*j*ornades
científiques

IEC

PHYSICS AND GEOMETRY

SECCIÓ DE CIÈNCIES I TECNOLOGIA

sèrie jornades científiques 2

jornades científiques
IEC

2
sèrie jornades científiques

Jornades
Científiques de
l'Institut d'Estudis
Catalans
SECCIÓ DE CIÈNCIES I TECNOLOGIA

Physics and Geometry

Editors

Sebastià Xambó
Universitat Politècnica de Catalunya
Societat Catalana de Matemàtiques

David Jou
Universitat Autònoma de Barcelona
Institut d'Estudis Catalans

Biblioteca de Catalunya. Dades CIP

Institut d'Estudis Catalans. Jornades Científiques *(2es : 1996)*
Physics and geometry. — (Sèrie jornades científiques ; 2)
Textos en anglès i francès, introducció en anglès i català
ISBN 84-7283-441-7
I. Xambó, Sebastià, ed. II. Jou i Mirabent, David, ed. III. Títol
IV. Col·lecció: Sèrie jornades científiques ; 2
1. Geometria — Congressos 2. Física matemàtica — Congressos
53:514(061.3)

Aquest treball ha comptat amb el suport de la Comissió Interdepartamental
de Recerca i Innovació Tecnològica (CIRIT) de la Generalitat de Catalunya

Disseny: Maria Casassas

Carrer del Carme, 47. 08001 Barcelona

Primera edició: març de 1999
Tiratge: 400 exemplars

Imprès a Limpergraf, SL
Polígon industrial Can Salvatella. Carrer de Mogoda, 29-31. 08210 Barberà del Vallès

ISBN: 84-7283-441-7
Dipòsit Legal: B. 6414-1999

Índex

Presentation

Every year, the Institut d'Estudis Catalans (an academic, scientific and cultural corporation founded in 1907, whose aim is to conduct scientific research, primarily into on all the elements of Catalan culture) organizes a number of interdisciplinary colloquia on different aspects of modern culture, in a broad perspective embracing both the sciences and the humanities. These colloquia aim to bring together several eminent figures in a particular area of culture to discuss the recent advances and main open problems in their field, in a rigorous but not highly specialized analysis. When we were invited by the President of the Institut, Prof. Manuel Castellet, to organize an interdisciplinary colloquium on physics and mathematics, we chose as a topic physics and geometry. Indeed, it was our impression that the relationship between both fields has been extremely lively and fruitful in latter years, and we wanted to stimulate the physicists and mathematicians of our country to participate in, and share some of the current excitement about the most dynamic interactions between these two fields.

The relationship between physics and geometry has a long history. We may recall, to mention only a few historical aspects, the use of conic sections in the physics of the Renaissance (namely, the ellipse for planetary orbits and the parabola for the trajectory of projectiles); the connection between Newton's *Principia* and the use of vectors and of differential methods; Einstein's general relativity, which introduced non-Euclidean geometries into physics, or the analysis of elementary particles by means of symmetry principles, as expressed by means of group theory or of principal vector bundles. These few examples suffice to stress that physics has been an inexhaustible source of problems and ideas for mathematics and, at the same time, that mathematics has often gone ahead in supplying frameworks for the formulation of physical theories. The relationship between physics and geometry has experienced, in the last decade, an especially brilliant period: topology and quantum field theory, fractal geometry and deterministic chaos, the geometric aspects of gravitation, both at the microscopic quantum level and on the cosmological scale, non-commutative geometry, the discovery and applications of materials with new kinds of symmetries, are some of the most fruitful aspects of this connection. Indeed, quantum field theory has proved to be very useful for obtaining new results in algebraic and differential topology which, in turn, are being helpful for the advance of quantum field theory. The lectures by Connes and Labastida deal with very different aspects (non-commutative geometry and topology, respectively) of this relationship. Since Einstein's times, gravitation has been closely linked to geometry. Now, some of the most compelling open problems in this connection arise at two extreme length scales: the microscopic quantum scale set by Planck's length and the cosmological scale. Ashtekar and Ellis deal respectively with these two extreme situations. There are also geometrical surprises in materials sciences: indeed, materials with new and surprising symmetries were discovered several years ago and they now offer interesting applications. In this direction, Janot's lecture provides an introduction to the geometry of quasicrystals. The physical and geometrical analysis of complex systems has been one of the areas where the greatest progress has been achieved in the last two decades: fractal geometry and deterministic chaos have been the basis of a revolution in our way of analyzing and interpreting the complexity of the world. Pietronero and Mandelbrot discuss the relations between fractal geometry and physics, and some open problems in the formulation of fractal geometry respectively.

After a careful examination of this far-reaching (but by no means exhaustive) relationship between physics and geometry, it still appears to many to be more mysterious than at first glance. In fact, along with Einstein's motto that "the most incomprehensible aspect of physics is that it is comprehensible", and Wigner's famous "unreasonable effectiveness of mathematics in physics", a no lesser "unreasonable effectiveness of physics in mathematics", and particularly in geometry and topology, has appeared in the last years. As phrased by Atiyah, "a somewhat surprising feature of the new developments is that quantum field theory seems to tie up with deep properties of low-dimensional geometry".

Finally, we give a brief outline of the lectures presented in this book. A. Ashtekar discusses the quantum mechanics of geometry. In this perspective, geometry is no longer inert, but has physical degrees of freedom of its own, very much like matter. One may thus discuss what these "atoms of geometry" are, at what scale they reveal themselves, or how they manage to cluster to form a smooth continuum at the laboratory scale. These issues are analyzed by combining the principles of general relativity and of quantum mechanics, at the Planck scale, where areas of surfaces and volumes of regions are quantized.

A. Connes, after introducing an operator-based infinitesimal calculus, develops a new notion of geometrical space. This notion abandons, among other things, the central role played by points in the ordinary Riemannian geometry and the commutativity of the product of two "functions", but affords a much greater freedom in the description of space-times at the Planck length scale. In fact, these ideas are used to show that space-time and the standard model fit nicely into one of these new geometries.

G. F. Ellis shows how phase plane techniques give illuminating information on how Robertson-Walker or Bianchi universe models evolve in time, and how attractors and unstable equilibrium points help to conjecture which are the most probable configurations of those models. Studies of the consistency of covariantly defined models give some ideas of the nature of evolution of more general inhomogeneous and anisotropic models.

C. Janot offers an introduction to the geometry of quasicrystals, a new form of the solid state which differs from the two other known forms (crystalline and amorphous) by possessing a new type of long-range translational order, quasiperiodicity, and a noncrystallographic orientational order. Beyond unconventional structures, quasicrystals exhibit very surprising physical properties which could be remarkably useful for technological applications.

J. M. F. Labastida presents a pespective on topological quantum field theory. Indeed, QFT has played a fundamental role in our understanding of the behaviour of elementary particles and, as was recently discovered, may also be a very useful tool for studying some aspects of low-dimensional topology, by a combination of some of the perturbative and non-perturbative methods of quantum field theory. From the results, a new picture emerges for some sets of topological invariants (for instance, Seiberg-Witten invariants and Vassiliev invariants) in which these are classified in terms of universality classes.

B. Mandelbrot reflects on diverse aspects of fractals (philosophy, science, mathematics, esthetics) and comments on some unanswered mathematical questions raised by fractal geometry in recent years.

L. Pietronero examines three levels of understanding of the impact of fractal geometry in physics. The first level refers to the phenomenological realization that a given structure manifests self-similar properties which may be characterized with fractal dimension and other exponents. The second level corresponds to the construction of physical models that lead spontaneously to fractal structures via their dynamic evolution and thus yield the basic physical elements for their generation. The third and deepest level is the construction of a physical theory which provides a complete understanding of the phenomenon and permits the analytical calculation of its exponents. Some important advances and open problems at these levels are discussed.

We thank the Institut d'Estudis Catalans for having provided the opportunity to share the excitement and stimulation of these discussions. We are also grateful for the financial support of the Direcció General de Investigación Científica y Técnica of the Spanish Ministry of Education, to the Centre de Recerca Matemàtica, which organized a specialized advanced course on physics and geometry in the summer of 1996, to Mrs. Neus Portet for her helpful and efficient administrative support, and to Mrs. Maria Julià for taking care of the preparation of the final version of the manuscripts.

Barcelona, Autumn 1997

David Jou
Fellow of the Sciences and Technology Section
of the Institut d'Estudis Catalans.
Full professor at the Faculty of Sciences
of the Universitat Autònoma de Barcelona

Sebastià Xambó
President of the Societat Catalana de Matemàtiques
Full professor at the Faculty of Mathematics and Statistics
of the Universitat Politècnica de Catalunya

Presentació

L’Institut d’Estudis Catalans organitza cada any diverses jornades interdisciplinàries sobre diferents aspectes de la cultura moderna, des d’una àmplia perspectiva científica i humanística. L’objectiu d’aquestes Jornades és reunir diversos protagonistes d’alguna de les àrees culturals per analitzar i discutir els progressos recents i els principals problemes oberts en el seu camp, en una anàlisi rigorosa però no altament especialitzada. Quan fórem invitats pel president de l’Institut, Prof. Manuel Castellet, a organitzar un col·loqui interdisciplinari sobre física i matemàtiques, vam triar com a tema “física i geometria”. En efecte, teníem la impressió que les relacions entre ambdues àrees ha estat extremament viva i fèrtil en els darrers anys, i volíem estimular els físics i els matemàtics del nostre país a compartir l’excitació actual sobre les interaccions més vives entre aquests dos camps i que s’engresquessin a participar-hi.

La relació entre la física i la geometria té una llarga història. Podem recordar, per esmentar tan sols unes poques fites històriques, l’ús de les corbes còniques en la física del Renaixement (l’el·lipse per a les òrbites planetàries i la paràbola per a la trajectòria de projectils); la connexió entre els *principia* de Newton i l’ús de vectors i de mètodes de diferenciació i integració; la relativitat general d’Einstein, que introduí en la física les geometries no euclidianes; l’anàlisi de partícules elementals segons principis de simetria, expressats mitjançant la teoria de grups o de fibrats principals. Aquests pocs exemples són suficients per subratllar que la física ha estat una font inexhaurible de problemes i d’idees per als matemàtics i que, alhora, les matemàtiques s’han avançat sovint a proporcionar a la física els instruments necessaris per a la formulació de teories.

La relació entre física i geometria ha tingut, en la darrera dècada, un període especialment brillant: topologia i teoria quàntica de camps, geometria fractal i caos determinista, els aspectes geomètrics de la gravitació, tant a escala microscòpica quàntica com a escala cosmològica, geometria no commutativa, el descobriment i les aplicacions de materials amb noves menes de simetries, són alguns dels aspectes més pròspers d'aquesta relació. En efecte, la teoria quàntica de camps ha resultat ser molt útil per obtenir nous resultats en topologia algebraica i diferencial, que, alhora, estan resultant molt fructífers per al progrés de la teoria quàntica de camps. Les conferències de Connes i de Labastida tracten aspectes molt diferents (geometria no commutativa i topologia, respectivament) d'aquesta relació. Des de l'època d'Einstein, la gravitació es considera íntimament relacionada amb la geometria. Actualment, alguns dels problemes més urgents en aquesta relació sorgeixen en les dues escales més extremes: l'escala microscòpica quàntica determinada per la longitud de Planck i l'escala cosmològica. Ashtekar i Ellis tracten, respectivament, aquestes situacions extremes. També trobem sorpreses geomètriques en les ciències de materials: en efecte, fa pocs anys foren descoberts materials amb simetries noves, i actualment presenten aplicacions molt interessants. La conferència de Janot constitueix una introducció a la geometria de quasicristalls. L'anàlisi física i geomètrica de sistemes complexos ha estat una de les àrees en què més progrés hi ha hagut en les dues darreres dècades: la geometria fractal i el caos determinista han constituït la base d'una revolució en la nostra manera d'analitzar i intepretar la complexitat del món. Pietronero i Mandelbrot discuteixen les relacions entre la geometria fractal i la física, i alguns dels problemes oberts en la formulació de la geometria fractal.

Després d'aquest examen detallat, ampli però exhaustiu, de les relacions entre física i geometria, aquesta relació encara sembla més misteriosa del que semblava a primera vista. De fet, amb les conegudes frases d'Einstein, segons el qual "l'aspecte més incomprensible de la física és que sigui comprensible", o de Wigner sobre "l'efectivitat, enllà del que és raonable, de les matemàtiques en la física", en els darrers anys ha aparegut una no menys irraonable "efectivitat de la física en les matemàtiques", i en particular en geometria i topologia. Tal com digué Atiyah, "una característica força sorprenent dels nous desenvolupaments és que la teoria quàntica de camps sembla estar vinculada a propietats molt profundes de la geometria en poques dimensions".

Agraïm a l'Institut d'Estudis Catalans que ens hagi proporcionat l'oportunitat de compartir l'excitació i l'estímul d'aquestes discussions. També agraïm a la Direcció General d'Investigació Científica i Tècnica del Ministeri d'Educació i Cultura l'ajut econòmic que ens concedí, al Centre de Recerca Matemàtica l'organització d'un curs especialitzat avançat sobre física i geometria durant l'estiu de 1996, a la senyora Neus Portet el seu suport administratiu amable i eficient i a la senyora Maria Julià la gran cura que ha tingut en l'edició final del llibre.

Barcelona, tardor 1997

David Jou
Membre de la Secció de Ciències i Tecnologia
de l'Institut d'Estudis Catalans
Catedràtic de la Facultat de Ciències
de la Universitat Autònoma de Barcelona

Sebastià Xambó
President de la Societat Catalana de Matemàtiques
Catedràtic de la Facultat de Matemàtiques i Estadística
de la Universitat Politècnica de Catalunya

Curriculum Vitae of the Contributors

Abhay Ashtekar is Eberly Professor of Physics and Director of the Center for Gravitational Physics and Geometry at Penn State University. He graduated in Physics at the University of Bombay in 1969, and received his Ph. D. at the University of Chicago in 1974. He was professor of Physics at the Syracuse University and at the Université de Paris VI. He is especially well-known for his work on quantum gravity.

Alain Connes is Professor of Analysis and Geometry at the Collège of France, and at the Institut des Hautes Études Scientifiques. Fields Medal for his work on non-commutative-geometry. He has also published the essay book *Matière à pensée*, together with the neurophysiologist Jean-Pierre Changeux (translated to Spanish in the collection Metatemas).

George F. R. Ellis is Full Professor of the Mathematics Department at the University of Cape Town. He is well-known for his wide-ranging research on different topics of cosmology. Amongst his many works on this subject stands the book *The large-scale structure of the Universe*, written in collaboration with S. Hawking.

Christian Janot is Professor at the University Joseph Fourier Grenoble, and is based at the Institut Laue-Langevin for his research work. He trained in metal physics and his main interests have been in non-crystalline structures (geometry and properties). He is author of the book *Quasicrystals*, a wide introduction to this topic. He is presently involved in defining the materials research programme for the CNRS.

José M. F. Labastida is currently Professor of Theoretical Physics in the University of Santiago de Compostela where he has been since 1991. He received his PhD. from the State University of New York in Stony Brook in 1985. From 1985 to 1988 he pursued postdoctoral study at the Institute for Advanced Study at Princeton and from 1988 to 1990 he joined CERN as a Fellow. He has held a research position at the Consejo Superior de Investigaciones Cientificas (CSIC) of Spain from 1986 to 1991. During the last few years his research has focused on the study of topological quantum field theory and its applications to low dimensional topology.

Benoît Mandelbrot is IBM Fellow at the J. J. Watson Research Center, at Yorktown Heights (New York). He has an extensive research work but he is especially known as the creator of the Fractal Geometry. His main works have been reflected in the book *The Fractal Geometry of Nature*. Prof. Mandelbrot is a member of the American Academy of Arts and Sciences and has received many awards. In Spanish one can find the books *Los objetos fractales* and *La geometria fractal de la naturaleza* (both in the collection Metatemas).

Luciano Pietronero is Professor of solid state theory at the University of Roma "La Sapienza". He worked as researcher at the scientific laboratories of Xerox (Rochester, NY) and of Brown Boveri (Baden, CH), and was professor of condensed matter theory at the University of Groningen. His research activity is focused on the use of fractal concepts in condensed matter theory and statistical physics, as well as in the analysis of cosmological data.

Quantum Mechanics of Riemannian Geometry

Abhay Ashtekar
Center for Gravitational Physics and Geometry
University of Pennsylvania, University Park, USA

Abstract

Over the past four years, a detailed framework has been constructed to unravel the quantum nature of the Riemannian geometry of physical space. A review of these developments is presented at a level which should be accessible to advanced undergraduate students in physics. As an illustrative application, I indicate how this micro-structure of geometry can have a direct impact on physical processes such as the evaporation of black holes through the Hawking process.

1 Introduction

During his Göttingen inaugural address in 1854, Riemann [1] suggested that geometry of space may be more than just a fiducial, mathematical entity serving as a passive stage for physical phenomena, and may in fact have direct physical meaning in its own right. General relativity provided a brilliant confirmation of this vision: curvature of space now encodes the physical gravitational field. This shift is profound. To bring out the contrast, let me recall the situation in Newtonian physics. There, space forms an inert arena on which the dynamics of physical systems —such as the solar system— unfolds. It is like a stage, an unchanging backdrop for all of physics. In general relativity, by contrast, the situation is very different. Einstein's equations tell us that matter curves space. Geometry is no longer immune to change. It reacts to matter. It is dynamical. It has "physical degrees of freedom" in its own right. Thus, in general relativity, the stage disappears and joins the troupe of actors. Geometry is a physical entity, very much like matter.

Now, the physics of this century has shown us that matter has constituents and the 3-dimensional objects we perceive as solids are in fact made of atoms. The continuum description of matter is an approximation which succeeds brilliantly in the macroscopic regime but fails hopelessly at the atomic scale. It is therefore natural to ask: Is the same true of geometry? If so, what is the analog of the 'atomic scale?' We know that a quantum theory of geometry will feature three fundamental constants of Nature, $c, G, \hbar$, the speed of light, Newton's gravitational constant and Planck's constant. Now, as Planck pointed out in his celebrated paper that marks the beginning of quantum mechanics, there is a unique combination, $\ell_P = \sqrt{\hbar G/c^3}$, of these constants which has dimension of length. ($\ell_P \approx 10^{-33}$cm.) It is now called the Planck length. Experience has taught us that the presence of a distinguished scale in a physical theory marks a potential transition; physics below the scale can be very different from that above the scale. Now, all of our well-tested physics occurs at length scales much bigger than than ℓ_P. In this regime, the continuum picture works well. A key question then is: Will it break down at the Planck length? Does geometry have constituents at this scale? If so, what are its atoms? Its elementary excitations? Is the space-time continuum only a 'coarse-grained' approximation? Is geometry quantized? If so, what is the nature of its quanta?

To probe such issues, it is natural to look for hints in the procedures that have been successful in describing matter. Let us begin by asking what we mean by quantization of physical quantities. Take a simple example —the hydrogen atom. In this case, the answer is clear: while the basic observables —energy and angular momentum— take on a continuous range of values classically, in quantum mechanics their eigenvalues are discrete; they are quantized. So, we can ask if the same is true of geometry. Classical geometrical quantities such as lengths, areas and volumes can take on continuous values on the phase space of general relativity. Are the eigenvalues of corresponding quantum operators discrete? If so, we would say that geometry is quantized and the precise eigenvalues and eigenvectors of geometric operators would reveal its detailed microscopic properties.

Thus, it is rather easy to pose the basic questions in a precise fashion. Indeed, they could have been formulated soon after the advent of quantum mechanics. Answering them, on the other hand, has proved to be surprisingly difficult. The main reason, I believe, is the inadequacy of the standard techniques. More precisely, to examine the microscopic structure of geometry, we must treat Einsteinian gravity quantum mechanically, i.e., construct at least the basics of a quantum theory of the gravitational field. Now, in the traditional approaches to quantum field theory, one *begins* with a continuum, background geometry. To probe the nature of quantum geometry, on the other hand, we should *not* begin by assuming the validity of this picture. We must let quantum gravity decide whether this picture is adequate; the theory itself should lead us to the correct microscopic model of geometry.

With this general philosophy, in this article I will summarize the picture of quantum geometry that has emerged from a specific approach to quantum gravity. This approach is non-perturbative. In perturbative approaches, one generally begins by assuming that space-time geometry is flat and incorporates gravity —and hence curvature— step by step by adding up small corrections. In the non-perturbative approach, by contrast, there is no background metric at all. All we have is a bare manifold to start with. All fields —matter as well as gravity/geometry— are treated as dynamical from the beginning. Consequently, the description cannot refer to a background metric. Technically this means that the full diffeomorphism group of the manifold is respected; the theory is generally covariant.

As we will see, this fact leads one to Hilbert spaces of quantum states which are quite different from the familiar Fock spaces of particle physics. Now gravitons —the three-dimensional wavy undulations on a flat metric— do not represent fundamental excitations. Rather, the fundamental excitations are *one*-dimensional. Microscopically, geometry is rather like a polymer. Recall that, although polymers are intrinsically one-dimensional, when densely packed in suitable configurations they approximate a three-dimensional system. Similarly, the familiar continuum picture of geometry arises as an approximation. Indeed, one can regard the fundamental excitations as 'quantum threads' and construct from them 'weave states' which approximate continuum geometries. Gravitons are no longer the basic mediators of the gravitational interaction. They now arise only as approximate notions; they represent perturbations of weave states. Because states are polymer-like, geometrical observables turn out to have discrete spectra. They provide a rather detailed picture of quantum geometry from which physical predictions can be made.

The article is divided into two parts. In the first, I will indicate how one can reformulate general relativity so that it resembles gauge theories. This formulation provides the starting point for the quantum theory. In particular, the one-dimensional excitations of geometry arise as the analogs of 'Wilson loops' which are themselves analogs of the line integrals $\exp i \oint A.d\ell$ of electro-magnetism. In the second part, I will indicate how this description leads us to a quantum theory of geometry. I will focus on area operators and show how the detailed information about the eigenvalues of these operators has interesting physical consequences, e.g., to the process of Hawking evaporation of black holes.

I should emphasize that this is *not* a technical review. Rather, the article is written at the level of colloquia in physics departments in the United States. Thus, I will purposely avoid technicalities and try to make the discussion intuitive. I will also make some historic detours of general interest. At the end, however, I will list some references where the details of the central results can be found.

