{ikx} + (-i/5\, kc/a \cdot \tau^3 - 1/5 \cdot \tau^2) \cdot \\ &\quad \cdot \exp[ik(x - c/a \cdot \tau)]\}, \\ W^{-1}A^2E_z &= \varepsilon_a B_0 \{-2/5 i\, a/kc\, J_0(kc/a \cdot \tau) e^{ikx} + (i/5\, kc/a \cdot \tau^2 - \\ &\quad - 2/5 \cdot \tau + 2/5\, ia/kc) \exp[ik(x - c/a \cdot \tau)]\}, \\ WA^2B_z &= WE_y = 0 \end{aligned}$$

and then

$$\begin{cases} F_3^1 = \varepsilon_a B_0 \{-2/5 J_1(kc/a \cdot \tau) \cdot \tau \cdot e^{ikx} + (i/5\, kc/a \cdot \tau^3 + \\ \quad +1/5\tau^2) \exp[ik(x - c/a \cdot \tau)]\}, \\ F^{03} = \varepsilon_a B_0 \{2/5\, i\, a/kc\, J_0(kc/a \cdot \tau) \cdot \tau^{-1} \cdot e^{ikx} + (2/5 - i/5\, kc/a \cdot \tau - \\ \quad - 2/5\, i\, a/kc\, \tau^{-1}) \exp[ik(x - c/a \cdot \tau)]\}, \\ F_2^1 = F^{02} = 0. \end{cases} \tag{4–3a'}$$

In the same way (3–1b') asymptotically become:

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_y) - \tau\partial/\partial x(W^{-1}A^2E_z) = 0, \\ a/c\partial/\partial\tau(W^{-1}A^2E_z) - \tau^{-1}\partial/\partial x(WA^2B_y) = 0, \end{cases} \tag{4–1b'}$$

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_z) + \tau^{-1}\partial/\partial x(WE_y) = 0, \\ a/c\partial/\partial\tau(WE_y) + \tau\partial/\partial x(WA^2B_z) = -ik\varepsilon_b \cdot B_0\tau \exp[ik(x - c/a \cdot \tau)], \end{cases}$$

with the solutions:

$$\begin{aligned} WA^2B_y &= (A_2e^{-ikx} + B_2e^{ikx})kc/a \cdot \tau \cdot J_1(kc/a \cdot \tau), \\ W^{-1}A^2E_z &= -i(A_2e^{-ikx} + B_2e^{ikx}) \cdot J_0(kc/a \cdot \tau), \\ WA^2B_z &= (C_2e^{-ikx} + D_2e^{ikx})J_0(kc/a \cdot \tau) + (-1/3 + ikc/a \cdot \tau)\varepsilon_b B_0 \cdot \\ &\quad \cdot \exp[ik(x - c/a \cdot \tau)], \\ WE_y &= -i(C_2e^{-ikx} + D_2e^{ikx})kc/a \cdot \tau \cdot J_1(kc/a \cdot \tau) + \\ &\quad + (-i/3kc/a \cdot \tau^2) \exp[ik(x - c/a \cdot \tau)] \end{aligned} \tag{4–2b'}$$

and, imposing the initial conditions, we get:

$$\begin{cases} F_3^1 = F^{03} = 0, \\ F_2^1 = \varepsilon_b B_0 \{1/3 J_0(kc/a \cdot \tau)e^{ikx} - 1/3(ikc/a \cdot \tau + 1) \exp[ik(x - c/a \cdot \tau)]\}, \\ F^{02} = \varepsilon_b B_0 \{i/3kc/a J_1(kc/a \cdot \tau)e^{ikx} + i/3kc/a \cdot \tau \exp[ik(x - c/a \cdot \tau)]\}. \end{cases} \tag{4–3b'}$$

[3]We assume the notations and definitions of V. I. Smirnov: A course of higher mathematics, Vol. III, part 2, Pergamon Press, Oxford — New York.

5. Isotropic point singularity

This case is characterized by

$$A(\tau) = W(\tau) = \tau^{\Gamma}, \text{ where } \Gamma = 2/3(1+\gamma) \text{ and } 1/3 < \gamma < 1.$$

(3–1a') become:

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_z) + \tau^{-\Gamma}\partial/\partial x(WE_y) = 0, \\ a/c\partial/\partial\tau(WE_y) + \tau^{-\Gamma}\partial/\partial x(WA^2B_z) = 0, \end{cases} \tag{5–1a'}$$

$$a/c\partial/\partial\tau(WA^2B_y) - \tau^{-\Gamma}\partial/\partial x(W^{-1}A^2E_z) = 0,$$
$$a/c\partial/\partial\tau(W^{-1}A^2E_z) - \tau^{-\Gamma}\partial/\partial x(WA^2B_y) = ik\varepsilon_a B_0\tau^{-\Gamma}\exp[ik(x - c/a\cdot\tau)],$$

and have the solutions

$$\begin{aligned}
WA^2B_z =& (H_1e^{-ikx} + L_1e^{ikx})\cdot \\
&\cdot\{M_1\exp[-ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}] + N_1\exp[ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\}, \\
WE_y =& (H_1e^{-ikx} - L_1e^{ikx})\cdot \\
&\cdot\{-M_1\exp[-ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}] + N_1\exp[ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\}, \\
WA^2B_y =& (P_1e^{-ikx} + Q_1e^{ikx})\cdot \\
&\cdot\{R_1\cdot\exp[-ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}] + S_1\exp[ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\} + \\
&+ \varepsilon_a B_0e^{ikx}\cdot\Big\{\Big\{-2e^{ikc/a\cdot\tau} + \exp[-ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}] + \\
&+ \exp[ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}] - ikc/a\cdot \\
&\cdot\Big\{\exp[-ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\cdot\int_0^{\tau}\exp[ikc/a(\xi^{1-\Gamma}/(1-\Gamma) - \xi)]d\xi - \\
&- \exp[ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\cdot\int_0^{\tau}\exp[-ikc/a(\xi^{1-\Gamma}/(1-\Gamma) + \xi)]d\xi\Big\}\Big\}\Big\}, \\
W^{-1}A^2E_z =& (-P_1e^{-ikx} + Q_1e^{ikx})\cdot \\
&\cdot\{-R_1\exp[-ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}] + S_1\exp[ikc/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\} + \\
&+ i\varepsilon_a B_0kc/a\Big\{\exp[ik(x - c/a(1-\Gamma)\cdot\tau^{1-\Gamma})]\cdot \\
&\cdot\int_0^{\tau}\exp[ikc/a(\xi^{1-\Gamma}/(1-\Gamma) - \xi)]d\xi - \exp[ik(x + c/a(1-\Gamma)\cdot\tau^{1-\Gamma})]\cdot \\
&\cdot\int_0^{\tau}\exp[-ikc/a(\xi^{1-\Gamma}/(1-\Gamma) + \xi)]d\xi\Big\}.
\end{aligned} \tag{5–2a'}$$

And, imposing the initial conditions:

$$WA^2B_z = WE_y = 0,$$

$$WA^2B_y = \varepsilon_a B_0 \{-2\exp[ik(x - c/a\cdot\tau)] + \exp[ik(x - c/a(1-\Gamma)\cdot\tau^{1-\Gamma})] +$$
$$+ \exp[ik(x + c/a(1-\Gamma)\cdot\tau^{1-\Gamma})]\} - i\varepsilon_a B_0 kc/a\cdot$$
$$\cdot\left\{\exp[ik(x - c/a(1-\Gamma)\cdot\tau^{1-\Gamma})]\cdot\int_0^\tau \exp[ikc/a(\xi^{1-\Gamma}/(1-\Gamma) - \xi)]d\xi -\right.$$
$$\left. - \exp[ik(x + c/a(1-\Gamma)\cdot\tau^{1-\Gamma})]\cdot\int_0^\tau \exp[ikc/a(\xi^{1-\Gamma}/(1-\Gamma) + \xi)]d\xi\right\},$$

$$W^{-1}A^2E_z = i\varepsilon_a B_0 kc/a\cdot$$
$$\cdot\left\{\exp[ik(x - c/a(1-\Gamma)\cdot\tau^{1-\Gamma})]\cdot\int_0^\tau \exp[ikc/a(\xi^{1-\Gamma}/(1-\Gamma) - \xi)]d\xi -\right.$$
$$\left. - \exp[ik(x + c/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\cdot\int_0^\tau \exp[-ikc/a(\xi^{1-\Gamma}/(1-\Gamma) + \xi)]d\xi\right\},$$

and then:

$$F_2^1 = F^{02} = 0,$$

$$F_3^1 = \varepsilon_a B_0\cdot\tau^{-2\Gamma}\{2\exp[ik(x - c/a\cdot\tau)] - \exp[ik(x + c/a(1-\Gamma)\cdot\tau^{1-\Gamma}] -$$
$$- \exp[ik(x - c/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\} + i\varepsilon_a B_0 kc/a\tau^{-2\Gamma}\cdot$$
$$\cdot\left\{\exp[ik(x - c/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\cdot\int_0^\tau \exp[ikc/a(\xi^{1-\Gamma}/(1-\Gamma) - \xi)]d\xi -\right.$$
$$\left. - \exp[ik(x + c/a(1-\Gamma)\cdot\tau^{1-\Gamma})]\cdot\int_0^\tau \exp[-ikc/a(\xi^{1-\Gamma}/(1-\Gamma) + \xi)]d\xi\right\},$$

$$F^{03} = -i\varepsilon_a B_0 kc/a\tau^{-3\Gamma}\cdot \qquad (5\text{–}3a')$$
$$\cdot\left\{\exp[ik(x - c/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\cdot\int_0^\tau \exp[ikc/a(\xi^{1-\Gamma}/(1-\Gamma) - \xi)]...\xi -\right.$$
$$\left. - \exp[ik(x + c/a(1-\Gamma)\cdot\tau^{1-\Gamma})]\cdot\int_0^\tau \exp[-ikc/a(\xi^{1-\Gamma}/(1-\Gamma) + \xi)]d\xi\right\}.$$

In the same way, from (3–1b') we get:

$$F_3^1 = F^{03} = 0,$$

$$F_2^1 = -\varepsilon_b B_0\tau^{-2\Gamma}\{2\exp[ik(x - c/a\cdot\tau)] -$$
$$- \exp[ik(x + c/a(1-\Gamma)\cdot\tau^{1-\Gamma}] - \exp[ik(x - c/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\} -$$
$$- i\varepsilon_b B_0 kc/a\cdot\tau^{-2\Gamma}\cdot$$
$$\cdot\left\{\exp[ik(x - c/a(1-\Gamma)\cdot\tau^{1-\Gamma}]\cdot\int_0^\tau \exp[ikc/a(\xi^{1-\Gamma}/(1-\Gamma) - \xi)]d\xi -\right.$$
$$\left. - \exp[ik(x + c/a(1-\Gamma)\cdot\tau^{1-\Gamma})]\cdot\int_0^\tau \exp[-ikc/a(\xi^{1-\Gamma}/(1-\Gamma) + \xi)]d\xi\right\},$$

$$F^{02} = i\varepsilon_b B_0 kc/a \cdot \tau^{-3\Gamma} \cdot \left\{ \exp\left[-ik(x - c/a(1-\Gamma)\tau^{1-\Gamma})\right] \cdot \int_0^\tau \exp\left[ikc/a(\xi^{1-\Gamma}/(1-\Gamma) - \xi)\right] d\xi - \right.$$

$$\left. - \exp\left[ik(x + c/a(1-\Gamma)\tau^{1-\Gamma})\right] \cdot \int_0^\tau \exp\left[-ikc/a(\xi^{1-\Gamma}/(1-\Gamma) + \xi)\right] d\xi \right\}. \tag{5-3b'}$$

6. Axisymmetric hard-magnetic solution

$A(\tau) = \tau^{1/2}$; $W(\tau) = \tau^{\Lambda}$ where $\Lambda = 1 - \gamma/1 + \gamma$ and $1/3 < \gamma < 1$. For this case, (3–1a') become

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_z) + \tau^{-\Lambda}\partial/\partial x(WE_y) = 0, \\ a/c\partial/\partial\tau(WE_y) + \tau^{\Lambda-1}\partial/\partial x(WA^2B_z) = 0, \end{cases} \tag{6-1a'}$$

$$\begin{cases} a/c\partial/\partial\tau(WA^2B_y) - \tau^{\Lambda-1}\partial/\partial x(W^{-1}A^2E_z) = 0 \\ a/c0/\partial\tau(W^{-1}A^2E_z) - \tau^{-\Lambda}\partial/\partial x(WA^2B_y) = ik\varepsilon_a B_0 \cdot \tau^{\Lambda-1} \cdot \exp\left[ik(x - c/a \cdot \tau)\right] \end{cases}$$

and the solutions are

$$WA^2B_z = (\bar{E}e^{-ikx} + \bar{F}e^{ikx})(2kc/a \cdot \tau^{1/2})^{1-\Lambda} \cdot \left[\bar{G}_1 \cdot J_{1-\Lambda}(2kc/a \cdot \tau^{1/2}) + \bar{H} \cdot J_{\Lambda-1}(2kc/a \cdot \tau^{1/2})\right],$$

$$WE_y = -i(-\bar{E}e^{-ikx} + \bar{F}e^{ikx})(2kc/a)^{1-\Lambda} \cdot \tau^{1/2} \cdot \left[\bar{H} \cdot J_{\Lambda}(2kc/a \cdot \tau^{1/2}) - \bar{G} \cdot J_{-\Lambda}(2kc/a \cdot \tau^{1/2})\right],$$

$$W^{-1}A^2E_z = (\bar{M}e^{-ikx} + \bar{N}e^{ikx})(2kc/a\tau^{1/2})^{1-\Lambda} \cdot \left[\bar{P} \cdot J_{1-\Lambda}(2kc/a \cdot \tau^{1/2}) + \bar{Q} \cdot J_{\Lambda-1}(2kc/a \cdot \tau^{1/2})\right] +$$

$$+ e^{ikx} \cdot \left\{ \sum_{n=0}^{\infty} I_{0_n} b_n \cdot \tau^{n+\Lambda} + \sum_{n=0}^{\infty} c_n \cdot \tau^{n+\Lambda+1} \right\},$$

$$b_n = \frac{i(2\Lambda-1)(-1)^n(kc/a)^{2n+1} \cdot \varepsilon_a B_0}{(\Lambda+1)_n \cdot (2\Lambda)_n} \cdot \left[1/\Lambda(2\Lambda-1) - \sum_{m=1}^{n} (ia/kc)^m 1/m!(\Lambda)_m \cdot (2\Lambda-1)_m\right]$$

$$c_n = \frac{(-1)^n(kc/a)^{2n+2} \cdot \varepsilon_a B_0}{[(\Lambda+2)_n]^2} \left\{ 1/(\Lambda+1)^2 - \sum_{m=1}^{n} (ia/kc)^m 1/m![(\Lambda+1)_m]^2 \right\},$$

$$WA^2B_y = i(-\bar{M}e^{-ikx} + \bar{N}e^{ikx})(2kc/a)^{1-\Lambda} \cdot \tau^{\Lambda/2} \cdot \left[\bar{Q} \cdot J_{\Lambda}(2kc/a \cdot \tau^{1/2}) - \bar{P} \cdot J_{-\Lambda}(2kc/a \cdot \tau^{1/2})\right] + e^{ikx} \cdot \sum_{n=0}^{\infty} a_n \tau^{n+2\Lambda},$$

$$a_n = \frac{(-1)^{n+1} \cdot (kc/a)^{2n+2} \cdot \varepsilon_a \cdot B_0}{(\Lambda+1)_n \cdot (2\Lambda+1)_n} \cdot \left[1/2\Lambda^2 + \sum_{m=1}^{n} (ia/kc)^m \cdot 1/m! \cdot (\Lambda)_m \cdot (2\Lambda)_m \right], \qquad (6\text{–}2a')$$

where J_p is the Bessel function of not integer subscript (see footnote 3) and, as usual $(\Lambda)_n = \Lambda \cdot (\Lambda+1) \ldots (\Lambda+n-1)$. Imposing the initial conditions, it remains:

$$WE_y = WA^2 B_z = 0,$$

$$W^{-1} A^2 E_z = e^{ikx} \cdot \left\{ \sum_{n=0}^{\infty} b_n \cdot \tau^{n+\Lambda} + \sum_{n=0}^{\infty} c_n \cdot \tau^{n+\Lambda+1} \right\},$$

$$WA^2 B_y = e^{ikx} \cdot \left\{ \sum_{n=0}^{\infty} a_n \cdot \tau^{n+2\Lambda} \right\},$$

and then

$$F_2^1 = F^{02} = 0,$$

$$F_3^1 = -e^{ikx} \cdot \tau^{2\Lambda-1} \cdot \sum_{n=0}^{\infty} a_n \cdot \tau^n, \qquad (6\text{–}3a')$$

$$F^{03} = e^{ikx} \left[\tau^{2\Lambda-1} \cdot \sum_{n=0}^{\infty} b_n \cdot \tau^n + \tau^{2\Lambda} \sum_{n=0}^{\infty} c_n \cdot \tau^n \right].$$

In the same way, we get from (3–1b'):

$$F_3^1 = F^{03} = 0,$$

$$F_2^1 = e^{ikx} \cdot \sum_{n=0}^{\infty} d_n \cdot \tau^n, \qquad (6\text{–}3b')$$

$$F^{02} = -e^{ikx} \cdot \sum_{n=0}^{\infty} e_n \cdot \tau^n,$$

where

$$d_n = \frac{(-1)^{n+1} \cdot (kc/a)^{2n+2} \cdot \varepsilon_b \cdot B_0}{(\Lambda+1)_n \cdot (n+1)!} \left[1/\Lambda + \sum_{m=0}^{n} (ia/kc)^m \cdot (\Lambda)_m \right],$$

$$e_n = \frac{(-1)^{n+1} \cdot (kc/a)^{2n+2} \cdot \varepsilon_b \cdot B_0}{(\Lambda+2)_n \cdot (n+1)!} \left[1/(\Lambda+1) + \sum_{m=0}^{n} (ia/kc)^m \cdot (\Lambda+1)_m \right].$$

7. Conclusions

As already mentioned in the first Section, the process we have studied and the inverse process, on the ground of general principles, must have the same rate.

Therefore our process, if working effectively, does not help to annihilate gravitons but, on the contrary, it supplies them with the same frequency of the photons, because in the mutual conversion the frequency is not altered. To estimate the effectiveness of the process we shall calculate for each case the coefficient η =(produced flux of e.m. waves)/(incident flux of g. wave). For the produced e.m. waves travelling in the positive x-direction we must calculate $c \cdot \overline{T^{01}_{em}}$ (-means average). For the first g. wave polarization (ε_a) we have:

$$c \cdot \overline{T^{01}_{em}} = -c/4\pi \overline{(ReF^1_3 \cdot ReF^{03})}$$

and for the polarization ε_b:

$$c \cdot \overline{T^{01}_{em}} = -c/4\pi \overline{(ReF^{12} \cdot ReF^{02})}.$$

Moreover, we remember that $c \cdot \overline{T^{01}_{\text{grav}}} = c^5/32\pi G \cdot K^2 \cdot \varepsilon^2_{a,b}$. The coefficient η must be evaluated each time, taking into account, when depending on the frequency of the incident g. wave of the value of the frequency at that time (considering the redshift of the particular model). Obviously, as we have inserted in the Maxwell equations asymptotic expressions of $A(\tau)$ and $W(\tau)$, the expressions we get for η are valid near the singularity. In the first case (axisymmetric pancake singularity), we have for the polarization ε_a

$$\eta \sim 4/25 \left(\frac{GB^2_o}{c^4}\right) \frac{c^2}{a^2} \tau^4.$$

η is growing with time and, for $B_0 \sim B_{\text{crit}}$: $\eta \to 1$ for $\tau \to 10^{-8}$ $(t = 10^9 s)$. For the polarization ε_b

$$\eta \sim 4/9 (\frac{GB^2_0}{c^4}) \cdot \frac{c^2}{a^2} \tau^2;$$

taking again $B_0 \sim B_{\text{crit}}$, $\eta \to 1$ for $\tau \to 10^{-16}$ $(t \sim 10\text{s})$. For both polarizations η does not depend on the frequency of the incident gravitational wave.

In the second case (isotropic point singularity) there is no difference between the two polarizations and we get

$$\eta \sim 16 (\frac{GB^2_0}{c^4}) \cdot \frac{1}{k^2} \tau^{-3\Gamma}.$$

Here we have a dependence on the frequency and the process is appreciable only at times inconceivably smaller than the Planck time. As an example, for $B_0 \sim 10^{-8}$ gauss and a frequency $\sim t_g^{-1}$ at $t = t_g$, one has $\eta \sim 1$ for $\tau^{-\Gamma} \sim 10^{70}$ $(1/3 < \Gamma < 1/2)$. The axisymmetric hard-magnetic solution gives η not depending on the frequency and for the polarization ε_a.

$$\eta \sim 4 (\frac{GB^2_0}{c^4}) \cdot \frac{c^2}{a^2} \cdot \tau^{4\Lambda - 3} = \frac{(1-\gamma)(3\gamma - 1)}{(1+\gamma)^2} \tau^{4\Lambda - 3}.$$

For the polarization ε_b, we have

$$\eta \sim 4\left(\frac{GB_0^2}{c^4}\right)\frac{c^2}{a^2}\tau^{-2} = \frac{(1-\gamma)93\gamma-1)}{(1+\gamma)^2}\tau^{-2}$$

$(1/3 < \gamma < 1)$, $\Lambda = (1-\gamma)/(1+\gamma)$. Through decreasing with time η, in this case, has unphysical values also at large τ.

If we consider the first case as the more realistic one (it has finite magnetic energy density at the early times and becomes isotropic at large τ), we are led to think that with sufficiently large primordial magnetic fields a continuous interchange between photons and gravitons can take place. In this way, the e.m. waves and the g. waves retain the same frequency.

References

1. M.E. Gerstsenshtein, Sov. Phys. JETP, *14*, 84, 1962;
G.A. Lupanov, Sov. Phys. JETP, *25*, 76, 1967.
2. D. Boccaletti, V. De Sabbata, P. Fortini and C. Gualdi, Nuovo Cimento, *70B*, 129, 1970;
D. Boccaletti and F. Occhionero, Lett. Nuovo Cimento, *2*, 549, 1971;
W.K. De Logi and A.R. Mickelson, Phys. Rev. D-*16*, 2915, 1977.
3. L.P. Grishchuk, M.V. Sazhin, Sov. Phys. JETP, *41*, 787, 1976;
L.P. Grishchuk, Sov. Phys. Usp., *20*, 319, 1977.
4. D. Boccaletti, F. Occhionero, LAS-Frascati Internal Report N^0 40, December 1978.
5. K.S. Thorne, Ap. J., *148*, 51, 1967.
6. Ya.B. Zel'dovich, Sov. Phys. JETP, *21*, 656, 1965;
A.G. Doroshkevich, Astrophysics, *1*, 138, 1965;
I.S. Shikin, Sov. Phys. Doklady, *11*, 944, 1967;
K.C. Jacobs, Ap. J., *153*, 661, 1968; *155*, 379, 1969.
7. M. Reinhardt, M. N. R. A. S. *156*, 151, 1972.;
M. Reinhardt, and M.S. Roberts, Astrophys. Lett., *12*, 201, 1972.

Acta Physica Hungarica 64 (4), pp. 339-351 (1988)

SOME ASTROPHYSICAL CONSEQUENCES DUE TO THE EXISTENCE OF MAGNETIC MONOPOLES*

M. ORLANDINI

Institute of Physics, University of Ferrara, Ferrara, Italy

(Received 29 July 1985)

This paper reviews the theory of the magnetic monopole as an astrophysical object. In particular we construct a model by which we are able to calculate the mass of the monopole by the unique assumption that the source of energy of quasars is due to monopole–antimonopole annihilation, without referring to particular GUTs. It results that the mass of the monopole must be 10^{16} GeV, i.e. the monopole predicted by the Georgi–Glashow model.

1. Introduction

Since 1931, when Dirac [1] introduced it for the first time, the magnetic monopole has been considered a mathematical object rather than a physical one, because the theory was not able to predict its physical properties and so the experimentalists had no clues to reveal it.

This situation was unchanged till 1974, when 't Hooft [2] and Polyakov [3] demonstrated the "necessity" of the existence of the magnetic monopole in the framework of a theory that predicts the unification of three of the four fundamental forces: the electromagnetic, the weak and the strong interactions.

The key point of the 't Hooft–Polyakov theory is that it is able to calculate the mass of the monopole in terms of free parameters of the Grand Unification Theories (GUTs). From this calculation it results that the monopole has a mass in the range of 10^4–10^{19} GeV [4,5], according to different theories, and the standard model of Georgi–Glashow [6] predicts a value of 10^{16} GeV. Because of the value of the mass, the monopole cannot be produced in any particle accelerators: an object of such energy can be produced only in the very early Universe, immediately after the "Big Bang".

So the magnetic monopole becomes a cosmological problem: in fact the matching between Cosmology and Particle Physics leads to the birth of the "new cosmology", which describes the GUTs in the framework of Einstein's General Theory of Relativity. From this matching a double benefit occurs: the cosmologists can now extrapolate the Big Bang theory to times less than one second without making

*Present address: International School for Advanced Studies - Strada Costiera 11 - 34014 Trieste - Italy

aprioristic assumptions (i.e. they can justify the initial conditions) while the theorists employ the Universe as a laboratory in which to search for the tests of their theories.

In Part 2 of this paper we will give a review of the theory of the magnetic monopole, both the classical and the GUT one. In Part 3 we will describe the inflationary Universe model; in Part 4 we will calculate the limits to the density and the flux of the monopoles from astrophysical observations; in Part 5 we will construct a model of energy production for quasars, from which we obtain, for the mass of the monopole, a value of 10^{16} GeV, i.e. the mass predicted from the standard model of grand unification, the Georgi–Glashow model. In this paper, except Part 5, we use units for which $\hbar = c = k$ (Boltzmann constant) $= 1$.

2. Theory of the magnetic monopole

More than fifty years ago Dirac [1] demonstrated that the quantization of the electric charge can be explained supposing the existence of at least one free magnetic charge: the magnetic monopole.

Dirac considered the monopole like a semi-infinite and infinitesimal thin solenoid; in fact if we consider a monopole of magnetic charge g, placed in the origin of a polar coordinate system, it will produce a magnetic field of strength

$$\bar{B} = g\frac{\hat{r}}{r^2}, \tag{1}$$

where $\hat{r}$ is the radial unit vector. In a suitable gauge the vector potential associated to this field has the form

$$\bar{A} = g\frac{1+\cos\vartheta}{\sin\vartheta}\frac{\hat{r}}{r}. \tag{2}$$

Although (1) is singular only for $r = 0$, (2) has a singularity also for $\vartheta = 0$. This "string" singularity does not depend on the chosen gauge: we will always find a string singularity from $r = 0$ to $r = \infty$.

Because the quantum mechanical equations of charged particles are expressed directly in terms of the vector potential, it is necessary to make the singularity unobservable.

So we consider the effect that a trip around the string produces on the wawe function of a particle with electric charge $-e$: generally the phase of the wave function will be modified because of the vector potential $\bar{A}$

$$\psi \to \psi' = \exp[-ie\int \bar{A}\cdot d\bar{x}]\psi. \tag{3}$$

Once the integral is computed, with the aid of (2), we obtain

$$\psi' = \exp[4\pi ige]\psi. \tag{4}$$

So the string singularity is not detectable if the magnetic charge of the monopole obeys the condition

$$g = \frac{n}{2e}, \tag{5}$$

when n is an integer. The smallest magnetic charge is the Dirac magnetic charge g_D, the value of which is

$$g_D = 1/2e \sim 68.5e \sim 3.310^{-8} \text{ CGS units.} \tag{6}$$

Now we reverse the reasoning: we suppose there is a particle of electric charge Q and zero magnetic charge. Because of the existence of the magnetic monopole we must have

$$\exp[4\pi i Q g_D] = 1 \tag{7}$$

and so

$$\frac{Q}{e} = \frac{1}{2eg_D} = m, \tag{8}$$

where m is an integer. The existence of the magnetic monopole involves the quantization of the electric charge. Someone could object that, in the calculation of g_D, we should have used $-e/3$, the charge of the quark d, and not $-e$ because it should be considered the smallest electric charge. This is not true because of the quark confinement: in fact it implies that the colour charge is screened for a distance greater than some fermis so that the previous discussion is still right. The Dirac's theory cannot predict the mass of the monopole because, from the Maxwell equation

$$div(\bar{B}) = 0 \tag{9}$$

it follows that the electromagnetic mass, considering a pointlike monopole, must be divergent.

As we have seen, the existence of the magnetic monopole implies the quantization of electric charge; in 1974 the inverse has been demonstrated [2,3]: every model of grand unification in which the electric charge is quantized (and so the electromagnetic group U(1) is embedded in the group which is broken by the mechanism of spontaneous symmetry breaking [7], contains necessarily magnetic monopoles.

So if we apply the GUTs we can determine the mass and the radius of a monopole. For example [8] we consider the electroweak theory which has (for simplicity) as gauge bosons the massless photon γ and the heavy $W^{\pm}$. As we have seen, (9) shows that the magnetic field of the electromagnetic interaction $\bar{B}^{\gamma}$ is everywhere divergent. Yet in this theory it is necessary to consider the "covariant" derivative [9] and not the usual one. So (9) becomes

$$\bar{D} \cdot \bar{B} = div(\bar{B}^{\gamma}) + ie\bar{A}^{+}\bar{B}^{-} - ie\bar{A}^{-}\bar{B}^{+} = 0, \tag{10}$$

where $\bar{A}^{\pm}$ and $\bar{B}^{\pm}$ are the fields associated with the bosons $W^{\pm}$. In this way (10) can be written in the form

$$div(\bar{B}^{\gamma}) = \rho, \tag{11}$$

where ρ is the source term due to the gauge bosons:

$$\rho = ie(\bar{A}^- \bar{B}^+ - \bar{A}^+ \bar{B}^-). \tag{12}$$

Because it is possible for (12) to have a constant value, different from zero, in a region of the size of the Compton wavelength of the gauge bosons, this implies that the source is no longer pointlike and so there is no problem of divergence: the physical properties of monopoles are well-defined.

Actually it is not possible to use the electroweak theory because we need a theory which considers also the strong interaction. The Georgi–Glashow model, based on the symmetry breaking from the group SU(5) to the group SU(3) × SU(2) × U(1), predicts that the physical properties of monopoles are

$$r \approx \frac{1}{M_x} \approx 10^{-28}\text{cm}, \qquad m \approx \frac{M_x}{\alpha_{\text{GUT}}} \approx 10^{16}\text{GeV}, \tag{13}$$

where M_x is the mass of one of the bosons that mediates the unified interactions and α_{GUT} the gauge coupling constant. As we can see (13) depends strictly from the particular GUT: in fact the mass of the monopole is predicted to be in the range from 10^4 to 10^{16} GeV, according as we consider particular phase transitions [10] or we introduce also the gravitational interaction (Kaluza–Klein theories).

Because, as we have shown, the Maxwell equations break down at very short distance, the magnetic monopole must have an internal structure, which is as follows [11, 12, 13]

a) the core, with radius $r_c \sim 10^{-28}$ cm, in which there are virtual X bosons;
b) the electroweak region, with radius $r_{ew} \sim 10^{-16}$ cm, containing virtual bosons $W^{\pm}, Z^0$ and γ;
c) the confinement region, with radius $r_{cf} \sim 10^{-13}$ cm, where there are strongly interacting objects (quarks, antiquarks and gluons);
d) a fermion–antifermion condensate, with radius of about one fermi for hadrons and few fermis for leptons, in which the baryonic number is not conserved;
e) for distances greater than few fermis the monopole acts like a Dirac monopole, yielding a magnetic field $|\bar{B}| = g_D/r^2$. From these results it follows that the monopole can act as a catalyst in the proton decay process [11, 12]

$$\begin{aligned} M + p &\to M + e^+ + \pi^0, \\ M + p &\to M + e^+, \end{aligned} \tag{14}$$

with cross-section of the order of 10^{-27}cm^2. Consequently, we have that the $M - \bar{M}$ annihilation cross section is larger than the classical one [14]; in fact from (14) it follows that the reaction

$$M + \bar{M} \to \bar{p} + e^+ + \pi^0 \tag{15}$$

has the same cross section of the catalysis process, i.e. $\sigma \sim 10^{-26}$ cm^2.

3. The inflationary Universe

The standard cosmological theory is the so-called "Big Bang" theory [15]. Starting from the assumption of an isotropic and homogeneous Universe and so described by a Robertson–Walker metric [16]

$$ds^2 = dt^2 - R^2(t)\left\{\frac{dr^2}{1-kr^2} + r^2(d\vartheta^2 + \sin^2\vartheta d\phi^2)\right\}, \tag{16}$$

where k assumes the values -1, 0, $+1$ for an open, flat or closed Universe, respectively, it describes the evolution of the scale factor $R(t)$ by means of the Einstein–Fridman equations

$$\left(\frac{dR}{dt}\right)^2 = \frac{8}{3}\pi G\rho R^2 - k, \tag{17}$$

$$\frac{d^2R}{dt^2} = -\frac{4}{3}\pi G(\rho + 3p)R. \tag{18}$$

This system of equations, plus the energy conservation, the state equation of matter and the additional assumption of an adiabatic expansion can explain a lot of observed features of our Universe: the red-shifted light from distant galaxies; the cosmic microwave background radiation and the abundance of the primordial elements [15,16]: Deuterium and Helium.

But if we extrapolate the model to times less than one second we cannot understand why the Universe is so uniform on a great scale (the "horizon problem") [17, 18]; why the density of the Universe is so close to the critical one (the "flatness problem") [19] and why the Universe is composed only of matter and not of antimatter (the "baryon number" problem) [20]. To solve these problems the cosmologists imposed ad hoc conditions, assumed as initial conditions. Among these the less acceptable was the assumption that the Universe had to be originally asymmetrical between baryons and antibaryons while it is extremely plausible that matter and antimatter should appear on the same footing.

Just this last problem has led to introduce theories that predict the non conservation of the baryonic number: the GUTs. The GUTs predict that for very high energies the fundamental interactions are unified, i.e. mediated by the same gauge boson. As the energy (and so the temperature) drops, the interactions decoupled by means of a mechanism called spontaneous symmetry breaking, which is driven by a potential which assumes different forms according as the particular type of GUT is taken into account. The result which we arrive in applying these theories to the Einstein–Fridman equations is that the scale factor $R(t)$ has, for a certain period, an exponential law of evolution ("inflation") [21]

$$R(t)\alpha\begin{cases} e^{Ht} & T \ll T_c, \\ t^{1/2} & T \gg T_c, \end{cases} \tag{19}$$

where $H = (8\pi G V_0/3)^{1/2}$, G is the Newton constant and V_0 the vacuum value of the potential. For a temperature larger than that at which the decoupling occurs T_c the scale factor resumes the power law predicted from the standard theory.

This theory solves in a very elegant way all the previous problems but a new problem comes in: during the phase transition that decouples the interactions some defects occur, defects that physically correspond to magnetic monopoles.

If we calculate the number density of these monopoles we obtain [14]

$$n \sim 10^{-10}\, T^3 \tag{20}$$

of the same order of the baryon one. This is naturally against the experimental observations because the monopole mass is much larger than the baryon one (of the order of 10^{16}!). To solve the "magnetic monopole" problem it has been considered a particular GUT, the Coleman–Weinberg theory [22]; but using this theory we obtain a monopole density extremely low [23,24,25]

$$n \sim 10^{-220}\, T^3 \tag{21}$$

so the direct observation of monopole becomes practically impossible. As we have seen the theory is not able to predict the magnetic monopole density unambiguously. In the next Section we will show how to calculate limits on the flux of monopoles from astrophysical observations.

4. Experimental limits to monopole density

Because the detectable monopoles come from cosmic rays, they are influenced by gravitational fields and/or magnetic fields. We can give an order of magnitude to the velocity acquired by a monopole free falling in the gravitational field of our galaxy; in fact

$$v \sim \left(\frac{GM_G}{R_G}\right)^{1/2} \sim 10^{-3} c, \tag{22}$$

where $M_G \sim 10^{11} \mathrm{M}_\odot \sim 10^{44} g$ and $R_G \sim 10^{22}$ cm are the mass and the radius of our galaxy [26]. As we can see [22] is mass independent.

To compute the effect of the galactic magnetic field on a monopole we remember that our galaxy has a magnetic field of 10^{-6} gauss uniform on a scale $L \sim 10^{21}$ cm; the velocity acquired by the monopole is

$$v = \left(\frac{2gBL}{m_M}\right)^{1/2} = \left(\frac{BL}{em_M}\right)^{1/2} \sim 3 \cdot 10^{-3} c \left(\frac{B}{3 \cdot 10^{-6}\mathrm{gauss}}\right)^{1/2} \left(\frac{L}{10^{21}\mathrm{cm}}\right)^{1/2} \left(\frac{10^{16}\mathrm{GeV}}{m_M}\right)^{1/2} \tag{23}$$

for a monopole of magnetic charge $g_D = 1/2e$.

From (23) we see that for $m_M \geq 10^{17}$ GeV the velocity acquired in the gravitational field is larger than that acquired in the magnetic one, moreover, that it is possible to consider the monopole as a classical object for $m_M \geq 10^{11}$ GeV.

For the interactions between monopoles and condensed matter it has been found [27] that, for rather slow monopoles ($\beta \sim 10^{-2}$), the stopping power is of the order of 10^{11} cm; in this way the Earth is trasparent to their motion.

A limit on the monopole density is first of all set by the critical density of matter in the Universe, i.e. $\rho \leq 10^{-29}$ g/cm so that

$$n \leq 10^{-21}\mathrm{cm}^{-3}\left(\frac{10^{16}\mathrm{GeV}}{m_M}\right). \tag{24}$$

Combining this expression with (23), valid for $m_M \leq 10^{17}$ GeV, we obtain

$$nv \leq 10^{-2}\mathrm{m}^{-2}\mathrm{yr}^{-1}\left(\frac{10^{16}\mathrm{GeV}}{m_M}\right)^{3/2}, \quad m_M \leq 10^{17}\mathrm{GeV}. \tag{25}$$

For a monopole mass greater than 10^{17} GeV the gravitational interaction prevails on the electromagnetic one so that the monopoles will be confined in the galactic halo, increasing their density. In this case, assuming a mass of the galactic halo of $10^{12}\mathrm{M}_\odot$ and a velocity of $10^{-3}c$ we obtain

$$nv \leq 10^{2}\mathrm{m}^{-2}\mathrm{yr}^{-1}\left(\frac{10^{17}\mathrm{GeV}}{m_M}\right), \qquad m_M \geq 10^{17}\mathrm{GeV}. \tag{26}$$

Moreover, is possible to obtain limits on the monopole flux by indirect reasoning, like the persistence of the galactic magnetic field [28,29,30] or the study of the X emission from neutron stars [31,32]. We see in detail these arguments. If the galactic magnetic field is due to the presence of persistent currents [33] (because rot $(\bar{B}_{\mathrm{gal}}) \neq 0$) then the monopoles that move in our galaxy acquired energy at the field expense, reducing consequently the currents. The actual observation of the galactic magnetic field implies that the rate at which the energy is dissipated is smaller than the regeneration rate of the currents by dynamo mechanism, with a period τ of about 10^8 years, so we have that

$$nv \leq 2 \cdot 10^{-4}\mathrm{m}^{-2}\mathrm{yr}^{-1}\left(\frac{B}{3 \cdot 10^{-6}\mathrm{gauss}}\right)\left(\frac{10^{8}\mathrm{yr}}{\tau}\right), \tag{27}$$

which is independent of the mass of the monopole. Actually, we have neglected the effects of the gravitational field and this assumption is valid only for monopoles of mass smaller than 10^{17} GeV. In fact the very massive monopoles "feel" less the influence of the galactic magnetic field and so they extract from it less energy; in this way we obtain a less stringent limit on their flux

$$nv \leq 2 \cdot 10^{-4}\mathrm{m}^{-2}\mathrm{yr}^{-1}\left(\frac{m_M}{10^{17}\mathrm{GeV}}\right)\left(\frac{B}{3 \cdot 10^{-6}\mathrm{gauss}}\right)\left(\frac{10^{8}\mathrm{yr}}{\tau}\right); m_M \geq 10^{17}\mathrm{GeV}. \tag{28}$$

Combining (28) with (26) we obtain a mass independent limit for the monopole flux (Parker limit)

$$nv \leq 10^{-1}\mathrm{m}^{-2}\mathrm{yr}^{-1}. \tag{29}$$

We can obtain another limit on the monopole flux from the observation of neutron stars. In fact because of their very high density the neutron stars capture all the monopoles which fall on them and if the number of monopoles present in the star at the moment of its formation is negligible, then the number N of monopoles present in a neutron star will be proportional to their flux. Because it has been found [11,12] that the monopoles act as catalyst in the nucleon decay process with cross section of the same order of that of strong interaction (of the order of $10^{-27}\mathrm{cm}^2$), it follows that also a little number of monopoles can cause a neutron star to emit in the ultraviolet and X. From observation in these frequency ranges it is possible to obtain a limit for σN and so limit for nv:

$$nv \leq 10^{-12}\mathrm{m}^{-2}\mathrm{yr}^{-1}\left(\frac{10^{-26}\mathrm{cm}^2}{\sigma}\right) \tag{30}$$

in clear contrast with (29).

5. $M - \bar{M}$ annihilation as "power supply" of quasars

The unsuccessful research of the magnetic monopoles shows that their number is now very little, below the sensibility of the detectors. But the theory predicts them so a mechanism that reduces their number is necessary, for example the $M - \bar{M}$ annihilation process. The energy emitted in this process, because of the mass of the monopole, would be enormous so we should observe "signs" of annihilation in very old astrophysical objects and therefore very close to the Big Bang. A good candidate as a seat of the $M - \bar{M}$ annihilation process seems to be quasars, quasi stellar objects characterized by having the lines of their spectra red-shifted of 16% and more with respect to the wavelength observed on the Earth. Explaining this shift as due to Doppler effect, they would be very far from us and so very close to the Big Bang. Once their distance is known, it is possible to calculate their intrinsic (absolute) luminosity, which is of the order of $10^{39}\mathrm{W}$ almost independent of the frequency of observation.

The physical properties of quasars are summarized in Table I [34]. From the analysis of the physical characteristics of quasars we deduce that the thermonuclear reactions cannot occur because the density is too low. The observed energy might be produced gravitationally but in this case the radiation pressure would prevail on the pressure of the gas, preventing the collapse. So our idea has been to consider a quasar like a cloud of mass $\mathcal{M}$ and volume V containing a number of monopoles and antimonopole, with the same density, so that their mass energy does not exceed the mass energy of the matter.

Because the major contribution to the mass energy of matter is due to protons, it will be

$$N_p E_p = 2N_M E_M, \tag{31}$$

Table I

Object	z	L_{OPT}	L_{IR}	L_{RADIO}
3CR 273	0.158	$1.5\ 10^{46}$	$6.7\ 10^{46}$	$5.0\ 10^{44}$
1400+16	0.244	$1.0\ 10^{45}$	$7.6\ 10^{44}$	
3C 279	0.536	$3.1\ 10^{46}$		$2.3\ 10^{45}$
3C 446	1.404	$2.1\ 10^{46}$	$5.4\ 10^{47}$	$6.4\ 10^{45}$
PKS2126-15	3.270	$1.0\ 10^{48}$		

The mass of a quasar is $M \approx 10^{41} - 10^{43}$ g
The radius of a quasar is $R \approx 10^{16}$ cm

where the factor 2 has been introduced to distinguish monopoles from antimonopoles. The number of protons in a quasar is

$$N_p = \mathcal{M}/m_p, \tag{32}$$

so that the monopole (antimonopole) number density, supposing their volume constant in time and equal to the volume of the quasar, is

$$n_M = \frac{N_M}{V} = \frac{\mathcal{M}}{2VM_M(g)}\text{cm}^{-3}, \tag{33}$$

where $M_M(g)$ stands for the mass of the monopole expressed in grams. Because of continuous annihilation we have that the number of monopoles (antimonopoles) will be a time-dependent quantity; in particular

$$\frac{d}{dt}n_M(t) = - < \sigma v > n_M^2(t), \tag{34}$$

where σ is the annihilation cross section, of order of 10^{-27}cm^2 and $v \sim 10^{-3}c$ is the monopole velocity. Integrating (34) with the condition that for $t = 0$ $n_M = n_{M_0}$ we have

$$n_M(t) = \frac{n_{M_0}}{1 + < \sigma v > n_{M_0} t}, \tag{35}$$

where n_{M_0} is given by (33).

The luminosity $\mathcal{L}$ of a quasar will become function of its age with the following expression

$$\mathcal{L} = 2M_M(g)c^2 n_M^2(t) < \sigma v > V = \frac{\mathcal{M}^2 c^2}{2VM_M(g)} < \sigma v > \left[1 + \frac{< \sigma v > \mathcal{M}t}{2VM_M(g)}\right]^{-2}. \tag{36}$$

The mass-luminosity diagram is shown in Fig. 1; as we can see there is a maximum luminosity, corresponding to the limit case of a quasar of infinite mass, given by

$$\mathcal{L}_{\max} = \frac{2VM_M(g)c^2}{<\sigma v> t^2}. \tag{37}$$

The inverse function, i.e. $\mathcal{M} = \mathcal{M}(\mathcal{L})$, of (36) is

$$\mathcal{M} = \frac{2VM_M(g)}{(2VM_M(g) <\sigma v> /\mathcal{L})^{1/2} - <\sigma v> t}. \tag{38}$$

From this expression we see that to have $\mathcal{M} > 0$ it is necessary for the denominator to be positive, i.e.

$$2VM_M/\mathcal{L} > <\sigma v> t^2. \tag{39}$$

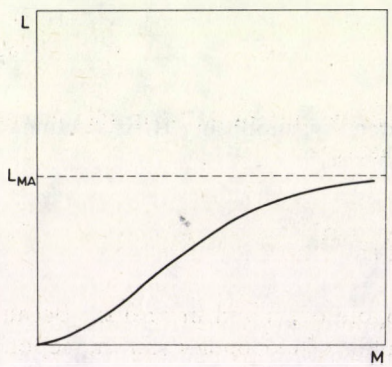


Fig. 1

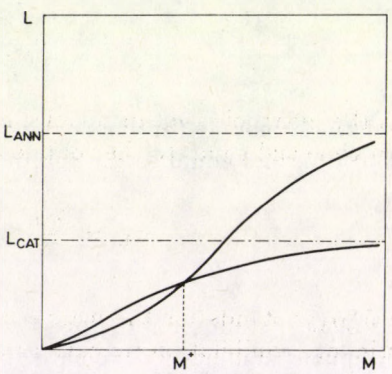


Fig. 2

This restriction on $\mathcal{L}$ implies also a restriction on the mass of the monopole. In fact if we consider $\mathcal{L} \sim \mathcal{L}_{\max} \sim 10^{39}$W we obtain Table II.

From this Table it is possible to see that the only monopole that allows for a luminosity smaller than $\mathcal{L}_{\max}$ and consistent with every observed red-shift is the monopole with mass of 10^{16} GeV. i.e. the monopole predicted by the Georgi–Glashow model.

Another mechanism which can increase the brightness of a quasar is the catalysis of the monopole on the proton decay. In this case the number density of the protons will vary in time with law

$$\frac{d}{dt} n_p(t) = - <\sigma v> n_M(t) n_p(t), \tag{40}$$

Table II

E_M(GeV)	z	$M(g)$	E_M(GeV)	z	$M(g)$
10^{16}	3	$7.83\ 10^{42}$	10^{16}	1	$8.69\ 10^{42}$
10^{15}	3	$2.80\ 10^{42}$	10^{15}	1	$4.34\ 10^{42}$
10^{14}	3	$1.53\ 10^{42}$	10^{14}	1	$-1.65\ 10^{42}$
10^{13}	3	$-3.78\ 10^{41}$	10^{13}	1	$-6.55\ 10^{40}$
10^{4}	3	$-1.45\ 10^{32}$	10^{4}	1	$-5.12\ 10^{31}$
10^{16}	2	$8.07\ 10^{42}$	10^{16}	0.5	$9.57\ 10^{42}$
10^{15}	2	$3.13\ 10^{42}$	10^{15}	0.5	$7.99\ 10^{42}$
10^{14}	2	$3.53\ 10^{42}$	10^{14}	0.5	$-6.33\ 10^{41}$
10^{13}	2	$-1.57\ 10^{41}$	10^{16}	0.1	$1.15\ 10^{43}$
10^{4}	2	$-9.42\ 10^{31}$	10^{15}	0.1	$-1.89\ 10^{43}$

where the catalysis cross section is equal to the annihilation one. Developing the same kind of calculations done for the annihilation process we obtain that the luminosity produced by the catalysis process is

$$\mathcal{L}_{\text{cat}} = \frac{\mathcal{M}^2 c^2}{V M_M(g)} < \sigma v > \left[1 + \frac{< \sigma v > \mathcal{M} t}{V M_M(g)}\right]^{-2}, \tag{41}$$

which compared with (36) shows that the two processes differ in the fact that in the annihilation there is $2V$ while in the catalysis there is only V. With the same arguments as the previous one we arrive at a constraint on the monopole mass, which must be necessarily 10^{16} GeV.

Now it is necessary to establish for which values of the physical parameters a process prevails on the other. In order that (36) is greater than (41) it has to be

$$\mathcal{M} > \frac{\sqrt{2} V M_M(g)}{< \sigma v > t} \equiv \mathcal{M}^*, \tag{42}$$

so for a certain age of the quasar the brightness due to annihilation prevails on that due to catalysis if $\mathcal{M} > \mathcal{M}^*$ (Fig. 2). From Table I we have, for a typical quasar

$$\mathcal{M}^* = 1.3 \cdot 10^{43}(1+z)^{3/2} \tag{43}$$

and so the observed luminosity is due to the catalysis process. This model can explain effectively the energy emitted from quasars, determining moreover the mass of the monopole, but it contains a "bug": from (35) we have that the time necessary to reduce the initial number of monopoles to half is much greater than the age of the Universe so it would be possible to observe them: of course it is not in agreement with the observational data. A way to solve this problem might be to assume that the volume occupied by the monopoles is a time-dependent quantity, but this is at the moment, not clear, yet.

6. Conclusion

The theoretical previsions about the number of monopoles that we should observe now are not in agreement, depending critically on the GUTs adopted. The mass itself of the monopole, in the framework of various theories, can vary widely: from 10^4 to 10^{19} GeV.

The only experimental data about monopoles we have at our disposal is that their number is actually below the sensitivity of the detectors, i.e. the present number of monopoles is practically zero. But because the theory predicts their existence a mechanism is necessary which is able to reduce their number in an effective manner; such a mechanism might be the monopole–antimonopole annihilation.

Because of the enormous mass of the monopole, the energy released in the process is very high and so it seems natural to us to think of their annihilation as taking place in quasars, which are very old and bright heavenly objects with not yet known source of energy.

The interest of this model resides in the fact that it can explain the energy emitted by the quasars and from the measurement of their luminosity it is obtained that the ONLY monopole that can produce such an energy must be the 10^{16} GeV monopole, i.e. the Georgi–Glashow monopole. One must point out that this result is obtained by the unique assumption that the source of energy of quasars is the $M - \bar{M}$ annihilation and so without referring to particular GUTs.

Moreover, if we suppose that the monopoles, inside the quasars, tend to gather towards the centre, then we have that the annihilation time is of the same order of magnitude of the age of the Universe and the volume occupied by them is very small: this makes them practically undetectable.

References

1. P.A.M. Dirac, Proc. Roy. Soc., *A133*, 60, 1931.
2. G. 't Hooft, Nucl. Phys. *B79*, 276, 1974; *B105*, 538, 1976.
3. A.M. Polyakov, JETP Lett., *20*, 194, 1974.
4. P. Goddard and D. Olive, Rep. Progr. Phys., *41*, 1357, 1978.
5. J.P. Preskill, Ann. Rev. Nucl. Part. Sc. *34*, 1984.
6. H. Georgi and S.L. Glashow, Phys. Rev. Lett., *33*, 438, 1974.
7. P. Langacker, Phys. Rep., *72C*, 185, 1981.
8. C.G. Callan, in Beyond the standard model, pp. 412 - 413, Frontieres, Gif sur Yvette, 1983.
9. J. Leite-Lopes, Gauge field theories, an introduction, Pergamon Press, London, 1983.
10. A.H. Guth and E. Weinberg, Phys. Rev.,*D23*, 876, 1981.
11. C.G. Callan, Phys. Rev., *D25*, 2141, 1982; *D26*, 2058, 1982.
12. V. Rubakov, JETP Lett., *33*, 644, 1982.
13. G. Giacomelli, CERN Preprint EP/84 - 17, 1984.
14. J.P. Preskill, Phys. Rev. Lett., *43*, 1365, 1979.
15. Ya.B. Zel'dovich and I.D. Novikov: Structure and evolution of the Universe, Riuniti, Roma, 1982.
16. S. Weinberg, Gravitation and cosmology, Wiley, New York, 1972.
17. W. Rindler, Mon. Not. R. Astron. Soc., *116*, 663, 1956.
18. C.W. Misner, K.S. Thorne and J.A. Wheeler, Gravitation, Freeman, S. Francisco, 1973.

19. R.H. Dicke and P.J. Peebles, General relativity: an Einstein centenary survey, Cambridge University Press, London, 1979.
20. F. Wilczek, Scientific American (Italian edition), *150*, 28, 1981.
21. A.H. Guth, Phys. Rev., *D23*, 347, 1981.
22. S. Coleman and E. Weinberg, Phys. Rev., *D7*, 1888, 1973.
23. A.D. Linde, Phys. Lett., *B108*, 389, 1982; *B114*, 431, 1982.
24. A. Albrecht and P.J. Steinhardt, Phys. Rev. Lett., *48*, 1220, 1982.
25. J.P. Preskill, in The very early Universe, Cambridge University Press, London 1983.
26. L. Gratton, Introduzione all'astrofisica, Zanichelli, Bologna, 1982.
27. S.P. Ahlen and K. Kinoshita, Phys. Rev., *D26*, 2347, 1982.
28. E.N. Parker, Ap. J. *160*, 383, 1970; *163*, 255, 1971; *166*, 295, 1971.
29. S.A. Bludman and M.A. Ruderman, Phys. Rev. Lett., *36*, 840, 1976.
30. M.S. Turner, E.N. Parker and T.J. Bogdan, Phys. Rev., *D26*, 1296, 1982.
31. S.Dimopoulos, J.P. Preskill and F. Wilczek, Phys. Lett., *B119*, 320, 1982.
32. E.W. Kolb, S. A. Colgate and J. A. Harvey, Phys. Rev. Lett., *49*, 1373, 1982.
33. E.N. Parker, Cosmological magnetic fields, Clarendon, Oxford, 1979.
34. G. Tanzella-Nitti, Coelum, *52*, 1, 1984.

Acta Physica Hungarica 64 (4), pp. 353-360 (1988)

A KERR-LIKE RADIATING METRIC IN THE EXPANDING UNIVERSE

L.K. PATEL and SHARDA S. KOPPAR

Department of Mathematics, Gujarat University
Ahmedabad, 380 009 India

(Received 3 June 1986)

Using the method of differential forms a Kerr-like metric is derived as an exact solution of Einstein's field equations corresponding to a perfect fluid distribution plus a pure radiation field. The solution is interpreted as a Kerr-like radiating metric in the cosmological background of an expanding universe. The radiating Kerr metric and the radiating associated Kerr metric are derived as particular cases. The details of the solution are also discussed.

1. Introduction

Vaidya [1] has discussed a metric which reduces to the metric of an expanding universe in the absence of the source and becomes the well known Kerr metric (Kerr [2]) in the absence of the expansion of the universe. The source for Vaidya's solution is an imperfect fluid (i.e. the pressures in three spatial directions are not equal). Guha Thakurta [3] has also discussed the Kerr metric in the background of expanding universe. He has used the field equations corresponding to a perfect fluid with heat flow for his discussion.

Kerr metric describes the exterior gravitational field of a rotating body. Many people have tried to construct the non-static generalization of Kerr metric. Vaidya and Patel [4] have obtained a radiating Kerr solution in terms of a Kerr–Schild (Kerr and Schild [5]) metric. Their solution describes the exterior gravitational field of a radiating rotating body. Vaidya, Patel and Bhatt [6] have also constructed Kerr-like radiating solutions of Einstein's equations. These solutions are simpler than the one reported earlier by Vaidya amd Patel [4]. The main purpose of the present investigation is to obtain an exact explicit solution of Einstein's equations which describes the radiating Kerr and the radiating associated Kerr solution, discussed by Vaidya, Patel and Bhatt [6], in the cosmological background of an expanding universe.

For the derivation of our solution we shall use Cartan's exterior calculus of differential forms. This method of differential forms is standard now. Therefore we shall not enter into the details here.

2. Equations of structure

For calculations leading to the equations of structure, we take the line element in the form

$$ds^2 = e^{2F}[2(du + g\sin\alpha d\beta)dt - M^2(d\alpha^2 + \sin^2\alpha d\beta^2) - 2L(du + g\sin\alpha d\beta^2], \quad (1)$$

were $F = F(t), M = M(u,\alpha,t), g = g(\alpha)$ and $L = L(u,\alpha,t)$. The metric (1) is conformal to the Kerr-NUT metric discussed by Vaidya, Patel and Bhatt [6], the conformal factor being a function of time t.