2 From metrics to connections

2.1 Gravity versus other fundamental forces

General relativity is normally regarded as a dynamical theory of metrics —tensor fields that define distances and hence geometry. It is this fact that enabled Einstein to code the gravitational field in the Riemannian curvature of the metric. Let me amplify with an analogy. Just as position serves as the configuration variable in particle dynamics, the three-dimensional metric of space can be taken to be the configuration variable of general relativity. Given the initial position and velocity of a particle, Newton's laws provide us with a trajectory of particle in the position space. Similarly, given a three-dimensional metric and its time derivative at an initial instant, Einstein's equations provide us with a four-dimensional space-time which can be regarded as a trajectory in the space of 3-metrics[1].

However, this emphasis on the metric sets general relativity apart from all other fundamental forces of Nature. Indeed, in the theory of electro-weak and strong interactions, the basic dynamical variable is a (matrix-valued) vector potential, or a connection. Like general relativity, these theories are also geometrical. The connection enables one to parallel-transport objects along curves. In electrodynamics, the object is a charged particle such as an electron; in chromodynamics, it is a particle with internal color, such as a quark. Generally, if we move the object around a closed loop, we find that its state does not return to the initial value; it is rotated by an unitary matrix. In this case, the connection is said to have curvature and the unitary matrix is a measure of the curvature

[1] Actually, only six of the ten Einstein's equations provide the evolution equations. The other four do not involve time-derivatives at all and are thus constraints on the initial values of the metric and its time derivative. However, if the constraint equations are satisfied initially, they continue to be satisfied at all times.

in a region enclosed by the loop. In the case of electrodynamics, the connection is determined by the vector potential and the curvature by the electro-magnetic field strength.

Since the metric also gives rise to curvature, it is natural to ask if there is a relation between metrics and connections. The answer is in the affirmative. Every metric defines a connection —called the Levi-Civita connection of the metric. The object that the connection enables one to parallel transport is a vector. (It is this connection that determines the geodesics, i.e. the trajectories of particles in absence of non-gravitational forces.) It is therefore natural to ask if one can not use this connection as the basic variable in general relativity. If so, general relativity would be cast in a language that is rather similar to gauge theories and the description of the (general relativistic) gravitational interaction would be very similar to that of the other fundamental interactions of Nature. It turns out that the answer is in the affirmative. Furthermore, both Einstein and Schrödinger gave such a reformulation of general relativity. Why is this fact then not generally known? Indeed, I know of no textbook on general relativity which even mentions it. One reason is that in this formulation the basic equations are somewhat complicated —but not much more complicated, I think, than the standard ones in terms of the metric. A more important reason is that we tend to think of distances, light cones and causality as fundamental. These are directly determined by the metric and in a connection formulation, the metric is a 'derived' rather than a fundamental concept. But in the last few years, I have come to the conclusion that the real reason why the connection formulation of Einstein and Schrödinger has remained so obscure lies in an interesting historical episode. I will return to this point at the end of this section.

2.2 Metrics versus connections

Modern day researchers re-discovered connection theories of gravity after the invention and successes of gauge theories for other interactions. Generally, however, these formulations lead one to theories which are quite distinct from general relativity and the stringent experimental tests of general relativity often suffice to rule them out. There is, however, a reformulation of standard general relativity whose basic equations, furthermore, are simpler than the standard ones: while Einstein's equations are non-polynomial in terms of the metric and its conjugate momentum, they turn out to be low order polynomials in terms of the new connection and its conjugate momentum. Furthermore, just as the simplest particle trajectories in space-time are given by geodesics, the 'trajectory' determined by the time evolution of this connection according to Einstein's equation turns out to be a geodesic in configuration space of connections.

In this formulation, the phase space of general relativity is identical to that of the Yang-Mills theory which governs weak interactions. Recall first that in electrodynamics, the (magnetic) vector potential constitutes the configuration variable and the electric field serves as the conjugate momentum. In weak interactions and general relativity, the configuration variable is a matrix-valued vector potential; it can be written as $\vec{A}_i\tau_i$ where $\vec{A}_i$ is a triplet of vector fields and τ_i are the Pauli matrices. The conjugate momenta are represented by $\vec{E}_i\tau_i$ where $\vec{E}_i$ is a triplet of vector fields[2]. Given a pair $(\vec{A}_i, \vec{E}_i)$ (satisfying appropriate conditions as noted in footnote 1), the field equations of the two theories determine the complete time-evolution, i.e., a dynamical trajectory.

The field equations —and the Hamiltonians governing them— of the two theories are of course very different. In the case of weak interactions, we have a background space-time and we can use its metric to construct the Hamiltonian. In general relativity, we do not have a background metric. On the one hand this makes life very difficult since we do not have a fixed notion of distances or causal structures; these notions are to arise from the solution of the equations we are trying to write down! On the other hand, there is also tremendous simplification: Because there is no background metric, there are very few mathematically meaningful, gauge invariant expressions of the Hamiltonian that one can write down. (As we will see, this theme repeats itself in the quantum theory.) It is a pleasant surprise that the simplest non-trivial expression one can construct from the connection and its conjugate momentum is in fact the correct one, i.e., is the Hamiltonian of general relativity! The expression is at most quadratic in $\vec{A}_i$ and at most quadratic in $\vec{E}_i$. The similarity with gauge theories opens up new avenues for quantizing general relativity and the simplicity of the field equations makes the task considerably easier.

What is the physical meaning of these new basic variables of general relativity? As mentioned before, connections tell us how to parallel transport various physical entities around curves. The Levi-Civita connection tells us how to parallel transport vectors. The new connection, $\vec{A}_i$, on the other hand, determines

[2] A summation over the repeated index i is assumed. Also, technically each $\vec{A}_i$ is a 1-form rather than a vector field. Similarly, each $\vec{E}_i$ is a vector density of weight one, i.e., natural dual of a 2-form

the parallel transport of *left handed spin-$\frac{1}{2}$ particles* (such as neutrinos) —the so called *chiral fermions*. These fermions are mathematically represented by spinors which, as we know from elementary quantum mechanics, can be roughly thought of as 'square roots of vectors'. Not surprisingly, therefore, this connection is not completely determined by the metric alone. It requires additional information which roughly is a square-root of the metric, or a tetrad. The conjugate momenta $\hat{E}_i$ represent restrictions of these tetrads to space. They can be interpreted as spatial triads, i.e., as 'square-roots' of the metric of the 3-dimensional space. Thus, information about the Riemannian geometry of space is coded directly in these momenta. The (space and) time-derivatives of the triads are coded in the connection.

To summarize, there *is* a formulation of general relativity which brings it closer to theories of other fundamental interactions. Furthermore, in this formulation, the field equations simplify greatly. Thus, it provides a natural point of departure for constructing a quantum theory of gravity and for probing the nature of quantum geometry non-perturbatively.

2.3 Historical detour

To conclude this section, let me return to the piece of history involving Einstein and Schrödinger that I mentioned earlier. In the forties, both men were working on unified field theories. They were intellectually very close. Indeed, Einstein wrote to Schrödinger saying that he was perhaps the only one who was not 'wearing blinkers' in regard to fundamental questions in science and Schrödinger credited Einstein for inspiration behind his own work that led to the Schrödinger equation. During the years 1946-47, they had periods of intense correspondence on unified field theory and, in particular, on the issue of whether connections should be regarded as fundamental or metrics. Einstein was in Princeton and Schrödinger in Dublin. But starting January 1946, they exchanged their ideas and latest results very frequently. In fact the dates on their letters often show that the correspondence was going back and forth with astonishing speed. It reveals how quickly they understood the technical material the other had sent, how they hesitated, how they teased each other. Here are a few quotes:

The whole thing is going through my head like a millwheel: To take Γ [the connection] *alone as the primitive variable or the* g*'s* [metrics] *and* Γ*'s ? ...*

—Schrödinger, May 1st, 1946.

How well I understand your hesitating attitude! I must confess to you that inwardly I am not so certain ... We have squandered a lot of time on this thing, and the results look like a gift from devil's grandmother.

—Einstein, May 20th, 1946

Einstein was expressing doubts about using the Levi-Civita connection alone as the starting point which he had advocated at one time. Schrödinger wrote back that he laughed very hard at the phrase 'devil's grandmother'. In another letter, Einstein called Schrödinger 'a clever rascal'. Schrödinger was delighted and took it to be a high honor. This continued all through 1946. Then, in the beginning of 1947, Schrödinger thought he had made a breakthrough. He wrote to Einstein:

Today, I can report on a real advance. Maybe you will grumble frightfully for you have explained recently why you don't approve of my method. But very soon, you will agree with me...

—Schrödinger, January 26th, 1947

Schrödinger sincerely believed that his breakthrough was revolutionary [3]. Privately, he spoke of a second Nobel prize. The very next day after he wrote to Einstein, he gave a seminar in the Dublin Institute of Advanced Studies. Both the Taoiseach (the Irish prime minister) and newspaper reporters were invited. The day after, the following headlines appeared:

Twenty persons heard and saw history being made in the world of physics. ... The Taoiseach was in the group of professors and students. ..[To a question from the reporter] *Professor Schrödinger replied "This is the generalization. Now the Einstein theory becomes simply a special case ..."*

—*Irish Press,* January 28th, 1947

Not surprisingly, the headlines were picked up by *New York Times* which obtained photocopies of Schrödinger's paper and sent them to prominent physicists —including of course Einstein— for comments. As Walter Moore, Schrödinger's biographer puts it, Einstein could hardly believe that such grandiose claims had been made based on a what was at best a small advance in an area of work that they both had been pursuing for some time along parallel lines. He prepared a carefully worded response to the request from *New York Times:*

[3] The 'breakthrough' was to drop the requirement that the (Levi-Civita) connection be symmetric, i.e., to allow for torsion.

> *It seems undesirable to me to present such preliminary attempts to the public. ... Such communiqués given in sensational terms give the lay public misleading ideas about the character of research. The reader gets the impression that every five minutes there is a revolution in Science, somewhat like a coup d'état in some of the smaller unstable republics. ...*

Einstein's comments were also carried by the international press. On seeing them, Schrödinger wrote a letter of apology to Einstein citing his desire to improve the financial conditions of physicists in the Dublin Institute as a reason for the exaggerated account. It seems likely that it only worsened the situation. Einstein never replied. He also stopped scientific communication with Schrödinger.

The episode must have been shocking to those few who were exploring general relativity and unified field theories at the time. Could it be that this episode effectively buried the desire to follow up on connection formulations of general relativity until an entirely new generation of physicists who were blissfully unaware of this episode came on the scene?

3 Quantum geometry

3.1 General setting

Now that we have a connection formulation of general relativity, let us consider the problem of quantization. Recall first that in the quantum description of a particle, states are represented by suitable wave functions $\Psi(\vec{x})$ on the configuration space of the particle. Similarly, quantum states of the gravitational field are represented by appropriate wave functions $\Psi(\vec{A}_i)$ of connections. Just as the momentum operator in particle mechanics is represented by $\hat{P} \cdot \Psi_I = -i\hbar\,(\partial\Psi/\partial x_I)$ (with $I = 1, 2, 3$), the triad operators are represented by $\hat{\vec{E}}_i \cdot \Psi = \hbar G(\delta\Psi/\delta\vec{A}_i)$. The task is take geometric quantities such as lengths of curves, areas of surfaces and volumes of regions, express them in terms of triads using ordinary differential geometry and then promote these expressions to well-defined operators on the Hilbert space of quantum states. In principle, the task is rather similar to that in quantum mechanics where we first express observables such as angular momentum or Hamiltonian, express them in terms of configuration and momentum variables, $\vec{x}, \vec{p}$ and then promote them to quantum theory as well-defined operators on the quantum Hilbert space.

In quantum mechanics, the task is relatively straightforward; the only potential problem is the choice of factor ordering. In the present case, by contrast, we are dealing with a *field theory*, i.e., a system with an infinite number of degrees of freedom. Consequently, in addition to factor ordering, we face the much more difficult problem of regularization. Let me explain qualitatively how this arises. A field operator, such as the triad mentioned above, excites infinitely many degrees of freedom. Technically, its expectation values are distributions rather than smooth fields. They don't take precise values at a given point in space. To obtain numbers, we have to integrate the distribution against a test function, which extracts from it a 'bit' of information. As we change our test or smearing field, we get more and more information. (Take the familiar Dirac δ-distribution $\delta(x)$; it does not have a well-defined value at $x = 0$. Yet, we can extract the full information contained in $\delta(x)$ through the formula: $\int \delta(x) f(x) dx = f(0)$ for all test functions $f(x)$.) Thus, in a precise sense, field operators are distribution-valued. Now, as is well known, product of distributions is not well-defined. If we attempt naively to give meaning to it, we obtain infinities, i.e., a senseless result. Unfortunately, all geometric operators involve rather complicated (in fact non-polynomial) functions of the triads. So, the naive expressions of the corresponding quantum operators are typically meaningless. The key problem is to regularize these expressions, i.e., to extract well-defined operators from the formal expressions in a coherent fashion.

3.2 Geometric operators

This problem is not new; it arises in all physically interesting quantum field theories. However, as I mentioned in the Introduction, in other theories one has a background space-time metric and it is invariably used in a critical way in the process of regularization. For example, consider the electro-magnetic field. We know that the energy of the Hamiltonian of the theory is given by $H = \int \vec{E} \cdot \vec{E} + \vec{B} \cdot \vec{B}\, d^3x$. Now, in the quantum theory, $\hat{\vec{E}}$ and $\hat{\vec{B}}$ are both operator-valued distributions and so their square is ill-defined. But then, using the background flat metric, one Fourier decomposes these distributions, identifies creation and annihilation operators and extracts a well-defined Hamiltonian operator by normal ordering, i.e., by physically moving all annihilators to the right of creators. This procedure removes the unwanted and unphysical infinite zero point energy from the formal expression and the subtraction makes the operator well-defined. In the present case, on the other hand, we are trying to construct a quantum theory of geometry/gravity and do not have a flat metric —or indeed, any metric— in the background. Therefore, many of the standard regularization techniques are no longer available.

Fortunately, however, between 1992 and 1995, a new functional calculus was developed on the space of connections $\vec{A}_i$ —i.e., on the configuration space of the theory. This calculus is mathematically rigorous and makes no reference at all to a background space-time geometry; it is generally covariant. It provides a variety of new techniques which make the task of regularization feasible. First of all, there is a well-defined integration theory on this space. To actually evaluate integrals and define the Hilbert space of quantum states, one needs a measure: given a measure on the space of connections, we can consider the space of square-integrable functions which can serve as the Hilbert space of quantum states. There is, however, a preferred measure, singled out by the physical requirement that the (gauge invariant versions of the) configuration and momentum operators be self-adjoint. This measure is diffeomorphism invariant and thus respects the underlying symmetries coming from general covariance. Thus, there is a natural Hilbert space of states to work with [4]. Let us denote it by $\mathcal{H}$. Differential calculus enables one to introduce physically interesting operators on this Hilbert space and regulate them in a generally covariant fashion. As in the classical theory, the absence of a background metric is both a curse and a blessing. On the one hand, because we have very little structure to work with, many of the standard techniques simply fail to carry over. On the other hand, at least for geometric operators, the choice of viable expressions is now severely limited, which greatly simplifies the task of regularization.

The general strategy is the following. The Hilbert space $\mathcal{H}$ is the space of square-integrable functions $\Psi(\vec{A}_i)$ of connections $\vec{A}_i$. A key simplification arises because it can be obtained as the (projective) limit of Hilbert spaces associated with systems with only a finite number of degrees of freedom. More precisely, given any graph γ (which one can intuitively think of as a 'floating lattice') in the physical space, using techniques which are very similar to those employed in lattice gauge theory, one can construct a Hilbert space $\mathcal{H}_\gamma$ for a quantum mechanical system with $3N$ degrees of freedom, where N is the number of edges of the graph. Roughly, these Hilbert spaces know only about how the connection parallel transports chiral fermions along the edges of the graph and not elsewhere. That

[4] This is called the kinematical Hilbert space; it enables one to formulate the quantum Einstein's (or supergravity) equations. The final, physical Hilbert space will consist of states which are solutions to these equations.

is, the graph is a mathematical device to extract $3N$ 'bits of information' from the full, infinite dimensional information contained in the connection, and $\mathcal{H}_\gamma$ is the sub-space of $\mathcal{H}$ consisting of those functions of connections which depend only on these $3N$ bits. (Roughly, it is like focussing on only $3N$ components of a vector with an infinite number of components and considering functions which depend only on these $3N$ components, i.e., are constants along the orthogonal directions.) To get the full information, we need all possible graphs. Thus, a function of connections in $\mathcal{H}$ can be specified by specifying a function in $\mathcal{H}_\gamma$ for *every* graph γ in the physical space. Of course, since two distinct graphs can share edges, the collection of functions on $\mathcal{H}_\gamma$ must satisfy certain consistency conditions. These lie at the technical heart of various constructions and proofs.

The fact that $\mathcal{H}$ is the (projective) limit of $\mathcal{H}_\gamma$ breaks up any given problem in quantum geometry into a set of problems in quantum mechanics. Thus, for example, to define operators on $\mathcal{H}$, it suffices to define a *consistent family of* operators on $\mathcal{H}_\gamma$ for each γ. This makes the task of defining geometric operators feasible. I want to emphasize, however, that the introduction of graphs is only for technical convenience. Unlike in lattice gauge theory, we are not *defining* the theory via a continuum limit (in which the lattice spacing goes to zero.) Rather, the full Hilbert space $\mathcal{H}$ of the continuum theory is already well-defined. Graphs are introduced only for practical calculations. Nonetheless, they bring out the one-dimensional character of quantum states/excitations of geometry. It is because 'most' states in $\mathcal{H}$ can be realized as elements of $\mathcal{H}_\gamma$ for some γ that quantum geometry can be regarded as polymer-like.

Let me now outline the result of applying this procedure for geometric operators. Suppose we are given a surface S, defined in local coordinates by $x_3 = \text{const.}$ The classical formula for the area of the surface is: $A_S = \int d^2x \sqrt{E_i^3 E_i^3}$, where E_i^3 are the third components of the vectors $\vec{E}_i$. As is obvious, this expression is non-polynomial in the basic variables $\vec{E}_i$. Hence, off-hand, it would seem very difficult to write down the corresponding quantum operator. However, thanks to the background independent functional calculus, the operator can in fact be constructed rigorously.

To specify its action, let us consider a state which belongs to $\mathcal{H}_\gamma$ for *some* γ. Then, the action of the final, regularized operator $\hat{A}_S$ is as follows. If the graph has no intersection with the surface, the operator simply annihilates the state. If there are intersections, it acts at each intersection via group theory. *This simple form is a direct consequence of the fact that we do not have a background* *geometry*: given a graph and a surface, the diffeomorphism invariant information one can extract lies in their intersections. To specify the action of the operator in detail, let me suppose that the graph γ has N edges. Then the state Ψ has the form: $\Psi(\vec{A}_i) = \psi(g_1, ...g_N)$ for some function ψ of the N variables $g_1, ..., g_N$, where g_k ($\in SU(2)$) denotes the spin-rotation that a chiral fermion undergoes if parallel transported along the k-th edge using the connection $\vec{A}_i$. Since g_k represent the possible rotations of spins, angular momentum operators have a natural action on them. In terms of these, we can introduce 'vertex operators' associated with each intersection point v between S and γ:

$$\hat{O}_v \cdot \Psi(A) = \sum_{I,J} k(I,L)\vec{J}_I \cdot \vec{J}_L \cdot \psi(g_1, ..., g_N) \tag{1}$$

where I, L run over the edges of γ at the vertex v, $k(I,J) = 0, \pm 1$ depending on the orientation of edges I, L at v, and $\vec{J}_I$ are the three angular momentum operators associated with the I-th edge. (Thus, $\vec{J}_I$ act only on the argument g_I of ψ and the action is via the three left invariant vector fields on $SU(2)$.) *Thus, the vertex operators resemble the Hamiltonian of a spin system, $k(I,L)$ playing the role of the coupling constant.* The area operator is just a sum of the square-roots of the vertex operators:

$$\hat{A}_S = \frac{G\hbar}{2c^3} \sum_v |O_v|^{\frac{1}{2}} \tag{2}$$

Thus, the area operator is constructed from angular momentum-like operators. Note that the coefficient in front of the sum is just $\frac{1}{2}\ell_P^2$, the square of the Planck length. This fact will be important later.

Because of the simplicity of these operators, their complete spectrum —i.e., full set of eigenvalues— is known explicitly: Possible eigenvalues a_S are given by

$$a_S = \frac{\ell_P^2}{2} \sum_v \left[2j_v^{(d)}(j_v^{(d)}+1) + 2j_v^{(u)}(j_v^{(u)}+1) - j_v^{(d+u)}(j_v^{(d+u)}+1)\right]^{\frac{1}{2}} \tag{3}$$

where v labels a finite set of points in S and $j^{(d)}, j^{(u)}$ and $j^{(d+u)}$ are non-negative half-integers assigned to each v, subject to the usual inequality

$$j^{(d)} + j^{(u)} \geq j^{(d+u)} \geq |j^{(d)} - j^{(u)}| \,. \tag{4}$$

Thus the entire spectrum is discrete; *areas are indeed quantized!* This discreteness holds also for the length and the volume operators. Thus the expectation that the continuum picture may break down at the Planck scale is borne out fully. Quantum geometry is *very* different from the continuum picture. This may be the fundamental reason for the failure of perturbative approaches to quantum gravity.

Let us now examine a few properties of the spectrum. The lowest eigenvalue is of course zero. The next lowest eigenvalue may be called the *area gap*. Interestingly, area-gap is sensitive to the topology of the surface S. If S is open, it is $\frac{\sqrt{3}}{4}\ell_P^2$. If S is a closed surface —such as a 2-torus in a 3-torus— which fails to divide the spatial 3-manifold into an 'inside' and an 'outside' region, the gap is larger, $\frac{2}{4}\ell_P^2$. If S is a closed surface —such as a 2-sphere in R^3— which divides space into an 'inside' and an 'outside' region, the area gap is even larger; it is $\frac{2\sqrt{2}}{4}\ell_P^2$. Another interesting feature is that in the large area limit, the eigenvalues crowd together. This follows directly from the form of eigenvalues given above. Indeed, one can show that for large eigenvalues a_S, the difference Δa_S between consecutive eigenvalues goes as $\Delta a_S \leq (exp - \sqrt{a_S/\ell_P^2})\ell_P^2$. Thus, Δa_S goes to zero very fast. (The crowding is noticeable already for low values of a_S. For example, in the case of trivial topology, there is only one non-zero eigenvalue with $a_S < 0.5\ell_P^2$, seven with $a_S < \ell_P^2$ and 98 with $a_S < 2\ell_P^2$.) Intuitively, this explains why the continuum limit works so well.