In the space-time manifold defined by the metric (1), let us introduce the following basic 1 - forms

$$\begin{aligned} \theta^1 &= e^F(du + g\sin\alpha d\beta), & \theta^2 &= e^F d\alpha, \\ \theta^4 &= e^F dt - L\theta^1, & \theta^3 &= e^F M\sin\alpha d\beta. \end{aligned} \quad (2)$$

Therefore the metric (1) becomes

$$ds^2 = 2\theta^1\theta^4 - (\theta^2)^2 - (\theta^3)^2 = g_{(ab)}\theta^a\theta^b. \quad (3)$$

Here and in what follows the bracketed indices indicate tetrad components with respect to the tetrad (2). Using (2) it is easy to compute the exterior derivatives $d\theta^a$, and Cartan's first structure equation $d\theta^a = -W^a_b\Lambda\theta^b$ will the give the connection 1 - forms w^a_b. A straightforward calculation gives

$$\begin{aligned}
e^F w^1_1 &= -e^F w^4_4 = (L_t + LF_t)\theta^1 - F_t\theta^4, \\
e^F w^1_2 &= e^F w^2_4 = \left(\frac{f}{m^2}\right)\theta^3 + \left(F_t + \frac{M_t}{M}\right)\theta^2, \\
e^F w^1_3 &= e^F w^3_4 = -\left(\frac{f}{M^2}\right)\theta^2 + \left(F_t + \frac{M_t}{M}\right)\theta^3, \\
e^F w^2_1 &= e^F w^4_2 = \frac{-L_\alpha}{M}\theta^1 + \left[\frac{M_u}{M} + L(F_t + \frac{M_t}{M})\right]\theta^2 - \frac{fL}{M^2}\theta^3, \\
e^F w^3_1 &= e^F w^4_3 = \frac{gL_u}{M}\theta^1 + \frac{fL}{M^2}\theta^2 + \left[\frac{M_u}{M} + L\left(F_t + \frac{M_t}{M}\right)\right]\theta^3, \\
e^F w^2_3 &= -e^F w^3_2 = -\frac{fL}{M^2}\theta^1 - g\frac{M_u}{M}\theta^2 - \frac{(M_\alpha + M_{cot\alpha})}{M^2}\theta^3 + \frac{f}{M^2}\theta^4
\end{aligned}$$

with $w_{ab} + w_{ba} = 0, f = \frac{1}{2}(g_\alpha + g\cot\alpha)$, and a suffix denoting partial derivatives (e.g. $g_\alpha = \frac{\partial g}{\partial\alpha}, F_t = \frac{\partial F}{\partial t}$ etc.) From the second equation of structure

$$dw^a_b + w^a_c\Lambda w^c_b = R^a_{bcd}\theta^c\Lambda\theta^d.$$

We can now compute the curvature components R^a_{bcd}. For the sake of brevity they are not listed here. If $R_{(ab)} = R^c abc$ denote the tetrad components of the Ricci tensor, we find that

$$
\begin{aligned}
R_{(23)} &= 0, \\
e^{2F} R_{(44)} &= 2\left(\frac{M_{tt}}{M} - \frac{f^2}{M^4}\right) + 2(F_{tt} - F_t^2), \\
e^{2F} R_{(24)} &= \frac{g}{M}\left[\left(\frac{M_t}{M}\right)_y - (f/M^2)_u\right], \\
e^{2F} R_{(34)} &= \frac{-g}{M}\left[\left(\frac{M_t}{M}\right)_u + (f/M^2)_y\right], \\
e^{2F} R_{(12)} &= L e^{2F} R_{(24)} + \frac{2g}{M} L_y F_t + \frac{g}{M}\left[\left(L_t + \frac{M_u}{M}\right)_y + \left(\frac{2fL}{M^2}\right)_u\right], \\
e^{2F} R_{(13)} &= L e^{2F} R_{(34)} - \frac{2g}{M} L_u F_t + \frac{g}{M}\left[-\left(L_t + \frac{M_u}{M}\right)_u + \left(\frac{2fL}{M^2}\right)_y\right], \\
e^{2F} R_{(14)} &= L_{tt} + \frac{2}{M}\left[M_{ut} + (LM_t)_t + \frac{Lf^2}{M^3}\right] + \\
&\quad + 4LF_{tt} + 2LF_t\left(F_t + \frac{2M_t}{M}\right) + 2F_t\left(2L_t + \frac{M_u}{M}\right). \\
e^{2F} R_{(22)} &= e^{2F} R_{(33)} = \frac{1}{M_2}\left[g^2\left(\frac{M_u}{M}\right)_u + g^2\left(\frac{M_y}{M}\right)_y + \right. \\
&\quad \left. + \frac{2fM_y}{M} + \frac{4f^2L^2}{M} - 1 - (M^2)_{ut} + \{L(M^2)_t\}_t\right] - \\
&\quad - 2F_t\left(L_t + \frac{2M_u}{M}\right) - 4LF_t\left(F_t + \frac{2M_t}{M}\right) - 2LF_{tt}, \\
e^2 F R_{(11)} &= L^2 e^{2F} R_{(44)} + \frac{1}{M^2}\left[g^2(L_{uu} + L_{yy}) + \right. \\
&\quad \left. + 2fL_y + 2L_u M M_t + 4LMM_{tu} - 2L_t M M_u + 2MM_{uu}\right] + 2L_u F_t.
\end{aligned}
\tag{5}
$$

In the above the variable y replaces the variable α in differentiation, the defining relation being

$$g d\alpha = dy. \tag{6}$$

3. The field equations

We shall try to solve the following field equations

$$R_{ik} - \frac{1}{2} g_{ik} R = -8\pi[(p+\rho) v_i v_k - p g_{ik} + \sigma w_i w_k] - \Lambda g_{ik}, \tag{7}$$

with

$$g^{ik}v_iv_k = 1, \quad g^{ik}w_iw_k = 0, \quad g^{ik}v_iw_k = 1. \tag{8}$$

The last condition in (8) is the normalizing condition.Here $\sigma w_i w_k$ is the tensor arising out of the flowing null radiation, v^i is the flow vector of the perfect fluid and Λ is the cosmological constant. The other symbols occurring in (7) have their usual meanings.

It is painless to see that the field equations (7) can be expressed in the tetrad form as

$$R_{(ab)} = \Lambda g_{(ab)} - 8\pi\sigma^w(a)^w(b) - 8\pi\left[(p+\rho)v_{(a)}v_{(b)} - \frac{1}{2}(\rho - p)g_{(ab)}\right]. \tag{9}$$

For the metric (1) and the tetrad (2) we take the tetrad components of the vectors v_i and w_i as

$$v_{(a)} = \left(\frac{1}{2\lambda}, 0, 0, \lambda\right), w_{(a)} = \left(\frac{1}{\lambda}, 0, 0, 0\right), \tag{10}$$

where λ is a function of the co-ordinates to be determined from the field equations. It is easy to see that v_i and w_i given by (10) satisfy the conditions (8). The equations (9) and (10) imply the following relations:

$$R_{(24)} = 0, \qquad R_{(34)} = 0, \tag{11}$$

$$R_{(12)} = 0, \qquad R_{(13)} = 0, \tag{12}$$

$$8\pi p = -R_{(14)} - \Lambda, \tag{13}$$

$$8\pi\rho = -2R_{(22)} - R_{(14)} - \Lambda, \tag{14}$$

$$2\lambda^2 = \frac{R_{(44)}}{R_{(22)} + R_{(14)}}, \tag{15}$$

$$16\pi\sigma = R_{(22)} + R_{(14)} - \frac{R_{(44)}R_{(11)}}{R_{(22)} + R_{(14)}}. \tag{16}$$

Here the tetrad components $R_{(ab)}$ of the Ricci tensor are given by (5).

4. The solution of the field equations

Take the two equations (11). These involve only one unknown function, M. One can easily verify that these equations admit the solution

$$M^2 = (f/Y)(X^2 + Y^2), \tag{17}$$

with

$$X_u = -Y_y, \qquad X_y = Y_u, \qquad X_t = -1, \qquad Y_t = 0. \tag{18}$$

For our purpose we shall take a particular case of the above solution. We assume $f = Y$. Therefore, the above solution becomes

$$M^2 = X^2 + Y^2, \tag{19}$$

with

$$Y = -ay + B, \quad X = au - t. \tag{20}$$

Here a and B are undetermined constants and no additive constant is shown explicitly in x, because such a constant can always be incorporated in the t coordinate. The two equations (12) can be explicitly written as.

$$-L_{tu} - \left(\frac{M_u}{M}\right)_u + \left(\frac{2fL}{M^2}\right)_y - 2L_u F_t = 0$$

and

$$L_{ty} + \left(\frac{M_u}{M}\right)_y + \left(\frac{2fL}{M^2}\right)_u + 2L_y F_t = 0.$$

Using the results (19) and (20) a solution of the above two differential equations can be expressed as

$$2L = a + \left[a - 1 + 2\frac{(F^*X + E^*Y)}{x^2 + y^2}\right] e^{-2F}, \tag{21}$$

with

$$E^* = -(a-1)Y, \quad F^* = -a(a-1)u - m, \tag{22}$$

where m is a constant of integration. The pressure p, the density ρ, λ^2 and the radiation density σ can be determined from (13), (14), (15) and (16). They are given by

$$8\pi p = \Lambda - e^{-2F}[a(2F_{tt} + F_t^2) + e^{-2F}(2L_0 - a)(F_{tt} - F_t^2)], \tag{23}$$

$$8\pi(p+\rho) = -2e^{-2F}\left[\frac{a-1}{X^2+Y^2} + a(F_{tt} - F_t^2)\right] - \\ - 2e^{-4F}\left[\frac{1-a}{X^2+Y^2} - (2L_o - a)F_t^2 - \frac{2F_t\{(a-1)t + m\}}{X^2+Y^2}\right], \tag{24}$$

$$\lambda^2 = \frac{2(F_{tt} - F_t^2)e^{-2F}}{-8\pi(p+\rho)}, \tag{25}$$

$$16\pi\sigma = -4\pi(p+\rho) + \frac{1}{4\pi(p+\rho)}\left[\frac{4ae^{-6F}}{X^2+Y^2}(F_{tt} - F_t^2)\{(a-1) + F^*F_t\} + \right. \\ \left. + (F_{tt} - F_t^2)e^{-4F}\{a + (2L_0 - a)e^{-2F}\}^2\right], \tag{26}$$

where $2L_0$ is given by

$$2L_0 = 2a - 1 + \frac{2(E^*Y + F^*X)}{X^2 + Y^2} \tag{27}$$

and F^* and E^* are given by (22).

With $2f = g_\alpha + g\cot\alpha, gd\alpha = dy$ and $f = Y$, the result (20) shows that the function Y satisfies the differential equation

$$Y_{zz}(1 - z^2) - 2zY_z + 2aY = 0, \quad z = \cos\alpha, \tag{28}$$

This differential equation admits a power series solution if $a \geq -\frac{1}{8}$. Given a satisfying this condition, one can find a real number n such that $2a = n(n+1)$, and so (28) becomes the Legendre equation

$$Y_{zz}(1 - z^2) - 2zY_z + n(n+1)y = 0. \tag{29}$$

The solution of (29) can be expressed as

$$Y = KP_n(\cos\alpha) + NQ_n(\cos\alpha), \tag{30}$$

where K and N are arbitrary constants and P and Q are respectively the Legendre polynomial and the associated Legendre polynomial of order n. It is easy to see that

$$g\sin\alpha = \frac{K}{a}\frac{dP_n}{dz}\sin^2\alpha + \frac{N}{a}\frac{dQ_n}{dz}\sin^2\alpha. \tag{31}$$

Thus the final form of the metric of our solution can be explicitly expressed as

$$\begin{aligned} ds^2 = e^{2F}\Bigg[&2\{du + \left(\frac{K}{a}\frac{dP_n}{dz} + \frac{N}{a}\frac{dQ_n}{dz}\right)\sin^2\alpha d\beta\}dt - \\ &- 2L\left\{du + \left(\frac{K}{a}\frac{dP_n}{dz} + \frac{N}{a}\frac{NQ_n}{dz}\right)\sin^2\alpha d\beta\right\}^2 - \\ &- (X^2 + Y^2)(d\alpha^2 + \sin^2\alpha d\beta^2)\Bigg], \end{aligned} \tag{32}$$

where

$$X = au - t, \quad Y = KP_n(\cos\alpha) + NQ_n(\cos\alpha),$$
$$2L = a + \left[(1-a) + \frac{2X\{(1-a)t - m\}}{X^2 + Y^2}\right]e^{-2F}.$$

When $F = 0, \Lambda = 0$, we have verified that $p = 0, \rho = 0$ but $\sigma \neq 0$. In this case the metric (32) reduces to the radiating Kerr-like metric discussed by Vaidya, Patel and Bhatt [6].

Let us study the situation when $a = 1$. In this case $n = 1$ and consequently Y is given by

$$Y = K\cos\alpha + N(\cos\alpha \log\tan\frac{\alpha}{2} + 1). \tag{33}$$

In this case the parameters p, ρ, λ^2 and σ are given by

$$\begin{aligned}
8\pi p =& \Lambda - e^{-2F}(2F_{tt} + F_t^2) + \frac{2mre^{-2F}}{X^2+Y^2}(F_{tt} - F_t^2),\\
8\pi(p+\rho) =& -2e^{2F}(F_{tt} - F_t^2) + \frac{4me^{-4F}(rF_t^2 + F_t)}{r^2+Y^2},\\
\lambda^2 =& \frac{2(F_{tt} - F_t^2)e^{-2F}}{-8\pi(p+\rho)},\\
16\pi\sigma =& \frac{1}{4\pi(p+\rho)}\left[-\frac{4mF_t}{r^2+Y^2}(F_{tt} - F_t^2)e^{-6F} + \right.\\
&\left. + (F_{tt} - F_t^2)e^{-4F}\left\{1 + \frac{2mr}{r^2+Y^2}e^{-2F}\right\}^2\right] - 4\pi(p+\rho),
\end{aligned} \tag{34}$$

where Y is given by (33) and $x = u - t = -r$. From (34) it is obvious that when $F = 0, \Lambda = 0$ we get $p = \rho = \sigma = 0$. Thus, we get an empty space time described by the metric

$$\begin{aligned}
ds^2 =& 2[du + \{K\sin^2\alpha + N(\sin^2\alpha \log\tan\frac{\alpha}{2} - \cos\alpha)\}d\beta]dt -\\
&- (r^2+Y^2)(d\alpha^2 + \sin^2\alpha d\beta^2) - \left[1 + \frac{2mr}{r^2+Y^2}\right] \times\\
&\times [du + \{K\sin^2\alpha + N(\sin^2\alpha \log\tan\frac{\alpha}{2} - \cos\alpha)d\beta\}]^2,
\end{aligned} \tag{35}$$

where Y is given by (33).

The metric (35) is the particular case of Kerr-like vacuum metric discussed by Demianski [7] with slight changes of notation. When $N = 0$, (35) reduces to the well-known Kerr metric. When $K = 0$, (35) reduces to the associated Kerr metric discussed by Vaidya [8]. Here it should be noted that when $m = 0$ the metric (35) reduces to the flat metric. The explicit transformations of coordinates for this purpose are given by Demianski [7]. Therefore when $a = 1$ and $m = 0$ the metric (32) reduces to the metric of an expanding universe.

Thus the metric (32) with $a = 1$ and $N = 0$ gives us Kerr metric in the background of an expanding universe. Also the metric (32) with $a = 1$ and $K = 0$ represents the field of the associated Kerr source embedded in an expanding universe.

We can interpret the metric (32) when $N = 0$ as the radiating Kerr metric in the cosmological background of an expanding universe. Similarly when $K = 0$, the metric (32) represents the radiating associated Kerr metric, in the background of an expanding universe.

When the background universe is pressure-free, then we have

$$2F_{tt} + F_t^2 = 0. \tag{36}$$

The Eq. (36) can be easily integrated to have

$$e^{2F} = (lt + q)^4, \tag{37}$$

where l and q are constants of integration. The metric (32) with e^{2F} given by (37) represents a radiating Kerr-like metric in the cosmological background of Einstein–de Sitter universe.

Acknowledgement

One of the authors (S.S.K) is highly indebted to the University Grants Commission, New Delhi, for the award of a Junior Research Fellowship.

References

1. P.C. Vaidya, Pramana, *8*, 512, 1977.
2. R.P. Kerr, Phys. Rev. Lett., *11*, 237, 1963.
3. S.N. Guha Thakurta, Indian J. Phys., *55B*, 304, 1981.
4. P.C. Vaidya and L.K. Patel, Phys Rev., *D7*, 3590, 1973.
5. R.P. Kerr and A. Schild., Atti del congresso sull a Relativita Generale, (ed.) G. Barkera, Firenze, 222, 1965.
6. P.C. Vaidya, L.K. Patel and P.V. Bhatt, Gen. Relativ. Gravitation, 7, 701, 1976.
7. M. Demianski, Phys. Lett., *42A*, 157, 1972.
8. P. C. Vaidya, Curr. Sci., *45*, 490, 1976.

Acta Physica Hungarica 64 (4), pp. 361-378 (1988)

ADDITION THEOREMS FOR THE SPHERICAL FUNCTIONS OF THE LORENTZ GROUP

M. HUSZÁR

Central Research Institute for Physics
1525 Budapest, Hungary

(Received 17 July 1986 — in revised form 4 June 1987)

Given a function $f(p_a p_b)$ of the scalar product of two timelike, lightlike or spacelike four-vectors it can be expanded in terms of the products $Y^\sigma_{\mathbf{P}}(p_a)^* Y^\sigma_{\mathbf{P}}(p_b)$ of spherical functions on the $(p_a)^2$ and $(p_b)^2$ hyperboloids. Simple formulas for the evaluation of the expansion coefficients are derived. The expansions are presented in the form of addition theorems by which the integrals $\int d^2\mathbf{P}\, Y^\sigma_{\mathbf{P}}(p_a)^* Y^\sigma_{\mathbf{P}}(p_b)$ are understood. There exist six types of them according to whether p_a or p_b are timelike, lightlike or spacelike four-vectors. When $p_a = q_a$ and $p_b = q_b$ are both spacelike vectors the above integral depends besides the scalar product $(q_a q_b)$ on an additional invariant, $I(q_a, q_b)$, defined whenever $(q_a q_b)^2 \geq 1$, as the sign of $q_b^- + (q_a q_b) q_a^-$.

1. Introduction

In a recent paper [1] spherical functions of the Lorentz group on the upper sheet of the double-sheeted hyperboloid, $u^2 = 1$, forward light cone, $k^2 = 0$, and the single-sheeted hyperboloid , $q^2 = -1$, have been derived. They are labelled by the two-dimensional horospheric momentum $\mathbf{P} = (P_1, P_2)$ within an irreducible representation. The above set is completed here by the spherical functions on the lower sheet of double-sheeted hyperboloid and the backward light cone.

The aim of the present paper is to give formulas for the expansion of a sufficiently well behaved invariant scalar function of two four-vectors in terms of spherical functions. Let e.g. $(u_a)^2 = 1$, $(u_b)^2 = 1$, $u_a^0 \geq 1$, $u_b^0 \geq 1$ be two timelike vectors then the expansion of a function of the scalar product looks like

$$f(u_a u_b) = \int_0^\infty \sigma^2 d\sigma \int_{-\infty}^\infty d^2\mathbf{P}\, a(\sigma)\, Y^\sigma_{\mathbf{P}}(u_a)^*\, Y^\sigma_{\mathbf{P}}(u_b), \tag{1.1}$$

where the expansion coefficient $a(\sigma)$ is determined by

$$\sigma a(\sigma) = 2\pi \int_0^\infty d\omega\ \mathrm{sh}\omega\ \sin(\sigma\omega)\, f(\ \mathrm{ch}\omega) \tag{1.2}$$

$$(\ \mathrm{ch}\omega = (u_a u_b) \geq 1,\ 0 \leq \omega < \infty).$$

The expansion (1.1) is conveniently presented in two steps. At first the addition theorem

$$\int d^2\mathbf{P}\, Y^{\sigma}_{\mathbf{P}}(u_a)^*\, Y^{\sigma}_{\mathbf{P}}(u_b) = \frac{1}{\pi^2\sigma}\frac{\sin(\sigma\omega)}{\mathrm{sh}\omega} \tag{1.3}$$

is stated. The right hand side of this equation will be called the addition function of the two spherical functions involved. It is clearly a Lorentz invariant quantity. At the subsequent step the expansion of a function $f(u_a u_b)$ in terms of the addition function is performed,

$$f(\,\mathrm{ch}\omega) = \frac{1}{\pi^2}\int_0^{\infty} \sigma^2 d\sigma\, a(\sigma)\frac{\sin(\sigma\omega)}{\sigma\,\mathrm{sh}\omega} \tag{1.4}$$

This is a sine Fourier expansion whose inverse is Eq. (1.2).

A kind of addition theorem has been derived in [1]. By addition theorem in the present context the one given by Eq. (1.3) and by its counterparts for spacelike and lightlike vectors are meant. Addition theorems like (1.3) evaluated in other e.g. angular momentum basis can be found in literature at several places (cf. [2], [3] Chapter X.§3/5). It is clear, however, that the addition functions are independent of the basis chosen.

The most general form of addition theorems on the Lorentz group is the group multiplication law for the irreducible matrix representations. Addition theorems for the spherical functions can be deduced from these by restricting the labels specifying states within an irreducible representation to certain particular values ([3] Chapter III. §4). Therefore, results obtained in the present paper could, at least in principle, be derived from the matrix elements of irreducible unitary representations expressed in an appropriate basis. As, however, these matrices are known to be rather complicated objects it is not the simplest way for obtaining the addition theorems in this manner. In the present paper a straightforward method of derivation is chosen.

In Eq. (1.3) it has been supposed that the integration with respect to $\mathbf{P}$ can be performed in advance and the addition function (1.3) exists. This is really so in the case considered, i. e. when both spherical functions are defined on the double-sheeted hyperboloid. However, in other cases like the lightlike–lightlike one the addition function may not exist at all the corresponding integral being divergent. Thus the method needs deeper foundations. While the analogous formulas for the spherical harmonics Y^l_m of the rotation group work without any trouble the situation in the Lorentz group is much more complicated. The reason is that the relevant spherical functions belong to the continuous spectrum not only through σ but also through the label $\mathbf{P}$. The commuting set of operators whose eigenfunctions are $Y^{\sigma}_{\mathbf{P}}$ have no common nonzero square integrable eigenfunctions, thus none of $Y^{\sigma}_{\mathbf{P}}$ is normalizable. A proper mathematical framework for treating such eigenfunctions is the theory of rigged Hilbert spaces. In [4] an attempt has been made to outline a more rigorous treatment on that basis. Here we recall the final result only. Let $\varphi(u)$ be an element of an appropriate test function space specified in [4], then the

action of an f operator with the kernel $f(u_a u_b)$ is given by

$$(f\varphi)(u_b) = \int \sigma^2 d\sigma \, d^2\mathbf{P} \, a(\sigma) Y^\sigma_{\mathbf{P}}(u_b)(Y^\sigma_{\mathbf{P}}, \varphi),$$

where $(Y^\sigma_{\mathbf{P}}, \varphi)$ is a scalar product with the test function $\varphi(u_a)$. The integration within the scalar product should be performed at first and the $\mathbf{P}$ integration subsequently. Whether these integrations may or may not be interchanged depends on the kind of the spherical functions concerned. The integrations are really interchangeable in the timelike–timelike case in contrast to some other cases implying that there is no addition theorem at all then since it is just the result of integration over $\mathbf{P}$. Hence if one requires the existence of addition functions the corresponding integrals should be regularized somehow. It turns out that if σ is extended to complex values the integrals acquire a meaning and the result can analytically be continued to real values of σ.

Out of the six addition theorems there are two of particular interest. The spacelike–spacelike addition function contains explicitly a Lorentz invariant quantity

$$I(q_a, q_b) = sg(q_b^- + (q_a q_b) q_a^-) \tag{1.5}$$

$$(q_a^2 = q_b^2 = -1, \quad q^- = q^0 - q^3)$$

defined whenever $(q_a q_b)^2 \geq 1$. It takes the values ± 1 independently of the scalar product provided $(q_a q_b)^2 \geq 1$. It is invariant under the simultaneous proper Lorentz transformation of q_a and q_b, $I(g^{-1}q_a, g^{-1}q_b) = I(q_a, q_b)$. In particular, if $(q_a q_b) < -1$ $I(q_a, q_b)$ is a Lorentz invariant scalar *antisymmetric* in q_a and q_b, $I(q_a, q_b) = -I(q_b, q_a)$. A function $f_I(q_a q_b)$ that depends on $I(q_a, q_b)$ in addition to the scalar product of two spacelike vectors is Lorentz invariant. Since I is defined only when $(q_a q_b)^2 \geq 1$ for the remaining $-1 < (q_a q_b) < 1$ values of the scalar product $f_+(q_a q_b)$ and $f_-(q_a q_b)$ are supposed to coincide. The expansion of $f_I(q_a q_b)$ in terms of spherical functions takes account of the I-dependence, too.

Another Lorentz invariant quantity that arises explicitly in an addition function is $g(\frac{k_b^-}{k_a^-})\delta(k_a k_b)$ which means simply that if in some frame the components of a lightlike vector are constant multiple of an other one, this constant is a Lorentz invariant quantity. Thus the function $F(k_a, k_b) = f(k_a k_b) + g(\frac{k_b^-}{k_a^-})\delta(k_a k_b)$ of two lightlike vectors is invariant, i.e. $F(g^{-1}k_a, g^{-1}k_b) = F(k_a, k_b)$, $(g \in L, k_a^2 = k_b^2 = 0)$. In particular, this implies the existence of an invariant, antisymmetric with respect to the interchange of the lightlike four-vectors, quantity

$$F(k_a, k_b) = h\left(\frac{k_b^-}{k_a^-}\right)\delta(k_a k_b) = -F(k_b, k_a), \tag{1.6}$$

where

$$h\left(\frac{k_b^-}{k_a^-}\right) = \frac{1}{2}\left[g\left(\frac{k_b^-}{k_a^-}\right) - g\left(\frac{k_a^-}{k_b^-}\right)\right].$$

Addition theorems provide the expansion of $F(k_a, k_b)$ in terms of spherical functions in the general form, including the delta function term.

In the next Section we recall from [1] the explicit form of spherical functions completed by suitable phase factors. Out of them $Y^{\sigma}_{\mathbf{P}}(q)$ and $y^{\sigma}_{\mathbf{P}}(k)$ defined on timelike and lightlike vectors, respectively, transform according to irreducible unitary representation of the principal series. The spherical functions on the single-sheeted hyperboloid as are given in [1] proved to be reducible since they can be decomposed into two irreducible components of even and odd parity [6] . Thus in order to get irreducible spaces one more label, r, the reflection parity is necessary and so $\mathbf{Y}^{\sigma,r}_{\mathbf{P}}(q)\,(r = \pm)$ transform according to irreducible unitary representations $T^{\sigma,r}_{g}$. It is an interesting consequence of the existence of the invariant $I(q_a, q_b)$ that the representations labelled by $r = +$ and $r = -$ are equivalent since an intertwining operator can be found between them (Section 3). Although the reflection parity is necessary for labelling the spherical functions and irreducible spaces the matrix elements of representations $T^{\sigma,+}_{g}$ and $T^{\sigma,-}_{g}$ in the irreducible basis coincide.

According to timelike, lightlike or spacelike character of four-vectors there exist essentially three types of spherical functions which can be combined with each other in six different ways, correspondingly, there exist six basic types of addition functions. Taking into account the backward light cone and the lower sheet of the double-sheeted hyperboloid there are five types of homogeneous spaces amounting to fifteen addition functions after all. However, these are absorbed into the six basic types in a concise form. At the next step a sufficiently well behaved invariant function is expanded in terms of addition functions. Inversion formulas for these expansions can be given easily, merely the spacelike-spacelike case proves slightly more involved.

2. Spherical functions

The homogeneous spaces considered are on the surface

$$u_\mu u^\mu = (u^0)^2 - (u^1)^2 - (u^2)^2 - (u^3)^2 = \xi \text{ with } \xi = 1, 0, -1$$

and can be specified as follows.

1. Double-sheeted hyperboloid.

a.) $H^{\uparrow}_{+}$: upper sheet, $\xi = 1$, $\quad u_\mu u^\mu = 1,\ u^0 \geq 1,\ (0 < l < \infty)$,

b.) $H^{\downarrow}_{+}$: lower sheet, $\xi = 1$, $\quad u_\mu u^\mu = 1,\ u^0 \leq -1,\ (-\infty < l < 0)$.

2. Light cone.

a.) $H^{\uparrow}_{0}$: forward light cone, $\xi = 0$, $\quad u_\mu u^\mu = 0,\ u^0 > 0,\ (0 < l < \infty)$,

b.) $H^{\downarrow}_{0}$: backward light cone, $\xi = 0$, $\quad u_\mu u^\mu = 0,\ u^0 < 0,\ (-\infty < l < 0)$.

3. Single-sheeted hyperboloid.

H_-, $\xi = -1$, $u_\mu u^\mu = -1$, $\quad (-\infty < l < \infty, l \neq 0)$.

The parametrization introduced in [1] can be written in a unified form for all the five homogeneous spaces as

$$u^+ = \xi l + \frac{|\mathbf{z}|^2}{l}, \qquad u^- = \frac{1}{l}, \qquad \mathbf{u} = \frac{\mathbf{z}}{l}, \qquad (\xi = 1, 0, -1), \tag{2.1}$$

where $u^+ = u^0 + u^3$, $u^- = u^0 - u^3$, $\mathbf{u} = (u^1, u^2)$, $\mathbf{z} = (x, y)$, $|\mathbf{z}|^2 = x^2 + y^2$. The components of $\mathbf{z}$ range in $-\infty < x < \infty$, $-\infty < y < \infty$ in each case and the limits of variation of l are indicated above at each homogeneous space.

The scalar product of two functions on either of these spaces is defined by

$$(\varphi, \psi) = \int \frac{dl}{2|l|^3} d^2\mathbf{z}\, \varphi^*(l, \mathbf{z})\, \psi(l, \mathbf{z}).$$

Denote the generators of spatial rotations and boosts by $\mathbf{M}$ and $\mathbf{N}$. The spherical functions on either of the above spaces satisfy the eigenvalue equation of the Casimir operator

$$(\mathbf{M}^2 - \mathbf{N}^2)Y = (j_0^2 - \sigma^2 - 1)Y, \tag{2.2}$$

where $j_0 = 0$, σ is real continuous for the spherical functions on $H_+^\uparrow$, $H_+^\downarrow$, $H_0^\uparrow$, $H_0^\downarrow$ and for the continuous part of the spectrum on H_-. On the latter space there is also a discrete spectrum for which $\sigma = 0$, $j_0 =$ integer.

The basis is defined by the eigenvalue equation of the horospheric momenta

$$(N_1 + M_2)Y = P_1 Y, \qquad (N_2 - M_1)Y = P_2 Y \tag{2.3}$$

which hold for all the three values of ξ. To specify states unambiguously one more label, the $\mathcal{R}$-reflection parity and its counterpart are necessary in spacelike and lightlike cases. It is discussed somewhat later in this Section.

The simultaneous eigenfunctions of (2.2) and (2.3) derived in [1] are as follows:

ad 1.a.),b.) Double-sheeted hyperboloid. (both sheets: $H_+^\uparrow$ or $H_+^\downarrow$, $\xi = 1$, $j_0 = 0$)

$$Y_{\mathbf{P}}^\sigma(l, \mathbf{z}) = i\varepsilon\varepsilon_\sigma \frac{2}{\pi}\sqrt{\frac{\mathrm{sh}\pi\sigma}{\sigma}}\, |l|\, K_{i\sigma}(P|l|)\left(\frac{1}{2\pi}e^{-i\mathbf{P}\mathbf{z}}\right), \tag{2.4}$$

where $\varepsilon = sg\, l$, $\varepsilon_\sigma = sg\, \sigma$, $K_{i\sigma}$ is the third kind modified Bessel function [5], $\mathbf{P} = (P_1, P_2)$, $\mathbf{z} = (x, y)$, $\mathbf{P}\mathbf{z} = P_1 x + P_2 y$, $P = \sqrt{(P_1)^2 + (P_2)^2}$. For a complete set it is sufficient to let σ vary between $0 < \sigma < \infty$ but it may vary between $-\infty < \sigma < 0$ on equal footing.

ad 2. Light cone.

a.) Forward light cone $(H_0^{\uparrow}, \xi = 0, j_0 = 0)$,

$$y_{\mathbf{P}}^{\sigma}(l,\mathbf{z}) = -\sqrt{\frac{\Gamma(i\sigma)}{\Gamma(-i\sigma)}}\left(\frac{P}{2}\right)^{-i\sigma}\frac{1}{i\sigma\sqrt{\pi}}\,l^{1-i\sigma}\left(\frac{1}{2\pi}e^{-i\mathbf{P}\mathbf{z}}\right) \tag{2.5}$$

$$(-\infty < \sigma < \infty).$$

b.) Backward light cone $(H_0^{\downarrow}, \xi = 0, j_0 = 0)$,

$$y_{\mathbf{P}}^{\sigma}(l,\mathbf{z}) = \sqrt{\frac{\Gamma(-i\sigma)}{\Gamma(i\sigma)}}\left(\frac{P}{2}\right)^{i\sigma}\frac{1}{i\sigma\sqrt{\pi}}\,|\,l\,|^{1+i\sigma}\left(\frac{1}{2\pi}e^{-i\mathbf{P}\mathbf{z}}\right) \tag{2.6}$$

$$(-\infty < \sigma < \infty).$$

These two cases can be written in a concise form

$$y_{\mathbf{P}}^{\sigma}(l,\mathbf{z}) = -\sqrt{\frac{\Gamma(i\varepsilon\sigma)}{\Gamma(-i\varepsilon\sigma)}}\left(\frac{P}{2}\right)^{-i\varepsilon\sigma}\frac{1}{i\varepsilon\sigma\sqrt{\pi}}|l|^{1-i\varepsilon\sigma}\left(\frac{1}{2\pi}e^{-i\mathbf{P}\mathbf{z}}\right) \tag{2.7}$$

$$(-\infty < \sigma < \infty,\ \varepsilon = sg\,l).$$

ad 3. Single-sheeted hyperboloid. $(H_-, \xi = -1)$

a.) Continuous spectrum ($j_0 = 0$, σ continuous)

$$\hat{\mathbf{Y}}_{\mathbf{P}}^{\sigma}(l,\mathbf{z}) = \frac{1}{\sqrt{\sigma\,\mathrm{sh}\pi\sigma}}\,|\,l\,|\,J_{-i\varepsilon\sigma}(P\,|\,l\,|)\left(\frac{1}{2\pi}e^{-i\mathbf{P}\mathbf{z}}\right) \tag{2.8}$$

$$(-\infty < \sigma < \infty,\ \varepsilon = sg\,l),$$

where $J_{-i\varepsilon\sigma}$ is the Bessel function.

b.) Discrete spectrum ($\sigma = 0$, $j_0 =$ integer)

$$\mathbf{Y}_{\mathbf{P}}^{j_0}(l,\mathbf{z}) = i^{|n|}\sqrt{\frac{2}{|\,n\,|}}\,|\,l\,|\,J_n(Pl)\left(\frac{1}{2\pi}e^{-i\mathbf{P}\mathbf{z}}\right) \tag{2.9}$$

$$(j_0 = n = 1,2,3,\ldots).$$

For a complete set it is sufficient to let $j_0 = n$ run over positive integers but it can run alternatively over negative integers. The notation $|n|$ is merely in order to cover both cases.

Spherical functions given in [1] have been extended to the lower sheet of the double-sheeted hyperboloid $H_+^{\downarrow}$ and to the backward light cone $H_0^{\downarrow}$. Furthermore, to obtain a coherent set they are endowed with suitable phases. In Eq. (2.8) the cap

is in order to distinguish $\hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}$ from the irreducible linear combination to be defined in Eqs. (2.11), (2.12).

With this minor alteration all formulas but the defining equations of [1] remain valid. They should be replaced by the present definitions.

Spherical functions given by (2.4), (2.7), (2.8) and (2.9) form complete orthonormal sets on the space of L_2 functions over the corresponding homogeneous space.

It is well known from the representation theory of the Lorentz group that the unitary representations labelled by $(j_0 = 0, \sigma)$, $(j_0 = 0, -\sigma)$ are equivalent ([8] p.178). It is possible to distinguish these function spaces ([6] p. 346) by defining one more discrete label, namely, let $\hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}(l,\mathbf{z})^{s} = \hat{\mathbf{Y}}^{s\sigma}_{\mathbf{P}}(l,\mathbf{z})\,(s = \pm 1)$ i.e.

$$\hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}(l,\mathbf{z})^{s} = \frac{1}{\sqrt{\sigma\ \mathrm{sh}\pi\sigma}}\,|\,l\,|\,J_{-is\sigma}(P\,|\,l\,|)\left(\frac{1}{2\pi}e^{-i\mathbf{P}\mathbf{z}}\right) \qquad (2.10)$$

$$(\sigma > 0,\ s = \pm 1).$$

Thus in order to label a state unambiguously the value of σ and two horospheric momenta are not sufficient since there appears an additional label s taking the values ± 1 . Then σ ranges over positive (negative) values only.

In Poincaré group space time reflection is necessarily represented by antilinear, antiunitary operator. Within the Lorentz group it can be represented either by linear or by antilinear operators. To see the meaning of the label s define the linear operator of q-reflection ($q^{\mu} \longrightarrow -q^{\mu}$) by

$$\mathcal{R}\hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}(q) = \hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}(-q).$$

(In terms of parameters this reflection can be performed simply by changing the sign of l and leaving $\mathbf{z}$ unaltered.)

It is obvious from the explicit form of $\hat{\mathbf{Y}}^{\sigma,s}_{\mathbf{P}}$ that

$$\hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}(-q)^{s} = \hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}(q)^{-s}.$$

Hence the $\mathcal{R}$-reflection eigenstates belonging to the eigenvalues ± 1 are the combinations

$$\mathbf{Y}^{\sigma}_{\mathbf{P}}(q)^{+} = \frac{1}{\sqrt{2}}\left[\hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}(q)^{+} + \hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}(q)^{-}\right] =$$

$$= \frac{1}{\sqrt{2\sigma\ \mathrm{sh}\pi\sigma}}|l|\,[J_{-i\sigma}(P|l|) + J_{i\sigma}(P|l|)]\left(\frac{1}{2\pi}e^{-i\mathbf{P}\mathbf{z}}\right), \qquad (2.11)$$

$$\mathbf{Y}^{\sigma}_{\mathbf{P}}(q)^{-} = \frac{1}{\sqrt{2}}\left[\hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}(q)^{+} - \hat{\mathbf{Y}}^{\sigma}_{\mathbf{P}}(q)^{-}\right] =$$

$$= \frac{\varepsilon}{\sqrt{2\sigma\ \mathrm{sh}\pi\sigma}}|l|\,[J_{-i\sigma}(P|l|) - J_{i\sigma}(P|l|)]\left(\frac{1}{2\pi}e^{-i\mathbf{P}\mathbf{z}}\right), \qquad (2.12)$$

$$(\varepsilon = sg\,l).$$

These satisfy the eigenvalue equation

$$R\mathbf{Y}_{\mathbf{P}}^{\sigma}(q)^{r} = r\mathbf{Y}_{\mathbf{P}}^{\sigma}(q)^{r} \qquad (r = \pm) \tag{2.13}$$

They are even and odd functions on H_- and can alternatively be derived by an integral transformation method suggested in [6] . Along with (2.9) they form a complete orthonormal set and satisfy the orthogonality relations

$$\left(\mathbf{Y}_{\mathbf{Q}}^{\sigma',s}, \mathbf{Y}_{\mathbf{P}}^{\sigma,r}\right) = \delta_{sr}\frac{1}{\sigma^2}\delta(\sigma' - \sigma)\delta^2(\mathbf{Q} - \mathbf{P}) \qquad (\sigma > 0, \quad \sigma' > 0; \quad s,r = \pm) \tag{2.14}$$

$$\left(\mathbf{Y}_{\mathbf{Q}}^{j_0}, \mathbf{Y}_{\mathbf{P}}^{\sigma,r}\right) = 0 \tag{2.15}$$

Since the linear operator $\mathcal{R}$ commutes with the operator of group action, $[\mathcal{R}, T_g] = 0$, the eigenstates of $\mathcal{R}$ form invariant, moreover, irreducible subspaces. Thus $\mathbf{Y}_{\mathbf{P}}^{\sigma,r}$ functions given by (2.11), (2.12) transform according to irreducible unitary representations of the principal series. When applied to spherical functions of the discrete spectrum the $\mathcal{R}$-reflection decomposes them into states of even and odd parity of the index j_0,

$$\mathcal{R}\mathbf{Y}_{\mathbf{P}}^{2n}(q) = \mathbf{Y}_{\mathbf{P}}^{2n}(q), \tag{2.16}$$

$$\mathcal{R}\mathbf{Y}_{\mathbf{P}}^{2n-1}(q) = -\mathbf{Y}_{\mathbf{P}}^{2n-1}(q). \qquad (n = 1,2,\ldots). \tag{2.17}$$

As distinct from the representations $(\sigma,+)$ and $(\sigma,-)$ of the continuous spectrum irreducible representations $n = j_0 > 0$ belonging to different eigenvalues of $\mathcal{R}$ are inequivalent.

Spherical functions on the cone can be treated in an analogous manner. Since, however, the reflection $k^{\mu} \longrightarrow -k^{\mu}$ is not a symmetry of either of the cones $H_0^{\uparrow}$, $H_0^{\downarrow}$ define the integral operator

$$(\mathcal{R}'y)(k_a) = \mathcal{P}\int dk_b \mathcal{R}'(k_b, k_a)y(k_b) \qquad (k_a \text{ and } k_b \text{ are both on } H_0^{\uparrow} \text{ or } H_0^{\downarrow}), \tag{2.18}$$

where

$$\mathcal{R}'(k_b, k_a) = \frac{2\varepsilon_b}{i\pi^2}\frac{1}{ln\left(\frac{l_a}{l_b}\right)}\delta(k_a k_b) \qquad (\varepsilon_b = sg\, l_b) \tag{2.19}$$

in order to distinguish spaces of functions σ and $-\sigma$. ($\mathcal{P}$ denotes the principal value of the integral). The kernel $\mathcal{R}'(k_a, k_b)$ is Lorentz invariant scalar, antisymmetric in k_a and k_b.

The operator $\mathcal{R}'$ commutes with the left group action

$$\left[T_g, \mathcal{R}'\right] = 0$$

therefore, its eigenvalue is a good quantum number for labelling the function spaces transforming according to equivalent representations σ, $-\sigma$. Replace in y_{P}^{σ} given by (2.7) $\sigma \longrightarrow r\sigma$ where $r = \pm$ and define

$$\begin{aligned} y_{\mathrm{P}}^{\sigma}(k)^r &= y_{\mathrm{P}}^{\sigma r}(k) = \\ &= -\sqrt{\frac{\Gamma(i\varepsilon r\sigma)}{\Gamma(-i\varepsilon r\sigma)}} \left(\frac{P}{2}\right)^{-i\varepsilon r\sigma} \frac{1}{i\varepsilon r\sigma\sqrt{\pi}} |l|^{1-i\varepsilon r\sigma} \left(\frac{1}{2\pi} e^{-i\mathbf{P}\mathbf{z}}\right) \\ &\qquad (0 < \sigma < \infty,\ r = \pm). \end{aligned} \tag{2.20}$$

It is easy to verify that the functions defined so satisfy the eigenvalue equation

$$\mathcal{R}' y_{\mathrm{P}}^{\sigma}(k)^r = r y_{\mathrm{P}}^{\sigma}(k)^r. \tag{2.21}$$

This completes the eigenvalue equations (2.2) and (2.3).

3. Irreducible representations and the invariant $I(q_a, q_b)$

Given two spacelike vectors $(q_a)^2 = -1$, $(q_b)^2 = -1$ then in addition to the scalar product there exists one more invariant under the proper Lorentz group provided $(q_a q_b)^2 \geq 1$, namely,[1)]

$$I(q_a, q_b) = sg[q_b^- + (q_a q_b) q_a^-] = \pm 1. \tag{3.1}$$

This is so because $u^\mu = q_b^\mu + (q_a q_b) q_a^\mu$ is a timelike or lightlike four-vector, therefore, the sign of its $u^- = u^0 - u^3$ component is a Lorentz invariant quantity taking the values $I(q_a, q_b) = \pm 1$ independently of the scalar product whenever $(q_a q_b)^2 \geq 1$. More explicitly,

$$I(g^{-1} q_a, g^{-1} q_b) = I(q_a, q_b),$$

where g is a proper Lorentz transformation[1)]. The value of $I(q_a, q_b)$ is independent of the scalar product in the sense that for a given $(q_a q_b)^2 \geq 1$ value of the scalar product $I(q_a, q_b)$ can take the values $+1$ or -1 and vice versa: for a fixed value

[1]This invariance holds also for space reflection, therefore, $I(q_a, q_b)$ is invariant under the orthochronous Lorentz group after all.

of $I(q_a, q_b)$ the scalar product may have different values. On the other hand both $I(q_a, q_b)$ and the scalar product are determined unambiguously by the components of q_a and q_b.

The invariant I is either symmetric or antisymmetric with respect to the interchange of the order of q_a and q_b,

$$I(q_b, q_a) = \zeta I(q_a, q_b), \tag{3.2}$$

where $\zeta = sg(q_a q_b) = \pm 1$. This means, in particular, that if $(q_a q_b) \leq -1$, $I(q_a, q_b)$ is an invariant scalar *antisymmmetric* in q_a and q_b, $I(q_b, q_a) = -I(q_a, q_b)$. Furthermore, it has the obvious properties,

$$I(-q_a, q_b) = I(q_a, q_b); \quad I(q_a, -q_b) = -I(q_a, q_b); \quad I(-q_a, -q_b) = -I(q_a, q_b). \tag{3.3}$$

(There exist invariants like this in the remaining cases, too. E.g. let $q^2 = -1$, $k^2 = 0 \in H_0^{\uparrow}$ then the four-vector $u^\mu = q^\mu + \frac{1}{(kq)} k^\mu$ defined when $(kq) \neq 0$ is timelike, $u^2 = 1$, nevertheless, the invariant $I(k, q) = sg(q^- + \frac{1}{(kq)} k^-)$ is not independent of the scalar product since $I(k, q) = sg(kq)$).

By means of $I(q_a, q_b)$ an invariant Hermitean form

$$\{\varphi, \psi\} = \{\psi, \varphi\}^* = (\varphi, G\psi) = \int dq_a dq_b \, \varphi(q_a)^* G(q_a, q_b) \psi(q_b) \tag{3.4}$$

with

$$G(q_a, q_b) = \frac{1}{2}[I(q_a, q_b) + I(q_b, q_a)] = I(q_a, q_b) \frac{1 + \zeta}{2} \tag{3.5}$$

can be defined where dq is the invariant measure on H_-, $dq = \frac{dl}{2|l|^3} d^2\mathbf{z}$.

Similarly, there is an invariant antisymmetric form

$$[\varphi, \psi] = -[\psi, \varphi]^* = (\varphi, S\psi) = \int dq_a dq_b \, \varphi(q_a)^* S(q_a, q_b) \psi(q_b), \tag{3.6}$$

where

$$S(q_a, q_b) = \frac{1}{2}[I(q_a, q_b) - I(q_b, q_a)] = I(q_a, q_b) \frac{1 - \zeta}{2}. \tag{3.7}$$

Both forms are invariant under the group action,

$$\{T_g \varphi, T_g \psi\} = \{\varphi, \psi\}, \quad [T_g \varphi, T_g \psi] = [\varphi, \psi], \tag{3.8}$$

where T_g is the operator of left displacement, $T_g \varphi(q) = \varphi(g^{-1} q)$.

Only some consequences of the existence of the invariant I are discussed here. One of them is the equivalence of $r = +$ and $r = -$ representations. To this end consider the representation spaces $\mathcal{H}_\sigma^+$ and $\mathcal{H}_\sigma^-$ spanned by $\mathbf{Y}_{\mathbf{P}}^{\sigma,+}$ and $\mathbf{Y}_{\mathbf{P}}^{\sigma,-}$ for a

fixed value of σ. This is the space of functions square integrable in $\mathbf{P}$. Then T_g^+ and T_g^-, the corresponding irreducible representations are equivalent.

To verify this consider the linear operator A in $\mathcal{H}_\sigma^-$,

$$(A\varphi)(q_a) = \int dq_b\, I(q_a, q_b) A(q_a q_b)\varphi(q_b), \tag{3.9}$$

where $A(q_a q_b)$ is an even function of the scalar product. It follows from the symmetry properties (3.3) of $I(q_a, q_b)$ that $\mathcal{R}(A\varphi) = A\varphi$ i.e. A is an $\mathcal{H}_\sigma^- \xrightarrow{A} \mathcal{H}_\sigma^+$ mapping, moreover, it is bijective. It follows from the relation $A\mathbf{Y}_{\mathbf{P}}^{\sigma,-}(q) = \alpha(\sigma)\mathbf{Y}_{\mathbf{P}}^{\sigma,+}(q)$ which can be derived from the addition theorems to be proved in the next Section. Here $\alpha(\sigma)$ is a constant for a definite value of σ . There exists a kernel $A(q_a q_b)$ that the constant $\alpha(\sigma)$ corresponding to it is nonzero and finite for $0 < \sigma < \infty$.

Furthermore, the symmetries (3.3) and the invariance of the measure dq_b imply, (cf. [3] Chapter I.§/4)

$$\begin{aligned}(T_g^+ A\varphi)(q_a) &= \int dq_b I(g^{-1}q_a, q_b) A(g^{-1}q_a q_b)\varphi(q_b) = \\ &= \int dq_b I(q_a, gq_b) A(q_a g q_b)\varphi(q_b) = \\ &= \int dq_b I(q_a, q_b) A(q_a q_b)\varphi(g^{-1}q_b),\end{aligned}$$

thus

$$T_g^+ A = A T_g^- \tag{3.10}$$

which proves the equivalence stated.

As a further consequence it can be easily shown that the matrix elements of irreducible representations T_g^+ and T_g^- in $\mathbf{Y}_{\mathbf{P}}^{\sigma,+}$ and $\mathbf{Y}_{\mathbf{P}}^{\sigma,-}$ bases coincide. Hence the matrix elements of irreducible representations, $D_{QP}(g)$, are independent of the label $r = \pm$, $D_{QP}^{\sigma,+}(g) = D_{QP}^{\sigma,-}(g) = D_{QP}^{\sigma}(g)$.

4. Addition theorems

The addition theorems stated in this Section refer to the spherical functions given by Eqs. (2.4), (2.9), (2.11), (2.12), (2.20).

1. Timelike–timelike case

Let $(u_a)^2 = 1$, $(u_b)^2 = 1$ be two timelike vectors and u_a, $u_b \in (H_+^\uparrow$ or $H_+^\downarrow)$ independently. Consider the addition function defined by

$$A^\sigma(u_a u_b) = \int d^2\mathbf{P} Y_{\mathbf{P}}^\sigma(u_a)^* Y_{\mathbf{P}}{}^\sigma(u_b), \tag{4.1}$$

where the integral is extended to the entire P_1, P_2 plane. It can be evaluated either directly or by noticing that $A^\sigma(u_a u_b)$ is a Lorentz invariant quantity thus depending on the scalar product $(u_a u_b)$ only. There exists a frame where $u_a = (\varepsilon_a, 0, 0, 0)$, $u_b = (u_b^0, 0, 0\, u_b^3)$. Denote the parameters of u_a and u_b by $(l_a, \mathbf{z}_a)$, $(l_b, \mathbf{z}_b)$ and let $\varepsilon_a = sgl_a$, $\varepsilon_b = sgl_b$ then this frame corresponds to the values of coordinates $l_a = \varepsilon_a$, $\mathbf{z}_a = 0$, $0 < \varepsilon_b l_b < \infty$, $\mathbf{z}_b = 0$.

The result of integration is

$$A^\sigma(u_a u_b) = \int d^2\mathbf{P}\, Y^\sigma_{\mathbf{P}}(u_a)^*\, Y^\sigma_{\mathbf{P}}(u_b) = \zeta \frac{1}{\pi^2 \sigma} \frac{\sin(\sigma\omega)}{\mathrm{sh}\omega} \tag{4.2}$$

$$(\zeta = \varepsilon_a \varepsilon_b = sg(u_a u_b)),$$

where

$$u_a u_b = \zeta\, \mathrm{ch}\omega = \frac{1}{2 l_a l_b}[l_a^2 + l_b^2 + (\mathbf{z}_a - \mathbf{z}_b)^2] \tag{4.3}$$

$$(\omega \geq 0).$$

Expand now a function $f(u_a u_b)$ of the scalar product in terms of the addition function as

$$f(u_a u_b) = f(\zeta\, \mathrm{ch}\omega) = \int_0^\infty \sigma^2 d\sigma\, a_\zeta(\sigma) A^\sigma(\zeta\, \mathrm{ch}\omega). \tag{4.4}$$

This can easily be reversed,

$$a_\zeta(\sigma) = 2\pi^3 \int_0^\infty d\omega\, \mathrm{sh}^2\omega A^\sigma(\zeta\, \mathrm{ch}\omega) f(\zeta\, \mathrm{ch}\omega). \tag{4.5}$$

Thus the expansion of $f(u_a u_b)$ in terms of spherical functions takes the form

$$f(u_a u_b) = \int_0^\infty \sigma^2 d\sigma \int_{-\infty}^\infty d^2\mathbf{P}\, a_\zeta(\sigma)\, Y^\sigma_{\mathbf{P}}(u_a)^*\, Y^\sigma_{\mathbf{P}}(u_b). \tag{4.6}$$

2. Timelike–lightlike case

Let $u^2 = 1$, $k^2 = 0\,(u \in (H_+^\uparrow$ or $H_+^\downarrow)$, $k \in (H_0^\uparrow$ or $H_0^\downarrow))$ two timelike and lightlike vectors. Then the addition theorem is

$$A^\sigma(ku)^r = \int d^2\mathbf{P}\, y^\sigma_{\mathbf{P}}(k)^{r*}\, Y^\sigma_{\mathbf{P}}(u) = \frac{i\varepsilon}{\pi^2 \sigma} |2ku|^{-1-i\varepsilon_0 r\sigma}, \tag{4.7}$$

$$(\varepsilon_0 = sg\, l_0,\ \varepsilon = sg\, l,\ r = \pm)$$

where the scalar product (ku) is expressed through the coordinates $u = u(l, \mathbf{z})$, $k = k(l_0, \mathbf{z}_0)$ as

$$(ku) = \zeta w = \frac{1}{2ll_0}[l^2 + (\mathbf{z}-\mathbf{z}_0)^2] \qquad (4.8)$$

$$(\zeta = sg(ku)).$$

Expansion of a function of (ku) in terms of the addition function is a Mellin transformation,

$$f(ku) = f(\zeta w) = \sum_r \int_0^\infty \sigma^2 d\sigma \, a_\zeta(\sigma)_r A^\sigma(\zeta w)^r \qquad (4.9)$$

which has the inverse

$$a_\zeta(\sigma)_r = 2\pi^3 \int_0^\infty w dw \, A^{-\sigma}(\zeta w)^r f(\zeta w). \qquad (4.10)$$

3. Timelike–spacelike case

Let now $u^2 = 1$, $q^2 = -1$ ($u \in (H_+^\uparrow$ or $H_+^\downarrow)$, $q \in H_-$) a timelike and spacelike vector whose coordinates are $u = u(l_a, \mathbf{z}_a)$, $q = q(l_b, \mathbf{z}_b)$ then two addition theorems exist,

$$A^\sigma(uq)^+ = \int d^2\mathbf{P} \, Y_{\mathbf{P}}^\sigma(u)^* \, \mathbf{Y}_{\mathbf{P}}^\sigma(q)^+ = \frac{-i\varepsilon_a}{\sqrt{2}\pi^2\sigma} \frac{\cos(\sigma\omega)}{\mathrm{ch}\omega}, \qquad (4.11)$$

$$A^\sigma(uq)^- = \int d^2\mathbf{P} \, Y_{\mathbf{P}}^\sigma(u)^* \, \mathbf{Y}_{\mathbf{P}}^\sigma(q)^- = \frac{\varepsilon_a}{\sqrt{2}\pi^2\sigma} \frac{\sin(\sigma\omega)}{\mathrm{ch}\omega}, \qquad (4.12)$$

$$(\varepsilon_a = sg\, l_a)$$

where the scalar product can be written as

$$(uq) = \varepsilon_a \, \mathrm{sh}\omega = \frac{1}{2l_a l_b}[l_a^2 - l_b^2 + (\mathbf{z}_a - \mathbf{z}_b)^2] \qquad (4.13)$$

$$(-\infty < \omega < \infty).$$

The expansion in terms of the addition functions is a Fourier transformation

$$f(uq) = f(\varepsilon_a \, \mathrm{sh}\omega) = \sum_r \int_0^\infty \sigma^2 d\sigma \, A^\sigma(\varepsilon_a \, \mathrm{sh}\omega)^r a(\sigma)_r, \qquad (4.14)$$

$$(r = \pm)$$

where the dependence of $a(\sigma)_r$ on ε_a is suppressed. Eq.(4.14) has the inverse

$$a(\sigma)_r = 2\pi^3 \int_{-\infty}^\infty d\omega \;\; \mathrm{ch}^2\omega f(\varepsilon_a \, \mathrm{sh}\omega) A^{-\sigma}(\varepsilon_a \, \mathrm{sh}\omega)^r. \qquad (4.15)$$

4. *Lightlike–lightlike case*

The addition theorem in this case slightly differs from the other ones. The reason for this is that the integrals $\int d^2\mathbf{P}\, y^\sigma_{\mathbf{P}}(k)^{r^*} y^\sigma_{\mathbf{P}}(k)^{\pm r}$ do not exists in the ordinary sense, some kind of regularization is necessary. For real σ the relation $y^\sigma_{\mathbf{P}}(k_a)^{r^*} = y^\sigma_{-\mathbf{P}}(k_a)^{-r}$ holds. Substituting this into the above integral it becomes an analytic function of σ which acquires a meaning for certain complex values of σ and can be continued to its real values.