3.3 Physical consequences: details matter!

We will now see that if Δa_S had failed to vanish sufficiently fast, one would have been forced to conclude that the semi-classical approximation to quantum gravity must fail in an important way. To bring out this point, let me backtrack a bit. Let us consider not the most general eigenstates of the area operator $\hat{A}_S$ but —as was first done chronologically— the simplest ones. These correspond to graphs which have simple intersections with S. For example, n edges of the graph may just pierce S, each one separately, so that at each vertex there is just a straight line passing through. For these states, the eigenvalues are $a_S = (\sqrt{3}/2)n\ell_P^2$. Thus, here, the level spacing is uniform, like that of the Hamiltonian of a simple harmonic oscillator. Even if we restrict ourselves to the simplest eigenstates, even for large eigenvalues, the level spacing does not go to zero. Suppose for a moment that this is the *full* spectrum of the area operator. Then, as I will indicate below, Hawking's semi-classical derivation of black hole evaporation would have been incorrect. That is, the effects coming from area quantization would have implied that even for large macroscopic black holes of, say, a thousand solar masses, we can not trust semi-classical arguments.

Let me explain this point in some detail. The original derivation of Hawking's was carried out in the framework of quantum field theory in curved space-times which assumes that there is a specific underlying continuum space-time and explores the effects of the curvature of this space-time on quantum matter fields. In this approximation, Hawking found that the black hole geometries are such that there is a spontaneous emission which has a Planckian spectrum at infinity. Thus, black holes, seen from far away, resemble black bodies and the associated temperature turns out to be inversely related to the mass of the hole. Now, physically one expects that, as it evaporates, the black hole must lose mass. Since the radius of the horizon is proportional to the the mass, the area of the horizon must decrease. If one uses a classical picture for the underlying space-time, one would conclude that the process is continuous. However, if in a more fundamental theory of quantum gravity area is quantized, one would expect that the black hole evaporates in discrete steps by making a transition from one area eigenvalue to another, smaller one. The process would be very similar to the way an excited atom descends to its ground state through a series of discrete transitions.

Let us look at this process in some detail. For simplicity let us use units with $c = 1$. *Suppose, to begin with, that the level spacing of eigenvalues of the area operator is the naive one, i.e. with* $\Delta a_S = \sqrt{3}/2\ell_P^2$. Then, the fundamental theory would have predicted that the smallest frequency, ω_o of emitted particles would be given by $\hbar\omega_o = \Delta M \sim (1/G^2M)\Delta a_H \sim \hbar/GM$, since the area A_H of the horizon goes as G^2M^2. Thus, the 'true' spectrum would have emission lines only at frequencies $\omega = N\omega_o \sim N\omega_p$, for $N = 1, 2, ...$ corresponding to transitions of the black hole through N area levels. How does this compare with the Hawking prediction? As I mentioned above, according to Hawking's semi-classical analysis, the spectrum would be the same as that of a black body at temperature T given by $kT \sim \hbar/GM$, where k is the Boltzmann constant. Hence, the peak of this spectrum would appear at ω_p given by $\hbar\omega_p \sim kT \sim \hbar/GM$. But this is precisely the order of magnitude of the minimum frequency ω_o that would be allowed if the area spectrum were the naive one. Thus, in this case, a more fundamental theory would predict that the spectrum would not resemble a black-body spectrum. The most probable transition would be for $N = 1$ and so the spectrum would be peaked at ω_p as in the case of a black body. However, there would be no emission lines at frequencies low compared with ω_p; this part of the black body spectrum would be simply absent. The part of the spectrum for $\omega > \omega_p$ would also not be faithfully reproduced since the discrete lines with frequencies $N\omega_o$, with $N = 1, 2, ...$ would *not* be sufficiently near each other —i.e. crowded— to yield an approximation to the continuous black-body spectrum.

The situation is completely different for the correct, full spectrum of the area operator if the black hole is macroscopic, i.e., large. Then, as I noted earlier, the area eigenvalues crowd and the level spacing goes as $\Delta a_H \leq (\exp - \sqrt{a_H/\ell_P^2})\ell_P^2$. As a consequence, as the black hole makes transition from one area eigenvalue to another, it would emit particles at frequencies equal to or larger than $\sim \omega_p \exp - \sqrt{a_H/\ell_P^2}$. Since for a macroscopic black-hole the exponent is very large (for a solar mass black-hole it is $\sim 10^{71}!$) the spectrum would be well-approximated by a continuous spectrum and would extend well below the peak frequency. Thus, the precise form of the area spectrum ensures that, for large black-holes, the potential problem with Hawking's semi-classical picture disappears. Note however that as the black hole evaporates, its area decreases, it gets hotter and evaporates faster. Therefore, a stage comes when the area is of the order of ℓ_P^2. Then, there *would* be deviations from the black body spectrum. But this is to be expected since in this extreme regime one does not expect the semi-classical picture to continue to be meaningful.

This argument brings out an interesting fact. Since the Planck length ℓ_P is so small, one would have thought that even if the area spectrum were the naive one —with equal level spacing $\Delta a_S = (\sqrt{3}/2)\ell_P^2$— one would not run in to a problem with classical or semi-classical approximations while dealing with large, macroscopic objects. Indeed, there are several iconoclastic approaches to quantum geometry in which one simply begins by postulating that geometric quantities should be quantized. Then, having no recourse to first principles from where to derive the eigenvalues of these operators, one simply postulates them to be multiples of appropriate powers of the Planck length. For area then, one would say that the eigenvalues are integral multiples of ℓ_P^2. The above argument shows how this innocent looking assumption can contradict semi-classical results even for large black holes. In our case, we did not begin by postulating the nature of quantum geometry. Rather, we *derived* the spectrum of the area operator from first principles. As we see, the form of these eigenvalues is rather complicated and could not have been guessed apriori. More importantly, the detailed form does carry rich information and in particular removes the conflict with semi-classical results in macroscopic situations.

3.4 Future directions

Exploration of quantum Riemannian geometry continues. Last year, it was found that geometric operators exhibit certain unexpected non-commutativity. This reminds one of the features explored by Alain Connes in his non-commutative geometry. Indeed, there are several points of contact between these two approaches. For instance, the Dirac operator that features prominently in Conne's theory is closely related to the connection $\vec{A}_i$ used here. However, at a fundamental level, the two approaches are rather different. In Conne's approach, one constructs a non-commutative analog of entire differential geometry. Here, by contrast, one focuses only on Riemannian geometry; the underlying manifold structure remains classical. In three space-time dimensions, it is possible to get rid of this feature in the final picture and express the theory in purely combinatorial fashion. Whether the same will be possible in four dimensions remains unclear. However, combinatorial methods continue to dominate the theory and it is quite possible that one would again be able to present the final picture without any reference to an underlying smooth manifold.

Another promising direction for further work is to construct better and better candidates for 'weave states' which can be regarded as non-linear analogs of coherent states approximating smooth, macroscopic geometries. Once one has an 'optimum' candidate to represent Minkowski space, one would develop quantum field theory on these weave quantum geometries. Because the underlying basic excitations are one-dimensional, the 'effective dimension of space' for these field theories would be less than three. Now, in the standard continuum approach, we know that quantum field theories in low dimensions tend to be better behaved because their ultra-violet problems are softer. Hence, there is hope that these theories will be free of infinities. If they are renormalizable in the continuum, their predictions at large scales cannot depend on the details of the behavior at very small scales. Therefore, theories based on weaves would not only be finite but their predictions may well agree with those of renormalizable theories at the laboratory scale.

Another major direction of research is devoted to formulating and solving quantum Einstein's equations using the new functional calculus. Over the past year, there have been some exciting developments in this area. The methods developed there seem to be applicable also to supergravity theories. In the coming years, therefore, there should be further work in this area. Finally, since this quantum geometry does not depend on a background metric, it provides a natural arena for other problem, in particular, that of obtaining a background independent formulation of string theory.

Acknowledgments

The work summarized here is based on contributions from many researchers especially John Baez, Alejandro Corichi, Roberto DePitri, Chris Isham, Jerzy Lewandowski, Renate Loll, Don Marolf, Jose Mourao, Thomas Thiemann, Carlo Rovelli, Lee Smolin and José-Antonio Zapata. Special thanks are due to Jerzy Lewandowski for long-range collaboration. This work was supported in part by the NSF Grant PHY95-14240 and by the Eberly fund of the Pennsylvania State University.

References

[1] B. Riemann, *Über die Hypothesen, welche der Geometrie zugrunde liegen* (1854).

Monographs and Reviews on Non-perturbative Quantum Gravity:

[2] A. Ashtekar, *Lectures on Non-perturbative Canonical Gravity*, Notes prepared in collaboration with R. S. Tate. (World Scientific, Singapore, 1991).

[3] R. Gambini and J. Pullin, *Loops, Knots, Gauge Theories and Quantum Gravity.* (Cambridge University Press, Cambridge, 1996).

[4] A. Ashtekar, in *Gravitation and Quantizations*, ed B. Julia and J. Zinn-Justin (Elsevier, Amsterdam, 1995).

Background-independent Functional Calculus:

[5] A. Ashtekar and C. J. Isham, *Class. Quant. Grav.* **9**, 1433 (1992).

[6] A. Ashtekar and J. Lewandowski, "Representation theory of analytic holonomy $C^\star$ algebras", in *Knots and quantum gravity*, J. Baez (ed), (Oxford University Press, Oxford 1994).

[7] J. Baez, *Lett. Math. Phys.* **31**, 213 (1994); "Diffeomorphism invariant generalized measures on the space of connections modulo gauge transformations", hep-th/9305045, in the Proceedings of the conference on quantum topology, D. Yetter (ed) (World Scientific, Singapore, 1994).

[8] A. Ashtekar and J. Lewandowski, *J. Math. Phys.* **36**, 2170 (1995).

[9] D. Marolf and J. M. Mourão, *Commun. Math. Phys.* **170**, 583 (1995).

[10] A. Ashtekar and J. Lewandowski, J. Geo. & Phys. **17**, 191 (1995).

[11] J. Baez, "Spin network states in gauge theory", Adv. Math. (in press); "Spin networks in non-perturbative quantum gravity," pre-print gr-qc/9504036.

[12] C. Rovelli and L. Smolin, *Phys. Rev.* D 52 (1995) 5743-5759

[13] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourão and T. Thiemann, *J. Math. Phys.* **36**, 6456 (1995).

[14] J. Baez and S. Sawin, *Functional integration on spaces of connections, preprint q-alg/9507023.*

[15] J. A. Zapata, *Combinatoric space from loop quantum gravity*, Penn State Pre-print (1997).

Geometric Operators

[16] A. Ashtekar, C. Rovelli and L. Smolin, *Phys. Rev. Lett.* **69**, 237 (1992).

[17] J. Iwasaki and C. Rovelli, *Int. J. Modern. Phys.* D **1**, 533 (1993); *Class. Quant. Grav.* **11**, 2899 (1994).

[18] C. Rovelli and L. Smolin, *Nucl. Phys.* B **442**, 593 (1995).

[19] A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourão and T. Thiemann, *J. Funct. Analysis*, **135**, 519 (1996).

[20] R. Loll, *Phys. Rev. Lett.* **75**, 3084 (1995).

[21] A. Ashtekar, J. Lewandowski, *Class. Quant. Grav.* **14**, A55-A81 (1997).

[22] R. Loll, "Spectrum of the volume operator in quantum gravity", *Nucl. Phys. B* (to appear).

[23] A. Ashtekar, A. Corichi, J. Lewandowski and J. A. Zapata, *Quantum theory of geometry II: Non-commutativity of Riemannian structures*, Penn State Pre-print (1997).

[24] A. Ashtekar and J. Lewandowski, *Quantum theory of geometry III: Volume operators*, Penn State Pre-print (1997).

Black Hole Evaporation:

[25] S. W. Hawking, *Comun. Math. Phys.* **43**, 199 (1975).

[26] J. Bekenstein and V. F. Mukhanov, "Spectroscopy of quantum black holes", *Physics Letters* B **360**, 7 (1995).

[27] S. Fairhurst, *Properties of the Spectrum of the Area Operator* (unpublished Penn State Report)

[28] A. Ashtekar, to appear in: *Geometric Issues in the Foundation of Science*, edited by L. Mason et al (Oxford University Press).

Brisure de Symétrie Spontanée et Géométrie du Point de Vue Spectral

Alain Connes
Institut des Hautes Études Scientifiques
Bures-sur-Yvette, France

1 Généralités

La géométrie de Riemann admet pour données préalables une *variété* M dont les points $x \in M$ sont localement paramétrés par un nombre fini de coordonnées réelles x^μ, et la *métrique* donnée par l'élément de longueur infinitésimal,

$$ds^2 = g_{\mu\nu}\, dx^\mu\, dx^\nu\,. \tag{1}$$

La distance entre deux points $x, y \in M$ est donnée par,

$$d(x,y) = \text{Inf Longueur } \gamma\,, \tag{2}$$

où γ varie parmi les arcs joignant x à y, et

$$\text{Longueur } \gamma = \int_x^y ds\,. \tag{3}$$

La théorie de Riemann est à la fois assez souple pour fournir (au prix d'un changement de signe) un bon modèle de l'espace temps de la relativité générale et assez restrictive pour mériter le nom de géométrie. Le point essentiel est que le calcul différentiel et intégral permet de passer du local au global et que les notions simples de la géométrie Euclidienne telle celle de droite continuent à garder un sens. L'équation des géodésiques,

$$\frac{d^2\, x^\mu}{dt^2} = -\Gamma^\mu_{\nu\rho}\, \frac{dx^\nu}{dt}\, \frac{dx^\rho}{dt} \tag{4}$$

(où $\Gamma^\mu_{\nu\rho} = \frac{1}{2}\, g^{\mu\alpha}(g_{\alpha\nu,\rho} + g_{\alpha\rho,\nu} - g_{\nu\rho,\alpha})$) pour la métrique $dx^2 + dy^2 + dz^2 - (1 + 2V(x,y,z))dt^2$ donne l'équation de Newton dans le potentiel V (cf. [W]

This text is a reproduction of the article published originally in Astérisque 241 (exposé 816, pp.313–349, 1996), with the kind permission of Séminaire Bourbaki and the Société Mathématique de France to reproduce the text.

pour un énoncé plus précis). Les données expérimentales récentes sur les pulsars binaires confirment [DT] la relativité générale et l'adéquation de la géométrie de Riemann comme modèle de l'espace temps à des échelles suffisamment grandes. La question ([R]) de l'adéquation de cette géométrie comme modèle de l'espace temps à très courte échelle est controversée mais la longueur de Planck

$$\ell_p = (G\hbar/c^3)^{1/2} \sim 10^{-33}\text{cm} \tag{5}$$

est considérée comme la limite naturelle sur la détermination précise des coordonnées d'espace temps d'un évènement. (Voir par exemple [F] ou [DFR] pour l'argument physique, utilisant la mécanique quantique, qui établit cette limite.)

Dans cet exposé nous présentons une nouvelle notion d'espace géométrique qui en abandonnant le rôle central joué par les *points* de l'espace permet une plus grande liberté dans la description de l'espace temps à courte échelle. Le cadre proposé est suffisamment général pour traiter les espaces discrets, les espaces Riemanniens, les espaces de configurations de la théorie quantique des champs et les duaux des groupes discrets non nécessairement commutatifs. Le problème principal est d'adapter à ce cadre général les notions essentielles de la géométrie et en particulier le calcul infinitésimal. Le formalisme opératoriel de la mécanique quantique joint à l'analyse des divergences logarithmiques de la trace des opérateurs donnent la généralisation cherchée du calcul différentiel et intégral (Section II). Nous donnons quelques applications directes de ce calcul (Théorèmes 1, 2, 4).

La donnée d'un espace géométrique est celle d'un *triplet spectral:*

$$(\mathcal{A}, \mathcal{H}, D) \tag{6}$$

où $\mathcal{A}$ est une algèbre involutive d'opérateurs dans l'espace de Hilbert $\mathcal{H}$ et D un opérateur autoadjoint non borné dans $\mathcal{H}$. L'algèbre involutive $\mathcal{A}$ correspond à la donnée de l'espace M selon la dualité Espace $\leftrightarrow$ Algèbre classique en géométrie algébrique. L'opérateur $D^{-1} = ds$ correspond à l'élément de longueur infinitésimal de la géométrie de Riemann.

Il y a deux différences évidentes entre cette *géométrie spectrale* et la géométrie Riemannienne. La première est que nous ne supposerons pas en général la commutativité de l'algèbre $\mathcal{A}$. La deuxième est que ds, étant un opérateur, ne commute pas avec les éléments de $\mathcal{A}$, même quand $\mathcal{A}$ est commutative.

Comme nous le verrons, des relations de commutation très simples entre ds et l'algèbre $\mathcal{A}$, jointes à la dualité de Poincaré caractérisent les triplets spectraux

(6) qui proviennent de variétés Riemanniennes (Théorème 6). Quand l'algèbre $\mathcal{A}$ est commutative sa fermeture normique dans $\mathcal{H}$ est l'algèbre des fonctions continues sur un espace compact M. Un point de M est un caractère de $\bar{\mathcal{A}}$, i.e. un homomorphisme de $\bar{\mathcal{A}}$ dans $\mathbb{C}$,

$$\chi : \bar{\mathcal{A}} \to \mathbb{C} \ , \ \chi(a+b) = \chi(a) + \chi(b) \ , \ \chi(\lambda a) = \lambda \, \chi(a) \ ,$$
$$\chi(ab) = \chi(a) \, \chi(b) \ , \forall \, a, b \in \bar{\mathcal{A}} \ , \ \forall \, \lambda \in \mathbb{C} . \tag{7}$$

Soit par exemple $\mathcal{A}$ l'algèbre $\mathbb{C}\Gamma$ d'un groupe discret Γ agissant dans l'espace de Hilbert $\mathcal{H} = \ell^2(\Gamma)$ de la représentation régulière (gauche) de Γ. Quand le groupe Γ et donc l'algèbre $\mathcal{A}$ sont commutatifs les caractères de $\bar{\mathcal{A}}$ sont les éléments du dual de Pontrjagin de Γ,

$$\hat{\Gamma} = \{ \chi : \Gamma \to U(1) \ ; \ \chi(g_1 \, g_2) = \chi(g_1) \, \chi(g_2) \qquad \forall g_1, g_2 \in \Gamma \} . \tag{8}$$

Les notions élémentaires de la géométrie différentielle pour l'espace $\hat{\Gamma}$ continuent à garder un sens dans le cas général où Γ n'est plus commutatif grâce au dictionnaire suivant dont la colonne de droite n'utilise pas la commutativité de l'algèbre $\mathcal{A}$,

Espace X	Algèbre $\mathcal{A}$
Fibré vectoriel	Module projectif de type fini
Forme différentielle de degré k	Cycle de Hochschild de dimension k
Courant de de Rham de dimension k	Cocycle de Hochschild de dimension k
Homologie de de Rham	Cohomologie cyclique de $\mathcal{A}$

L'intérêt de la généralisation ci-dessus au cas non commutatif est illustré par exemple par la preuve de la conjecture de Novikov pour les groupes Γ qui sont hyperboliques [CM1].

Dans le cas général la notion de point, donnée par (7) est de peu d'intérêt, par contre celle de mesure de probabilité garde tout son sens. Une telle mesure φ est une forme linéaire positive sur $\mathcal{A}$ telle que $\varphi(1) = 1$,

$$\varphi : \bar{\mathcal{A}} \to \mathbb{C} \ , \ \varphi(a^* a) \geq 0 \ , \quad \forall \, a \in \bar{\mathcal{A}} \ , \ \varphi(1) = 1 . \tag{9}$$

Au lieu de mesurer les distances entre les points de l'espace par la formule (2) nous mesurons les distances entre états φ, ψ sur $\bar{\mathcal{A}}$ par une formule duale qui implique un *sup* au lieu d'un *inf* et n'utilise pas les arcs tracés dans l'espace,

$$d(\varphi, \psi) = \operatorname{Sup} \{ |\varphi(a) - \psi(a)| \ ; \ a \in \mathcal{A} \ , \ \|[D, a]\| \leq 1 \} . \tag{10}$$

Vérifions que cette formule redonne la distance géodésique dans le cas Riemannien. Soit M une variété Riemannienne compacte munie d'une K-orientation, i.e. d'une structure spinorielle. Le triplet spectral $(\mathcal{A}, \mathcal{H}, D)$ associé est donné par la représentation,

$$(f\xi)(x) = f(x)\,\xi(x) \qquad \forall x \in M \ , \ f \in \mathcal{A} \ , \ \xi \in \mathcal{H} \tag{11}$$

de l'algèbre des fonctions sur M dans l'espace de Hilbert

$$\mathcal{H} = L^2(M, S) \tag{12}$$

des sections de carré intégrable du fibré des spineurs.

L'opérateur D est l'opérateur de Dirac (cf. [L-M]). On vérifie immédiatement que le commutateur $[D, f]$, $f \in \mathcal{A}$ est l'opérateur de multiplication de Clifford par le gradient ∇f de f et que sa norme hilbertienne est,

$$\|[D, f]\| = \operatorname*{Sup}_{x \in M} \|\nabla f\| = \text{Norme Lipschitzienne de } f\,. \tag{13}$$

Soient $x, y \in M$ et φ, ψ les caractères correspondants: $\varphi(f) = f(x)$, $\psi(f) = f(y)$ $\forall f \in \mathcal{A}$ la formule (10) donne le même résultat que la formule (2), i.e. donne la distance géodésique entre x et y.

Contrairement à (2) la formule duale (10) garde un sens en général et en particulier pour les espaces discrets ou totalement discontinus.

La notion usuelle de *dimension* d'un espace est remplacée par un *spectre de dimension* qui est un sous-ensemble de $\mathbb{C}$ dont la partie réelle est bornée supérieurement par $\alpha > 0$ si

$$\lambda_n^{-1} = O(n^{-\alpha}) \tag{14}$$

où λ_n est la n-ième valeur propre de $|D|$.

La relation entre le local et le global est donnée par la formule locale de l'indice (Théorème 4) ([CM2]).

La propriété caractéristique des *variétés différentiables* qui est transposée au cas non commutatif est la *dualité de Poincaré*. La dualité de Poincaré en homologie ordinaire est insuffisante pour caractériser le type d'homotopie des variétés ([Mi-S]) mais les résultats de D. Sullivan ([S2]) montrent (dans le cas simplement connexe, de dimension ≥ 5 et en ignorant le nombre premier 2) qu'il suffit de remplacer l'homologie ordinaire par la KO-homologie.

De plus la K-homologie admet grâce aux résultats de Brown Douglas Fillmore, Atiyah et Kasparov une traduction algébrique très simple, donnée par:

Espace X	Algèbre $\mathcal{A}$
$K_1(X)$	Classe d'homotopie stable de triplet spectral $(\mathcal{A}, \mathcal{H}, D)$
$K_0(X)$	Classe d'homotopie stable de triplet spectral $\mathbb{Z}/2$ gradué

(i.e. pour K_0 on suppose que $\mathcal{H}$ est $\mathbb{Z}/2$ gradué par γ, $\gamma = \gamma^*$, $\gamma^2 = 1$ et que $\gamma a = a\gamma \quad \forall\, a \in \mathcal{A}$, $\gamma D = -D\gamma$).

Cette description suffit pour la K-homologie complexe qui est périodique de période 2.

Dans le cas non commutatif la *classe fondamentale* d'un espace est une classe μ de KR-homologie pour l'algèbre $\mathcal{A} \otimes \mathcal{A}^0$ munie de l'involution,

$$\tau(x \otimes y^0) = y^* \otimes (x^*)^0 \qquad \forall\, x, y \in \mathcal{A} \tag{15}$$

où $\mathcal{A}^0$ désigne l'algèbre opposée de $\mathcal{A}$. Le produit intersection de Kasparov [K] permet de formuler la dualité de Poincaré par l'invertibilité de μ. La KR-homologie est périodique de période 8 et la dimension modulo 8 est spécifiée par les règles de commutation suivantes, où J est une isométrie antilinéaire dans $\mathcal{H}$ qui implémente l'involution τ,

$$JxJ^{-1} = \tau(x) \qquad \forall\, x \in \mathcal{A} \otimes \mathcal{A}^0\,. \tag{16}$$

On a $J^2 = \varepsilon$, $JD = \varepsilon' DJ$, $J\gamma = \varepsilon''\gamma J$ où $\varepsilon, \varepsilon', \varepsilon'' \in \{-1, 1\}$ et si n désigne la dimension modulo 8,

$n =$	0	1	2	3	4	5	6	7
ε	1	1	−1	−1	−1	−1	1	1
ε'	1	−1	1	1	1	−1	1	1
ε''	1		−1		1		−1	

L'isométrie antilinéaire J est donnée dans le cas Riemannien par l'opérateur de conjugaison de charge et dans le cas non commutatif par l'opérateur de Tomita [Ta] qui dans le cas où une algèbre d'opérateurs $\mathcal{A}$ admet un vecteur cyclique qui est cyclique pour le commutant $\mathcal{A}'$, établit un antiisomorphisme,

$$a \in \mathcal{A}'' \to Ja^*J^{-1} \in \mathcal{A}'\,. \tag{17}$$

La donnée de μ ne spécifie que la classe d'homotopie stable du triplet spectral $(\mathcal{A}, \mathcal{H}, D)$ muni de l'isométrie J (et de la $\mathbb{Z}/2$ graduation γ si n est pair). La non trivialité de cette classe d'homotopie est visible dans la forme d'intersection,

$$K(\mathcal{A}) \times K(\mathcal{A}) \to \mathbb{Z} \tag{18}$$

donnée par l'indice de Fredholm de D à coefficient dans $K(\mathcal{A} \otimes \mathcal{A}^0)$.

Pour comparer les triplets spectraux dans la classe μ nous utiliserons la fonctionnelle spectrale suivante

$$\mathrm{Trace}\,(\varphi(D)) \tag{19}$$

où $\varphi : \mathbb{R} \to \mathbb{R}_+$ est une fonction positive convenable.

L'algèbre $\mathcal{A}$ une fois fixée, une géométrie spectrale est déterminée par la classe d'équivalence unitaire du triplet spectral $(\mathcal{A}, \mathcal{H}, D)$ avec l'isométrie J. Si l'on note π la représentation de $\mathcal{A}$ dans $\mathcal{H}$ l'équivalence unitaire entre $(\pi_1, \mathcal{H}_1, D_1, J_1)$ et $(\pi_2, \mathcal{H}_2, D_2, J_2)$ signifie qu'il existe un unitaire $U : \mathcal{H}_1 \to \mathcal{H}_2$ tel que

$$U\pi_1(a)U^* = \pi_2(a) \quad \forall\, a \in \mathcal{A}\ ,\ UD_1U^* = D_2\ ,\ UJ_1U^* = J_2 \tag{20}$$

(et $U\gamma_1U^* = \gamma_2$ dans le cas où n est pair).

Le groupe $\mathrm{Aut}(\mathcal{A})$ des automorphismes de l'algèbre involutive $\mathcal{A}$ agit sur l'ensemble des géométries spectrales par composition,

$$\pi'(a) = \pi(\alpha^{-1}(a)) \qquad \forall\, a \in \mathcal{A}\ ,\ \alpha \in \mathrm{Aut}(\mathcal{A})\,. \tag{21}$$

Le sous-groupe $\mathrm{Aut}^+(\mathcal{A})$ des automorphismes qui préservent la classe μ agit sur la classe d'homotopie stable déterminée par μ et préserve par construction la fonctionnelle d'action (19). En général ce groupe est non compact, et il coïncide par exemple dans le cas Riemannien avec le groupe $\mathrm{Diff}^+(M)$ des difféomorphismes qui préservent la K-orientation, i.e. la structure spinorielle de M. A l'inverse le groupe d'isotropie d'une géométrie donnée, est automatiquement *compact* (pour $\mathcal{A}$ unifère). Ceci montre que la fonctionnelle d'action (19) donne automatiquement naissance au phénomène de brisure de symétrie spontanée (Figure 1).

Nous montrerons que pour un choix convenable de l'algèbre $\mathcal{A}$ la fonctionnelle d'action (19) ajoutée au terme $\langle \xi, D\xi \rangle$, $\xi \in \mathcal{H}$ donne le modèle standard de Glashow-Weinberg-Salam couplé à la gravitation. L'algèbre $\mathcal{A}$ est le produit tensoriel de l'algèbre des fonctions sur un espace Riemannien M par une algèbre non commutative de dimension finie dont les données phénoménologiques spécifient la géométrie spectrale.

2 Un calcul infinitésimal.

Nous montrons comment le formalisme opératoriel de la mécanique quantique permet de donner un sens précis à la notion de variable infinitésimale. La notion d'infinitésimal est sensée avoir un sens intuitif évident. Elle résiste cependant fort bien aux essais de formalisation donnés par exemple par l'analyse non standard. Ainsi, pour prendre un exemple précis ([B-W]), soit $dp(x)$ la probabilité pour qu'une fléchette lancée au hasard sur la cible Ω termine sa course au point $x \in \Omega$ (Figure 2). Il est clair que $dp(x) < \varepsilon \quad \forall \varepsilon > 0$ et que néanmoins la réponse $dp(x) = 0$ n'est pas satisfaisante. Le formalisme usuel de la théorie de la mesure ou des formes différentielles contourne le problème en donnant un sens à l'expression

$$\int f(x)\, dp(x) \qquad f : \Omega \to \mathbb{C} \tag{1}$$

mais est insuffisant pour donner un sens par exemple à $e^{-\frac{1}{dp(x)}}$. La réponse, à savoir un réel non standard, fournie par l'analyse non standard, est également décevante: tout réel non standard détermine canoniquement un sous-ensemble non Lebesgue mesurable de l'intervalle $[0,1]$ de sorte qu'il est impossible ([Ste]) d'en exhiber un seul. Le formalisme que nous proposons donnera une réponse substantielle et calculable à cette question.

Le cadre est fixé par un espace de Hilbert séparable $\mathcal{H}$ décomposé comme somme directe de deux sous-espaces de dimension infinie. Donner cette décomposition revient à donner l'opérateur linéaire F dans $\mathcal{H}$ qui est l'identité, $F\xi = \xi$, sur le premier sous-espace et moins l'identité, $F\xi = -\xi$ sur le second, on a

$$F = F^* \ , \ F^2 = 1 \, . \tag{2}$$

Le cadre ainsi déterminé est unique à isomorphisme près. Le début du dictionnaire qui traduit les notions classiques en language opératoriel est le suivant :

Classique	Quantique
Variable complexe	Opérateur dans $\mathcal{H}$
Variable réelle	Opérateur autoadjoint
Infinitésimal	Opérateur compact
Infinitésimal d'ordre α	Opérateur compact dont les valeurs caractéristiques μ_n vérifient $\mu_n = O(n^{-\alpha})$, $n \to \infty$
Différentielle d'une variable réelle ou complexe	$đf = [F, f] = Ff - fF$
Intégrale d'un infinitésimal d'ordre 1	$⨍ T =$ Coefficient de la divergence logarithmique dans la trace de T .

Les deux premières lignes du dictionnaire sont familières en mécanique quantique. L'ensemble des valeurs d'une variable complexe correspond au *spectre* d'un opérateur. Le calcul fonctionnel holomorphe donne un sens à $f(T)$ pour toute fonction holomorphe f sur le spectre de T. Les fonctions holomorphes sont les seules à opérer dans cette généralité ce qui reflète la différence entre l'analyse complexe et l'analyse réelle où les fonctions boréliennes arbitraires opèrent. Quand $T = T^*$ est autoadjoint $f(T)$ a un sens pour toute fonction borélienne f. Notons que toute variable aléatoire usuelle X sur un espace de probabilité, (Ω, P) peut être trivialement considérée comme un opérateur autoadjoint. On prend $\mathcal{H} = L^2(\Omega, P)$ et

$$(T\xi)(p) = X(p)\,\xi(p) \qquad \forall\, p \in \Omega \ , \ \xi \in \mathcal{H} \, . \tag{3}$$

La mesure spectrale de T redonne la probabilité P.

Passons à la troisième ligne du dictionnaire. Nous cherchons des "variables infinitésimales", i.e. des opérateurs T dans $\mathcal{H}$ tels que

$$||T|| < \varepsilon \qquad \forall\, \varepsilon > 0 \, , \tag{4}$$

où $||T|| = \mathrm{Sup}\, \{||T\xi|| \ ; \ ||\xi|| = 1\}$ est la norme d'opérateur. Bien entendu si l'on prend (4) au pied de la lettre on obtient $||T|| = 0$ et $T = 0$ est la seule solution. Mais on peut affaiblir (4) de la manière suivante,

$$\forall\, \varepsilon > 0 \ , \ \exists \text{ sous-espace de dimension finie } E \subset \mathcal{H} \text{ tel que } ||T/E^{\perp}|| < \varepsilon \tag{5}$$

où $E^\perp$ désigne l'orthogonal de E dans $\mathcal{H}$,

$$E^\perp = \{\xi \in \mathcal{H} \; ; \; \langle \xi, \eta \rangle = 0 \quad \forall \eta \in E\} \tag{6}$$

qui est un sous-espace de codimension finie de $\mathcal{H}$. Le symbole $T/E^\perp$ désigne la restriction de T à ce sous-espace,

$$T/E^\perp : E^\perp \to \mathcal{H}. \tag{7}$$

Les opérateurs qui satisfont la condition (5) sont les *opérateurs compacts*, i.e. sont caractérisés par la compacité normique de l'image de la boule unité de $\mathcal{H}$. L'opérateur T est compact ssi $|T| = \sqrt{T^*T}$ est compact, et ceci a lieu ssi le spectre de $|T|$ est une suite de valeurs propres $\mu_0 \geq \mu_1 \geq \mu_2 \ldots$, $\mu_n \downarrow 0$.

Ces valeurs propres sont les valeurs caractéristiques de T et on a,

$$\mu_n(T) = \operatorname{Inf} \{\|T - R\| \; ; \; R \text{ opérateur de rang } \leq n\} \tag{8}$$

$$\mu_n(T) = \operatorname{Inf} \{\|T/E^\perp\| \; ; \; \dim E \leq n\}. \tag{9}$$

Les opérateurs compacts forment un idéal bilatère $\mathcal{K}$ dans l'algèbre $\mathcal{L}(\mathcal{H})$ des opérateurs bornés dans $\mathcal{H}$ de sorte que les règles algébriques élémentaires du calcul infinitésimal sont vérifiées.

La taille d'un infinitésimal $T \in \mathcal{K}$ est gouvernée par l'ordre de décroissance de la suite $\mu_n = \mu_n(T)$, quand $n \to \infty$. En particulier pour tout réel positif α la condition,

$$\mu_n(T) = O(n^{-\alpha}) \qquad \text{quand } n \to \infty \tag{10}$$

(i.e. il existe $C > 0$ tel que $\mu_n(T) \leq C n^{-\alpha} \quad \forall n \geq 1$) définit les infinitésimaux d'ordre α. Ils forment de même un idéal bilatère, comme on le voit en utilisant (8), (cf. [Co]) et de plus,

$$T_j \text{ d'ordre } \alpha_j \Rightarrow T_1 T_2 \text{ d'ordre } \alpha_1 + \alpha_2 . \tag{11}$$

(Pour $\alpha < 1$ l'idéal correspondant est un idéal normé obtenu par interpolation réelle entre l'idéal $\mathcal{L}^1$ des opérateurs traçables et l'idéal $\mathcal{K}$ ([Co]).) Ainsi, hormis la commutativité, les propriétés intuitives du calcul infinitésimal sont vérifiées.

Comme la taille d'un infinitésimal est mesurée par une suite $\mu_n \to 0$ il pourrait sembler inutile d'utiliser le formalisme opératoriel. Il suffirait de remplacer l'idéal

$\mathcal{K}$ de $\mathcal{L}(\mathcal{H})$ par l'idéal $c_0(\mathbb{N})$ des suites convergeant vers 0 dans l'algèbre $\ell^\infty(\mathbb{N})$ des suites bornées. Cette version commutative ne convient pas car tout élément de $\ell^\infty(\mathbb{N})$ a un spectre ponctuel et une mesure spectrale discrète. Ce n'est que la *non commutativité* de $\mathcal{L}(\mathcal{H})$ qui permet la coexistence de variables ayant un spectre de Lebesgue avec des variables infinitésimales.

En fait la ligne suivante du dictionnaire utilise de manière cruciale la non commutativité de $\mathcal{L}(\mathcal{H})$. La différentielle df d'une variable réelle ou complexe,

$$df = \Sigma \frac{\partial f}{\partial x^\mu} dx^\mu \tag{12}$$

est remplacée par le commutateur,

$$đf = [F, f]\,. \tag{13}$$

Le passage de (11) à (12) est semblable à la transition du crochet de Poisson $\{f, g\}$ de deux observables f, g de la mécanique classique, au commutateur $[f, g] = fg - gf$ d'observables quantiques.

Etant donnée une algèbre $\mathcal{A}$ d'opérateurs dans $\mathcal{H}$ la *dimension* de l'espace correspondant (au sens du dictionnaire 1) est gouvernée par la taille des différentielles $đf$, $f \in \mathcal{A}$. En dimension p on a

$$đf \text{ d'ordre } \frac{1}{p}\,, \ \forall f \in \mathcal{A}\,. \tag{14}$$

Nous verrons très vite des exemples concrets où p est la dimension de Hausdorff d'un ensemble de Julia. Des manipulations algébriques très simples sur la fonctionnelle

$$\varphi(f^0, \ldots, f^n) = \mathrm{Trace}\,(f^0\, đf^1 \ldots đf^n) \qquad n \text{ impair, } n > p \tag{15}$$

montrent que φ est un cocycle cyclique et permettent de transposer les idées de la topologie différentielle en exploitant *l'intégralité* du cocycle φ, i.e. $\langle \varphi, K_1(\mathcal{A})\rangle \subset \mathbb{Z}$.

Si le dictionnaire s'arrêtait là, il nous manquerait un outil vital du calcul infinitésimal, la *localité*, i.e. la possibilité de négliger les infinitésimaux d'ordre > 1 dans un calcul. Dans notre cadre les infinitésimaux d'ordre > 1 sont contenus dans l'idéal bilatère suivant,

$$\left\{T \in \mathcal{K}\ ;\ \mu_n(T) = o\left(\frac{1}{n}\right)\right\} \tag{16}$$

où le petit o à la signification usuelle.

Ainsi, si nous utilisons la trace comme dans (15) pour intégrer nous rencontrons deux problèmes,

a) Les infinitésimaux d'ordre 1 ne sont pas dans le domaine de la trace,

b) La trace des infinitésimaux d'ordre > 1 n'est pas nulle.

Le domaine naturel de la trace est l'idéal bilatère $\mathcal{L}^1(\mathcal{H})$ des opérateurs traçables

$$\mathcal{L}^1 = \left\{ T \in \mathcal{K} \; ; \; \sum_o^\infty \mu_n(T) < \infty \right\} . \tag{17}$$

La trace d'un opérateur $T \in \mathcal{L}^1(\mathcal{H})$ est donnée par la somme,

$$\mathrm{Trace}\,(T) = \sum \langle T\xi_i, \xi_i \rangle \tag{18}$$

indépendamment du choix de la base orthonormale (ξ_i) de $\mathcal{H}$. On a

$$\mathrm{Trace}\,(T) = \sum_o^\infty \mu_n(T) \qquad \forall\, T \geq 0 . \tag{19}$$

Soit $T \geq 0$ un infinitésimal d'ordre 1, le seul contrôle sur $\mu_n(T)$ est

$$\mu_n(T) = O\left(\frac{1}{n}\right) \tag{20}$$

ce qui ne suffit pas pour assurer la finitude de (19). Ceci précise la nature du problème a) et de même pour b) puisque la trace ne s'annule pas pour le plus petit idéal de $\mathcal{L}(\mathcal{H})$, l'idéal $\mathcal{R}$ des opérateurs de rang fini.

Ces deux problèmes sont résolus par la trace de Dixmier [Dx], i.e. par l'analyse suivante de la divergence logarithmique des traces partielles,

$$\mathrm{Trace}_N(T) = \sum_o^{N-1} \mu_n(T) \ , \ T \geq 0 . \tag{21}$$

Il est utile de définir $\mathrm{Trace}_\Lambda(T)$ pour tout $\Lambda > 0$ par la formule d'interpolation

$$\mathrm{Trace}_\Lambda(T) = \mathrm{Inf}\, \{ ||A||_1 + \Lambda ||B|| \ ; \ A + B = T \} \tag{22}$$

où $||A||_1$ est la norme $\mathcal{L}^1$ de A, $||A||_1 = \mathrm{Trace}\,|A|$. Cette formule coïncide avec (21) pour Λ entier et donne l'interpolation affine par morceaux. On a de plus, ([Co])

$$\mathrm{Trace}_\Lambda(T_1 + T_2) \leq \mathrm{Trace}_\Lambda(T_1) + \mathrm{Trace}_\Lambda(T_2) \qquad \forall\, \Lambda \tag{23}$$

$$\mathrm{Trace}_{\Lambda_1+\Lambda_2}(T_1+T_2) \geq \mathrm{Trace}_{\Lambda_1}(T_1) + \mathrm{Trace}_{\Lambda_2}(T_2) \qquad \forall\, \Lambda_1, \Lambda_2 \tag{24}$$

où T_1, T_2 sont *positifs* pour (24).

Soit $T > 0$ infinitésimal d'ordre 1 on a alors

$$\mathrm{Trace}_\Lambda(T) \leq C \log \Lambda \tag{25}$$

et la propriété remarquable *d'additivité asymptotique* du coefficient de la divergence logarithmique (25) est la suivante: $(T_j \geq 0)$,

$$|\tau_\Lambda(T_1+T_2) - \tau_\Lambda(T_1) - \tau_\Lambda(T_2)| \leq 3C\, \frac{\log(\log \Lambda)}{\log \Lambda} \tag{26}$$

où pour tout $T \geq 0$ on pose,

$$\tau_\Lambda(T) = \frac{1}{\log \Lambda} \int_e^\Lambda \frac{\mathrm{Trace}_\mu(T)}{\log \mu}\, \frac{d\mu}{\mu} \tag{27}$$

qui est la moyenne de Cesaro sur le groupe $\mathbb{R}_+^*$ des échelles, de la fonction $\frac{\mathrm{Trace}_\mu(T)}{\log \mu}$.

Il résulte facilement de (26) que toute limite simple τ des fonctionnelles non linéaires τ_Λ définit une trace positive et linéaire sur l'idéal bilatère des infinitésimaux d'ordre 1,

$$\tau(\lambda_1 T_1 + \lambda_2 T_2) = \lambda_1\, \tau(T_1) + \lambda_2\, \tau(T_2) \qquad \forall\, \lambda_j \in \mathbb{C}$$

$$\tau(T) \geq 0 \qquad \forall\, T \geq 0 \tag{28}$$

$$\tau(ST) = \tau(TS) \qquad \forall\, S \text{ borné}$$

$$\tau(T) = 0 \text{ si } \mu_n(T) = o\left(\frac{1}{n}\right).$$

En pratique le choix du point limite τ est sans importance car dans tous les exemples importants (et en particulier comme corollaire des axiomes dans le cadre général, cf. Section IV) la condition suivante de *mesurabilité* est satisfaite:

$$\tau_\Lambda(T) \text{ est convergent quand } \Lambda \to \infty\,. \tag{29}$$

Pour les opérateurs mesurables la valeur de $\tau(T)$ est indépendante de τ et est notée

$$⨍ T\,. \tag{30}$$

Le premier exemple intéressant est celui des opérateurs pseudodifférentiels T sur une variété différentiable M. Quand T est d'ordre 1 (au sens de (20)) il est mesurable et $⨍ T$ est le résidu non commutatif de T ([Wo], [Ka]). Ce résidu a une expression locale très simple en terme du noyau distribution $k(x,y)$, $x,y \in M$. Quand T est d'ordre 1 (au sens de (20)) le noyau $k(x,y)$ admet une divergence logarithmique au voisinage de la diagonale,

$$k(x,y) = a(x) \log |x-y| + 0(1) \tag{31}$$

où $|x-y|$ est la distance Riemannienne dont le choix n'affecte pas la 1-densité $a(x)$. On a alors (à normalisation près),

$$⨍ T = \int_M a(x) \tag{32}$$

et le terme de droite de cette formule se prolonge de manière quasi évidente à tous les opérateurs pseudodifférentiels (cf. [Wo]) si l'on note que le noyau d'un tel opérateur admet un développement asymptotique de la forme,

$$k(x,y) = \sum a_k(x, x-y) + a(x) \log |x-y| + 0(1) \tag{33}$$

où $a_k(x,\xi)$ est homogène de degré $-k$ en la variable ξ, et où la 1-densité $a(x)$ est définie de manière intrinsèque.