Let $k_a^2 = 0$, $k_b^2 = 0$ be two lightlike vectors ($k_a, k_b \in (H_0^\uparrow \text{or} H_0^\downarrow)$) with the coordinates $k_a = k_a(l_a, \mathbf{z}_a)$, $k_b = k_b(l_b, \mathbf{z}_b)$ and denote by $\varepsilon_a = sg\, l_a$, $\varepsilon_b = sg\, l_b$. These signs determine which light cone are k_a and k_b on.

The scalar product is simply

$$(k_a k_b) = \zeta w = \frac{(\mathbf{z}_a - \mathbf{z}_b)^2}{2 l_a l_b}, \qquad (\zeta = sg(k_a k_b)). \tag{4.16}$$

Let k_a and k_b be situated on the same light cone then

$$A^\sigma(k_a, k_b)^{sr} = \int d^2\mathbf{P}\, y^\sigma_{-\mathbf{P}}(k_a)^{-s}\, y^\sigma_{\mathbf{P}}(k_b)^r \tag{4.17}$$

with

$$A^\sigma(k_a, k_b) = \begin{pmatrix} \dfrac{1}{\pi^2\sigma^2}\left|\dfrac{l_a}{l_b}\right|^{i\varepsilon\sigma} \delta(k_a k_b) & \dfrac{1}{\pi^2 i\varepsilon\sigma}|2k_a k_b|^{-1-i\varepsilon\sigma} \\ -\dfrac{1}{\pi^2 i\varepsilon\sigma}|2k_a k_b|^{-1+i\varepsilon\sigma} & \dfrac{1}{\pi^2\sigma^2}\left|\dfrac{l_a}{l_b}\right|^{-i\varepsilon\sigma} \delta(k_a k_b) \end{pmatrix} \tag{4.18}$$

($\frac{1}{4} < \varepsilon s\, Im\sigma < 1$ for the off-diagonal elements, $\varepsilon_a = \varepsilon_b = \varepsilon$; $s, r = \pm$).

Elements of this and subsequent matrices are arranged so that in the first row the elements $(s, r) = (+, +); (+, -)$ are situated. The addition theorem is valid in the indicated range but it can be analytically continued to real values of σ.

Addition theorems for $\varepsilon_a = -\varepsilon_b = \varepsilon$ are obtained by interchanging the columns of the matrix (4.17).

As mentioned in the introduction the function

$$F(k_a, k_b) = f(k_a k_b) + g\left(\frac{l_a}{l_b}\right)\delta(k_a k_b)$$

is Lorentz invariant. Its expansion in terms of addition functions is

$$F(k_a, k_b) = \sum_{s,r} \int \sigma^2 d\sigma\, a_{rs}(\sigma) A^\sigma(k_a, k_b)^{sr}, \tag{4.19}$$

where the path of integration on the complex σ -plane for the off-diagonal elements $(s,r)=(r,-r)$ runs from $ir\varepsilon\sigma_2$ to $\infty+ir\varepsilon\sigma_2$, $(\sigma_2=Im\sigma\,,\ \frac{1}{4}<\varepsilon\sigma_2<1)$ and for the diagonal elements $(s,r)=(r,r)$ from 0 to ∞. If the coefficients $a_{-r,r}(\sigma)$ have no singularities in the strip $[\sigma_1+ir0,\sigma_1+i\varepsilon r\sigma_2]$, $0\leq\sigma_1<\infty$ the path of integration can be shifted to the real axis. Then the limits of integration in (4.19) are $0,\infty$. For the expansion coefficients one gets

$$a_{-r,r}(\sigma)=2\pi^3\int_0^\infty wdwA^{-\sigma}(w)^{r,-r}f(w), \tag{4.20}$$

$$a_{r,r}(\sigma)=\frac{\pi}{2}\int_0^\infty dw\,w^{-1\ -i\varepsilon r\sigma}g(w). \tag{4.21}$$

5. *Lightlike–spacelike case*

Let $k^2=0(k\in(H_0^\uparrow$ or $H_0^\downarrow))$ be a lightlike and $q^2=-1$ a spacelike vector $(q\in H_-)$. In terms of the parameters $k=k(l_0,\mathbf{z}_0)$, $q=q(l,\mathbf{z})$ the scalar product can be written as

$$(kq)=\varepsilon_0 w=\frac{1}{2l_0l}[-l^2+(\mathbf{z}-\mathbf{z}_0)^2]. \tag{4.22}$$

Taking into account the relation $y^\sigma_{\mathbf{P}}(k)^{r^*}=y^\sigma_{-\mathbf{P}}(k)^{-r}$ again and adding an imaginary part to σ the following addition theorem can be derived

$$A^\sigma(kq)^{sr}=\int d^2\mathbf{P}\,y^\sigma_{-\mathbf{P}}(k)^{-s}\mathbf{Y}^\sigma_{\mathbf{P}}(q)^r \tag{4.23}$$

$$(s,r=\pm),$$

with

$$A^\sigma(kq)=\frac{\varepsilon_0}{\sqrt{2}\pi^2 i\sigma}\begin{pmatrix}|2kq|^{-1-i\varepsilon_0\sigma} & -\zeta|2kq|^{-1-i\varepsilon_0\sigma}\\ -|2kq|^{-1+i\varepsilon_0\sigma} & -\zeta|2kq|^{-1+i\varepsilon_0\sigma}\end{pmatrix}, \tag{4.24}$$

$$(0<\varepsilon_0 s\sigma_2<1;\quad \sigma_2=Im\sigma,\quad \zeta=sg(kq),\quad \varepsilon_0=sgl_0).$$

The expansion in terms of the addition functions takes the form

$$f(kq)=\sum_{s,r}\int\sigma^2 d\sigma\,a_{rs}(\sigma)A^\sigma(kq)^{sr}. \tag{4.25}$$

The path of integration on the complex σ-plane runs from $is\varepsilon_0\sigma_2$ to $\infty+is\varepsilon_0\sigma_2$ but it can be shifted to the real axis provided $a_{rs}(\sigma)$ is free of singularities in the strip $[\sigma_1+is0,\sigma_1+is\sigma_2]$, $0\leq\sigma_1<\infty$. The integration in (4.25) runs then from 0 to ∞. The expansion can be reversed as

$$a_{rs}(\sigma)=2\pi^3\int_{-\infty}^\infty dw|w|\,A^{-\sigma}(w)^{sr}f(w). \tag{4.26}$$

6. *Spacelike–spacelike case*

Let $q_a^2 = -1$, $q_b^2 = -1$ be two spacelike vectors with the parameters $q_a = q_a(l_a, \mathbf{z}_a)$, $q_b = q_b(l_b, \mathbf{z}_b)$ then in addition to the scalar product

$$(q_a q_b) = \frac{1}{2l_a l_b}[-l_a^2 - l_b^2 + (\mathbf{z}_a - \mathbf{z}_b)^2] \tag{4.27}$$

the addition functions depend also on the invariant $I(q_a, q_b) = sg(q_b^- + (q_a q_b) q_a^-)$ defined formerly in (3.1),

$$A^\sigma(q_a q_b)^{sr} = \int d^2\mathbf{P}\, \mathbf{Y}^\sigma_{\mathbf{P}}(q_b)^{s*} \mathbf{Y}^\sigma_{\mathbf{P}}(q_a)^r =$$

$$= \begin{cases} A^\sigma_-(\Theta)^{sr}_I & \text{if } (q_a q_b) = -\operatorname{ch}\Theta < -1 \\ A^\sigma_0(\varphi)^{sr} & \text{if } -1 < (q_a q_b) = \cos\varphi < 1 \\ A^\sigma_+(\Theta)^{sr}_I & \text{if } (q_a q_b) = \operatorname{ch}\Theta > 1 \end{cases} \tag{4.28}$$

$$(s, r = \pm, \Theta > 0,\ 0 < \varphi < \pi, \sigma > 0)$$

where matrices $A^\sigma_-, A^\sigma_0, A^\sigma_+$ have the explicit form

$$A^\sigma_-(\Theta)_I = -\frac{1}{2\pi^2 \sigma \operatorname{sh}\Theta} \begin{pmatrix} \sin\sigma\Theta & iI\cos\sigma\Theta \\ iI\cos\sigma\Theta & \sin\sigma\Theta \end{pmatrix}, \tag{4.29}$$

$$(I = I(q_a, q_b))$$

$$A^\sigma_+(\Theta)_I = -\frac{1}{2\pi^2 \sigma \operatorname{sh}\Theta} \begin{pmatrix} \sin\sigma\Theta & -iI\cos\sigma\Theta \\ iI\cos\sigma\Theta & -\sin\sigma\Theta \end{pmatrix}, \tag{4.30}$$

$$A^\sigma_0(\varphi) = \frac{1}{2\pi^2 \sigma \operatorname{sh}\pi\sigma \sin\varphi} \begin{pmatrix} \operatorname{ch}\sigma + & \operatorname{ch}\sigma(\pi - \varphi) & 0 \\ 0 & & \operatorname{ch}\sigma\varphi - \operatorname{ch}\sigma(\pi - \varphi) \end{pmatrix}. \tag{4.31}$$

The addition theorem for the spherical functions of the discrete spectrum reads

$$B^n(q_a q_b) = \int d^2\mathbf{P}\, \mathbf{Y}^n_{\mathbf{P}}(q_b)^* \mathbf{Y}^n_{\mathbf{P}}(q_a) =$$

$$= \begin{cases} 0 & \text{if } (q_a q_b) < -1, \\ \dfrac{(-)^n \cos n\varphi}{\pi^2 |n| \sin\varphi} & \text{if } -1 < (q_a q_b) = \cos\varphi < 1, \\ 0 & \text{if } (q_a q_b) > 1, \end{cases} \tag{4.32}$$

$$(0 < \varphi < \pi).$$

An invariant function may depend on $I(q_a, q_b)$ in addition to the scalar product. Since I is defined only for $(q_a q_b)^2 \geq 1$, the value of an invariant function in the intermediate range $-1 < (q_a q_b) < 1$ is supposed to be independent of I.

The expansion in the two outer ranges reads

$$f_I(-ch\Theta) = \sum_{s,r} \int_0^\infty \sigma^2 d\sigma\, a_{rs}(\sigma) A_-^\sigma(\Theta)_I^{sr} \qquad (4.33)$$

$$((q_a q_b) = -ch\Theta < -1),$$

$$f_I(\text{ch}\,\Theta) = \sum_{s,r} \int_0^\infty \sigma^2 d\sigma\, a_{rs}(\sigma) A_+^\sigma(\Theta)_I^{sr} \qquad (4.34)$$

$$(ch\Theta = (q_a q_b) > 1).$$

Eqs (4.33), (4.34) can be reversed by knowing the values of $f_I(q_a q_b)$ merely in the outer ranges,

$$a_{rs}(\sigma) = 2\pi^3 \int_0^\infty d\Theta\, \text{sh}^2\Theta \sum_I [f_I(-\text{ch}\Theta) A_-^{-\sigma}(\Theta)_I^{sr} + f_I(\text{ch}\Theta) A_+^{-\sigma}(\Theta)_I^{sr}]. \quad (4.35)$$

In the intermediate range $-1 < (q_a q_b) < 1$ both the addition functions of the continuous and discrete spectra are needed,

$$f(\cos\varphi) = \sum_{s,r} \int_0^\infty \sigma^2 d\sigma\, a_{rs}(\sigma) A_0^\sigma(\varphi)^{sr} + \sum_{n=1}^\infty n^2 b_n B^n(\cos\varphi) \qquad (4.36)$$

$$((q_a q_b) = \cos\varphi,\ 0 < \varphi < \pi).$$

Since the coefficients $a_{rs}(\sigma)$ have been determined through Eq. (4.35) the integral part of the expansion can be considered as a known function. There is, however, an ambiguity in the value of the expansion coefficients $a_{rs}(\sigma)$, disregarded in (4.33), (4.34). On substituting (4.29) and (4.30) into these equations it is seen that the expansion coefficients are written in the form $\sigma a_{rs}(\sigma)$ which means that $a_{rs}(\sigma)$ itself is determined only up to an additive term proportional to $\delta(\sigma)$ i.e. the replacement $a_{rs}(\sigma) \longrightarrow a_{rs}(\sigma) + \pi^3 a_0 \delta(\sigma)$ yields the same $f_I(\pm\,\text{ch}\Theta)$ functions in (4.33) and (4.34). This ambiguity, irrelevant in the outer ranges, becomes relevant in the intermediate range where the expansion (4.36) holds since under this replacement the integral term transforms as

$$\sum_{s,r} \int_0^\infty \sigma^2 d\sigma (a_{rs}(\sigma) + \pi^3 a_0 \delta(\sigma)) A_0^\sigma(\varphi)^{sr} = \sum_{s,r} \int_0^\infty \sigma^2 d\sigma\, a_{rs}(\sigma)\, A_0^\sigma(\varphi)^{sr} + \frac{a_0}{2} \frac{1}{\sin\varphi}.$$

The coefficient a_0 can be determined by multiplying Eq. (4.36) by $\sin\varphi$ and integrating from 0 to π. The value of a_0 thus obtained is $a_0 = f_0 + g_0$ where f_0 and g_0 are given by Eqs (4.38), (4.39).

Substituting (4.35) into (4.36) the Chebyshev expansion of the integral term can be given explicitly. Accordingly, the coefficient b_n in (4.36) consists of two terms, the contribution from the integral term and from $f(\cos\varphi)$ itself,

$$\frac{(-)^n}{\pi^2} n b_n = f_n + g_n, \; a_0 = f_0 + g_0 \qquad (n = 1, 2, 3, \ldots), \tag{4.37}$$

with

$$f_n = \frac{2}{\pi} \int_0^{\pi} d\varphi \, \sin\varphi \, \cos n\varphi \, f(\cos\varphi) \qquad (n = 0, 1, 2, \ldots), \tag{4.38}$$

$$g_n = \frac{2}{\pi} \int_0^{\infty} d\Theta \, \mathrm{sh}\Theta \, e^{-n\Theta} \frac{1}{2} \sum_I [f_I \, (\mathrm{ch}\, \Theta) + (-)^n f_I(-\mathrm{ch}\, \Theta)] \qquad (n = 0, 1, 2, \ldots). \tag{4.39}$$

Thus, finally, the expansion in terms of spherical functions takes the form

$$f_I(q_a q_b) = \sum_{s,r} \int_0^{\infty} \sigma^2 d\sigma \, d^2\mathbf{P} [a_{rs}(\sigma) + a_0 \pi^3 \delta(\sigma)] \mathbf{Y}^{\sigma}_{\mathbf{P}}(q_b)^{*s} \mathbf{Y}^{\sigma}_{\mathbf{P}}(q_a)^r + \\ + \sum_{n=1}^{\infty} n^2 b_n \int d^2\mathbf{P} \, \mathbf{Y}^n_{\mathbf{P}}(q_b)^* \mathbf{Y}^n_{\mathbf{P}}(q_a), \tag{4.40}$$

where the expansion coefficients are given by (4.35), (4.37).

References

1. M. Huszár, Acta Phys. Hung., *58*, 175, 1985.
2. Yi.A. Verdiyev, Ann. of Phys., *72*, 139, 1972, p. 166.
3. N.Y. Vilenkin, Special Functions and Theory of Group Representations AMS Transl. Providence, Rhode Island. 1968.
4. M. Huszár, Report KFKI, 1986-100/A,1985-78.
5. H. Bateman and A. Erdélyi, Higher Transcendental Functions,McGraw-Hill, New-York, 1953, Vol. 2,7. 10. 5(75).
6. R.S. Strichartz, Journ. of Func. Anal., *12*, 341, 1973.
7. A.P. Prudnikov, Y. A. Brichkov and O. I. Marichev, Integrals and Series, Vol. II. Special Functions, "Nauka", Moscow 1983 (in Russian).
8. I. M. Gel'fand, M. I. Graev and N. Y. Vilenkin, Generalized Functions Vol. 5, Integral Geometry and Representation Theory, Academic Press, New-York, London, 1986.

Acta Physica Hungarica 64 (4), pp. 379–384 (1988)

EFFECT OF PERTURBATION TO THE THERMODYNAMIC SYSTEM

P. SAMU

Department for Low Temperature Physics, Roland Eötvös University
1088 Budapest, Hungary

(Received 30 July 1986)

According to the Le Chatelier's principle the spontaneous processes induced by a deviation from equilibrium are in a direction to restore the system to equilibrium. We give this principle in detailed and explicit form.

1. Introduction

It is important to know about a physical system whether it is stable or not. If it is in equilibrium state then it is known that the conditions of stability are $ds = 0$ and $d^2S < 0$ where S is the entropy function of the system [1]. Now we consider a system that is taken out of equilibrium by some imposed perturbation. According to the Le Chatelier–Braun principle the perturbation directly includes a process that reduces the perturbation.

The most generalized thermostatic form of this principle is due to Fényes and Tisza and is in close connection with the following inequality series [2, 3]

$$\frac{\partial y_1}{\partial x_1}\Big|_{x_2,\ldots,x_n} > \frac{\partial y_1}{\partial x_1}\Big|_{y_2,x_3,\ldots,x_n} > \frac{\partial y_1}{\partial x_1}\Big|_{y_2,\ldots,y_{n-1},x_n} > \frac{\partial y_1}{\partial x_1}\Big|_{y_2,\ldots,y_n} \tag{1}$$

$$\frac{\partial x_1}{\partial y_1}\Big|_{y_2,\ldots,y_n} > \frac{\partial x_1}{\partial y_1}\Big|_{x_2,y_3,\ldots,y_n} > \frac{\partial x_1}{\partial y_1}\Big|_{x_2,\ldots,x_{n-1},y_n} > \frac{\partial x_1}{\partial y_1}\Big|_{x_2,\ldots,x_n}, \tag{2}$$

which refer to the initial state of the perturbed system, where x_i are the extensive parameters and y_i the intensive ones. These are consequences of the entropy maximum principle.

We characterize the departure from equilibrium state with parameters α_i[4,5]

$$\alpha_i = \frac{x_i - x_{io}}{x_{n+1}}, \qquad i = 1, 2 \ldots, n, \tag{3}$$

then the thermodynamic forces are:

$$X_i = \frac{\partial \alpha_s}{\partial \alpha_i} - \left.\frac{\partial \alpha_s}{\partial \alpha_i}\right|_0, \qquad i = 1, 2 \ldots, n, \tag{4}$$

where α_s is the α parameter of the entropy as the function of the other α_i parameters.

By means of parameters α_i and forces X_i I. Kirschner gave another way to characterize the perturbed system initial state [6].

$$|X_1|_{\alpha_2,\ldots,\alpha_n} > |X_1|_{X_2,\alpha_3,\ldots,\alpha_n} > |X_1|_{X_2,X_3\alpha_4,\ldots,\alpha_n} > \cdots > |X_1|_{X_2,\ldots,X_n} \tag{5}$$

$$|\alpha_1|_{X_2,\ldots,X_n} > |\alpha_1|_{a_2,x_3,\ldots,X_n} > |\alpha_1|_{\alpha_2,\alpha_3,X_4,\ldots,X_n} > \cdots > |\alpha_1|_{\alpha_2,\ldots,\alpha_n} \tag{6}$$

and by fluxes: $\dot{\alpha}_i = \frac{d\alpha_i}{dt}$

$$|\dot{\alpha}_1|_{X_2,\ldots,X_n} > |\dot{\alpha}_1|_{\dot{\alpha}_2,X_3,\ldots,X_n} > |\dot{\alpha}_1|_{\dot{\alpha}_2,\dot{\alpha}_3,X_4,\ldots,X_n} > \cdots > |\dot{\alpha}_1|_{\dot{\alpha}_2,\ldots,\dot{\alpha}_n} \tag{7}$$

where $X_1 \neq 0$ and

$$|X_1|_{\dot{\alpha}_2,\ldots,\dot{\alpha}_n} > |X_1|_{X_2,\dot{\alpha}_3,\ldots,\dot{\alpha}_n} > |X_1|_{X_2,X_3\dot{\alpha}_4,\ldots,\dot{\alpha}_n,} > \cdots > |X_1|_{X_2,\ldots,X_n}, \tag{8}$$

where $\dot{\alpha}_1 \neq 0$.

2. Effect of perturbation to the thermodynamic system

The parameter α of the entropy is:

$$\alpha_s = \alpha_s(\alpha_1\alpha_2\ldots,\alpha_n). \tag{9}$$

We can expand α_s in powers of the parameters α_i

$$\alpha_s = \alpha_s(0) + \sum_i \frac{\partial\alpha_s}{\partial\alpha_i}\alpha_i + \frac{1}{2}\sum_{i,k}\frac{\partial^2\alpha_s}{\partial\alpha_i\partial\alpha_{k_0}}\alpha_i\alpha_k + \ldots. \tag{10}$$

We introduce the positive definite symmetric matrix g_{ik}

$$g_{ik} = -\frac{\partial^2\alpha_s}{\partial\alpha_k\partial\alpha_i} \tag{11}$$

and by inserting Eq. (10) in Eq. (4), we find in second-order approximation

$$x_i = -\sum_k g_{ik}\alpha_k,$$
$$i = 1,2,\ldots,n. \tag{12}$$

From Eq. (4) it follows in second-order approximation

$$\Delta\alpha_s = -\frac{1}{2}\sum_{i,k} g_{ik}\alpha_i\alpha_k = -\frac{1}{2}\hat{g}\underline{\alpha}\underline{\alpha}. \tag{13}$$

The entropy production is the sum of products of each force with its conjugated affinity [7, 8] and in this case, we have

$$\sigma = \Delta\dot{\alpha}_s = \sum_{i,k} g_{ik}\dot{\alpha}_i\alpha_k = \sum_i \dot{\alpha}_i X_i = \underline{\dot{\alpha}}\mathbf{X}. \tag{14}$$

Differentiating the entropy production σ with respect to the time, we have:

$$\dot{\sigma} = \underline{\ddot{\alpha}}\mathbf{X} + \underline{\dot{\alpha}}\dot{\mathbf{X}}, \tag{15}$$

$$\underline{\dot{\alpha}} = \hat{L}\mathbf{X}, \tag{16}$$

where the Onsager theorem states that L_{ik} is a positive definite symmetric matrix [9].

Assuming that $\hat{L}$ is constant the acceleration can be written as follows

$$\underline{\ddot{\alpha}} = \hat{L}\dot{\mathbf{X}}. \tag{17}$$

Inserting this Eq. (15), we obtain

$$\dot{\sigma} = \hat{L}\dot{\mathbf{X}}\mathbf{X} + \hat{L}\mathbf{X}\dot{\mathbf{X}}, \tag{18}$$

where the two terms on the right hand side are equal. Thus

$$\dot{\sigma} = 2\hat{\mathbf{L}}\mathbf{X}\dot{\mathbf{X}} = 2\underline{\dot{\alpha}}\dot{\mathbf{X}}. \tag{19}$$

It is remarkable that this result was obtained already in 1963 by Gyarmati and Oláh, however, in a different way [10]. From Table I substituting any initial condition into Eq. (19) we find

$$\dot{\sigma} = 2\dot{X}_1\dot{\alpha}_1 \tag{20}$$

and substituting the K-th initial condition into Eqs. (12) we have

$$\begin{aligned}
X_1 &= -g_{11}\dot{\alpha}_1 - g_{22}\dot{\alpha}_2 - \cdots - g_{1,n-k}\dot{\alpha}_{n-k},\\
0 &= -g_{21}\dot{\alpha}_1 - g_{22}\dot{\alpha}_2 - \cdots - g_{2,n-k}\dot{\alpha}_{n-k},\\
0 &= -g_{31}\dot{\alpha}_1 - g_{32}\dot{\alpha}_2 - \cdots - g_{n-k,n-k}\dot{\alpha}_{n-k},\\
&\cdot\\
&\cdot\\
&\cdot\\
0 &= -g_{n-k,1}\dot{\alpha}_1 - g_{n-k,2}\dot{\alpha}_2 - \cdots - g_{3,n-k}\dot{\alpha}_{n-k},
\end{aligned} \tag{21}$$

Table I

Perturbation possibility		Parameters equal to zero	Parameters unequal to zero
	1.	$x_2, x_3, \ldots, x_{k-1}, x_k, \ldots, x_{n-1}, x_n$	$\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{k-1}, \alpha_k, \ldots, \alpha_{n-1}, \alpha_n$
	2.	$x_2, x_3, \ldots, x_{k-1}, x_k, \ldots, x_{n-1}, \alpha_n$	$\alpha_1, \alpha_2, \ldots, \alpha_{k-1}, \alpha_k, \ldots, \alpha_{n-1}, x_n$
	3.	$x_2, x_3, \ldots, x_{k-1}, \ldots, x_{n-2}, \alpha_{n-1}, \alpha_n$	$\alpha_1, \alpha_2, \ldots, \alpha_{k-1}, \ldots, \alpha_{n-2}, x_{n-1}, x_n$
	.		
$x_1 \neq 0$	.		
	.		
	/k+1/	$x_2, x_3, \ldots, x_{n-k}, \alpha_{n-k+1}, \alpha_{n-k+2}, \ldots, \alpha_n$	$\alpha_1, \alpha_2, \ldots, \alpha_{n-k}, x_{n-k+1}, x_{n-k+2}, \ldots, x_n$
	.		
	.		
	.		
	n	$\alpha_2, \alpha_3, \ldots, \alpha_{k-1}, \ldots, \alpha_{n-1}, \alpha_n$	$\alpha_1, x_2, x_3, \ldots, x_{n-k}, x_{n-k+1}, \ldots, x_n$

where we have written out the first $n-k$ equations explicitly.

$$\alpha_1 = \frac{D^1_{n-k}}{D_{n-k}} X_1, \tag{22}$$

where notation D_{n-k} stands for the principal minor of the g_{ik} matrix with $n-k$ rows and columns and D^1_{n-k} is obtainable from D_{n-k} omitting the first row and column. Because:

$$\frac{D_k}{D^1_k} = \frac{D_{k-1}}{D^1_{k-1}} - \frac{D^1_{k-1}}{D^1_k} DD', \tag{23}$$

where

$$DD' = \begin{vmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,k-1} \\ g_{2,k} & & & \\ \cdot & & & \\ \cdot & & D^1_{k-1} & \\ g_{k-1,k} & & & \end{vmatrix} \begin{vmatrix} & & & g_{k,2} \\ & & & g_{k,3} \\ & & D^1_{k-1} & \cdot \\ & & & \cdot \\ & & & g_{k,k-1} \\ g_{k-2,k} & \cdots & g_{k,k} & g_{k,1} \end{vmatrix}. \tag{24}$$

Thus by the help of positive definitivity of matrix g_{ik} from Eq. (22) it follows:

$$|\dot{\alpha}_1|_{\dot{X}_2,\ldots,\dot{X}_n} > |\dot{\alpha}_1|_{\dot{X}_2,\ldots,\dot{X}_{n-1},\dot{\alpha}_n} > |\dot{\alpha}_1|_{\dot{X}_2,\ldots,\dot{X}_{n-2},\dot{\alpha}_{n-1},\dot{\alpha}_n} > \cdots > |\dot{\alpha}_1|_{\dot{\alpha}_2,\ldots,\dot{\alpha}_n}. \tag{25}$$

Inserting this in Eq. (19), we have:

$$|\dot{\sigma}|_{\dot{X}_2,\ldots,\dot{X}_n} > |\dot{\sigma}|_{\dot{X}_2,\ldots,\dot{X}_{n-1},\dot{\alpha}_n} > |\dot{\sigma}|_{\dot{X}_2,\ldots,\dot{X}_{n-2},\dot{\alpha}_{n-1},\dot{\alpha}_n} > \cdots > |\dot{\sigma}|_{\dot{\alpha}_2,\ldots,\dot{\alpha}_n}. \tag{26}$$

Similarly, if the initial state is $\alpha_1 \neq 0$ and the bound conditions are as in Table I, we obtain:

$$|\dot{X}_1|_{\dot{\alpha}_2,\ldots,\dot{\alpha}_n} > |\dot{X}_1|_{\dot{X}_2,\dot{\alpha}_3,\ldots,\dot{\alpha}_n} > |\dot{X}_1|_{\dot{X}_2,\dot{X}_3\dot{\alpha}_4,\ldots,\dot{\alpha}_n} > \cdots > |\dot{X}_1|_{\dot{X}_2,\ldots,\dot{X}_n} \tag{27}$$

and from Eq. (19) it follows:

$$|\dot{\sigma}|_{\dot{\alpha}_2,\ldots,\dot{\alpha}_n} > |\dot{\sigma}|_{\dot{X}_2\dot{\alpha}_3,\ldots,\dot{\alpha}_n} > |\dot{\sigma}|_{\dot{X}_2\dot{X}_3\dot{\alpha}_4,\ldots,\dot{\alpha}_n} > \cdots > |\dot{\sigma}|_{\dot{X}_2\ldots,\dot{X}_n}. \tag{28}$$

We would like to state that the inequality series (25), (26), (27) and (28) refer to the initial state of the perturbed system, and give information about the velocity of decrease of the deviation from equilibrium.

Acknowledgement

I would like to thank Professor I. Kirschner for useful discussions.