En fait le même principe de prolongement de $⨍$ à des infinitésimaux d'ordre < 1 s'applique aux opérateurs hypoelliptiques et plus généralement (cf. Théorème 4) aux triplets spectraux dont le spectre de dimension est simple.

Après cette description passons à des exemples. La variable infinitésimale $dp(x)$ qui donne la probabilité dans le jeu de fléchettes (Figure 2) est donnée par l'opérateur,

$$dp = \Delta^{-1} \tag{34}$$

où Δ est le Laplacien de Dirichlet pour le domaine Ω. Il agit dans l'espace de Hilbert $L^2(\Omega)$ ainsi que l'algèbre des fonctions $f(x_1,x_2)$, $f : \Omega \to \mathbb{C}$, qui agissent par opérateurs de multiplication (cf. (3)). Le théorème de H. Weyl montre immédiatement que dp est d'ordre 1, que $f\,dp$ est mesurable et que

$$⨍ f\,dp = \int_\Omega f(x_1,x_2)\, dx_1 \wedge dx_2 \tag{35}$$

donne la probabilité usuelle.

Montrons maintenant comment utiliser notre calcul infinitésimal pour donner un sens à des expressions telles que l'aire d'une variété de dimension 4, qui sont dépourvues de sens dans le calcul usuel.

Il y a, à équivalence unitaire et multiplicité près, une seule quantification du calcul infinitésimal sur $\mathbb{R}$ qui soit invariante par translations et dilatations. Elle est donnée par la représentation de l'algèbre des fonctions f sur $\mathbb{R}$ comme opérateurs de multiplication dans $L^2(\mathbb{R})$ (cf. (3)), alors que l'opérateur F dans $\mathcal{H} = L^2(\mathbb{R})$ est la transformation de Hilbert ([St])

$$(f\xi)(s) = f(s)\,\xi(s) \qquad \forall s \in \mathbb{R} \, , \ \xi \in L^2(\mathbb{R}) \, , \ (F\xi)(t) = \frac{1}{\pi i} \int \frac{\xi(s)}{s-t}\, ds \, . \tag{36}$$

On a une description unitairement équivalente pour $S^1 = P_1(\mathbb{R})$ avec $\mathcal{H} = L^2(S^1)$ et

$$F\, e_n = \mathrm{Sign}\,(n)\, e_n \, , \ e_n(\theta) = \exp\,(in\theta) \qquad \forall \theta \in S^1 \, , \ (\mathrm{Sign}\,(0) = 1) \, . \tag{37}$$

L'opérateur $\text{đ} f = [F, f]$, pour $f \in L^\infty(\mathbb{R})$, est représenté par le noyau $\frac{1}{\pi i}\, k(s,t)$, avec

$$k(s,t) = \frac{f(s) - f(t)}{s-t} \, . \tag{38}$$

Comme f et F sont des opérateurs bornés il en est de même de $\text{đ} f = [F, f]$ pour toute f mesurable bornée sur S^1, ce qui donne un sens à $|\text{đ} f|^p$ pour tout $p > 0$. Soient par exemple $c \in \mathbb{C}$ et J l'ensemble de Julia associé à l'itération de la transformation

$$\varphi(z) = z^2 + c \, , \ J = \partial B \, , \ B = \{z \in \mathbb{C} \ ; \ \mathop{\mathrm{Sup}}_{n \in \mathbb{N}} |\varphi^n(z)| < \infty\} \, . \tag{39}$$

Pour c petit J est une courbe de Jordan et B la composante bornée de son complément. Soit $Z : S^1 \to J$ la restriction à $S^1 = \partial D$, $D = \{z \in \mathbb{C}, \ |z| < 1\}$ d'une équivalence conforme $D \sim B$. Comme (par un résultat de D. Sullivan) la dimension de Hausdorff p de J est > 1 (pour $c \neq 0$) la fonction Z n'est nulle part à variation bornée et la valeur absolue $|Z'|$ de la dérivée de Z au sens des distributions n'a pas de sens. Cependant $|\text{đ} Z|$ est bien défini et on a:

Théorème 1. *a) $|\text{đ}Z|$ est un infinitésimal d'ordre $\frac{1}{p}$. b) Pour toute fonction continue h sur J, l'opérateur $h(Z)\,|\text{đ}Z|^p$ est mesurable. c)* $\exists\,\lambda > 0$,

$$\fint h(Z)\,|\text{đ}Z|^p = \lambda \int h\,d\Lambda_p \qquad \forall\,h \in C(J)$$

où $d\Lambda_p$ désigne la mesure de Hausdorff sur J.

L'énoncé a) utilise un résultat de V.V. Peller qui caractérise les fonctions f pour lesquelles $\mathrm{Trace}\,(|\text{đ}f|^\alpha) < \infty$. La constante λ gouverne le développement asymptotique de la distance dans $L^\infty(S^1)$ entre Z et les fonctions rationnelles ayant au plus n-pôles hors du disque unité. Cette constante est de l'ordre de $\sqrt{p-1}$ et s'annule pour $p = 1$. Cela tient à une propriété spécifique de la dimension 1, à savoir que pour $f \in C^\infty(S^1)$ $\text{đ}f$ n'est pas seulement d'ordre $(\dim S^1)^{-1} = 1$ mais est traçable, avec,

$$\mathrm{Trace}\,(f^0\,\text{đ}f^1) = \frac{1}{\pi i}\int_{S_1} f^0\,df^1 \qquad \forall\,f^0, f^1 \in C^\infty(S^1)\,. \tag{40}$$

En fait par un résultat classique de Kronecker $\text{đ}f$ est de rang fini ssi f est une fraction rationnelle (cf. [P]).

Le calcul différentiel quantique s'applique de la même manière à l'espace projectif $P_1(K)$ sur un corps local arbitraire K (i.e. un corps localement compact non discret) et est invariant par le groupe des transformations projectives. Les cas spéciaux $K = \mathbb{C}$ et $K = \mathbb{H}$ (corps des quaternions) sont des cas particuliers du calcul sur les variétés compactes conformes orientées de dimension paire, $M = M_{2n}$, qui se définit ainsi:

$$\mathcal{H} = L^2(M, \Lambda^n T^*)\ ,\ (f\xi)(p) = f(p)\,\xi(p) \qquad \forall\,f \in L^\infty(M)\ ,\ F = 2P - 1 \tag{41}$$

où le produit scalaire sur l'espace de Hilbert des formes différentielles de degré $n = \frac{1}{2}\dim M$ est donné par $\langle \omega_1, \omega_2\rangle = \int \omega_1 \wedge *\,\omega_2$ et ne dépend que de la structure conforme de M. L'opérateur P est le projecteur orthogonal sur le sous-espace des formes exactes.

Prenons d'abord $n = 1$. Un calcul immédiat donne

$$\fint \text{đ}f\,\text{đ}g = -\frac{1}{\pi}\int df \wedge *\,dg \qquad \forall\,f, g \in C^\infty(M)\,. \tag{42}$$

Soit alors X une application (C^∞) de M dans l'espace $\mathbb{R}^N$ muni de la métrique Riemannienne $g_{\mu\nu}\,dx^\mu\,dx^\nu$, on a

$$\fint g_{\mu\nu}\,\text{đ}X^\mu\,\text{đ}X^\nu = -\frac{1}{\pi}\int_M g_{\mu\nu}\,dX^\mu \wedge *\,dX^\nu \tag{43}$$

où le terme de droite est l'action de Polyakov de la théorie des cordes. Pour $n = 4$ l'égalité (43) n'a pas lieu, l'action définie par le terme de droite n'est pas intéressante car elle n'est pas invariante conforme. Le terme de gauche est parfaitement défini par le calcul quantique et est invariant conforme, on a,

Théorème 2. *Soit X une application C^∞ de M_4 dans $(\mathbb{R}^N, g_{\mu\nu}\, dx^\mu\, dx^\nu)$,*

$$\fint g_{\mu\nu}(X)\, đX^\mu\, đX^\nu = (16\pi^2)^{-1} \int_M g_{\mu\nu}(X)$$

$$\left\{ \frac{1}{3}\, r\langle dX^\mu, dX^\nu\rangle - \Delta\langle dX^\mu, dX^\nu\rangle + \langle \nabla dX^\mu, \nabla dX^\nu\rangle - \frac{1}{2}\,(\Delta X^\mu)\,(\Delta X^\nu) \right\} dv$$

où pour écrire le terme de droite on utilise sur M une structure Riemannienne η quelconque compatible avec la structure conforme. Ainsi la courbure scalaire r, le Laplacien Δ et la connection de Levi Civita ∇ se réfèrent à η, mais le résultat n'en dépend pas.

Le Théorème 2 est à rapprocher de la formule suivante qui exprime l'action de Hilbert Einstein comme l'aire d'une variété de dimension 4 (cf. [Kas] [K-W])

$$\fint ds^2 = \frac{-1}{96\pi^2} \int_{M_4} r\,\sqrt{g}\; d^4x \tag{44}$$

($dv = \sqrt{g}\; d^4x$ est la forme volume et $ds = D^{-1}$ est l'élément de longueur, i.e. l'inverse de l'opérateur de Dirac).

Quand la métrique $g_{\mu\nu}\, dx^\mu\, dx^\nu$ sur $\mathbb{R}^N$ est invariante par translations, la fonctionnelle d'action du Théorème 2 est donnée par l'opérateur de Paneitz sur M. C'est un opérateur d'ordre 4 qui joue le rôle du Laplacien en géométrie conforme ([B-O]). Son anomalie conforme a été calculée par T. Branson [B].

Reprenons le cas $n = 2$ et modifions la structure conforme de M par une différentielle de Beltrami $\mu(z,\bar z)\, d\bar z/dz$, $|\mu(z,\bar z)| < 1$ en utilisant pour définir les angles en $z \in M$

$$X \in T_z(M) \to \langle X, dz + \mu(z,\bar z)\, d\bar z\rangle \in \mathbb{C} \tag{45}$$

au lieu de $\langle X, dz\rangle$. Le calcul quantique sur M associé à la nouvelle structure conforme s'obtient simplement en remplaçant l'opérateur F par l'opérateur F',

$$F' = (\alpha F + \beta)(\beta F + \alpha)^{-1}\ ,\ \alpha = (1 - m^2)^{-1/2}\ ,\ \beta = m(1 - m^2)^{-1/2} \tag{46}$$

où m est l'opérateur dans $\mathcal{H} = L^2(M, \Lambda^1 T^*)$ donné par l'endomorphisme du fibré vectoriel $\Lambda^1 T^* = \Lambda^{(1,0)} \oplus \Lambda^{(0,1)}$ de matrice,

$$m(z,\bar{z}) = \begin{bmatrix} 0 & \bar{\mu}(z,\bar{z})\, d\bar{z}/dz \\ \mu(z,\bar{z})\, dz/d\bar{z} & 0 \end{bmatrix} . \tag{47}$$

Les propriétés cruciales de l'opérateur $m \in \mathcal{L}(\mathcal{H})$ sont

$$\|m\| < 1 \ , \ m = m^* \ , \ m f = f m \qquad \forall f \in \mathcal{A} = C^\infty(M) \tag{48}$$

et la déformation (46) de F est un cas particulier de

Proposition 3. *Soient $\mathcal{A}$ une algèbre involutive d'opérateurs dans $\mathcal{H}$ et $N = \mathcal{A}' = \{T \in \mathcal{L}(\mathcal{H})\,;\, Ta = aT \quad \forall\, a \in \mathcal{A}\}$ l'algèbre de von Neumann commutant de $\mathcal{A}$. a) L'égalité suivante définit une action du groupe $G = GL_1(N)$ des éléments inversibles de N sur les opérateurs F, $F = F^*$, $F^2 = 1$*

$$g(F) = (\alpha F + \beta)\,(\beta F + \alpha)^{-1} \qquad \forall g \in G$$

où $\alpha = \frac{1}{2}(g - (g^{-1})^*)$, $\beta = \frac{1}{2}(g + (g^{-1})^*)$.
b) On a $[g(F), a] = Y[F, a]\,Y^* \quad \forall\, a \in \mathcal{A}$, où $Y = (\beta F + \alpha)^{*-1}$.

L'égalité b) montre que pour tout idéal bilatère $J \subset \mathcal{L}(\mathcal{H})$ la condition

$$[F, a] \in J \tag{49}$$

est préservée par la déformation $F \to g(F)$. Comme seule la *mesurabilité* de la différentielle de Beltrami μ est requise pour que m vérifie (48), seule la mesurabilité de la structure conforme sur M est requise pour que le calcul quantique associé soit défini. De plus b) montre que la condition de régularité sur $a \in L^\infty(M)$ définie par (49) ne dépend que de la structure quasiconforme de la variété M ([CST]). Un homéomorphisme local φ de $\mathbb{R}^n$ est *quasiconforme* ssi il existe $K < \infty$ tel que

$$H_\varphi(x) = \operatorname*{Lim\,sup}_{r \to 0} \ \frac{\max|\varphi(x) - \varphi(y)|\,;\, |x-y| = r}{\min|\varphi(x) - \varphi(y)|\,;\, |x-y| = r} \leq K \ , \ \forall\, x \in \text{Domaine}\,\varphi . \tag{50}$$

Une structure quasiconforme sur une variété topologique M_n est donnée par un atlas quasiconforme. La discussion ci-dessus s'applique au cas général (n pair) ([CST]) et montre que le calcul quantique est bien défini pour toute variété quasiconforme. Le résultat de D. Sullivan [S] et S. Novikov [N] montre que toute variété topologique M_n, $n \neq 4$ admet une structure quasiconforme. En utilisant le calcul quantique et la cohomologie cyclique à la place du calcul différentiel et de la théorie de Chern Weil on obtient ([CST]) une formule locale pour les classes de Pontrjagin topologiques de M_n.

3 La formule de l'indice locale et la classe fondamentale transverse

Nous montrons dans cette section que le calcul infinitésimal ci-dessus permet le passage du local au global dans le cadre général des triplets spectraux $(\mathcal{A}, \mathcal{H}, D)$. Nous appliquons ensuite le résultat général au produit croisé d'une variété par le groupe des difféomorphismes.

Nous ferons l'hypothèse de régularité suivante sur $(\mathcal{A}, \mathcal{H}, D)$

$$a \text{ et } [D,a] \in \cap \operatorname{Dom} \delta^k \ , \ \forall \, a \in \mathcal{A} \tag{1}$$

où δ est la dérivation $\delta(T) = [|D|, T]$.

Nous désignerons par $\mathcal{B}$ l'algèbre engendrée par les $\delta^k(a)$, $\delta^k([D,a])$. La *dimension* d'un triplet spectral est bornée supérieurement par $p > 0$ ssi $a(D+i)^{-1}$ est un infinitésimal d'ordre $\frac{1}{p}$ pour tout $a \in \mathcal{A}$. Quand $\mathcal{A}$ est unifère cela ne dépend que du spectre de D.

La notion précise de dimension est définie comme le sous-ensemble $\Sigma \subset \mathbb{C}$ des singularités des fonctions analytiques

$$\zeta_b(z) = \operatorname{Trace}(b|D|^{-z}) \qquad \operatorname{Re} z > p \ , \ b \in \mathcal{B} \, . \tag{2}$$

Nous supposerons que Σ est discret et simple, i.e. que les ζ_b se prolongent à $\mathbb{C}/\Sigma$ avec des pôles simples en Σ.

Nous renvoyons à [CM2] pour le cas de spectre multiple.

L'indice de Fredholm de l'opérateur D détermine une application additive, $K_1(\mathcal{A}) \xrightarrow{\varphi} \mathbb{Z}$ donnée par l'égalité

$$\varphi([u]) = \operatorname{Indice}(PuP) \ , \ u \in GL_1(\mathcal{A}) \tag{3}$$

où P est le projecteur $P = \frac{1+F}{2}$, $F = \operatorname{Signe}(D)$.

Cette application est calculée par l'accouplement entre $K_1(\mathcal{A})$ et la classe de cohomologie du cocycle cyclique suivant

$$\tau(a^0, \ldots, a^n) = \operatorname{Trace}(a^0[F,a^1] \ldots [F,a^n]) \qquad \forall \, a^j \in \mathcal{A} \tag{4}$$

où $F =$ Signe D et où n est un entier impair $n \geq p$.

Le problème est que τ est difficile à déterminer en général car la formule (4) implique la trace ordinaire au lieu de la trace locale $⨍$.

Ce problème est résolu par la formule suivante,

Théorème 4. ([CM2]) *Soit $\mathcal{A}, \mathcal{H}, D)$ un triplet spectral vérifiant les hypothèses (1) et (2). a) L'égalité $\fint P = \mathrm{Res}_{z=0}\,\mathrm{Trace}\,(P|D|^{-z})$ définit une trace sur l'algèbre engendrée par $\mathcal{A}$, $[D, \mathcal{A}]$ et $|D|^z$, $z \in \mathbb{C}$.*

b) La formule suivante n'a qu'un nombre fini de termes non nuls et définit les composantes $(\varphi_n)_{n=1,3,\ldots}$ d'un cocycle dans le bicomplexe (b, B) de $\mathcal{A}$,

$$\varphi_n(a^0, \ldots, a^n) = \sum_k c_{n,k} \fint a^0 [D, a^1]^{(k_1)} \ldots [D, a^n]^{(k_n)}\, |D|^{-n-2|k|} \qquad \forall\, a^j \in \mathcal{A}$$

où l'on note $T^{(k)} = \nabla^k(T)$ et $\nabla(T) = D^2T - TD^2$, et où k est un multiindice, $c_{n,k} = (-1)^{|k|}\sqrt{2i}(k_1! \ldots k_n!)^{-1}\,((k_1+1)\ldots(k_1+k_2+\ldots+k_n+n))^{-1}\,\Gamma\left(|k|+\frac{n}{2}\right)$, $|k| = k_1 + \ldots + k_n$.

c) L'accouplement de la classe de cohomologie cyclique $(\varphi_n) \in HC^*(\mathcal{A})$ avec $K_1(\mathcal{A})$ donne l'indice de Fredholm de D à coefficient dans $K_1(\mathcal{A})$.

Rappelons que le bicomplexe (b, B) est donné par les opérateurs suivants agissant sur les formes multilinéaires sur l'algèbre $\mathcal{A}$,

$$(b\varphi)(a^0, \ldots, a^{n+1}) = \sum_0^n (-1)^j\, \varphi(a^0, \ldots, a^j a^{j+1}, \ldots, a^{n+1}) + (-1)^{n+1}\, \varphi(a^{n+1}a^0, a^1, \ldots, a^n) \qquad (5)$$

$$\begin{aligned} &B = AB_0\ , \\ &B_0\, \varphi(a^0, \ldots, a^{n-1}) = \varphi(1, a^0, \ldots, a^{n-1}) - (-1)^n\, \varphi(a^0, \ldots, a^{n-1}, 1) \\ &(A\psi)(a^0, \ldots, a^{n-1}) = \sum_0^{n-1} (-1)^{(n-1)j}\, \psi(a^j, a^{j+1}, \ldots, a^{j-1})\,. \end{aligned} \qquad (6)$$

Nous renvoyons à [Co] pour la normalisation de l'accouplement entre HC^* et $K(\mathcal{A})$.

Remarques. a) L'énoncé du Théorème 4 reste valable si l'on remplace dans toutes les formules l'opérateur D par $D|D|^\alpha$, $\alpha \geq 0$.

b) Dans le cas pair, c'est-à-dire si l'on suppose que $\mathcal{H}$ et $\mathbb{Z}/2$ gradué par γ, $\gamma = \gamma^*$, $\gamma^2 = 1$, $\gamma a = a\gamma \ \ \forall\, a \in \mathcal{A}$, $\gamma D = -D\gamma$, on a une formule analogue pour un cocycle (φ_n), n pair qui donne l'indice de Fredholm de D à coefficient dans K_0. Cependant la composante φ_0 ne s'exprime pas en terme du résidu $\fint$ car elle est non locale pour $\mathcal{H}$ de dimension finie (cf. [CM2]).

c) Quand le spectre de dimension Σ a de la multiplicité on a une formule analogue mais qui implique un nombre fini de termes correctifs, dont le nombre est borné indépendamment de la multiplicté (cf. [CM2]).

Le spectre de dimensions d'une variété V est $\{0, 1, \ldots, n\}$, $n = \dim V$, et est simple. La multiplicité apparait pour les variétés singulières et les ensembles de Cantor donnent des exemples de points complexes, $z \notin \mathbb{R}$ dans ce spectre. Nous discutons maintenant une construction géométrique générale pour laquelle les hypothèses (1) et (2) sont vérifiées. Il s'agit de construire la classe fondamentale en K-homologie d'une variété K-orientée M sans briser la symétrie du groupe $\mathrm{Diff}^+(M)$ des difféomorphismes de M qui préservent la K-orientation. De manière plus précise nous cherchons un triplet spectral, $(C^\infty(M), \mathcal{H}, D)$ de la même classe de K-homologie que l'opérateur de Dirac associé à une métrique Riemannienne (cf. I (11) et (12)) mais qui soit équivariant par rapport au groupe $\mathrm{Diff}^+(M)$ au sens de [K]. Cela signifie que l'on a une représentation unitaire $\varphi \to U(\varphi)$ de $\mathrm{Diff}^+(M)$ dans $\mathcal{H}$ telle que

$$U(\varphi)\, f\, U(\varphi)^{-1} = f \circ \varphi^{-1} \qquad \forall\, f \in C^\infty(M)\ ,\ \varphi \in \mathrm{Diff}^+(M) \tag{7}$$

et que

$$U(\varphi)\, D\, U(\varphi)^{-1} - D \text{ est borné pour tout } \varphi \in \mathrm{Diff}^+(M)\,. \tag{8}$$

Lorsque D est l'opérateur de Dirac associé à une structure Riemannienne le symbole principal de D détermine cette métrique et les seuls difféomorphismes qui vérifient (8) sont les isométries.

La solution de ce problème est essentielle pour définir la géométrie transverse des feuilletages et elle est effectuée en 2 étapes. La première est l'utilisation ([Co1]) de la métrique de courbure négative de l'espace $GL(n)/0(n)$ et de l'opérateur "dual Dirac" de Miscenko et Kasparov pour se ramener à l'action de $\mathrm{Diff}^+(M)$ sur l'espace total P du fibré des métriques de M. La deuxième, dont l'idée est due à Hilsum et Skandalis ([HS]) est l'utilisation des opérateurs hypoelliptiques pour construire l'opérateur D sur P.