References

1. H.B. Callen, Thermodynamics, John Wiley, New York, 1960.
2. I. Fényes, Acta Phys. Hung., *8*, 419, 1958.
3. L. Tisza, Ann. of Phys., *13*, 1, 1961.
4. I. Kirschner, Acta Phys. Hung., *30*, 61, 1971.
5. I. Kirschner, Acta Phys. Hung., *29*, 319, 1970.
6. I. Kirschner, Magyar Fizikai Folyóirat, *XVII*, 71, 1969 (in Hungarian).
7. I. Gyarmati, Non-equilibrium Thermodynamics, Field Theory and Variational Principles, Springer-Verlag, Berlin, 1970.
8. I. Kirschner and R. Tőrös, Non-equilibrium Thermodynamics, *9*, 156, 1984.
9. L. Onsager, Phys. Rev., *37*, 405, 1931.
10. I. Gyarmati and K. Oláh, Acta Chimica Academiae Scientiarium Hungaricae, *35*, 95, 1963.

Acta Physica Hungarica 64 (4), pp. 385-393 (1988)

SURFACE SCATTERING POTENTIAL FOR ELECTRON DIFFRACTION*

K. STACHULEC

Technical University,
25-314 Kielce, Poland

(Received 30 July 1986)

A simple dynamical scattering potential for electron diffraction by a free surface of a solid state sample is proposed. A few monoatomic layers parallel to the surface of the scattering sample are treated as a thin film while the other part of the sample is considered as a substrate. The scattering potential of the sample is expressed in terms of the mean square displacement of atoms and the electron density distribution at a scattering surface of the thin film sample. The obtained formula for the scattering potential in special cases leads to results of Dvoriankin's paper [3]. The proposed surface scattering potential can be used to describe low energy electron diffraction (LEED) as well as high energy electron diffraction (HEED) experiments.

1. Introduction

In the present paper we propose a simple dynamical scattering potential for electron diffraction by a solid state sample with a free surface. The sample is treated as a thin film evaporated on a substrate. We regard a thin film as a system of a few monoatomic layers parallel to the surface of the sample, while the other part of the sample is considered as the substrate.

For such samples we have calculated a scattering potential including the exchange of the incident electrons with the electrons of the sample. Including the exchange part of the scattering potential gives a possibility to use of the scattering potential to describe a polarized electron diffraction experiment in the cases of low energy electron diffraction (LEED) as well as of high energy electron diffraction (HEED).

The proposed surface scattering potential in its analytical form is a temperature and thickness dependent quantity and it is expressed in terms of the mean square displacement of atoms and the electron density distribution of the thin film. The division of the scattering film into the thin film and the substrate allows one to compute the dynamical parameters of the scattering potential by means of the field theories developed for thin films [2].

*Supported by the Polish Goverment Program, CPBP 01. 08. B3

2. The scattering potential for electron diffraction by a sample with free surface

We shall assume that a spin polarized electron beam is incident on a perfectly clean, well ordered surface of a sample. Our interest will focus on the elastically scattering electrons, because they produce almost all the structure in the diffraction patterns.

To construct the scattering potential we divide the scattering sample into a thin film and a substrate. By the thin film we understand n monoatomic layers parallel to the surface. They are numbered by ν, starting with $\nu = 1$ for the free surface of the film and finishing with $\nu = n$ for the atomic layer evaporated directly on the substrate. Any monoatomic layer ν we divide into two-dimensional elementary cells and the position of any cell inside the ν-th atomic layer related to the cell chosen as the origin of the coordinate system is given by a two-dimensional vector $\mathbf{j}_\nu = \mathbf{a}j_x^\nu + \mathbf{b}j_y^\nu$, where $\mathbf{a}$ and $\mathbf{b}$ are lattice vectors and j_x^ν, j_y^ν denote integer numbers.

Let the z-axis of the coordinate system be perpendicular to the surface and directed into the inside of the film. A position of any atom in the film can be described by the vector

$$\mathbf{R}_{\nu\mathbf{j}_\nu s} = \mathbf{j}_\nu + \hat{e}_z c_\nu \nu + \rho_s^\nu = \mathbf{R}_{\nu\mathbf{j}_\nu} + \rho_s^{\nu'}, \tag{1}$$

where ρ_S^ν describes the position of an s-th atom belonging to the $(\nu\mathbf{j}_\nu)$-th cell related to the origin of the local coordinate system bounded with the $(\nu\mathbf{j}_\nu)$-th cell inside the film.

From the point of view of the diffraction problem the thin film is regarded as a system of $n \cdot N^2$ bounded atoms, where N^2 denotes the number of atoms in any ν-th layer parallel to the surface.

Such a system produces a suitable effective scattering potential localized round about each lattice node of the film. In general case, the scattering potential of the $(\nu\mathbf{j}_\nu s)$-th atom of the thin film depends on the position $(\mathbf{r} - \mathbf{R}_{\nu\mathbf{j}_\nu s})$ and spin coordinates s_0 of the diffracted electrons. Denoting this potential by $V(\mathbf{r} - \mathbf{R}_{\nu\mathbf{j}_\nu s}, s_0)$, a total scattering potential on the position $\mathbf{r}$ inside the film and spin s_0 $V(\mathbf{r}, s_0)$ can be treated in a first approximation as the superposition of the effective potentials of the individual atoms from which the film is built:

$$V(\mathbf{r}, s_0) = \sum_{\nu\mathbf{j}_\nu s} V(\mathbf{r} - \mathbf{R}_{\nu\mathbf{j}_\nu s}, s_0). \tag{2}$$

By the effective potential of the individual atom we understand the potential produced by a given atom in the presence of other atoms of the thin film and the substrate on which the film is evaporated. This potential can be constructed by a modification of the scattering potential of a free atom.

Now, let us consider the scattering of an electron by free atom. Denoting by $\mathbf{r}_0$ and s_0 the position and spin coordinates of the incident electron, and by $\mathbf{r}_1 s_1$,

$\mathbf{r}_2 s_2, \ldots \mathbf{r}_N s_N$, the position and spin coordinates of the atomic electrons, where N denotes the total number of electrons in the atom, we should describe the scattering problem by means of a wave function $\Phi = \Phi(\mathbf{r}_0 s_0, \mathbf{r}_1 s_1, \ldots \mathbf{r}_N s_N)$ which depends on all coordinates of the $(N+1)$ electrons.

The many body function Φ has to satisfy an equation which we write in the form:

$$\left[\sum_{j=0}^{N}\left(-\frac{\hbar^2}{2m}\nabla^2 - \frac{Ze^2}{r_j}\right) + \sum_{j=0}^{N}\sum_{i=j+1}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|}\right]\Phi = E_t\Phi, \tag{3}$$

where $(-\hbar^2\nabla^2/2m)$ is the operator of the kinetic enegy of the j-th electron, so that $\sum(-\hbar^2\nabla^2/2m)$ describes the kinetic energy of the system of N atomic electrons plus an incident electron. The term $\sum(-Ze^2/r_j)$ is the potential energy of all the eletrons in the field of nucleus, the third term in Eq. (3) stands for the Coulomb interaction between electrons. E_t denotes the total energy of the system and Z_e the nuclear charge while j numbers the electrons of the system.

The difficulty in Eq. (3) is that electrons influence one another via Coulomb repulsion and the incident electron may distort wave functions of the atomic electrons by its own electrostatic field correlating their motion with its own and changing the potential seen by the incident electron itself.

If the free atom wave function is known, we can express the many body function by means of the free atom functions $\psi_1(\mathbf{r}_1 s_1)\psi_2(\mathbf{r}_2 s_2)\ldots\psi_N(r_N s_N)$ as follows [1]

$$\Phi = \sum_p \varepsilon_p \phi(\mathbf{r}_0 s_0)\psi_1(\mathbf{r}_1 s_1)\ldots\psi_N(\mathbf{r}_N s_N), \tag{4}$$

where $\phi(\mathbf{r}_0 s_0)$ is the incident electron's function.

The sum is over permutations of the particle coordinates and ε_p takes the value of $+1$ if permutation p can be achieved by exchanging an even number of particle coordinates, -1 for an odd number. There are $(N+1)!$ possible permutations.

Since the atom state part of Φ is known, we want to eliminate this part from our equations and concentrate on $\phi(\mathbf{r}_0 s_0)$ being the incident electron wave function.

Multiplying Eq. (3) by

$$\chi^* = \psi_1^*(\mathbf{r}_1 s_1)\psi_2^*(\mathbf{r}_2 s_2)\ldots\psi_N^*(\mathbf{r}_N s_N), \tag{5}$$

integrating over space, and summing over spin coordinates of χ we obtain the equation for $\phi(\mathbf{r}_0 s_0)$

$$\left[-\frac{\hbar^2}{2m}\nabla_0^2 - \frac{Ze^2}{r_0} + \sum_j \int \sum_{s_j} \frac{|\psi_j(\mathbf{r}_j s_j)|^2}{|\mathbf{r}_0 - \mathbf{r}_j|} d^3 r_j\right]\phi(\mathbf{r}_0 s_0) -$$

$$-e^2 \sum_j \int \sum_{s_j} \frac{\psi_j^*(\mathbf{r}_j s_j)\phi(\mathbf{r}_j s_j)}{|\mathbf{r}_0 - \mathbf{r}_j|} d^3 r_j \psi_j(\mathbf{r}_0 s_0) = E_0\phi(\mathbf{r}_0 s_0), \tag{6}$$

where E_0 denotes the energy of the incident electron.

The last equation is a one-electron equation for the wave function of the incident electron and it defines the effective scattering potential of the free atom $V_0(\mathbf{r}_0 s_0)$

$$V_0(\mathbf{r}_0 s_0) = \frac{Ze}{r_0} - e \sum_j \int \sum_{s_j} \frac{|\psi_j^*(\mathbf{r}_j s_j)|^2}{|\mathbf{r}_0 - \mathbf{r}_j|} d^3 r + V_{\text{exc}}(\mathbf{r}_0 s_0), \tag{7}$$

where by $V_{\text{exc}}(\mathbf{r}_0 s_0)$ we denote the so called exchange potential, which is defined by the equation:

$$V_{\text{exc}}(\mathbf{r}_0 s_0)\phi(\mathbf{r}_0 s_0) = \sum_j \sum_{s_j} \int d^3 r_j \frac{\psi^*(\mathbf{r}_j s_j)\phi(\mathbf{r}_j s_j)}{|\mathbf{r}_0 - \mathbf{r}_j|} \psi_j(\mathbf{r}_0 s_0). \tag{8}$$

The two first terms on the right hand side of the equation (7) have a simple interpretation as the electrostatic potentials produced by the nucleus and the electron density distribution of the atom, respectively. The exchange potential defined by (8) is non-local potential and it arises out of considerations of antisymmetry of the wave function under exchange of particle coordinates, because no two electrons can be in the same place at the same time, hence each electron is surrounded by a region depleted of other electrons and hence lower potential.

Writing the one-electron wave function $\psi_j(\mathbf{r}_j s_j)$ as $\psi_j(\mathbf{r}_j s_j) = \psi_j(r_j)\chi_j(s_j)$, where $\chi_j(s_j)$ is the spin coordinate dependent part of the wave function of the j-th atomic electron while $\psi(\mathbf{r}_j)$ is space coordinate dependent only, we can express the electrostatic potential of atomic electrons by means of the electron distribution in the atom $\rho_j(\mathbf{r})$ as:

$$e \sum_j \sum_{s_j} \int d^3 r_j \frac{|\psi_j(\mathbf{r}_j s_j)|^2}{|\mathbf{r}_0 - \mathbf{r}_j|} = -e \sum_j \int d^3 r \frac{\rho_j(\mathbf{r})}{|\mathbf{r}_0 - \mathbf{r}|}. \tag{9}$$

Analogically, writing $\phi(\mathbf{r}_0 s_0) = \phi(\mathbf{r}_0)\eta(s_0)$, where $\eta(s_0)$ is the spin wave function of an incident electron, the exchange potential can be written as

$$V_{\text{exc}}(\mathbf{r}_0 s_0) = \sum_j \sum_{s_j} \frac{\chi_j^*(s_j)\eta(s_j)\chi_j(s_0)\eta^*(s_0)}{|\eta(s_0)|^2} \cdot I_j(\mathbf{r}_0) = \sum_j \delta_{s_0, s_j} I_j(\mathbf{r}_0), \tag{10}$$

where

$$I_j(r_0) = \frac{\psi_j(\mathbf{r}_0)\phi^*(\mathbf{r}_0)}{|\phi(\mathbf{r}_0)|^2} \int d^3 r_j \frac{\psi^*(\mathbf{r}_j)\phi(\mathbf{r}_j)}{|\mathbf{r}_0 - \mathbf{r}_j|} \tag{11}$$

and the sum is over the atomic electrons which have the same spin as the incident electron.

In this way a full scattering potential of a free atom takes the form:

$$V_{\text{at}}(\mathbf{r}_0 s_0) = \frac{eZ}{r_0} - e \sum_j \int \frac{\rho_j(\mathbf{r})}{|\mathbf{r}_0 - \mathbf{r}|} d^3 r + \sum_j \delta_{s_0 s_j} I_j(\mathbf{r}_0). \tag{12}$$

The above formula will be a basis to construct the scattering potential for the electrons scattered by thin films.

It is known that crystal lattice atoms at a given temperature take part in the temperature vibrations around their equilibrium position and the influence of the temperature on the scattering potential must be taken into consideration defining a dynamic scattering potential for an atom as [3, 7]

$$V_T(\mathbf{r}, s_0) = \int V_{\mathrm{at}}(\mathbf{r}', s_0') T(\mathbf{r} - \mathbf{r}') d^3 r', \tag{13}$$

where $T(\mathbf{r})$ describes a distribution of the mass centre of an atom in its vibrations around the equilibrium position. For the atom bounded in crystal $T(\mathbf{r})$ is determined by the lattice dynamics.

Other thing we would like to stress is that placing an atom in a lattice site of an infinite size sample does not change the mean square displacement which characterizes the dynamics of the atom in the sample and it is the same for every atom. However, in the thin film crystal case the surface atoms of any solid body are in a situation which is different from that of atoms situated in the inside of the film. The surface atoms feel the changes in the geometry of the neighbours surrounding them caused by the missing neighbours, by the spontaneous deformation of the lattice near the surface. As a result, the scattering potential near the surface must be different from that inside of a bulk material. It is to be expected that the changes of the electronic structure near the surface must cause some changes of the physical properties related to the surface.

To take above into consideration we introduce the effective numbers of electrons per j-th orbital of the $(\nu \mathbf{j}_\nu s)$ atom in the film $\langle n_{\nu \mathbf{j}_\nu s} \rangle$ which are the same for the atoms in the ν-th monoatomic layer but they create a distribution in the direction perpendicular to the surface. The redistribution of the electronic charge in the film creates some new boundary conditions for lattice vibrations which must influence the temperature dependence of the mean square displacements of the atoms in other atomic layers parallel to the surface. Denoting the mean square displacement of the $(\nu \mathbf{j}_\nu s)$-th atom in the film by $B_{\nu \mathbf{j}_\nu s}$ we can write

$$B_{\nu \mathbf{j}_\nu s} \equiv \langle (\delta \mathbf{R}_{\nu \mathbf{j}_\nu s})^2 \rangle = B_{\nu s}, \tag{14}$$

as well as for the electron density distribution of the j-th orbital

$$\langle n^j_{\nu \mathbf{j}_\nu s} \rangle = \langle n^j_{\nu s} \rangle, \tag{15}$$

where the symbol $\langle \ldots \rangle$ stands for the thermodynamical average, and the above relations are the consequence of the translation symmetry of the film.

By means of the above dynamical parameters $B_{\nu s}$ and $\langle n^j_{\nu s} \rangle$ we propose the following modification of the scattering potential of the atom in the $(\nu \mathbf{j}_\nu s)$-th site of the film:

$$V_T(\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}, s_0) = (2\pi B_{\nu s})^{-\frac{3}{2}} \int d^3 r' V_{\mathrm{at}}(\mathbf{r}' - \mathbf{R}_{\nu \mathbf{j}_\nu s}, s_0) e^{-\frac{|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s} - \mathbf{r}'|^2}{2B_{\nu s}}}, \tag{16}$$

where $T(\mathbf{r}') = (2\pi B_{\nu s})^{-\frac{3}{2}} \exp(-|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s} - \mathbf{r}'|^2/2B_{\nu s})$ describes the distribution of the mass centre of the $(\nu \mathbf{j}_\nu s)$-th atom in its temperature vibrations [3] and in place of the free atom potential V_{at} given by (12) we take $\tilde{V}_{\text{at}}$ which we define as

$$\tilde{V}_{\text{at}}(\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}, s_0) = \frac{Ze}{|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} - e \sum_j \langle n^j_{\nu s} \rangle \int \frac{\rho_j(\mathbf{r}_j - \mathbf{R}_{\nu \mathbf{j}_\nu s})}{|\mathbf{r}_0 - \mathbf{r}_j|^0} d^3 r_j + \\ + e \sum_j \langle n^j_{\nu s} \rangle I_j(\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}) \delta_{s_0 s_j}. \tag{17}$$

To obtain an analytic formula for the scattering potential of the spin polarized electrons scattered by thin films using the formulae (2), (16) and (17) we have to know the atomic orbitals of a free atom. By means of these orbitals the quantities ρ_j and I_j in formula (17) are expressed. For that we restrict our next considerations to the case when the free atom orbitals are spherically symmetric and we take them in the Slater form [1, 4]:

$$\psi_j(r) = A_j r^{\mu_j} \exp(\lambda_j r), \tag{18}$$

where μ_j, λ_j are the numerical parameters given for all free atoms in [4], and A_j is the normalization constant. If we now write the incident electron wave function in the form of the plane wave:

$$\phi(\mathbf{r}) = \exp(i\mathbf{k}\,\mathbf{r}) \tag{19}$$

where $\mathbf{k}$ is the wave vector of the incident electron, we can insert expressions (18) and (19) into the equation (17) and calculate the static scattering potential of $(\nu \mathbf{j}_\nu s)$ atom $\tilde{V}_{\text{at}}(\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}, s_0)$ of the film. We have:

$$\tilde{V}_{\text{at}}((\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}), s_0) = \frac{Ze}{|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} - \\ -4\pi e \sum_j A_j^2 \langle n^j_{\nu s} \rangle \left[\frac{1}{|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} \int_0^{|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} |\mathbf{r}_j - \mathbf{R}_{\nu \mathbf{j}_\nu s}|^{2\mu_j + 2} e^{-2\lambda_j |\mathbf{r}_j - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} dr_j + \right. \\ \left. + \int_{|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|}^{\infty} |\mathbf{r}_j - \mathbf{R}_{\nu \mathbf{j}_\nu s}|^{2\mu_j + 1} e^{-2\lambda_j |\mathbf{r}_j - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} dr_j \right] + \\ + 4\pi e \sum_j A_j^2 \langle n^j_{\nu s} \rangle |\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|^{\mu_j} e^{-(\lambda_j + ik')|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} * \\ * \left[\frac{1}{|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} \int_0^{|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} |\mathbf{r}_j - \mathbf{R}_{\nu \mathbf{j}_\nu s}|^{\mu_j + 2} e^{-(\lambda_j - ik')|\mathbf{r}_j - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} dr_j + \right. \\ \left. + \int_{|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|}^{\infty} |\mathbf{r} - \mathbf{R}_{\nu \mathbf{j}_\nu s}|^{\mu_j + 1} e^{-(\lambda_j - ik')|\mathbf{r}_j - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} dr_j \right]. \tag{20}$$

Integrating over r_j by means of the standard formulae [5]:

$$\int_0^u x^{\nu-1} \exp(-\mu x)dx = \mu^{-\nu}\gamma(\nu, \mu u),$$

$$\int_u^\infty x^{\nu-1} \exp(-\mu x)dx = \mu^{-\nu}\Gamma(\nu, \mu u),$$

where $\gamma(\nu, \mu u)$ and $\Gamma(\nu, \mu u)$ are so-called incomplete gamma functions, as the result we obtain

$$\begin{aligned}\tilde{V}_{\text{at}} = &\frac{Ze}{|\mathbf{r} - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} - 4\pi e \sum_j A_j^2 \frac{\langle n_{\nu s}^j \rangle}{(2\lambda_j)^{2\mu_j+2}} * \\ &* \left[\frac{\gamma(\mu_j + 3, |\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|2\lambda_j)}{|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|2\lambda_j} + \Gamma(\mu_j + 2, |\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|2\lambda_j)\right] + \\ &+ 4\pi e \sum_j A_j^2 \delta_{s_0 s_j} \langle n_{\nu s}^j \rangle e^{-(\lambda_j + ik')|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|} \cdot |\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|^{\mu_j} (\lambda_j - ik')^{(\mu_j+2)} * \\ &* \left[\frac{\gamma(\mu_j + 3, (\lambda_j - ik')|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|)}{|r_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|(\lambda_j - ik')} + \Gamma(\mu_j + 2, (\lambda_j - ik')|\mathbf{r}_0 - \mathbf{R}_{\nu \mathbf{j}_\nu s}|)\right].\end{aligned} \tag{21}$$

Next, the calculation for the dynamical scattering potential by means of the formula (16) and explicit form of $\tilde{V}_{\text{at}}$ given by (21) leads to the expression

$$V_T(\mathbf{r} - \mathbf{R}_{\nu \mathbf{j}_\nu s}, s_0) = e^{-\frac{|\mathbf{r} - \mathbf{R}_{\nu \mathbf{j}_\nu s}|^2}{2B_{\nu s}}} \sum_{k=0}^{\infty} \frac{(1/B_{\nu s})^{2k}}{(2k+1)!} \tilde{V}_T^{(k)}(B_{\nu s}, s_0), \tag{22}$$

where

$$\begin{aligned}\tilde{V}_T^{(k)}(B_{\nu s}, s_0) = (2\pi B_{\nu s})^{-\frac{3}{2}} \int_0^\infty 4\pi r'^{2k+2} e^{-\frac{r'^2}{2B_{\nu s}}} \cdot \tilde{V}_{\text{at}}(\mathbf{r}', s_0)dr' = \\ = Ze\langle F_{0,-1}\rangle_k - 4\pi e \sum_j A_j^2 \langle n_{\nu s}^j \rangle \frac{(2\mu_j+1)!}{(2\lambda_j)^{2\mu_j+2}} * \\ * \left[\frac{2\mu_j + 2}{2\lambda_j}\langle F_{0,-1}\rangle_k - (2\mu_j + 2)\sum_{m=0}^{2\mu_j+2} \frac{(2\lambda_j)^{m-1}}{m!}\langle F_{2\lambda_j, m-1}\rangle_k + \right. \\ \left. + \sum_{m=0}^{2\mu_j+1} \frac{(2\lambda_j)^m}{m!}\langle F_{2\lambda_j, m}\rangle_k\right] + 4\pi e \sum_j A_j^2 \langle n_{\nu s}^j \rangle \delta_{s_0 s_j} \frac{(2\mu_j+1)!}{\beta_j^{\mu_j+2}} * \\ * \left[(\mu_j + 2)\langle\left(F_{\beta_j', \mu_j - 1}\rangle_k \frac{1}{\beta_j} - \sum_{m=0}^{\mu_j+2} \frac{\beta_j^{m-1}}{m!}\langle F_{\beta_j', m+\mu_j-1}\rangle_k\right) + \sum_{m=0}^{\mu_j+1} \frac{\beta_j^m}{m!}\langle F_{\beta_j', m+\mu_j}\rangle_k\right],\end{aligned} \tag{23}$$

where $F_{xn} = r^n e^{-xr}$ and $\langle F_{x,n} \rangle$ stands for the following integral for the $(\nu \mathbf{j}_\nu s)$ - site of the lattice

$$\langle F_{x,n} \rangle_k = (2\pi B_{\nu s})^{-\frac{3}{2}} \int 4\pi r^{2k+2} \exp(-r^2/2B_{\nu s}) F_{xn}(r) dr =$$
$$= 4(2\pi B_{\nu s})^{-\frac{3}{2}} (B_{\nu s})^{\frac{2k+n+3}{2}} (2k+n+2)! e^{\frac{1}{4}B_{\nu s}x^2} D_{-(2k+n+3)}\left(x\sqrt{B_{\nu s}}\right). \tag{24}$$

In the last expression $D_n(x)$ denotes the standard function of the parabolic cylinder [5], while the quantity β_j' which appears in (23) is defined as

$$\beta_j' = 2\lambda_j - ik(1 - \cos\Theta) = 2\left(\lambda_j - ik\sin^2\left(\frac{\theta}{2}\right)\right), \tag{25}$$

where Θ is the scattering angle of the electrons and $\beta_j = \lambda_j - ik$.

The final result for the total scattering potential produced by thin films for the spin polarized electrons takes the form

$$V_T(\mathbf{r}, s_0) = \sum_{\nu \mathbf{j}_\nu s} e^{\frac{-|\mathbf{r}-\mathbf{R}_{\nu \mathbf{j}_\nu s}|^2}{2B_{\nu s}}} \sum_{k=0}^{\infty} \frac{(1/b_{\nu s})^2}{(2k+1)!} \quad \tilde{V}_T^{(k)}(B_{\nu s}, s_0) \tag{26}$$

and $\tilde{V}_T^{(k)}(B_{\nu s}, s_0)$ is given by expressions (23,24) and (25).

3. Conclusions

The formulae obtained for the scattering potential have interesting properties. At first this potential is finite at every lattice point $\mathbf{R}_{\nu \mathbf{j}_\nu s}$ contrary to the effective potential of the free atom which is infinite at the middle of the atom. Another thing we point out is that one can obtain from it the forms used in literature [3,7]. Namely, if we ignore the exchange part of this potential and restrict a consideration to the high temperature limit, which means the high values of $B_{\nu s}$ the formula (26) can be approximated and for simple cubic lattice it takes the form:

$$V_T(r) \cong \sum_{\nu \mathbf{j}_\nu} \quad e^{-\frac{|\mathbf{r}-\mathbf{R}_{\nu \mathbf{j}}|^2}{B_\nu}} \quad V_T^{(0)}(B_\nu) \tag{27}$$

where $V_T^{(0)}(B_\nu)$ is given by

$$V_T^{(0)}(B_\nu) = Ze\langle F_{0,-1} \rangle_0 - 4\pi \sum_j eA_j^2 \langle n_{\nu s}^j \rangle \frac{(2\mu_j+1)!}{(2\lambda_j)^{2\mu_j+2}} \left[\frac{2\mu_j+2}{2\lambda_j}\langle F_{0.-1}\rangle_0 - \right.$$
$$\left. - (2\mu_j+2) \sum_{m=0}^{2\mu_j+2} \frac{(2\lambda_j)^{m-1}}{m!} \langle F_{2\lambda_j, m-1} \rangle_0 + \sum_{m=0}^{2\mu_j+1} \frac{(2\lambda_j)^m}{m!} \langle F_{2\lambda_j, m} \rangle_0 \right] =$$

$$=4\pi e \sum_j A_j^2 \langle n_\nu^j \rangle \frac{(2\mu_j+1)!}{(2\lambda_j)^{2\mu_j+2}} *$$

$$* \left[(2\mu_j+2) \sum_{m=0}^{2\mu_j+2} \frac{(2\lambda_j)^{m-1}}{m!} \langle F_{2\lambda_j,m-1} \rangle_0 + \sum_{m=0}^{2\mu_j+1} \frac{(2\lambda_j)^m}{m!} \langle F_{2\lambda_j,m} \rangle_0 \right]$$

and

$$\langle F_{2\lambda_j,n} \rangle_0 = 4\pi (2\pi B_\nu)^{-\frac{3}{2}} (B)^{\frac{n+3}{2}} (n+2)! e^{B_\nu \lambda_j^2} D_{-(n+3)} \left(2\lambda_j \sqrt{B_\nu} \right). \quad (28)$$

The last two formulae in the hydrogen atom case, for which the wave function has the simplest form lead to the result of Dworiankin [3]. Instead of the Slater type of function by means of which the electrostatic potential of free atoms is calculated one can use the electrostatic potential of a free atom in the Strand and Tietz form [6]:

$$V(r) = \frac{Ze}{r} \left[\frac{Zp(r)}{Z} \right] = \frac{Ze}{r} \sum_\lambda a_\lambda e^{-b_\lambda r}, \quad (29)$$

where a_λ and b_λ are numerical parameters given in [6]. In this case the formulae (27–29) give the results we have obtained in the papers [7, 8].

Let us remark that the surface scattering potential given by the formulae (26, 23–25) due to its general and analytical form can be very useful for the description of the LEED as well as of the HEED experiments. Due to including the exchange part of the scattering potential it can be used to search the magnetic sample surfaces by means of the spin polarized low energy electron diffraction, which introduces a new dimension to surface physics. The last problem is presented in our recent paper [9].