On notera qu'alors que la géométrie équivariante obtenue pour P est de dimension finie et vérifie les hypothèses (1) (2) la géométrie obtenue sur M en utilisant le produit intersection avec le "dual Dirac" est de dimension infinie et θ-sommable,

$$\mathrm{Trace}\,(e^{-\beta D^2}) < \infty \qquad \forall\, \beta > 0\,. \tag{9}$$

Par construction, le fibré $P \xrightarrow{\pi} M$ est le quotient $F/0(n)$ du $GL(n)$ fibré principal F des repères sur M par l'action du groupe orthogonal $0(n) \subset GL(n)$. L'espace P admet la structure canonique suivante: le feuilletage vertical $V \subset TP$, $V = \operatorname{Ker} \pi_*$ et les structures Euclidiennes suivantes sur les fibrés V et $N = (TP)/V$. Le choix d'une métrique Riemannienne $GL(n)$-invariante sur $GL(n)/0(n)$ détermine la métrique sur V et celle de N est la métrique tautologique: $p \in P$ détermine une métrique sur $T_{\pi(p)}(M)$ qui grâce à π_* est isomorphe à N_p.

Cette construction est fonctorielle pour les difféomorphismes de M.

Le calcul hypoelliptique adapté à cette structure est un cas particulier du calcul pseudodifférentiel sur les variétés de Heisenberg ([BG]). Il modifie simplement l'homogénéité des symboles $\sigma(p, \xi)$ en utilisant les homotéties:

$$\lambda \cdot \xi = (\lambda \xi_v, \lambda^2 \xi_n) \ , \ \forall \lambda \in \mathbb{R}_+^* \tag{10}$$

où ξ_v, ξ_n sont les composantes verticales et normales du covecteur ξ. La formule (10) dépend de coordonnées locales (x_v, x_n) adaptées au feuilletage vertical mais le calcul pseudodifférentiel correspondant n'en dépend pas. Le symbole principal d'un opérateur hypoelliptique d'ordre k est une fonction, homogène de degré k pour (10), sur le fibré $V^* \oplus N^*$. Le noyau distribution $k(x, y)$ d'un opérateur pseudodifférentiel T dans le calcul hypoelliptique admet un développement au voisinage de la diagonale de la forme,

$$k(x, y) \sim \sum a_j(x, x - y) + a(x) \log |x - y|' + 0(1) \tag{11}$$

où a_j est homogène de degré $-j$ en $x - y$ pour (10) et où la métrique $|x - y|'$ est localement de la forme

$$|x - y|' = ((x_v - y_v)^4 + (x_n - y_n)^2)^{1/4} \, . \tag{12}$$

Comme dans le calcul pseudodifférentiel ordinaire, le résidu se prolonge aux opérateurs de tout degré et est donné par l'égalité,

$$⨍ T = \frac{1}{v + 2m} \int a(x) \tag{13}$$

où la 1-densité $a(x)$ ne dépend pas du choix de la métrique $| \ |'$ et où $v = \dim V$, $m = \dim N$ de sorte que $v + 2m$ est la dimension de Hausdorff de l'espace métrique $(P, | \ |')$.

L'opérateur D est défini par l'équation $D|D| = Q$ où Q est l'opérateur différentiel hypoelliptique de degré 2 obtenu en combinant (quand v est pair) l'opérateur $d_V d_V^* - d_V^* d_V$ de signature où d_V est la différentiation verticale, avec l'opérateur de Dirac transverse. (On utilise le revêtement métaplectique $M\ell(n)$ de $GL(n)$ pour définir la structure spinorielle sur M.) La formule explicite de Q utilise une connection affine sur M mais le choix de cette connection n'affecte pas le *symbole principal hypoelliptique* de Q et donc de D ce qui assure l'invariance (8) de D par rapport aux difféomorphismes de M.

Donnons la formule explicite de Q dans le cas $n = 1$, i.e. pour $M = S^1$. On remplace P par la suspension $SP = \mathbb{R} \times P$ pour se ramener au cas où la dimension verticale v est paire. Un point de $SP = \mathbb{R} \times P$ est paramétré par 3 coordonnées $\alpha \in \mathbb{R}$ et $p = (s, \theta)$ où $\theta \in S^1$ et où $s \in \mathbb{R}$ définit la métrique $e^{2s}(d\theta)^2$ en $\theta \in S^1$. On munit SP de la mesure $\nu = d\alpha\, ds\, d\theta$ et l'on représente l'algèbre $C_c^\infty(SP)$ par opérateurs de multiplication dans $\mathcal{H} = L^2(SP, \nu) \otimes \mathbb{C}^2$. La fonctorialité de la construction ci-dessus donne la repésentation unitaire suivante du groupe $\mathrm{Diff}^+(S^1)$,

$$(U(\varphi)^{-1}\xi)(\alpha, s, \theta) = \varphi'(\theta)^{1/2}\, \xi(\alpha, s - \log\varphi'(\theta), \varphi(\theta)) \,. \qquad (14)$$

Enfin l'opérateur Q est donné par la formule,

$$Q = -2\partial_\alpha \partial_s \sigma_1 + \frac{1}{i}\, e^{-s} \partial_\theta \sigma_2 + \left(\partial_s^2 - \partial_\alpha^2 - \frac{1}{4} \right) \sigma_3 \qquad (15)$$

où $\sigma_1, \sigma_2, \sigma_3 \in M_2(\mathbb{C})$ sont les 3 matrices de Pauli.

L'opérateur ∂_θ est de *degré 2* dans le calcul hypoelliptique et l'on vérifie que Q est hypoelliptique.

Un long calcul donne le résultat suivant ([CM3]):

Théorème 5. *Soit $\mathcal{A}$ l'algèbre produit croisé de $C_c^\infty(SP)$ par $\mathrm{Diff}^+(S^1)$. a) Le triplet spectral $(\mathcal{A}, \mathcal{H}, D)$ (où $\mathcal{A}$ agit dans $\mathcal{H}$ par (14) et $D|D| = Q$) satisfait les hypothèses (1) et (2) et son spectre de dimension est $\Sigma = \{0, 1, 2, 3, 4\}$.*

b) La seule composante non nulle du cocycle associé (Théorème 4) est φ_3 et elle est cohomologue à 2ψ où ψ est le 3-cocycle cyclique classe fondamentale transverse du produit croisé.

L'intégralité de 2ψ, i.e. de l'accouplement $\langle 2\psi, K_1(\mathcal{A})\rangle$ résulte alors du Théorème 4.

Le 3-cocycle ψ est donné par (cf. [Co])

$$\begin{aligned}
&\psi(f^0U(\varphi_0)\,,\ f^1U(\varphi_1)\,,\ f^2U(\varphi_2)\,,\ f^3U(\varphi_3)) = 0 && (16)\\
&\text{sauf si } \varphi_0\varphi_1\varphi_2\varphi_3 = 1 \quad \text{et} \quad = \int h^0\,dh^1 \wedge dh^2 \wedge dh^3 \text{ si } \varphi_0\varphi_1\varphi_2\varphi_3 = 1\\
&\text{avec} \quad h^0 = f^0\,,\ h^1 = (f^1)^{\varphi_0}\,,\ h^2 = (f^2)^{\varphi_0\varphi_1}\,,\ h^3 = (f^3)^{\varphi_0\varphi_1\varphi_2}\,.
\end{aligned}$$

L'homogie entre φ_3 et 2ψ met en évidence l'action sur l'algèbre $\mathcal{A}$ de l'algèbre de Hopf engendrée par les transformations linéaires suivantes (pour la relation de δ_3 avec l'invariant de Godbillon Vey, voir [Co]) de $\mathcal{A}$

$$\begin{aligned}
\delta_1(fU(\varphi)) = (\partial_\alpha f)\,U(\varphi)\,,\ \delta_2(fU(\varphi)) = (\partial_s f)\,U(\varphi)\,, \qquad (17)\\
\delta_3(fU(\varphi)) = f\,e^{-s}\,\partial_\theta \log(\varphi^{-1})'\,U(\varphi)\,,\ X(fU(\varphi)) = e^{-s}(\partial_\theta f)\,U(\varphi)
\end{aligned}$$

dont la compatibilité avec la multiplication de $\mathcal{A}$ est régie par le coproduit

$$\Delta\delta_j = \delta_j \otimes 1 + 1 \otimes \delta_j \qquad j = 1,2,3 \qquad (18)$$

(i.e. les δ_j sont des dérivations de $\mathcal{A}$)

$$\Delta X = X \otimes 1 + 1 \otimes X - \delta_3 \otimes \delta_2 \qquad (19)$$

où (19) montre que X est de degré 2.

4 La notion de variété et les axiomes de la géométrie.

Commençons par spécifier la place de la géométrie Riemannienne dans notre cadre en caractérisant (Théorème 6) les triplets spectraux correspondants. Soit $n \in \mathbb{N}$ la dimension, le triplet $(\mathcal{A}, \mathcal{H}, D)$ est supposé $\mathbb{Z}/2$ gradué par γ, $\gamma = \gamma^*$, $\gamma^2 = 1$ quand n est pair.

Les axiomes *commutatifs* sont les suivants:

1) (Dimension) $ds = D^{-1}$ est infinitésimal d'ordre $\frac{1}{n}$.

2) (Ordre un) $[[D, f], g] = 0 \quad \forall\, f, g \in \mathcal{A}$.

3) (Régularité) Pour tout $f \in \mathcal{A}$, f et $[D, f]$ appartiennent à $\bigcap\limits_k$ Domaine δ^k, où δ est la dérivation $\delta(T) = [|D|, T]$.

4) (Orientabilité) Il existe un cycle de Hochschild $c \in Z_n(\mathcal{A}, \mathcal{A})$ tel que $\pi(c) = 1$ (n impair) ou $\pi(c) = \gamma$ (n pair), où π: $\mathcal{A}^{\otimes(n+1} \to \mathcal{L}(\mathcal{H})$ est l'unique application linéaire telle que $\pi(a^0 \otimes a^1 \otimes \cdots \otimes a^n) = a^0[D, a^1]\ldots[D, a^n] \quad \forall\, a^j \in \mathcal{A}$.

5) (Finitude) Le $\mathcal{A}$-module $\mathcal{E} = \bigcap\limits_k$ Domaine D^k est projectif de type fini et l'égalité suivante définit une structure hermitienne sur $\mathcal{E}$,

$$\langle a\xi, \eta\rangle = \int\!\!\!\!\!- a(\xi,\eta)\, ds^n \qquad \forall\, \xi,\eta \in \mathcal{E}\ ,\ a \in \mathcal{A}\,.$$

6) (Dualité de Poincaré) La forme d'intersection $K_*(\mathcal{A}) \times K_*(\mathcal{A}) \to \mathbb{Z}$ donnée par la composition de l'indice de Fredholm de D avec la diagonale, $m_* : K_*(\mathcal{A}) \times K_*(\mathcal{A}) \to K_*(\mathcal{A} \otimes \mathcal{A}) \to K_*(\mathcal{A})$, est *inversible.*

7) (Réalité) Il existe une isométrie antilinéaire J sur $\mathcal{H}$ telle que $Ja^*J^{-1} = a$ $\forall a \in \mathcal{A}$ et $J^2 = \varepsilon$, $JD = \varepsilon' DJ$, $J\gamma = \varepsilon''\gamma J$ où la table des valeurs de $\varepsilon, \varepsilon', \varepsilon'' \in \{-1, 1\}$ en fonction de n modulo 8 est donnée en (16).

Les axiomes 2) et 4) donnent la présentation de l'algèbre abstraite notée $(\mathcal{A}, ds)$ engendrée par $\mathcal{A}$ et $ds = D^{-1}$.

Théorème 6. *Soit $\mathcal{A} = C^\infty(M)$ où M est une variété compacte de classe C^∞.*
a) Soit π une représentation unitaire de $(\mathcal{A}, ds)$ satisfaisant les conditions 1) à 7). Il existe alors une unique structure Riemannienne g sur M telle que la distance géodésique soit donnée par,

$$d(x,y) = \mathrm{Sup}\,\{|a(x) - a(y)|\ ;\ a \in \mathcal{A}\ ,\ \|[D,a]\| \leq 1\}\,.$$

b) La métrique $g = g(\pi)$ ne dépend que de la classe d'équivalence unitaire de π et les fibres de l'application {classe d'équivalence unitaire} $\to g(\pi)$ forment un nombre fini d'espaces affines $\mathcal{A}_\sigma$ paramétrés par les structures spinorielles σ de M.
c) La fonctionnelle $\int\!\!\!\!- ds^{n-2}$ est quadratique et positive sur chaque $\mathcal{A}_\sigma$ où elle admet un unique minimum π_σ.
d) π_σ est la représentation de $(\mathcal{A}, ds)$ dans $L^2(M, S_\sigma)$ donnée par les opérateurs de multiplication et l'opérateur de Dirac associé à la connection de Levi Civita de la métrique g.
e) La valeur de $\int\!\!\!\!- ds^{n-2}$ en π_σ est l'action de Hilbert Einstein de la métrique g,

$$\int\!\!\!\!\!- ds^{n-2} = -c_n \int r\sqrt{g}\; d^n x\ ,\ c_n = \tfrac{n-2}{12}\,(4\pi)^{-n/2}\,2^{[n/2]}\,\Gamma\left(\tfrac{n}{2}+1\right)^{-1}.$$

L'exemple le plus simple pour comprendre la signification du théorème et de vérifier que la géométrie du cercle S^1 de longueur 2π est entièrement spécifiée par la présentation:

$$U^{-1}[D, U] = 1 \ , \text{ où } UU^* = U^*U = 1 \, . \tag{1}$$

L'algèbre $\mathcal{A}$ étant celle des fonctions C^∞ de l'opérateur unitaire U. On a $S^1 =$ Spectre $(\mathcal{A})$ et l'égalité (1) est le cas le plus simple de l'axiome 4.

Remarques. a) L'hypothèse $\mathcal{A} = C^\infty(M)$ devrait résulter des axiomes 1) - 7) (et de la commutativité de $\mathcal{A}$). Il résulte de 3) et 5) que si $\mathcal{A}''$ est l'algèbre de von Neumann engendrée par $\mathcal{A}$ on a

$$\mathcal{A} = \left\{ T \in \mathcal{A}'' \ ; \ T \in \bigcap_{k>0} \operatorname{Dom} \delta^k \right\} \tag{2}$$

ce qui montre que $\mathcal{A}$ est uniquement spécifiée dans $\mathcal{A}''$ par la donnée de D. Cela montre que $\mathcal{A}$ est stable par calcul fonctionnel C^∞ dans sa fermeture normique $A = \bar{\mathcal{A}}$ et en particulier que

$$\operatorname{Spectre} \mathcal{A} = \operatorname{Spectre} A \, . \tag{3}$$

Soit X cet espace compact, on devrait déduire des axiomes que l'application de X dans $\mathbb{R}^N$ donnée par les $a_i^j \in \mathcal{A}$ qui interviennent dans le cycle de Hochschild c de (4) est un plongement de X comme sous-variété C^∞ de $\mathbb{R}^N$ (cf. Proposition 15 p.312 de [Co]).

b) Rappelons qu'un cycle de Hochschild $c \in Z_n(\mathcal{A}, \mathcal{A})$ est un élément de $\mathcal{A}^{\otimes(n+1)}$, $c = \sum a_i^0 \otimes a_i^1 \ldots \otimes a_i^n$ tel que $bc = 0$, où b est l'application linéaire $b : \mathcal{A}^{\otimes n+1} \to \mathcal{A}^{\otimes n}$ telle que

$$b(a^0 \otimes \ldots \otimes a^n) =$$
$$\sum_0^{n-1} (-1)^j \, a^0 \otimes \ldots \otimes a^j a^{j+1} \otimes \ldots \otimes a^n + (-1)^n \, a^n a^0 \otimes a^1 \otimes \ldots \otimes a^{n+1} \, .$$

La classe de Hochschild du cycle c détermine la *forme volume.*

c) Nous utilisons la convention selon laquelle la courbure scalaire r est positive pour la sphère S^n, en particulier le signe de l'action $⨍ ds^{n-2}$ est le bon pour la formulation Euclidienne de la gravitation. Par exemple pour $n = 4$ l'action de Hilbert Einstein $-\frac{1}{16\pi G} \int r \sqrt{g} \, d^4x$ coïncide avec l'aire $\frac{1}{\ell_p^2} ⨍ ds^2$ en unité de Planck.

d) Quand M est une variété spinorielle l'application $\pi \to g(\pi)$ du théorème est surjective et si l'on fixe le cycle $c \in Z_n(\mathcal{A},\mathcal{A})$ son image est l'ensemble des métriques dont la forme volume est fixée – (b)).

e) Si l'on supprime l'axiome 7 on a un résultat analogue au théorème en remplaçant les structures spinorielles par les structures spin^c ([LM]), mais l'on n'a plus unicité dans c) à cause de la liberté dans le choix de la connection spinorielle.

f) Il résulte de l'axiome 4 et de ([Co] Théorème 8 p.309) que les opérateurs $a\, ds^n$, $a \in \mathcal{A}$ sont automatiquement mesurables de sorte que le symbole $⨍$ qui apparait dans 5 est bien défini.

Passons au cas général non commutatif. Étant donnée une algèbre involutive $\mathcal{A}$ d'opérateurs dans l'espace de Hilbert $\mathcal{H}$ la théorie de Tomita [Ta] associe à tout vecteur $\xi \in \mathcal{H}$ cyclique pour $\mathcal{A}$ et pour son commutant $\mathcal{A}'$,

$$\overline{\mathcal{A}\xi} = \mathcal{H} \ , \ \overline{\mathcal{A}'\xi} = \mathcal{H} \tag{4}$$

une involution antilinéaire isométrique $J : \mathcal{H} \to \mathcal{H}$ obtenue à partir de la décomposition polaire de l'opérateur

$$S\, a\xi = a^*\xi \qquad \forall\, a \in \mathcal{A} \tag{5}$$

et qui vérifie la propriété de commutativité suivante,

$$J\mathcal{A}''J^{-1} = \mathcal{A}' . \tag{6}$$

On a donc en particulier $[a, b^0] = 0 \quad \forall\, a, b \in \mathcal{A}$ où

$$b^0 = Jb^*J^{-1} \qquad \forall b \in \mathcal{A} \tag{7}$$

de sorte que $\mathcal{H}$ devient un $\mathcal{A}$-bimodule en utilisant la représentation de l'algèbre opposée $\mathcal{A}^0$ donnée par (7). Dans le cas commutatif on a $a^0 = a \quad \forall\, a \in \mathcal{A}$ de sorte que l'on ne perçoit pas la nuance entre module et bimodule.

Le théorème de Tomita est l'outil nécessaire pour assurer la substance des axiomes dans le cas général. Les axiomes 1) 3) et 5) son inchangés, dans l'axiome de réalité 7) on remplace l'égalité $Ja^*J^{-1} = a \quad \forall\, a \in \mathcal{A}$ par

$$[a, b^0] = 0 \qquad \forall\, a, b \in \mathcal{A} \text{ où } b^0 = Jb^*J^{-1} \tag{7'}$$

et l'axiome 2) (ordre un) se formule ainsi

$$[[D, a], b^0] = 0 \qquad \forall\, a, b \in \mathcal{A} . \tag{2'}$$

(On notera que comme a et b^0 commutent $2'$ équivaut à $[[D,a^0],b]=0 \quad \forall\, a,b \in \mathcal{A}$.)

L'axiome $(7')$ fait de $\mathcal{H}$ un $\mathcal{A}$-bimodule et donne une classe μ de KR^n-homologie pour l'algèbre $\mathcal{A}\otimes\mathcal{A}^0$ munie de l'automorphisme antilinéaire τ,

$$\tau(x\otimes y^0)=y^*\otimes x^{*0}\,.$$

Le produit intersection de Kasparov [K] permet alors de formuler la dualité de Poincaré, comme l'invertibilité de μ,

$$\exists\,\beta\in KR_n(\mathcal{A}^0\otimes\mathcal{A})\ ,\ \beta\otimes_{\mathcal{A}}\mu=\mathrm{id}_{\mathcal{A}^0}\ ,\ \mu\otimes_{\mathcal{A}^0}\beta=\mathrm{id}_{\mathcal{A}}\ . \qquad (6')$$

Ceci implique l'isomorphisme $K_*(\mathcal{A})\xrightarrow{\cap\mu}K^*(\mathcal{A})$. La forme d'intersection,

$$K_*(\mathcal{A})\times K_*(\mathcal{A})\to\mathbb{Z}$$

est obtenue à partir de l'indice de Fredholm de D à coefficient dans $K_*(\mathcal{A}\otimes\mathcal{A}^0)$ et n'utilise plus l'application diagonale $m:\mathcal{A}\otimes\mathcal{A}\to\mathcal{A}$ qui n'est un homomorphisme que dans le cas commutatif. Cette forme d'intersection est quadratique ou symplectique selon la valeur de n modulo 8.

L'homologie de Hochschild à coefficient dans un bimodule garde tout sons sens dans le cas général et l'axiome 4) prend la forme suivante,

$(4')$ Il existe un cycle de Hochschild $c\in Z_n(\mathcal{A},\mathcal{A}\otimes\mathcal{A}^0)$ tel que $\pi(c)=1$ (n impair) ou $\pi(c)=\gamma$ (n pair).

(Où $\mathcal{A}\otimes\mathcal{A}^0$ est le $\mathcal{A}$ bimodule obtenu par restriction à la sous-algèbre $\mathcal{A}\otimes 1\subset\mathcal{A}\otimes\mathcal{A}^0$ de la structure de $\mathcal{A}\otimes\mathcal{A}^0$ bimodule de $\mathcal{A}\otimes\mathcal{A}^0$, i.e.

$$a(b\otimes c^0)d=abd\otimes c^0 \qquad \forall\, a,b,c,d\in\mathcal{A}\,.)$$

Les axiomes (1), (3) et (5) sont inchangés dans le cas noncommutatif et la démonstration de la mesurabilité des opérateurs $a(ds)^n$, $a\in\mathcal{A}$ reste valable en général.

Nous adopterons les axiomes (1), $(2')$, (3), $(4')$, (5), $(6')$ et $(7')$ dans le cas général comme définition d'une *variété spectrale* de dimension n. L'algèbre $\mathcal{A}$ étant fixée nous parlerons de géométrie spectrale sur $\mathcal{A}$ comme dans 1.20 et 1.21. On démontre que l'algèbre de von Neumann $\mathcal{A}''$ engendrée par $\mathcal{A}$ dans $\mathcal{H}$ est automatiquement finie et hyperfinie et on a la liste complète de ces algèbres à isomorphisme près [Co]. L'algèbre $\mathcal{A}$ est stable par calcul fonctionnel C^∞ dans sa fermeture normique $A = \bar{\mathcal{A}}$ de sorte que $K_j(\mathcal{A}) \simeq K_j(A)$, i.e. $K_j(\mathcal{A})$ ne dépend que de la topologie sous-jacente (définie par la C^* algèbre A). L'entier $\chi = \langle \mu, \beta \rangle \in \mathbb{Z}$ donne la caractéristique d'Euler sous la forme

$$\chi = \operatorname{Rang} K_0(\mathcal{A}) - \operatorname{Rang} K_1(\mathcal{A})$$

et le Théorème 4 en donne une formule locale.

Le groupe $\operatorname{Aut}(\mathcal{A})$ des automorphismes de l'algèbre involutive $\mathcal{A}$ joue en général le rôle du groupe $\operatorname{Diff}(M)$ des difféomorphismes d'une variété M. (On a un isomorphisme canonique $\operatorname{Diff}(M) \overset{\alpha}{\to} \operatorname{Aut}(C^\infty(M))$ donné par

$$\alpha_\varphi(f) = f \circ \varphi^{-1} \qquad \forall f \in C^\infty(M) \ , \ \varphi \in \operatorname{Diff}(M) \, .)$$

Dans le cas général non commutatif, parallèlement au sous-groupe normal $\operatorname{Int} \mathcal{A} \subset \operatorname{Aut} \mathcal{A}$ des automorphismes intérieurs de $\mathcal{A}$,

$$\alpha(f) = u f u^* \qquad \forall f \in \mathcal{A} \tag{8}$$

où u est un élément unitaire de $\mathcal{A}$ (i.e. $uu^* = u^*u = 1$), il existe un feuilletage naturel de l'espace des géométries spectrales sur $\mathcal{A}$ en classes d'équivalences formées des *déformations intérieures* d'une géométrie donnée. Une telle déformation est obtenue sans modifier ni la représentation de $\mathcal{A}$ dans $\mathcal{H}$ ni l'isométrie antilinéaire J par la formule

$$D \to D + A + JAJ^{-1} \tag{9}$$

où $A = A^*$ est un opérateur autoadjoint arbitraire de la forme

$$A = \Sigma \, a_i[D, b_i] \ , \ a_i, b_i \in \mathcal{A} \, . \tag{10}$$

Le nouveau triplet spectral obtenu continue à vérifier les axiomes (1) – (7′).

L'action du groupe $\operatorname{Int}(\mathcal{A})$ sur les géométries spectrales (cf. 1.21) se réduit à une transformation de jauge sur A, donnée par la formule

$$\gamma_u(A) = u[D, u^*] + uAu^* \, . \tag{11}$$

L'équivalence unitaire est implémentée par la représentation suivante du groupe unitaire de $\mathcal{A}$ dans $\mathcal{H}$,

$$u \to uJuJ^{-1} = u(u^*)^0 \, . \tag{12}$$

La transformation (9) se réduit à l'identité dans le cas Riemannien usuel. Pour obtenir un exemple non trivial il suffit d'en faire le produit par l'unique géométrie spectrale sur l'algèbre de dimension finie $\mathcal{A}_F = M_N(\mathbb{C})$ des matrices $N \times N$ sur $\mathbb{C}$, $N \geq 2$. On a alors $\mathcal{A} = C^\infty(M) \otimes \mathcal{A}_F$, $\mathrm{Int}(\mathcal{A}) = C^\infty(M, PSU(N))$ et les déformations intérieures de la géométrie sont paramétrées par les potentiels de jauge pour une théorie de jauge de groupe $SU(N)$. L'espace $P(\mathcal{A})$ des états purs de l'algèbre $\mathcal{A}$ est le produit $P = M \times P_{N-1}(\mathbb{C})$ et la métrique sur $P(\mathcal{A})$ déterminée par la formule 1.10 dépend du potentiel de jauge A. Elle coïncide avec la métrique de Carnot [G] sur P définie par la distribution horizontale de la connection associée à A (cf. [Co3]). Le groupe $\mathrm{Aut}(\mathcal{A})$ des automorphismes de $\mathcal{A}$ est le produit semi direct,

$$\mathrm{Aut}(\mathcal{A}) = \mathcal{U} \rtimes \mathrm{Diff}(M) \tag{13}$$

du groupe $\mathrm{Int}(\mathcal{A})$ des transformations de jauges locales par le groupe des difféomorphismes. En dimension $n = 4$, les fonctionnelles d'action de Hilbert Einstein pour la métrique Riemannienne et de Yang-Mills pour le potentiel vecteur A apparaissent simplement, et avec les bons signes, dans le développement asymptotique en $\frac{1}{\Lambda}$ du nombre $N(\Lambda)$ de valeurs propres de D qui sont $\leq \Lambda$. On régularise cette expression en la remplaçant par

$$\mathrm{Trace}\ \varphi\left(\frac{D}{\Lambda}\right) \tag{14}$$

où $\varphi \in C_c^\infty(\mathbb{R})$ est une fonction paire qui vaut 1 sur l'intervalle $[-1, 1]$, (cf. [CC]). Les seuls autres termes non nuls du développement asymptotique sont un terme cosmologique, un terme de gravité de Weyl et un terme topologique.

Un exemple plus élaboré de variété spectrale est le tore non commutatif $\mathbb{T}_\theta^2$. Le paramètre $\theta \in \mathbb{R}/\mathbb{Z}$ définit la déformation suivante de l'algèbre des fonctions C^∞ sur le tore $\mathbb{T}^2$, de générateurs U, V. Les relations

$$VU = \exp 2\pi i\theta\, UV \quad \text{et} \quad UU^* = U^*U = 1\ ,\ VV^* = V^*V = 1 \tag{15}$$

définissent la structure d'algèbre involutive de $\mathcal{A}_\theta = \{\Sigma\, a_{n,m} U^n V^n\ ;\ a = (a_{n,m}) \in \mathcal{S}(\mathbb{Z}^2)\}$ où $\mathcal{S}(\mathbb{Z}^2)$ est l'espace de Schwartz des suites à décroissance rapide. Comme pour les courbes elliptiques on utilise comme paramètre pour définir la géométrie de $\mathbb{T}_\theta^2$ un nombre complexe τ de partie imaginaire positive et, à isométrie près, cette géométrie ne dépend que de l'orbite de τ pour $PSL(2, \mathbb{Z})$

[Co]. Le phénomène nouveau qui apparait est *l'équivalence de Morita* qui relie entre elles les algèbres $\mathcal{A}_{\theta_1}, \mathcal{A}_{\theta_2}$ lorsque θ_1 et θ_2 sont dans la même orbite de l'action de $PSL(2,\mathbb{Z})$ sur $\mathbb{R}$ [Ri].

Etant donné une variété spectrale $(\mathcal{A}, \mathcal{H}, D)$ et une équivalence de Morita entre $\mathcal{A}$ et une algèbre $\mathcal{B}$ donnée par

$$\mathcal{B} = \mathrm{End}_{\mathcal{A}}(\mathcal{E}) \tag{16}$$

où $\mathcal{E}$ est une $\mathcal{A}$-module à droite , projectif de type fini et hermitien, on obtient une géométrie spectrale sur $\mathcal{B}$ par le choix d'une *connection hermitienne* sur $\mathcal{E}$. Une telle connection ∇ est une application linéaire ∇ : $\mathcal{E} \otimes_{\mathcal{A}} \Omega_D^1$ vérifiant les règles ([Co])

$$\nabla(\xi a) = (\nabla \xi) a + \xi \otimes da \qquad \forall \xi \in \mathcal{E} \ , \ a \in \mathcal{A} \tag{17}$$

$$(\xi, \nabla \eta) - (\nabla \xi, \eta) = d(\xi, \eta) \qquad \forall \xi, \eta \in \mathcal{E} \tag{18}$$

où $da = [D, a]$ et où $\Omega_D^1 \subset \mathcal{L}(\mathcal{H})$ est le $\mathcal{A}$-bimodule formé par les opérateurs de la forme (10).

Toute algèbre $\mathcal{A}$ est Morita équivalente à elle même (avec $\mathcal{E} = \mathcal{A}$) et quand on applique la construction ci-dessus on obtient les déformations intérieures de la géométrie spectrale.

5 La géométrie spectrale de l'espace temps.

L'information expérimentale et théorique dont on dispose sur la structure de l'espace temps est résumée par la fonctionnelle d'action suivante, $\mathcal{L} = \mathcal{L}_E + \mathcal{L}_G + \mathcal{L}_{G\varphi} + \mathcal{L}_\varphi + \mathcal{L}_{\varphi f} + \mathcal{L}_f$ où $\mathcal{L}_E = -\frac{1}{16\pi G} \int r \sqrt{g}\, d^4x$ est l'action de Hilbert-Einstein et les 5 autres termes constituent le modèle standard de la physique des particules, couplé de manière minimale à la gravitation. Outre la métrique $g_{\mu\nu}$ ce Lagrangien implique plusieurs champs de bosons et de fermions. Les bosons de spin 1 sont le photon γ, les bosons médiateurs $W^\pm$ et Z et les huit gluons. Les bosons de spin 0 sont les champs de Higgs φ qui sont introduits pour briser la parité et pour que le mécanisme de brisure de symétrie spontanée confère une masse aux diverses particules sans contredire la renormalisabilité des champs de jauge non abéliens. Tous les fermions sont de spin $\frac{1}{2}$ et forment 3 familles de quarks et leptons.

Les champs impliqués dans le modèle standard ont a priori un statut très différent de celui de la métrique $g_{\mu\nu}$. Le groupe de symétrie de ces champs, à savoir le groupe des transformations de jauge locales,

$$\mathcal{U} = C^\infty(M, U(1) \times SU(2) \times SU(3)) \qquad (1)$$

est a priori très différent du groupe $\mathrm{Diff}(M)$ de symétries de $\mathcal{L}_E$. Le groupe de symétrie naturel de $\mathcal{L}$ est le produit semidirect $\mathcal{U} \rtimes \mathrm{Diff}(M) = G$. La première question à résoudre si l'on veut donner une signification purement géométrique à $\mathcal{L}$ est de trouver un espace géométrique X tel que $G = \mathrm{Diff}(X)$. Ceci détermine, en tenant compte du relèvement des difféomorphismes aux spineurs, l'algèbre $\mathcal{A}$,

$$\mathcal{A} = C^\infty(M) \otimes \mathcal{A}_F \ , \ \mathcal{A}_F = \mathbb{C} \oplus \mathbb{H} \oplus M_3(\mathbb{C}) \qquad (2)$$

où l'algèbre involutive $\mathcal{A}_F$ est la somme directe des algèbres $\mathbb{C}$, $\mathbb{H}$ des quaternions et $M_3(\mathbb{C})$ des matrices 3×3 complexes.

L'algèbre $\mathcal{A}_F$ correspond à un espace *fini* dont les fermions du modèle standard et les paramètres de Yukawa (masses des fermions et matrice de mélange de Kobayashi Maskawa) déterminent la géométrie spectrale de la manière suivante. L'espace de Hilbert $\mathcal{H}_F$ est de dimension finie et admet pour base la liste des fermions élémentaires. Par exemple pour la 1ère génération de leptons cette liste est

$$e_L, e_R, \nu_L, \bar{e}_L, \bar{e}_R, \bar{\nu}_L \, . \qquad (3)$$

L'algèbre $\mathcal{A}_F$ admet une représentation naturelle dans $\mathcal{H}_F$ (cf. [Co3]) et en désignant par J_F l'unique involution antilinéaire qui échange f et $\bar{f}$ pour tout vecteur de la base, on a la commutation,

$$[a, Jb^* J^{-1}] = 0 \qquad \forall \, a, b \in \mathcal{A}_F \, . \qquad (4)$$

L'opérateur D_F est simplement donné par la matrice $\begin{bmatrix} Y & 0 \\ 0 & \bar{Y} \end{bmatrix}$ où Y est la matrice de couplage de Yukawa. De plus les propriétés particulières de Y assurent la commutation,

$$[[D_F, a], b^0] = 0 \qquad \forall \, a, b \in \mathcal{A}_F \, . \qquad (5)$$

La $\mathbb{Z}/2$ graduation naturelle de $\mathcal{H}_F$ vaut 1 pour les fermions gauches $(e_L, \nu_L \ldots)$ et -1 pour les fermions droits, on a

$$\gamma_F = \varepsilon \, \varepsilon^0 \text{ où } \varepsilon = (1, -1, 1) \in \mathcal{A}_F \, . \qquad (6)$$

Nous renvoyons à [Co3] pour les vérifications des axiomes (1) – (7′). Le seul défaut est que le nombre de générations introduit une multiplicité dans la forme d'intersection, $K_0(\mathcal{A}) \times K_0(\mathcal{A}) \to \mathbb{Z}$, donnée par un multiple entier de la matrice 3×3

$$\begin{bmatrix} -1 & 1 & -1 \\ 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \tag{7}$$

Nous reviendrons à la fin de cet exposé sur la signification de la géométrie spectrale $(\mathcal{A}_F, \mathcal{H}_F, D_F) = F$.

Le pas suivant consiste à calculer les déformations intérieures (formule 3.9) de la géométrie produit $M \times F$ où M est une variété Riemannienne spinorielle de dimension 4. Le calcul donne les bosons de jauge du modèle standard, $\gamma, W^{\pm}, Z$, les huits gluons et les champs de Higgs φ avec les bons nombres quantiques et montre que,

$$\mathcal{L}_{\varphi f} + \mathcal{L}_f = \langle \psi, D\psi \rangle \tag{8}$$

où $D = D_0 + A + JAJ^{-1}$ est la déformation intérieure de la géométrie produit (donnée par l'opérateur $D_0 = \not{\partial} \otimes 1 + \gamma_5 \otimes D_F$).

La structure de produit de $M \times F$ donne une bigraduation de Ω_D^1 et une décomposition $A = A^{(1,0)} + A^{(0,1)}$ de A qui correspond à la décomposition (8). Le terme $A^{(1,0)}$ rassemble tous les bosons de spin 1 et le terme $A^{(0,1)}$ les bosons de Higgs qui apparaissent comme des termes de différence finie sur l'espace F. Cette bigraduation existe sur l'analogue Ω_D^2 des 2-formes ([Co]) et décompose la courbure $\theta = dA + A^2$ en trois termes $\theta = \theta^{(2,0)} + \theta^{(1,1)} + \theta^{(0,2)}$ 2 à 2 orthogonaux pour le produit scalaire,

$$\langle \omega_1, \omega_2 \rangle = ⨍ \omega_1 \, \omega_2^* \, ds^4 \, . \tag{9}$$

Ainsi l'action de Yang-Mills, $\langle \theta, \theta \rangle = ⨍ \theta^2 \, ds^4$ se décompose comme somme de 3 termes et on démontre que ces termes sont respectivement $\mathcal{L}_G$, $\mathcal{L}_{G\varphi}$ et $\mathcal{L}_\varphi$ pour $(2,0)$, $(1,1)$ et $(0,2)$ respectivement [Co].

L'action de Yang-Mills $⨍ \theta^2 \, ds^4$ utilise la décomposition $D = D_0 + A + JAJ^{-1}$ et n'est donc pas, a priori, une fonction ne dépendant que de la géométrie définie par D. Nous avons vu en 3.14 que, dans un cas plus simple, la combinaison $\mathcal{L}_E + \mathcal{L}_G$ apparait directement dans le développement asymptotique du nombre de valeurs propres inférieures à Λ de D. Le même principe (cf. [CC]) s'applique au modèle standard et conduit à la fonctionnelle suivante

$$\mathrm{Trace}\left(\varphi\left(\frac{D}{\Lambda}\right)\right)+\langle\psi, D\psi\rangle \tag{10}$$

dont le développement asymptotique ([CC]) donne $\mathcal{L}+$ un terme de gravité de Weyl et un terme en $r\varphi^2$ qui est le seul terme que l'on peut rajouter à $\mathcal{L}$ sans altérer le modèle standard. Nous renvoyons à [CC] pour l'interprétation physique de ces résultats.

La géométrie finie F ci-dessus était dictée par les résultats expérimentaux et il reste à en comprendre la signification conceptuelle à partir de l'analogue des groupes de Lie en géométrie non commutative, i.e. la théorie des groupes quantiques. Le fait simple (cf. [M]) est que le revêtement spinoriel $\mathrm{Spin}(4)$ de $SO(4)$ n'est pas un revêtement maximal parmi les groupes quantiques. On a $\mathrm{Spin}(4)=SU(2)\times SU(2)$ et même le groupe $SU(2)$ admet grâce aux résultats de Lusztig des revêtements finis de la forme (Frobenius à l'∞),

$$1\to H\to SU(2)_q\to SU(2)\to 1 \tag{11}$$

où q est une racine de l'unité, $q^m=1$, m impair. Le cas le plus simple est $m=3$, $q=\exp\left(\frac{2\pi i}{3}\right)$. Le groupe quantique fini H a une algèbre de Hopf de dimension finie très voisine de $\mathcal{A}_F$, et la représentation spinorielle de H définit un bimodule sur cette algèbre de Hopf de structure très voisine du bimodule $\mathcal{H}_F$ sur $\mathcal{A}_F$. Cela suggère d'étendre la géométrie spinorielle ([LM]) aux revêtements quantiques du groupe spinoriel, ce qui nécessite même pour parler de G-fibré principal, d'introduire un minimum de non commutativité (du style $C^\infty(M)\otimes\mathcal{A}_F$) dans l'algèbre des fonctions.

Mentionnons enfin que nous avons négligé dans cet exposé la nuance importante entre les signatures Riemanniennes et Lorentziennes.

References

[At] M.F. Atiyah: K-theory and reality, *Quart. J. Math. Oxford* (2), **17** (1966), 367-386.

[B-G] R. Beals and P. Greiner: Calculus on Heisenberg manifolds, *Annals of Math. Studies* **119**, Princeton Univ. Press, Princeton, N.J., 1988.

[B] T.P. Branson: An anomaly associated with 4-dimensional quantum gravity, to appear.

[B-O] T.P. Branson and B. Ørsted: Explicit functional determinants in four dimensions, *Proc. Amer. Math. Soc.*, **113** (1991), 669-682.

[B-W] A.R. Bernstein and F. Wattenberg: Non standard measure theory. In *Applications of model theory to algebra analysis and probability*, Edited by W.A.J. Luxenburg Halt, Rinehart and Winstin (1969).

[C-J-G] A.H. Chamseddine, J. Fröhlich and O. Grandjean: The gravitational sector in the Connes-Lott formulation of the Standard model, *J. Math. Phys.*, **36** n.11 (1995).

[C-C] A. Chamseddine and A. Connes: The spectral action principle, to appear.

[Co] A. Connes: Noncommutative geometry, Academic Press (1994).

[Co1] A. Connes: Cyclic cohomology and the transverse fundamental class of a foliation, Geometric methods in operator algebras, (Kyoto, 1983), pp. 52-144, *Pitman Res. Notes in Math.* **123** Longman, Harlow (1986).

[Co2] A. Connes: Noncommutative geometry and reality, *Journal of Math. Physics* **36** n.11 (1995).

[Co3] A. Connes: Gravity coupled with matter and the foundation of noncommutative geometry.

[Co-L] A. Connes and J. Lott: Particle models and noncommutative geometry, *Nuclear Phys. B*, **18B** (1990), suppl. 29-47 (1991).

[C-M1] A. Connes and H. Moscovici: Cyclic cohomology, the Novikov conjecture and hyperbolic groups, *Topology* **29** (1990), 345-388.

[C-M2] A. Connes and H. Moscovici: The local index formula in noncommutative geometry, GAFA, **5** (1995), 174-243.

[C-M3] A. Connes and H. Moscovici: Hypoelliptic Dirac operator, diffeomorphisms and the transverse fundamental class.

[Co-S] A. Connes and G. Skandalis: The longitudinal index theorem for foliations, *Publ. Res. Inst. Math. Sci. Kyoto* **20** (1984), 1139-1183.

[C-S-T] A. Connes, D. Sullivan and N. Teleman: Quasiconformal mappings, operators on Hilbert space, and local formulae for characteristic classes, *Topology*, Vol.33 n.4 (1994), 663-681.

[D-T] T. Damour and J.H. Taylor: Strong field tests of relativistic gravity and binary pulsars, *Physical Review D*, Vol.45 n.6 (1992), 1840-1868.

[Dx] J. Dixmier: Existence de traces non normales, *C.R. Acad. Sci. Paris*, Ser. A-B **262** (1966).

[D-F-R] S. Doplicher, K. Fredenhagen and J.E. Roberts: Quantum structure of space time at the Planck scale and Quantum fields, to appear in CMP.

[F] J. Fröhlich: The noncommutative geometry of two dimensional supersymmetric conformal field theory, *Preprint ETH* (1994).

[G-F] K. Gawedzki and J. Fröhlich: Conformal Field theory and Geometry of Strings, *CRM Proceedings and Lecture Notes*, Vol.7 (1994), 57-97.

[Gh] E. Ghys: L'invariant de Godbillon Vey, *Séminaire Bourbaki, Exposé* 706, *Astérisque* Vol. 88-89.

[Gi] P. Gilkey: Invariance theory, the heat equation and the Atiyah-Singer index theorem, *Math. Lecture Ser.* **11**, Publish or Perish, Wilmington, Del., 1984.

[G] M. Gromov: Carnot-Caratheodory spaces seen from within, Preprint IHES/M/94/6.

[G-K-P] H. Grosse, C. Klimcik and P. Presnajder: On finite 4 dimensional quantum field theory in noncommutative geometry, CERN Preprint TH/96 – 51 Net Hep-th/9602115.

[H-S] M. Hilsum, G. Skandalis: Morphismes K-orientés d'espaces de feuilles et fonctorialité en théorie de Kasparov, *Ann. Sci. Ecole Norm. Sup.* (4) **20** (1987), 325-390.

[I-K-S] B. Iochum, D. Kastler and T. Schücker: Fuzzy mass relations for the Higgs, *J. Math. Phys.* **36** n.11 (1995).

[K-W] W. Kalau and M. Walze: Gravity, noncommutative geometry and the Wodzicki residue, *J. of Geom. and Phys.* **16** (1995), 327-344.

[K] G. Kasparov: The operator K-functor and extensions of C^*-algebras, *Izv. Akad. Nauk. SSSR Ser. Mat.*, **44** (1980), 571-636.