References

1. J.C. Slater, The Self-consistent Field for Molecules and Solids, McGraw Hill Books Co., New York, Vol. 4, 1974.
2. L. Valenta and L. Wojtczak, Czech. J. Phys., *B30*, 1980.
3. F. Dworiankin, Kristalografia, *10*, 242, 1965.
4. E. Clementi, Tables of Atomic Functions, IMBJ. Res. and Develop. Vol. 9, Suppl. pp. 2-70, 1965.
5. I.S. Gradshteyn and I.M. Rizhik, Tables of Integrals Series and Products, Academic Press, New York, 1965.
6. T.G. Strand and T. Tietz, Nuovo Cimento, *41*B, 89, 1966.
7. K. Stachulec, Acta Phys. Hung., *54*, 267, 1983.
8. K. Stachulec, Acta Phys. Hung., *57*, 55, 1985.
9. K. Stachulec, Physica B (in print), *B 472*, 1986.

Acta Physica Hungarica 64 (4), pp. 395-403 (1988)

APPLICATION OF COMMERCIAL FIELD-EFFECT TRANSISTORS IN Si(Li) SPECTROMETERS

J. PÁLVÖLGYI

Institute for Nuclear Research of the Hungarian Academy of Sciences
H-4001 Debrecen, Hungary

(Received 10 November 1986)

The first stage of the charge sensitive preamplifier of the high energy resolution Si(Li) X-ray spectrometers is a cooled junction Field-Effect Transistor (FET) exhibiting low level of noise. The spectrometer manufacturers use selected FET chips mounted into low dielectic loss material. Noise measurements have been performed for selection of low cost commercial FETs of type 2N4416 from different series of Texas Instruments factories. The FETs were encapsulated and mounted into a teflon block. The measurements have been carried out in a high impedance gate mode in time domain. The noise measurement arrangement, the results of selection of FETs and some observations on generation-recombination (GR) noise of FETs are described.

Introduction

For the first stage of the Si(Li) spectrometer application the 2N4416 FET chips have been found most suitable [1], [2]. In order to eliminate the noise of borosilicate glass header spectrometer manufacturers mount the preselected chips into low dielectric loss material into boron nitride or teflon. For the first stage of static or pulsed drain feedback charge sensitive preamplifier commercial FETs are used in our laboratory encapsulated and mounted into a teflon block. The feedback capacitor is also mounted in the same block [3]. The selection of FETs can be carried out in time domain or frequency domain. Detailed analytical and experimental discussion, further their advantages and disadvantages are given by Llacer [4]. The noise of the FETs is influenced by several factors: the geometry of the chip, the number and activitation energy of impurities in gate–channel junction and the quality of the surface of the chip, the material of the capsulation. The noise of the FET strongly depends also on the device working parameters: on drain current, drain voltage, chip temperature and input capaticance. Our measurements have been carried out in high impedance gate mode, as the first stage of a static drain-feedback preamplifier. The working parameters of the FET (the drain current, drain voltage, input capacitance) were chosen close to the working condition in the Si(Li) spectrometer. In order to eliminate the noise of the test Si(Li) detector it was replaced by a cooled teflon capacitor.

The noise measurement configuration

Figure 1 shows the noise measurement configuration. The FET and the charge injector capacitor (C_i) is mounted into a detector cyrostat. The capacitance of C_i is 1.1 pF which value corresponds to the capacitance of a single grooved 20 mm^2 active area 5 mm thick Si(Li) detector at high voltage bias. The noise of the cooled teflon has been found low [5], therefore this arrangement allows to measure the noise of the FET without important additional noise. To keep the microphonics at low level the connecting wires must be short and rigid. The 10^{-5} Pa pressure in the cryostat is maintained by a 1 l/s titanium getter ion pump.

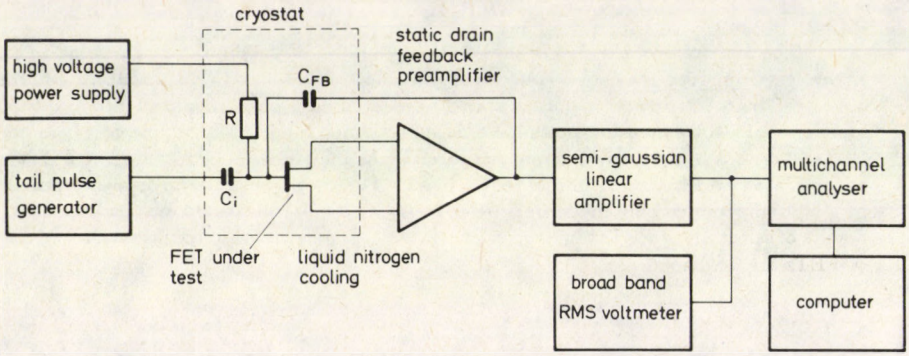


Fig. 1. Schematic diagram of the noise measurement configuration

The temperature of the capacitor and the teflon block without FET power dissipation is 87.5 K. The chip temperature was varied with its selfheating. The drain current was settled with a current generator in the preamplifier and the drain voltage was trimmed with the gate current forced by the high voltage (0–1 kV) connected on the teflon block (R). The noise was filtered with a semi-gaussian linear-amplifier of [9] peaking time constans in the range of 0.46–80 μs. The transfer function of the linear amplifier and its noise indices are [6], [3]:

$$H(p) = \frac{p\left(\frac{p}{\tau} + \frac{9}{4\tau^2}\right)^2}{\left(p + \frac{1}{\tau}\right)\left(p^2 + \frac{9}{4\tau} + \frac{9}{4\tau^2}\right)^2};$$

$$\text{delta noise index: } < N_\Delta^2 > = 1.92/\tau_\rho,$$

$$\text{step noise index: } < N_s^2 > = 0.98\tau_\rho$$

$$f^\alpha, \alpha = \pm 1 \text{ noise index: } < N_{-1_s}^2 > = 6.81.$$

For the calibration of the system long decay time constant (1000 μs) pulses were fed to the capacitor C_i. The amplitude of pulses measured by the multichannel analyser corresponds to the pulses of energy $E = uC_i\bar{\varepsilon}/q$, where u is the amplitude of the pulses, $\bar{\varepsilon}$ the average energy required to produce one hole–electron pair. The calibration and the measurement of the line width was carried out at 11.8 μs peaking time, for the other peaking time constant settlings, the FWHM were calculated from the RMS voltage readings. The FWHM^2 versus peaking time has the general form of:

$$\mathrm{FWHM}^2 = A\tau_p + B/\tau_p + C + D(2E/\tau_p + \tau_p/2E),$$

where $A = 2.355^2\bar{\varepsilon}^2/q < N_s^2 > I_g$, I_g is the gate current, $B = 2.355^2\bar{\varepsilon}^2/q^2 2kTr_s < N_\Delta^2 > C_{in}^2$, r_s is the equivalent resistance in series with the FET gate, which generates noise equivalent to the white noise in the channel, T is the temperature of the chip, C_{in} is the input capacitance,

$$C = 2.355^2\bar{\varepsilon}^2/q^2 A_\alpha C_{in}^2,$$

where $2A_\alpha$ is the value of the power spectrum of the $1/f$ noise at 1 Hz,

$$D = 2.355^2\bar{\varepsilon}^2/q^2 K_g < N_{GR}^2 > C_{in}^2,$$

where K_g is a proportionality constant for the generation recombination noise, $< N_{GR}^2 >$ is the GR noise index, $E = \tau_g$ is the GR noise generation time.

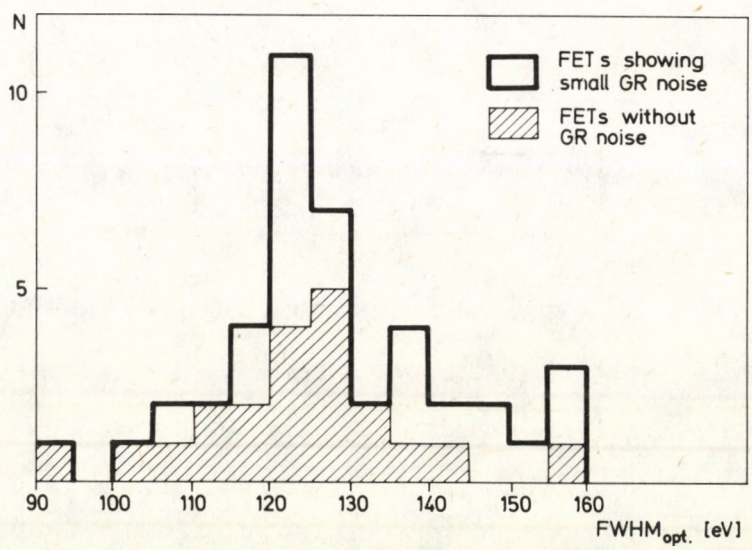


Fig. 2. Distribution of the minimum FWHM of FETs measured at optimum working parameters

140 FETs have been tested. For searching the optimum working parameters of the FETs FWHM versus peaking time curves were measured at different drain currents in the range of 3 mA–9 mA while the drain voltage was kept between 5–7 V. The $A \ldots E$ noise parameters have been calculated using least squares fit.

Results of the measurements

Figure 2 shows the distribution of the minimum FWHM measured at optimum chip dissipation and peaking time. 30% of the FETs exhibited relatively low noise, their minimum FWHM was less than 170 eV, 7% of the FET belonged to the group of FWHM less than 130 eV. 55% of the relatively low noise FETs showed small generation–recombination noise.

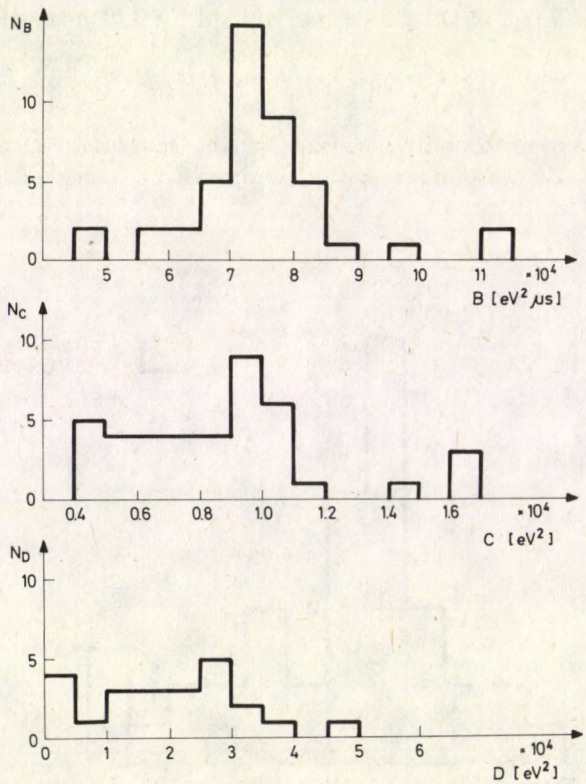


Fig. 3. Distribution of the noise parameters B, C and D of the FETs showing low noise

Figure 3 shows the distribution of the calculated most important noise parameters: B, C and D of the FETs having FWHM less than 170 eV at optimum. In 2 cases the least squares fit gave too low values for $B(< 5000\text{eV}^2)$ because of the fold back of the $\text{FWHM}^2(\tau_p)$ curve at low peaking time constants. The GR noise of FETs is discussed in the next Section.

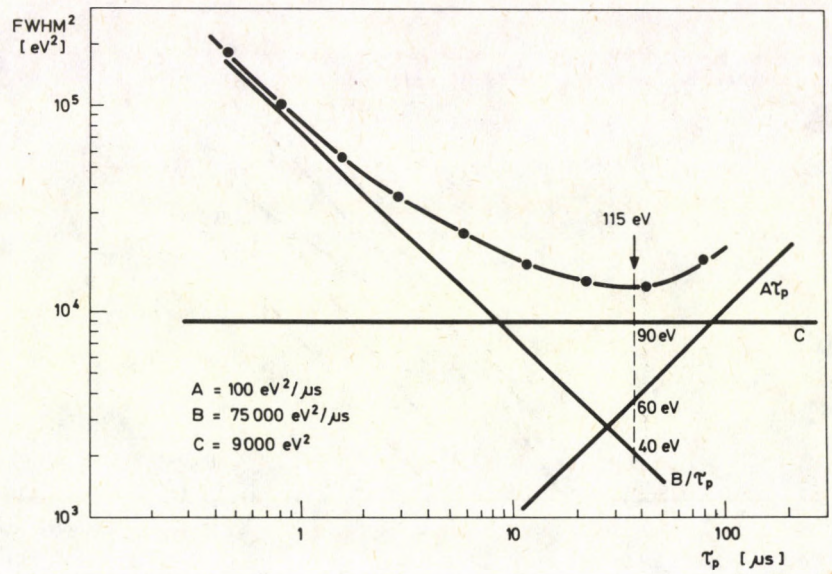


Fig. 4. FWHM^2 versus peaking time and the calculated noise components of a FET without GR noise

Figure 4 shows that the FWHM^2 versus peaking time and the noise components of one of the best FET did not exhibit GR noise. The largest influence on the noise performance of the FET at optimum dissipation and peaking time is made by the f^α noises: 61%, while the contribution of the step noise is 27%, the shot noise is 12%. The values of the noise parameters are: $A = 100\ \text{eV}^2\mu\text{s}$, $B = 7400\ \text{eV}^2\mu\text{s}$, $C = 9000\ \text{eV}^2$.

The generation-recombination noise of the FETs

A great part (84%) of the tested FETs exhibited small or large GR bumps. The typical value for noise parameter D proportional to the number of trap centres in the gate–channel junction is 100 000-200 000 eV^2. In some cases the FWHM versus peaking time curve exhibited a regular figure and it was possible to determine

the activation energy of the GR centres. The measurement of the activation energy (E_{gr}) of the GR centres is based on the variation of the position of the GR bump (parameter E) versus chip temperature: $d(\ln \tau g)/d(1/T) = -E_{gr}/k$, k is the Boltzmann factor [5]. The chip temperature was calculated from the power dissipation of the FET and the thermal resistance of the teflon capsulation and from the zero power temperature of the FET.

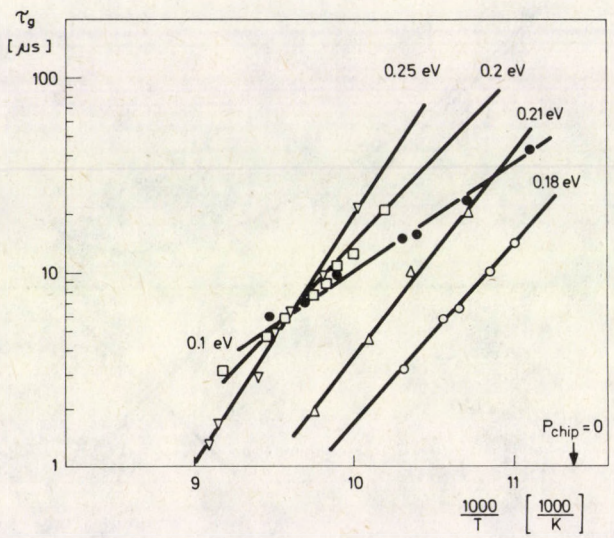


Fig. 5. GR noise generation time versus $1000/T$ of the groups of FETs exhibiting GR noise

Figure 5 shows that five different groups of GR noise were observed. In a series of FETs made before 1980 in the USA, two groups have been found. Most of the FETs exhibiting low noise showed small GR bump. The typical value for parameter D at optimum working parameters is 20000 eV^2. By proper chip temperature the GR bump is shifted to the low peaking time constant, and its contribution at the optimum peaking time is low, about 50 eV. The GR noise of this type is caused by GR centres of activation energy 0.18 eV. Figure 6 shows the FWHM$^2(\tau_p)$ and the noise components of a FET belonging to this group. In this case the value for parameter D is larger than the average, $D = 55000$ eV^2.

FETs from the same series showing poor noise performance (200–500 eV) exhibited frequently large GR bump from GR centres of activation energy 0.1 eV. The typical value for parameter D is 150 000 eV^2 (see Fig. 7). In another series marked E8035, one part of the FETs showed no or very little GR noise, in the other part the activation energies of 0.2 eV and 0.25 eV have been observed. The values

for parameter D were between 20 000 and 50 000 eV^2. Figures 5 and 8 show that in this case the position and the value of parameter D varies strongly with chip temperature. The GR centre levels of 0.18 eV and 0.25 eV were observed earlier by J. Llacer [5], while Hiatt identified activation energy 0.19 eV in specially made transistors [9].

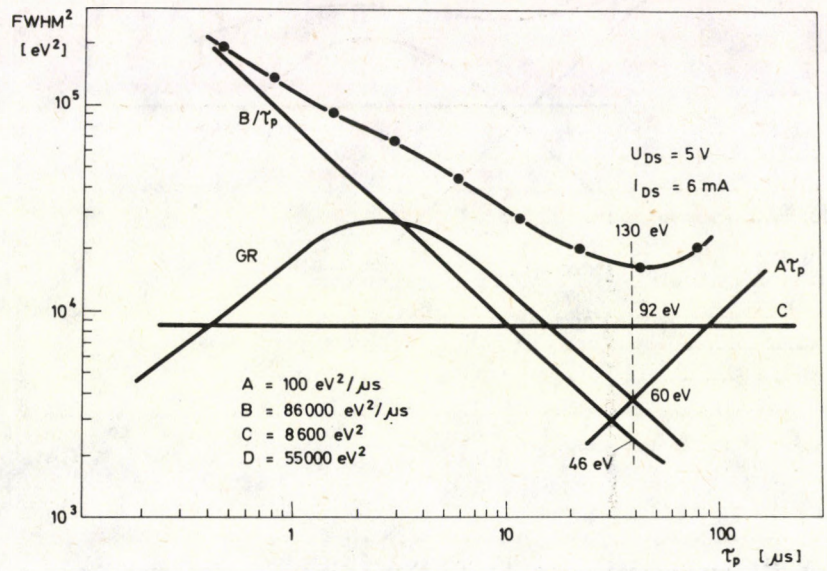


Fig. 6. FWHM^2 versus peaking time of a FET exhibiting GR noise from 0.18 eV activation energy centers

Applicability of the commercial FETs in Si(Li) spectrometers

Si(Li) X-ray spectrometers have been built with detector cryostat [7] and Si(Li) detectors of grooved type, with 20 mm^2 active area and 4–5 mm thickness prepared in our laboratory [3], and with the selected FETs under 135 eV noise level at optimum. The energy resolution of the spectrometers at 5.9 keV measured with NV-809 pulsed drain-feedback charge sensitive preamplifier and NZ-870 time variant filter signal processor [7], [8] with noise indices: $< N_s^2 >= 1.9/\tau_p$, $< N_\Delta^2 >= 0.6\tau_p$, $< N_{-1s}^2 >= 4$ is between 150–180 eV; the electronic noise is 90–130 eV. FETs not exhibiting GR noise with noise parameters $A \approx 100$ eV$^2/\mu s$, $B \approx 75000$ eV$^2\mu s$, $C \approx 10000$ eV^2 have been found most suitable. Some of the FETs belonging to the group of 0.18 eV, 0.21 eV, 0.25 eV activation energy with noise parameter $D \leq 30000$ eV^2 were also used. FETs belonging to the group of 0.1 eV

and 0.2 eV or to the other groups with large value of D are unusable for the first stage of Si(Li) spectrometers.

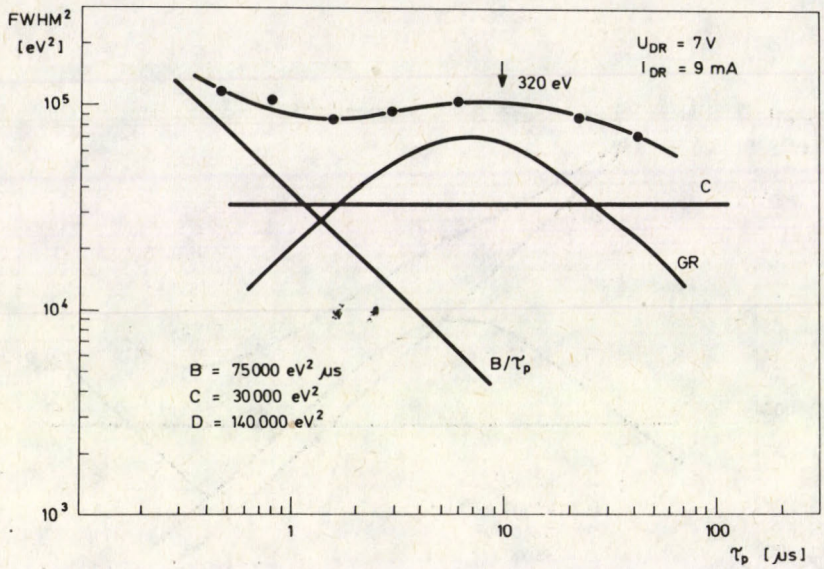


Fig. 7. FWHM² versus peaking time of a FET showing large GR noise from GR centres of 0.1 eV

Conclusions

Time domain noise measurements were performed in high impedance gate mode, without Si(Li) test detector for selection of commercial field-effect transistor chips mounted into teflon block. This arrangement has the advantage that the FET noises were measured close to the working condition of the FET in Si(Li) spectrometer without important additional noise. With the measurement of the FWIIM² versus peaking time at different FET working parameters the optimum working conditions were searched. Less than 15% of the FETs has been found useful for high energy resolution Si(Li) spectrometer application.

Acknowledgements

The author would like to thank G. Kalinka and P. Kovács for their assistance and helpful discussions in the course of the work.

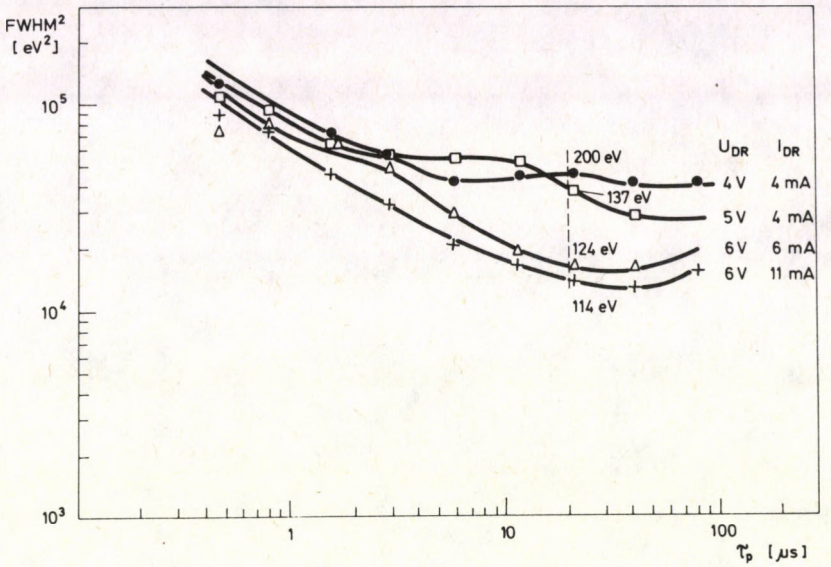


Fig. 8. FWHM² versus peaking time curves measured at different power dissipation of a FET having GR noise 0.2 eV activation energy centers

References

1. V. Radeka, IEEE Trans. Nucl. Sci., NS-*20*, No.1, 182, 1973.
2. F.S. Goulding, Nucl. Instr. and Meth., *142*, 213. 1977.
3. G. Kalinka, Phys. Doctor Thesis, 1980.
4. J. Llacer, IEEE Trans. Nucl. Sci., NS-*22*, No.5, 2033, 1975.
5. J. Llacer and D. Meier, IEEE Trans. Nucl. Sci., NS-*24*, No.1, 317, 1977.
6. J. Gál, ATOMKI Bull., *15*, 277, 1973.
7. J. Bacsó, G. Kalinka, Zs. Kertész, P. Kovács and T. Lakatos, ATOMKI Bull., *24*, 133, 1982.
8. T. Lakatos, Phys. Doctor Thesis, 1982.
9. L.F. Hiatt, A. Van der Ziel and K.M. Van Vliet, IEEE Trans. Elec. Devices, ED-*22*, 614, 1975.

Acta Physica Hungarica 64 (4), pp. 405-416 (1988)

THE FIRST IONIZATION ENERGY, ELECTRON AFFINITY AND ELECTRONEGATIVITY CALCULATED BY THE $X\alpha$ METHOD WITH AB INITIO SELF-CONSISTENT EXCHANGE PARAMETER

R. GÁSPÁR and Á. NAGY

Institute of Theoretical Physics, Kossuth Lajos University
4010 Debrecen, Hungary

(Received 18 December 1986)

Ab initio self-consistent exchange parameters α_{SCF} have been introduced into the $X\alpha$ method. Slater's transition-state method and the definition of electronegativity given by Iczkowski and Margrave have been applied. First ionization energy, electron affinity and electronegativity have been calculated.

Introduction

Important development has been made in the theory of the $X\alpha$ method since it was first advanced by Slater [1]. All the same it is still not precisely known how the exchange parameter α is to be determined. In 1976 Gopinathan, Whitehead and Bogdanovic [2] derived an analytical expression for the α parameter assuming a linear variation of the Fermi-hole density. However, they applied an adjusting parameter, too.

Earlier, in 1974, one of the authors [3] proposed a method to determine ab initio exchange parameters in the $X\alpha$ method. Recently a comparison has been presented [4] for the spin orbitals of a few atoms calculated by the $X\alpha$ method with several values of exchange parameters.

Here, the first ionization energy, electron affinity, electronegativity and chemical potential calculated by the $X\alpha$ method using ab initio self-consistent exchange parameter are presented. Slater's transition-state method has been applied.

The electronegativity determined by the formula of Mulliken [5] and the transition-state method is equal to the one obtained using the definition of electronegativity by Iczkowski and Margrave [6] disregarding the third order terms.

The $X\alpha$ method with self-consistent ab initio exchange parameters

One of the authors has suggested a method [3] which rests on the free-electron gas theory. The exchange potential of an electron with spin up can be written in the form [7]

$$V_{X\uparrow}(1) = -8F(\eta)\left[\frac{3}{4\pi}\rho_{\uparrow}(1)\right]^{1/3}, \tag{1}$$

where

$$F(\eta) = \frac{1}{2} + \frac{1-\eta^2}{4\eta} \ln \left| \frac{1+\eta}{1-\eta} \right| . \tag{2}$$

$\rho_\uparrow$ is the total electron density of the electrons with spin up and

$$\eta = \frac{p}{p_F}, \tag{3}$$

where p is the momentum of the electron considered and p_F is the Fermi momentum. (Energies are in Rydbergs, other quantities are in atomic units.) This exchange potential can be averaged in the momentum space near the Fermi surface for a thin shell containing $\nu_\uparrow$ electrons in the unit volume, where

$$\nu_\uparrow(1) = u_i^*(1) u_i(1) \tag{4}$$

is the density of the electron considered. We then find for the exchange potential the following

$$V_{X\alpha_{\uparrow\text{shell}}} = -6\alpha_{\uparrow\text{shell}} \left[\frac{3}{4\pi} \rho_\uparrow(1) \right]^{1/3}, \tag{5}$$

where

$$\alpha_{\uparrow\text{shell}} = \frac{\rho_\uparrow}{\nu_\uparrow} \left[1 - \frac{1}{2}\eta - \frac{1}{2}\eta^3 + \frac{1}{4}(\eta^2 - 1)^2 \ln \left| \frac{1+\eta}{1-\eta} \right| \right] \tag{6}$$

and

$$\eta = \left(1 - \frac{\nu_\uparrow}{\rho_\uparrow} \right)^{1/3} . \tag{7}$$

As this exchange potential is different for the different spin orbitals, i.e. for different shells, it is averaged over the shells

$$\alpha_\uparrow(r) = \frac{\sum_{j\uparrow} n_j \alpha_{\uparrow\text{shell}}}{\sum_{j\uparrow} n_j} . \tag{8}$$

This exchange parameter $\alpha_\uparrow(r)$ is still a function. To get a constant parameter $\alpha_\uparrow$ the mean squared deviation of the original $X\alpha$ exchange potential

$$V_{X\alpha_\uparrow}(1) = -6\alpha_\uparrow \left[\frac{3}{4\pi} \rho_\uparrow(1) \right]^{1/3} \tag{9}$$

and the exchange potential

$$V_{X\alpha_\uparrow(1)}(1) = -6\alpha_\uparrow(1) \left[\frac{3}{4\pi} \rho_\uparrow(1) \right]^{1/3} \tag{10}$$

with the exchange parameter (8) is minimised. This exchange parameter $\alpha_\uparrow$ can be determined self-consistently. That is why the method is denoted by $X\alpha_{\text{SCF}}$.

Calculation of ionization energy, electron affinity and chemical potential in the $X\alpha$ method

The first ionization energies of atoms have been calculated by the Slater's transition-state method [8]. The ionization energy can be given by the negative of the corresponding orbital energy with an occupation number reduced by half

$$I = -\epsilon_{X\alpha}\left(n - \frac{1}{2}\right). \tag{11}$$

Similarly, the electron affinity is

$$A = -\epsilon_{X\alpha}\left(n + \frac{1}{2}\right). \tag{12}$$

The electronegativity was given by Mulliken [5] with the expression

$$\chi = \frac{A + I}{2}. \tag{13}$$

As the ionization energy of an atom with N electrons is

$$I = \langle EX\alpha(N-1)\rangle - \langle EX\alpha(N)\rangle \tag{14}$$

and the electron affinity is

$$A = \langle EX\alpha(N)\rangle - \langle EX\alpha(N+1)\rangle, \tag{15}$$

for the electronegativity we get

$$\chi = \frac{1}{2}\{\langle EX\alpha(N-1)\rangle - \langle EX\alpha(N+1)\rangle\}. \tag{16}$$

Two cases are treated. First, the electron leaves the same shell (i) as it enters in case of ionization. Expanding around the neutral atomic state (as a transition-state) the total $X\alpha$ energy is given by

$$\langle EX\alpha(N-1)\rangle = \langle EX\alpha(N)\rangle - \frac{\partial\langle EX\alpha\rangle}{\partial n_i}\bigg|_0 + \frac{1}{2}\frac{\partial^2\langle EX\alpha\rangle}{\partial n_i^2}\bigg|_0 - \frac{1}{3!}\frac{\partial^3\langle EX\alpha\rangle}{\partial n_i^3}\bigg|_0 + \ldots \tag{17}$$

and

$$\langle EX\alpha(N+1)\rangle = \langle EX\alpha(N)\rangle + \frac{\partial\langle EX\alpha\rangle}{\partial n_i}\bigg|_0 + \frac{1}{2}\frac{\partial^2\langle EX\alpha\rangle}{\partial n_i^2}\bigg|_0 + \frac{1}{3!}\frac{\partial^3\langle EX\alpha\rangle}{\partial n_i^3}\bigg|_0 + \ldots . \tag{18}$$

For the electronegativity we get

$$\chi = -\frac{\partial\langle EX\alpha\rangle}{\partial n_i}\bigg|_0 - \frac{1}{3!}\frac{\partial^3\langle EX\alpha\rangle}{\partial n_i^3}\bigg|_0 \approx -\epsilon_{iX\alpha} \tag{19}$$

using the well-known

$$\epsilon_{iX\alpha} = \frac{\partial\langle EX\alpha\rangle}{\partial n_i} \tag{20}$$

formula of the $X\alpha$ method.

Secondly, the electron leaves the ith shell forming a positive ion and enters the kth shell forming a negative ion. The occupation numbers of the ith shell are denoted by n_i^{N-1}, and n_i^{N+1} and n_i^0 for the negative and positive ions and for a suitably chosen transition-state, respectively. The same notation is used for the kth shell.

$$\begin{aligned}\langle EX\alpha(N-1)\rangle = \langle EX\alpha\rangle_0 + \frac{\partial\langle EX\alpha\rangle}{\partial n_i}\bigg|_0 (n_i^{N-1} - n_i^0) + \\ + \frac{\partial\langle EX\alpha\rangle}{\partial n_k}\bigg|_0 (n_k^{N-1} - n_k^0) + \frac{1}{2}\frac{\partial\langle EX\alpha\rangle}{\partial n_i^2}\bigg|_0 (n_i^{N-1} - n_i^0)^2 + \\ + \frac{\partial^2\langle EX\alpha\rangle}{\partial n_i \partial n_k}\bigg|_0 (n_i^{N-1} - n_i^0)(n_k^{N-1} - n_k^0) + \frac{1}{2}\frac{\partial^2\langle EX\alpha\rangle}{\partial n_k^2}\bigg|_0 (n_k^{N-1} - n_k^0)^2 + \dots ,\end{aligned} \tag{21}$$

$$\begin{aligned}\langle EX\alpha(N-1)\rangle = \langle EX\alpha\rangle_0 + \frac{\partial\langle EX\alpha\rangle}{\partial n_i}\bigg|_0 (n_i^{N+1} - n_i^0) + \\ + \frac{\partial\langle EX\alpha\rangle}{\partial n_k}\bigg|_0 (n_k^{N+1} - n_k^0) + \frac{1}{2}\frac{\partial\langle EX\alpha\rangle}{\partial n_i^2}\bigg|_0 (n_i^{N+1} - n_i^0)^2 + \\ + \frac{\partial^2\langle EX\alpha\rangle}{\partial n_i \partial n_k}\bigg|_0 (n_i^{N+1} - n_i^0)(n_k^{N+1} - n_k^0) + \frac{1}{2}\frac{\partial^2\langle EX\alpha\rangle}{\partial n_k^2}\bigg|_0 (n_k^{N+1} - n_k^0)^2 + \dots .\end{aligned} \tag{22}$$

The third-order terms are neglected. If the transition-state is given by

$$n_i^0 = (n_i^{N+1} + n_i^{N-1})/2$$

and

$$n_k^0 = (n_k^{N+1} + n_k^{N-1})/2 \tag{23}$$

the second-order terms will vanish in the expression of electronegativity. Taking into consideration that

$$n_i^{N-1} - n_i^{N+1} = n_k^{N-1} - n_k^{N+1} = -1, \tag{24}$$

we arrive at the results

$$\chi \approx -\frac{1}{2}\epsilon_{iX\alpha}^0 - \frac{1}{2}\epsilon_{kX\alpha}^0, \tag{25}$$

where the superscript 0 shows that the one-electron energies are calculated for the transition-state.

Recently the electronegativity χ of an atom has been defined by the formula [7]

$$\chi = -\mu = -\left.\frac{\partial E}{\partial N}\right|_Z, \tag{26}$$

where E is the total energy of the atom, N is the number of electrons and Z is the atomic number. So the electronegativity is the negative of the chemical potential μ. The partial derivative of the $X\alpha$ total energy with respect to the number of electrons is given by

$$\left.\frac{\partial \langle EX\alpha\rangle}{\partial N}\right|_Z = \sum_j \int \left\{ \left.\frac{\partial \langle EX\alpha\rangle}{\partial u_j^*}\right|_Z \left.\frac{\partial u_j^*}{\partial N}\right|_Z + \left.\frac{\partial \langle EX\alpha\rangle}{\partial u_j}\right|_Z \left.\frac{\partial u_j}{\partial N}\right|_Z \right\} dv + \\ + \sum_j \left.\frac{\partial \langle EX\alpha\rangle}{\partial n_j}\right|_Z \frac{\partial n_j}{\partial N}. \tag{27}$$

The summation goes for the spin orbitals with the occupation number n_j. The first term in Eq. (27) containing the partial derivative of the spin orbitals vanishes as the $X\alpha$ one-electron equations come from the variation principle. We have

$$\sum_j \int \left\{ \left.\frac{\partial \langle EX\alpha\rangle}{\partial u_j^*}\right| \left.\frac{\partial u_j^*}{\partial N}\right|_Z + \left.\frac{\partial \langle EX\alpha\rangle}{\partial u_j}\right|_Z \left.\frac{\partial u_j}{\partial N}\right|_Z \right\} dv = \\ = \sum_j \int (n_j \epsilon_j u_j \left.\frac{\partial u_j^*}{\partial N}\right|_Z + n_j \epsilon_j u_j^* \left.\frac{\partial u_j}{\partial N}\right|_Z) dv = \sum_j n_j \epsilon_j \frac{\partial}{\partial N} \int u_j^* u_j \, dv = 0, \tag{28}$$

because the spin orbitals are normalised.

To evaluate the second term of expression (27), two cases are considered. Provided only one spin orbital (i) changes its occupation number when going from the negative ion to the positive one, we have

$$\sum_j \left.\frac{\partial \langle EX\alpha\rangle}{\partial n_j}\right|_Z \frac{\partial n_j}{\partial N} = \epsilon_{iX\alpha}. \tag{29}$$

So the electronegativity equals the negative of the highest occupied orbital energy, i.e.,

$$\chi = -\mu = -\epsilon_{iX\alpha}. \tag{30}$$

In the second case two spin orbitals (i and k) change their occupation numbers during a transition from the negative to the positive ion. These occupation numbers n_i and n_k are not a continuous function of N at the ground state of the atom and hence the formula (29) cannot be applied. As it has been shown the Slater's

transition-state method can be used in this case, too. Nevertheless, it is worth mentioning that the formula (25) obtained by the transition-state method contains derivatives at the transition-state which differ from the derivatives calculated at the ground state of the atom.

It is easy to calculate the difference between the Mulliken's expression of electronegativity and the formula given by Margrave and Iczkowski. The electron affinity and the first ionization energy ($I = I_1$) can be written in the power series of N

$$A = \langle EX\alpha(N)\rangle - \langle EX\alpha(N+1)\rangle = -\frac{\partial\langle EX\alpha\rangle}{\partial N}\bigg|_Z - \\ - \frac{1}{2}\frac{\partial^2\langle EX\alpha\rangle}{\partial N^2}\bigg|_Z - \frac{1}{3!}\frac{\partial^3\langle EX\alpha\rangle}{\partial N^3} + \dots , \tag{31}$$

$$I = I_1 = \langle EX\alpha(N-1)\rangle - \langle EX\alpha(N)\rangle = -\frac{\partial\langle EX\alpha\rangle}{\partial N}\bigg|_Z + \\ + \frac{1}{2}\frac{\partial^2\langle EX\alpha\rangle}{\partial N^2}\bigg|_Z - \frac{1}{3!}\frac{\partial^3\langle EX\alpha\rangle}{\partial N^3}\bigg|_Z + \dots . \tag{32}$$

Similarly, the second ionization energy is given by

$$I_2 = \langle EX\alpha(N-2)\rangle - \langle EX\alpha(N-1)\rangle = -I_1 + \langle EX\alpha(N-2)\rangle - \\ -\langle EX\alpha(N)\rangle = -I_1 - 2\frac{\partial\langle EX\alpha\rangle}{\partial N}\bigg|_Z + 2\frac{\partial^2\langle EX\alpha\rangle}{\partial N^2}\bigg|_Z + \frac{4}{3}\frac{\partial^3\langle EX\alpha\rangle}{\partial N^3}\bigg|_Z + \dots . \tag{33}$$

Combining the formulae (31), (32) and (33) the electronegativity is given by

$$\chi = -\frac{\partial E}{\partial N}\bigg|_Z = \frac{1}{2}(I_1 + A) - \frac{A + I_2 - 2I_1}{6}. \tag{34}$$

Results and discussion

The first ionization energies have beeen determined by the transition-state method using the formula (11). Though, in the case of a few atoms the occupation number of two shells changes during the ionization. (For example the electron configuration for Ni is $4s^2 3d^8$ and for Ni^+ $3d^9$, so the transition-state is $4s^1 3d^{8.5}$.) Denoting the occupation numbers of these two shells in the initial state by n_1^i, n_2^i, in the final state by n_1^f, n_2^f and in the transition-state by n_1^0, n_2^0, the expansion of

the total energy around the transition-state is given by

$$\langle EX\alpha(n_1^i, n_2^i)\rangle = \langle EX\alpha(n_1^0, n_2^0)\rangle + \left.\frac{\partial\langle EX\alpha\rangle}{\partial n_1}\right|_0 (n_1^i - n_1^0) + \left.\frac{\partial\langle EX\alpha\rangle}{\partial n_2}\right|_0 (n_2^i - n_2^0) + \dots , \tag{35}$$

$$\langle EX\alpha(n_1^f, n_2^f)\rangle = \langle EX\alpha(n_1^0, n_2^0)\rangle + \left.\frac{\partial\langle EX\alpha\rangle}{\partial n_1}\right|_0 (n_1^f - n_1^0) + \left.\frac{\partial\langle EX\alpha\rangle}{\partial n_2}\right|_0 (n_2^f - n_2^0) + \dots . \tag{36}$$

Neglecting the third-order terms we arrive at the ionization energy

$$\langle EX\alpha(n_1^f, n_2^f)\rangle - \langle EX\alpha(n_1^i, n_2^i)\rangle = \left.\frac{\partial\langle EX\alpha\rangle}{\partial n_1}\right|_0 (n_1^f - n_1^i) + \left.\frac{\partial\langle EX\alpha\rangle}{\partial n_2}\right|_0 (n_2^f - n_2^i). \tag{37}$$

For example the ionization energy of the Ni is given by

$$I = \epsilon_{3d}^0 - 2\epsilon_{4s}^0, \tag{38}$$

where ϵ_{3d}^0 and ϵ_{4s}^0 are the $3d$ and $4s$ one-electron energies in the transition-state mentioned above. The same expression has been used for V and Co.

Table I contains the ionization energies calculated by the $X\alpha$ method with the exchange parameters $\alpha = 1$, α_{vt} [9] and α_{SCF}. For comparison the experimental [10] and the Hartree–Fock data [11] are also presented. Of course, the Hartree–Fock ionization energies are the closest to the experimental data. The $X\alpha$ results with $\alpha = 1$ are the furthest from the experimental ionization energies. The $X\alpha$ ionization energies gained by α_{vt} and α_{SCF} are often very close together, though sometimes there is considerable difference between them.

Table II contains the electronegativity of atoms calculated by the $X\alpha$ method with the exchange parameters α_{SCF} and α_{vt} [12]. The expressions (19) and (25) are used for this calculation. Generally, there is only small difference between the results gained by the exchange parameter α_{SCF} and α_{vt}.

Table III includes the electron affinity of atoms calculated by the $X\alpha$ method with the exchange parameters α_{SCF} and α_{vt} [12]. The electron affinity may be given by the Mulliken formula (13)

$$A = 2\chi - I. \tag{39}$$

Table I

First ionization energies of atoms calculated by the Hartree–Fock and the $X\alpha$ methods with $\alpha = 1$ [9] and α_{vt} [9] and the self-consistently determined exchange parameters α_{SCF}. Experimental ionization energies are also presented [10] (in Rydbergs)

Z	HF	$X\alpha$			Exp
		$\alpha = 1$	α_{vt}	α_{SCF}	
3	0.392	0.434	0.370	0.318	0.396
4	0.619	0.811	0.675	0.638	0.685
5	0.620	0.750	0.582	0.471	0.610
6	0.814	0.992	0.781	0.809	0.828
7	1.017	1.255	0.984	1.012	1.069
8	1.233	1.522	1.194	1.223	1.000
9	1.460	1.798	1.412	1.442	1.280
10	1.701	2.086	1.635	1.671	1.585
11	0.365	0.430	0.345	0.338	0.378
12	0.500	0.683	0.547	0.547	0.562
13	0.420	0.519	0.384	0.378	0.440
14	0.557	0.699	0.528	0.525	0.599
15	0.701	0.881	0.673	0.747	0.775
16	0.853	1.068	0.824	0.818	0.761
17	1.013	1.260	0.979	0.973	0.956
18	1.182	1.458	1.140	1.280	1.158
19	0.295	0.362	0.286	0.279	0.319
20	0.391	0.540	0.427	0.417	0.449
21	0.420	0.584	0.467	0.310	0.482
22	0.444	0.620	0.499	0.493	0.502
23	0.466	0.406	0.474	0.475	0.495
24	0.415	0.553	0.464	0.463	0.497
25	0.503	0.711	0.572	0.477	0.546
26	0.520	0.738	0.594	0.588	0.514
27	0.537	0.293	0.400	0.403	0.578
28	0.553	0.266	0.383	0.386	0.561
29	0.477	0.660	0.542	0.646	0.567
30	0.585	0.836	0.670	0.664	0.690
31	0.417	0.523	0.377	0.521	0.441
32	0.539	0.681	0.503	0.495	0.579
33	0.660	0.834	0.625	0.616	0.721
34	0.785	0.986	0.747	0.737	0.717
35	0.914	1.140	0.872	0.859	0.870
36	1.048	1.296	0.997	0.984	1.029

Table II
Electronegativity of atoms calculated by the $X\alpha$ method with the exchange parameters α_{SCF} and α_{vt} [12] (in Rydbergs)

Z	α_{SCF}	α_{vt}
3	0.165	0.190
4	0.253	0.279
5	0.264	0.250
6	0.399	0.377
7	0.535	0.512
8	0.680	0.656
9	0.836	0.808
10	0.791	0.758
11	0.164	0.171
12	0.223	0.223
13	0.160	0.165
14	0.262	0.265
15	0.365	0.368
16	0.474	0.479
17	0.590	0.596
18	0.519	0.523
19	0.135	0.141
20	0.128	0.137
21	0.209	0.185
22	0.217	0.224
23	0.299	0.245
24	0.314	0.254
25	0.311	0.318
26	0.338	0.346
27	0.351	0.276
28	0.362	0.284
29	0.373	0.290
30	0.228	0.269
31	0.149	0.155
32	0.241	0.248
33	0.332	0.340
34	0.425	0.434
35	0.521	0.532

Table III

Electron affinity of atoms calculated by the $X\alpha$ method with the exchange parameter α_{SCF} and α_{vt} [12] using the formula (13) and (12). In a few cases the Hartree–Fock [14] and the experimental [13] electron affinities are also presented (in Rydbergs)

Z	$X\alpha$				Exp	HF
	α_{SCF} Eq. (39)	α_{SCF} Eq. (12)	α_{vt} Eq. (39)	α_{vt} Eq. (12)		
3	0.012	0.014	0.009	0.032		0.043
4	-0.132		-0.115			
5	0.057		-0.082			0.022
6	-0.010	0.065	-0.027	0.050	0.092	0.086
7	0.058	0.135	0.040	0.118		-0.020
8	0.137	0.216	0.118	0.198	0.108	0.090
9	0.229	0.309	0.206	0.287	0.253	0.248
10	-0.089		-0.120			
11	-0.010	0.017	-0.004			0.057
12	-0.101		-0.101			
13	-0.057		-0.053			0.036
14	-0.001	0.082	0.001	0.041		0.102
15	-0.017	0.142	0.063	0.101		0.057
16	0.130	0.167	0.134	0.171	0.152	0.156
17	0.208	0.244	0.213	0.248	0.266	0.262
18	-0.242		-0.095			
19	-0.010	0.012	-0.004			
20	-0.161		-0.153			
21	0.108		-0.097			
22	-0.059		-0.051			
23	0.123		0.016			
24	0.150		0.043			
25	0.144	0.097	0.065	0.107		
26	0.088		0.099	0.142		
27	0.094		0.154	0.176		
28	0.339		0.185	0.210		
29	0.101		0.038			
30	-0.209		-0.132			
31	0.066	0.061	-0.067			
32	-0.013		-0.007			
33	0.048	0.083	0.054			
34	0.113		0.121			
35	0.182	0.213	0.193			

Table IV
The electron affinity, the first and second ionization energies and the chemical potential for a few atoms (in Rydbergs)

Z	A Eq. (12)	I_1 Eq. (11)	I_2 Eq. (11)	χ Eq. (30)	$\chi_M = \frac{1}{2}(I_1 + A)$	χ Eq. (34)
6	0.065	0.809	1.817	0.399	0.437	0.393
9	0.309	1.442	2.872	0.836	0.876	0.826
17	0.244	0.973	1.827	0.590	0.609	0.588
35	0.213	0.860	1.608	0.521	0.536	0.519

In a few cases the electron affinity has been determined by using the expression (12), too. For comparison the experimental [13] and Hartree–Fock [14] data are also presented for a few atoms. With a few exceptions there is quite good agreement between the experimental and the $X\alpha$ data. In the majority of the results the electron affinities determined by the exchange parameter α_{SCF} and α_{vt} are close together. There is considerable disagreement between them in a few cases. However, it is not surprising that there is a difference between the electron affinity calculated by expressions (12) and (39). The Eq. (34) presents a more precise expression. To illustrate the correctness of Eq. (34), the electronegativity has been calculated using the expressions (30) and (34) for a few atoms. Table IV presents the electron affinity calculated by Eq. (12), the first and second ionization energies calculated by Eq. (11), the electronegativity using the expression (30), the Mulliken's electronegativity and the electronegativity given by Eq. (34).

References

1. J.C. Slater, Phys. Rev., *81*, 385, 1951.
2. M. S. Gopinathan, M.A. Whitehead and R. Bogdanovic, Phys. Rev., *14*, 1, 1976.
3. R. Gáspár, Acta Phys. Hung., *35*, 213, 1974; Acta Phys. et Chim. Debr., *19*, 7, 1974.
4. R. Gáspár and Á. Nagy, Acta Phys. Hung., *53*, 247, 1982.
5. R.S. Mulliken, J. Chem. Phys., *2*, 782, 1934.
6. R.P. Iczkowski and J.L. Margrave, J. Am. Chem. Soc., *83*, 3547, 1961.
7. P. Gombás, Die statistische Theorie des Atoms und ihre Anwendungen, Springer Verlag, Wien, 1949.
8. J.C. Slater, The Self-Consistent Field for Molecules and Solids, McGraw-Hill, New-York, 1974.
9. K. Schwarz, J. Phys., B*11*, 1339, 1978.
10. C.E. Moore, Atomic Energy Levels NBS Circular No 467 Washington DC: VS Govt Printing Office.
11. J.B. Mann, Atomic Structure Calculations I, Los Alamos Scientific Laboratory Report LA-3690.

12. L.J. Bartolotti, S.R. Gadre and R.G. Parr, J. Am. Chem. Soc., *102*, 2945, 1980.
13. Electronic Structure of Atoms and Molecules, Ed. D. Henderson, Academic Press, New-York, 1969.
14. E. Clementi, A.D. McLean, Phys. Rev., *133*, A 419, 1964.

NOTES TO CONTRIBUTORS

I. PAPERS will be considered for publication in Acta Physica Hungarica only if they have not previously been published or submitted for publication elsewhere. They may be written in English, French, German or Russian.

Papers should be submitted to

Prof. I. Kovács, Editor
Department of Atomic Physics, Technical University
1521 Budapest, Budafoki út 8, Hungary

Papers may be either articles with abstracts or short communications. Both should be as concise as possible, articles in general not exceeding 25 typed pages, short communications 8 typed pages.

II. MANUSCRIPTS

1. Papers should be submitted in three copies.
2. The text of papers must be of high stylistic standard, requiring minor corrections only.
3. Manuscripts should be typed in double spacing on good quality paper, with generous margins.
4. The name of the author(s) and of the institutes where the work was carried out should appear on the first page of the manuscript.
5. Particular care should be taken with mathematical expressions. The following should be clearly distinguished, e.g. by underlining in different colours: special founts (italics, script, bold type, Greek, Gothic, etc.); capital and small letters; subscripts and superscripts, e.g. x^2, x_3; small *l* and l; zero and capital O; in expressions written by hand: *e* and *l*, *n* and *u*, *v* and v, etc.

A List of Symbols on a separate sheet should be attached to each paper.

6. References should be numbered serially and listed at the end of the paper in the following form: J. Ise and W. D. Fretter, Phys. Rev., *76*, 933, 1949.

For books, please give the initials and family name of the author(s), title, name of publisher, place and year of publication, e.g.: J. C. Slater, Quantum Theory of Atomic Structures, I. McGraw-Hill Book Company, Inc., New York, 1960.

References should be given in the text in the following forms: Heisenberg [5] or [5].

7. Captions to illustrations should be listed on a separate sheet, not inserted in the text.
8. In papers submitted to Acta Physica all measures should be expressed in SI units.

III. ILLUSTRATIONS AND TABLES

1. Each paper should be accompanied by three sets of illustrations, one of which must be ready for the blockmaker. The other sets attached to the copies of the manuscript may be rough drawings in pencil or photocopies.
2. Illustrations must not be inserted in the text.
3. All illustrations should be identified in blue pencil by the author's name, abbreviated title of the paper and figure number.
4. Tables should be typed on separate pages and have captions describing their content. Clear wording of column heads is advisable. Tables should be numbered in Roman numerals (I, II, III, etc.).

IV. RETURN OF MATERIAL

Owing to high postage costs, the Editorial Office cannot undertake to return *all* material not accepted for any reason for publication. Of papers to be revised (for not being in conformity with the above Notes or other reasons) only *one* copy will be returned. Material rejected for lack of space or on account of the Referees' opinion will not be returned to authors outside Europe.

Periodicals of the Hungarian Academy of Sciences are obtainable at the following addresses:

AUSTRALIA
C.B.D. LIBRARY AND SUBSCRIPTION SERVICE
Box 4886, G.P.O., *Sydney N.S.W. 2001*
COSMOS BOOKSHOP, 145 Ackland Street
St. Kilda (Melbourne), Victoria 3182

AUSTRIA
GLOBUS, Höchstädtplatz 3, *1206 Wien XX*

BELGIUM
OFFICE INTERNATIONAL DES PERIODIQUES
Avenue Louise, 485, *1050 Bruxelles*
E. STORY-SCIENTIA P.V.B.A.
P. van Duyseplein 8, *9000 Gent*

BULGARIA
HEMUS, Bulvar Ruszki 6, *Sofia*

CANADA
PANNONIA BOOKS, P.O. Box 1017
Postal Station "B", *Toronto, Ont. M5T 2T8*

CHINA
CNPICOR, Periodical Department, P.O. Box 50
Peking

CZECHOSLOVAKIA
MAD'ARSKA KULTURA, Národni třida 22
115 66 Praha
PNS DOVOZ TISKU, Vinohradská 46, *Praha 2*
PNS DOVOZ TLAČE, *Bratislava 2*

DENMARK
EJNAR MUNKSGAARD, 35, Nørre Søgade
1370 Copenhagen K

FEDERAL REPUBLIC OF GERMANY
KUNST UND WISSEN ERICH BIEBER
Postfach 46, *7000 Stuttgart 1*

FINLAND
AKATEEMINEN KIRJAKAUPPA, P.O. Box 128
00101 Helsinki 10

FRANCE
DAWSON-FRANCE S.A., B.P. 40, *91121 Palaiseau*
OFFICE INTERNATIONAL DE DOCUMENTATION ET LIBRAIRIE, 48 rue Gay-Lussac
75240 Paris, Cedex 05

GERMAN DEMOCRATIC REPUBLIC
HAUS DER UNGARISCHEN KULTUR
Karl Liebknecht-Straße 9, *DDR-102 Berlin*

GREAT BRITAIN
BLACKWELL'S PERIODICALS DIVISION
Hythe Bridge Street, *Oxford OX1 2ET*
BUMPUS, HALDANE AND MAXWELL LTD.
Cowper Works, *Olney, Bucks MK46 4BN*
COLLET'S HOLDINGS LTD., Denington Estate,
Wellingborough, Northants NN8 2QT
WM DAWSON AND SONS LTD., Cannon House
Folkstone, Kent CT19 5EE
H. K. LEWIS AND CO., 136 Gower Street
London WC1E 6BS

GREECE
KOSTARAKIS BROTHERS INTERNATIONAL BOOKSELLERS, 2 Hippokratous Street, *Athens-143*

HOLLAND
FAXON EUROPE, P.O. Box 167
1000 AD Amsterdam
MARTINUS NIJHOFF B. V.
Lange Voorhout 9–11, *Den Haag*
SWETS SUBSCRIPTION SERVICE
P.O. Box 830, 2160 Sz Lisse

INDIA
ALLIED PUBLISHING PVT. LTD.
750 Mount Road, *Madras 600002*
CENTRAL NEWS AGENCY PVT. LTD.
Connaught Circus, *New Delhi 110001*
INTERNATIONAL BOOK HOUSE PVT. LTD.
Madame Cama Road, *Bombay 400039*

ITALY
D. E. A., Via Lima 28, *00198 Roma*
INTERSCIENTIA, Via Mazzé 28, *10149 Torino*
LIBRERIA COMMISSIONARIA SANSONI
Via Lamarmora 45, *50121 Firenze*
SANTO VANASIA, Via M. Macchi 58
20124 Milano

JAPAN
KINOKUNIYA COMPANY LTD.
Journal Department, P.O. Box 55
Chitose, *Tokyo 156*
MARUZEN COMPANY LTD., Book Department
P.O. Box 5050 Tokyo International, *Tokyo 100-31*
NAUKA LTD., Import Department
2-30-19 Minami Ikebukuro, Toshima-ku, *Tokyo 171*

KOREA
CHULPANMUL, *Phenjan*

NORWAY
TANUM-TIDSKRIFT-SENTRALEN A.S.
Karl Johansgata 43, *1000 Oslo*

POLAND
WĘGIERSKI INSTYTUT KULTURY
Marszalkowska 80, *00-517 Warszawa*
CKP I W, ul. Towarowa 28, *00-958 Warszawa*

ROUMANIA
D. E. P., *Bucuresti*
ILEXIM, Calea Grivitei 64–66, *Bucuresti*

SOVIET UNION
SOYUZPECHAT — IMPORT, *Moscow*
and the post offices in each town
MEZHDUNARODNAYA KNIGA, *Moscow G-200*

SPAIN
DIAZ DE SANTOS Lagasca 95, *Madrid 6*

SWEDEN
ESSELTE TIDSKRIFTSCENTRALEN
Box 62, *101 20 Stockholm*

SWITZERLAND
KARGER LIBRI AG, Petersgraben 31, *4011 Basel*

USA
EBSCO SUBSCRIPTION SERVICES
P.O. Box 1943, *Birmingham, Alabama 35201*
F. W. FAXON COMPANY, INC.
15 Southwest Park, *Westwood Mass. 02090*
MAJOR SCIENTIFIC SUBSCRIPTIONS
1851 Diplomat, P.O. Box 819074,
Pallas, Tx. 75381-9074
READ-MORE PUBLICATIONS, INC.
140 Cedar Street, *New York, N. Y. 10006*

YUGOSLAVIA
JUGOSLOVENSKA KNJIGA, Terazije 27, *Beograd*
FORUM, Vojvode Mišića 1, *21000 Novi Sad*

Index: 26.022