[Ka] C. Kassel: Le résidu non commutatif, *Séminaire Bourbaki, exposé* 708, *Astérisque* Vol. 88-89.

[Kas] D. Kastler: The Dirac operator and gravitation, *Commun. Math. Phys.* **166** (1995), 633-643.

[L-M] B. Lawson and M.L. Michelson: *Spin Geometry*, Princeton 1989.

[M] Y. Manin: Quantum groups and noncommutative geometry, *Centre Recherche Math. Univ. Montréal* (1988).

[Mi-S] J. Milnor and D. Stasheff: Characteristic classes, *Ann. of Math. Stud.*, **76** Princeton University Press, Princeton, N.J. (1974).

[N] S.P. Novikov: Topological invariance of rational Pontrjagin classes, *Doklady A.N. SSSR*, **163** (1965), 921-923.

[P] S. Power: Hankel operators on Hilbert space, *Res. Notes in Math.*, **64** Pitman, Boston, Mass. (1982).

[Ri1] M.A. Rieffel: Morita equivalence for C^*-algebras and W^*-algebras, *J. Pure Appl. Algebra* **5** (1974), 51-96.

[Ri2] M.A. Rieffel: C^*-algebras associated with irrational rotations, *Pacific J. Math.* **93** (1981), 415-429; MR 83b:46087.

[R] B. Riemann: Mathematical Werke, Dover, New York (1953).

[St] E. Stein: Singular integrals and differentiability properties of functions, *Princeton Univ. Press, Princeton, N.J.* (1970).

[Ste] J. Stern: Le problème de la mesure, *Séminaire Bourbaki*, Vol. 1983/84, Exp. 632, pp. 325-346, *Astérisque* N. 121/122, *Soc. Math. France, Paris* (1985).

[S1] D. Sullivan: Hyperbolic geometry and homeomorphisms, in *Geometric Topology, Proceed. Georgia Topology Conf. Athens*, Georgia (1977), 543-555.

[S2] D. Sullivan: Geometric periodicity and the invariants of manifolds, *Lecture Notes in Math.* **197**, Springer (1971).

[Ta] M. Takesaki: Tomita's theory of modular Hilbert algebras and its applications, *Lecture Notes in Math.* **128**, Springer (1970).

[W] S. Weinberg: Gravitation and Cosmology, John Wiley, New York (1968).

Studying the Evolution of Cosmological Models

George F. R. Ellis
University of Cape Town
Capetown, South Africa

Abstract

This paper discusses geometric issues arising in the study of relativistic cosmology, particularly as seen by their evolution in the state-space of models. Two main approaches are via space-time symmetries, and by imposing conditions on covariantly defined variables. At present these two approaches are not satisfactorily related to each other.

1 Specifying models

A cosmological model represents the universe at a particular scale. It is defined by specifying (Ehlers 1961, 1993, Ellis 1971, 1973):

* the *space-time geometry* (determined by the metric), which —because of the requirement of compatibility with observations— must either have some expanding Robertson-Walker ('RW') geometries as a regular limit (see Krasinksi 1993), or else be demonstrated to have observational properties compatible with the major features of current astronomical observations of the universe;

* the *matter present* and its behaviour (the stress tensor of each matter component, the equations governing the behaviour of each such component, and the interaction terms between them), which must represent physically plausible matter; and

* the *interaction of the geometry and matter* —how matter determines the geometry, which in turn determines the motion of the matter. Usually we assume this is through the Einstein gravitational field equations ('EFE')

$$G_{ab} \equiv R_{ab} - \frac{1}{2}Rg_{ab} = \kappa T_{ab}\,, \tag{1}$$

which guarantee the conservation of total energy-momentum because of the *contracted Bianchi identities*

$$G^{ab}{}_{;b} = 0 \;\Rightarrow\; T^{ab}{}_{;b} = 0\,. \tag{2}$$

The usual choices for the matter description will be

* a fluid with given equation of state, for example a perfect fluid with 4-velocity u^a, energy density μ, and pressure p, where $p = p(\mu)$, $\mu + p > 0$ (beware of imperfect fluids, unless they have well-defined and motivated physical properties);

* a mixture of fluids, with the same or different 4-velocities;

* a set of particles represented by a kinetic theory description;

* a scalar field ϕ, with a given potential $V(\phi)$ (at early times);

* possibly an electromagnetic field described by Maxwell's equations.

To be useful in an explanatory role, a cosmological model must be easy to describe —that means they have symmetries or special properties of some kind or other. However we are interested in the *full* state space of solutions, allowing us to see how more realistic models are related to each other and to higher symmetry models.

2 Covariant description and equations

It should be emphasized that the equations considered here are exact, generic, and describe a cosmological context.

2.1 Variables

2.1.1 The average 4-velocity of matter

In a cosmological space-time $(\mathcal{M}, \mathrm{g})$ there will be a family of 'fundamental observers' moving with the average motion of matter at each point. Their 4-velocity is

$$u^a = \frac{dx^a}{d\tau}, \; u^a u_a = -1 \tag{3}$$

where τ is proper time measured along the fundamental worldlines. We assume this 4-velocity is unique: that is, there is a preferred motion of matter at each space-time event. At recent times this is taken to be the 4-velocity defined by the dipole of the Cosmic Blackbody Radiation ('CBR'): for there is precisely one

4-velocity which will set this dipole to zero. It is usually assumed that this is the same as the average 4-velocity of matter in a suitably sized volume (Ellis 1971).

Given u^a, there are defined unique projection tensors

$$U^a_b = -u^a u_b \quad \Rightarrow \quad U^a{}_b U^b{}_c = U^a{}_c,\ U^a{}_a = 1,\ U_{ab}u^b = u_a\,, \tag{4}$$

$$h_{ab} = g_{ab} + u_a u_b \quad \Rightarrow \quad h^a{}_b h^b{}_c = h^a{}_c,\ h^a{}_a = 3,\ h_{ab}u^b = 0. \tag{5}$$

The first projects parallel to the velocity vector u^a, and the second determines the metric properties of the instantaneous rest-space of observers moving with 4-velocity u^a. There is also defined a volume element for the rest-spaces

$$\eta^{abc} = \eta^{abcd}u_d \ \Rightarrow\ \eta^{abc} = \eta^{[abc]},\ \eta^{abc}u_c = 0 \tag{6}$$

where η^{abcd} is the 4-dimensional volume element ($\eta^{abcd} = \eta^{[abcd]}$, $\eta^{0123} = 1/\sqrt{|\det g_{ab}|}$.)

Two derivatives are also defined: the time derivative $\dot{}$ along the fundamental world lines, where for any tensor T

$$\dot{T}^{ab}{}_{cd} = T^{ab}{}_{cd;e}u^a\,, \tag{7}$$

and the orthogonal spatial derivative $\hat{\nabla}$, where for any tensor T

$$\hat{\nabla}_e T^{ab}{}_{cd} = h^a{}_s h^b{}_t h_c{}^v h_d{}^w \nabla_p T^{st}{}_{vw} h^p{}_e \tag{8}$$

with total projection on all free indices (note that we interchangeably use a semicolon and ∇_a for the covariant derivative: $T^a{}_{b;c} \equiv \nabla_c T^a{}_b$).

2.1.2 Kinematic quantities

We split the first covariant derivative of u_a into its irreducible parts, defined by their symmetry properties:

$$u_{a;b} = \omega_{ab} + \sigma_{ab} + \frac{1}{3}\Theta h_{ab} - \dot{u}_a u_b \tag{9}$$

where ω_{ab} is the vorticity tensor ($\omega_{ab} = \omega_{[ab]}$, $\omega_{ab}u^b = 0$), σ_{ab} is the shear tensor ($\sigma_{ab} = \sigma_{(ab)}$, $\sigma_{ab}u^b = 0$, $\sigma^a{}_a = 0$), $\Theta = u^a{}_{;a} = 3H$ is the (volume) expansion (and H the Hubble parameter), and $\dot{u}_a = u_{a;b}u^b$ is the acceleration.

2.1.3 Matter tensor

The matter stress tensor can be decomposed relative to u^a in the form

$$T_{ab} = \mu u_a u_b + q_a u_b + u_a q_b + p h_{ab} + \pi_{ab}, \tag{10}$$
$$q_a u^a = 0, \ \pi_{ab} = \pi_{ba}, \ \pi_{ab} u^b = 0, \ \pi^a{}_a = 0$$

where $\mu = T_{ab}u^a u^b$ is the relativistic energy density, $q_a = -T_{ab}u^b$ is the relativistic momentum density, which is also the energy flux relative to u^a, $p = \frac{1}{3}T^a{}_a$ is the isotropic pressure, and π_{ab} is the trace-free anisotropic stresses.

The physics of the situation is in the equations of state relating these quantities, for example the commonly imposed restrictions

$$q_a = 0 = \pi_{ab} \ \Leftrightarrow \ T_{ab} = \mu u_a u_b + p h_{ab} \tag{11}$$

characterize a 'perfect fluid'. If in addition we assume that $p = 0$, we have the simplest case: pressure-free matter ('dust' or 'baryonic matter'). Otherwise we must specify an equation of state determining p from μ and possibly other thermodynamic variables. Whatever these relations may be, we usually require that various 'energy conditions' hold: one or all of

$$\mu > 0, \ \mu + p > 0, \ \mu + 3p > 0 \tag{12}$$

and additionally demand the speed of sound c_s obeys

$$0 \leq c_s^2 \leq 1 \ \Leftrightarrow \ 0 \leq dp/d\mu \leq 1.$$

2.1.4 The Weyl tensor

The Weyl conformal curvature tensor C_{abcd} is split relative to u^a into 'electric' and 'magnetic' parts:

$$E_{ac} = C_{abcd}u^b u^d \ \Rightarrow \ E^a{}_a = 0, \ E_{ab} = E_{(ab)}, \ E_{ab}u^b = 0\,, \tag{13}$$

$$H_{ac} = \frac{1}{2}\eta_{ab}{}^{ef}C_{efcd}u^b u^d \ \Rightarrow \ H^a{}_a = 0, \ H_{ab} = H_{(ab)}, \ H_{ab}u^b = 0\,. \tag{14}$$

These represent the 'free gravitational field', enabling gravitational action at a distance (tidal forces, gravitational waves). Together with the Ricci tensor R_{ab} (determined locally at each point by the matter tensor through the EFE (1)), these quantities completely represent the space-time Riemann curvature tensor R_{abcd}.

2.1.5 Auxiliary quantities

It is useful to define some associated kinematic quantities: the vorticity vector

$$\omega^a = \frac{1}{2}\eta^{abcd}u_b\omega_{cd} \;\Rightarrow\; \omega^a u_a = 0,\; \omega^a\omega_{ab} = 0\,, \tag{15}$$

the magnitudes

$$\omega^2 = \frac{1}{2}\omega^{ab}\omega_{ab} \geq 0\,,\; \sigma^2 = \frac{1}{2}\sigma^{ab}\sigma_{ab} \geq 0\,, \tag{16}$$

and the average length scale ℓ determined by

$$\dot{\ell}/\ell = \frac{1}{3}\Theta\,. \tag{17}$$

Further it is helpful to define particular spatial gradients orthogonal to u^a, characterizing the inhomogeneity of space-time:

$$X_a = \hat{\nabla}_a\mu,\; Y_a = \hat{\nabla}_a p,\; Z_a = \hat{\nabla}_a\Theta\,. \tag{18}$$

These satisfy the important identity

$$\hat{\nabla}_{[a}\hat{\nabla}_{b]}\mu = 2\omega_{ab}\dot{\mu}\,. \tag{19}$$

The latter shows that if $\omega_{ab}\dot{\mu} \neq 0$ in an open set then $X_a \neq 0$ there.

2.2 Equations

There are three sets of equations to be considered, resulting from the EFE (1).

2.2.1 The Ricci identity

The first set arise from the *Ricci identity* for the vector field u^a, i.e.

$$u^a{}_{;bc} - u^a{}_{;cb} = R_d{}^a{}_{bc}u^d\,.$$

We obtain three propagation equations and three constraint equations. The *propagation equations* are,

1. The Raychaudhuri equation

$$\dot{\Theta} + \frac{1}{3}\Theta^2 + 2(\sigma^2 - \omega^2) - \dot{u}^a{}_{;a} + \frac{1}{2}\kappa(\mu + 3p) = 0\,, \tag{20}$$

which is the basic equation of gravitational attraction,

2. The vorticity propagation equation

$$h^{f}{}_{\epsilon}(\ell^{2}\omega^{\epsilon})^{\cdot} = \ell^{2}\sigma^{f}{}_{d}\omega^{d} + \ell^{2}\frac{1}{2}\eta^{fcbd}u_{c}\dot{u}_{b;d} \tag{21}$$

showing how vorticity conservation follows if there is a perfect fluid with acceleration potential,

3. The shear propagation equation

$$\begin{aligned} h_{a}{}^{f}h_{b}{}^{g}(\ell^{-2}(\ell^{2}\sigma_{fg})^{\cdot} - \dot{u}_{(f;g)}) - \dot{u}_{a}\dot{u}_{b} + \omega_{a}\omega_{b} + \sigma_{a}{}^{f}\sigma_{fg} + \\ + h_{ab}(-\frac{1}{3}\omega^{2} + \frac{2}{3}\sigma^{2} + \dot{u}^{a}{}_{;a}) - \frac{1}{2}\kappa\pi_{ab} + E_{ab} = 0, \end{aligned} \tag{22}$$

showing how E_{ab} induces shear.

The *constraint equations* are,

1. The $(0, \nu)$ equations

$$h^{ab}(\omega_{b}{}^{c}{}_{;d}h_{c}^{d} - \sigma_{b}{}^{c}{}_{;d}h_{c}^{d} + \frac{2}{3}\Theta_{,b}) + (\omega^{a}{}_{b} + \sigma^{a}{}_{b})\dot{u}^{b} = \kappa q^{a} \;, \tag{23}$$

2. The vorticity divergence identity

$$\omega^{a}{}_{;b}h^{b}{}_{a} = \omega^{a}\dot{u}_{a} \;, \tag{24}$$

3. The H_{ab} equation

$$H_{ad} = 2\dot{u}_{(a}\omega_{d)} - h_{a}{}^{t}h_{d}{}^{s}(\omega_{(t}{}^{b;c} + \sigma_{(t}{}^{b;c})\eta_{s)fbc}u^{f} \;. \tag{25}$$

2.2.2 The contracted Bianchi identities

The second set of equations arise from the *contracted Bianchi identities* (2). We obtain one propagation equation:

$$\dot{\mu} + (\mu + p)\Theta = 0 \;, \tag{26}$$

the energy conservation equation, and one constraint equation:

$$(\mu + p)\dot{u}_{a} + h_{a}{}^{c}p_{,c} = 0 \;, \tag{27}$$

the momentum conservation equation, where for simplicity we have given only the perfect fluid form.

2.2.3 The other Bianchi identities

If one attains a consistent solution to the equations given so far, that is all one requires. However often it is useful to additionally explicitly consider the integrability conditions for the equations listed so far. These are the *Bianchi identities*

$$R_{ab[cd;e]} = 0\,.$$

Double contraction gives (2), already considered. Apart from these equations, the full Bianchi identities give two further propagation equations and two constraint equations, which are similar in form to Maxwell's equations.

The *propagation equations* are,

$$\begin{aligned} & h^m{}_a h^t{}_c \dot{E}^{ac} + J^{mt} - 2H_a{}^{(t}\eta^{m)bpq} u_b \dot{u}_p + h^{mt}\sigma^{ab}E_{ab} + \\ & \quad +\Theta E^{mt} - 3E_s{}^{(m}\sigma^{t)s} - E_s{}^{(m}\omega^{t)s} = -\frac{1}{2}(\mu + p)\sigma^{tm}\,, \end{aligned} \tag{28}$$

the '$\dot{E}$' equation, and

$$\begin{aligned} & h^m{}_a h^t{}_c \dot{H}^{ac} - I^{mt} + 2E_a{}^{(t}\eta^{m)bpq} u_b \dot{u}_p + h^{mt}\sigma^{ab}H_{ab} + \\ & \quad +\Theta H^{mt} - 3H_s{}^{(m}\sigma^{t)s} - H_s{}^{(m}\omega^{t)s} = 0 \end{aligned} \tag{29}$$

the '$\dot{H}$' equation, where again we have given only the perfect fluid form, and we have defined

$$J^{mt} = h_a{}^{(m}\eta^{t)rsd} u_r H^a{}_{s;d} = \text{`}curl\ H\text{'}\,,$$

$$I^{mt} = h_a{}^{(m}\eta^{t)rsd} u_r E^a{}_{s;d} = \text{`}curl\ E\text{'}\,.$$

The *constraint equations* are

$$h^t{}_a E^{as}{}_{;d} h^d{}_s - \eta^{tbpq} u_b \sigma^d{}_p H_{qd} + 3H^t{}_s\omega^s = \frac{1}{3}h^{tb}\mu_{;b}\,, \tag{30}$$

the 'div E' equation, and

$$h^t{}_a H^{as}{}_{;d} h^d{}_s + \eta^{tbpq} u_b \sigma^d{}_p E_{qd} - 3E^t{}_s\omega^s = (\mu + p)\omega^t\,, \tag{31}$$

the 'div H' equation.

2.3 *The set of equations*

Altogether we have six propagation equations and six constraint equations; considered as a set of evolution equations for the covariant variables, they are a first-order system of equations. This set is determinate once the fluid equations of state are given; together they then form a complete set of equations that we can regard as an infinite dimensional dynamical system (the system closes up, but is essentially infinite dimensional because of the spatial derivatives that occur).

Useful solutions are defined by considering appropriate restrictions on the kinematic quantities, Weyl tensor, or space-time geometry for a specified plausible matter content. In many cases these define a finite dimensional subset of the full system. Given such restrictions,

(a) we need to check *consistency* of the constraints with the evolution equations. It is believed that they are *generally consistent*, i.e. they are consistent if no restrictions are placed on their evolution other than implied by the evolution equations (this has not been proved, but is very plausible). Once we impose further restrictions, they may or may not be consistent. This is what we have to investigate.

(b) we need to understand the *dynamical evolution* that results, particularly fixed points, attractors, etc., in terms of suitable variables,

(c) we particularly seek to determine and characterize *involutive subsets* of the space of space-times: that is regions that are mapped into themselves by the dynamical evolution of the system, and so are left invariant by that evolution.

As far as possible we aim to do this for the exact equations. We are also concerned with

(d) *linearization* of the equations about known simple solutions, and determination of properties of the resulting linearized solutions, in particular considering whether they accurately represent the behaviour of the full non-linear theory in a neighborhood of the background solution (the issue of *linearization stability*).

The idea is to relate the different models, if possible by determining the dynamic flows in the state space of models.

3 Classification by symmetries

Symmetries of a space or a space-time (generically, 'space') are transformations of the space into itself that leave the metric tensor and all physical and geometrical properties invariant. We deal here only with continuous symmetries, characterized by a continuous group of transformations and associated vector fields (Eisenhart 1933).

3.1 Killing vectors

A space or space-time *symmetry* or *isometry* is a transformation that drags the metric into itself. The generating vector field ξ_i is called a *Killing vector (field)* (or 'KV'), and obeys Killing's equations,

$$(L_\xi g)_{ij} = 0 \;\Leftrightarrow\; \xi_{(i;j)} = 0 \;\Leftrightarrow\; \xi_{i;j} = -\xi_{j;i} \tag{32}$$

where L_X is the *Lie derivative*. By the Ricci identity for the KV, this implies the curvature equation:

$$\xi_{i;jk} = R_{ijkl}\xi^l \tag{33}$$

and so the infinite series of further equations that follows by taking covariant derivatives of this one, e.g.

$$\xi_{i;jkm} = R_{ijkl;m}\xi^l + R_{ijk}{}^l\xi_{l;m} \tag{34}$$

The Killing vector fields form a Lie algebra with a basis ξ_a $(a = 1, 2, .., r)$ with components ξ^i_a with respect to a local coordinate basis where a,b,c label the KV basis, i j k the coordinate components, $r \leq \frac{1}{2}n(n-1)$ is the dimension of the algebra. Any KV can be written in terms of this basis, with *constant coefficients*. Hence: if we take the commutator $[\xi_a, \xi_b]$ of two of the basis KV's, this is also a KV, and so can be written in terms of its components relative to the Killing vector basis, which will be constants. We can write the constants as $C^c{}_{ab}$, obtaining

$$[\xi_a, \xi_b] = C^c{}_{ab}\,\xi_c, \; C^a{}_{bc} = C^a{}_{[bc]}\,. \tag{35}$$

By the Jacobi identities for the basis vectors, these structure constants must satisfy

$$C^d{}_{s[c}C^s{}_{ab]} = 0\,. \tag{36}$$

These are the integrability conditions that must be satisfied in order that the Lie Algebra exist in a consistent way. The transformations generated by the Lie Algebra form a Lie group (Eisenhart 1933, Cohn 1961) of the same dimension.

Arbitrariness of the basis: we can change the basis of KV's in the usual way;

$$\xi_{a'} = \Lambda_{a'}{}^{a}\xi_{a} \;\Leftrightarrow\; \xi^{i}_{a'} = \Lambda_{a'}{}^{a}\xi^{i}_{a} \tag{37}$$

where the $\Lambda_{a'}{}^{a}$ are constants with $det(\Lambda_{a'}{}^{a}) \neq 0$, so unique inverse matrices $\Lambda^{a'a}$ exist. Then the structure constants transform as tensors:

$$C^{c'}{}_{a'b'} = C^{c}{}_{ab}\Lambda^{c'}{}_{c}\Lambda_{a'}{}^{a}\Lambda_{b'}{}^{b}\,. \tag{38}$$

Thus the (non)-equivalence of two Lie Algebras is not obvious, as they may be given in quite different bases.

3.2 Groups of isometries

The isometries of a space of dimension n must be a group, as the identity is an isometry, the inverse of an isometry is an isometry, and the composition of two isometries is an isometry. Continuous isometries are generated by the Lie Algebra of Killing Vector fields. The group structure is determined locally by the Lie algebra, in turn characterized by the structure constants (Cohn, 1961). The action of the group is characterized by the nature of its orbits in space; this is only partially determined by the group structure (indeed the same group can act as a space-time symmetry group in quite different ways).

3.2.1 Dimensionality of groups and orbits

Most spaces have no Killing vectors, but special spaces (with symmetries) have some. The group action defines orbits in the space where it acts, and the dimensionality of these orbits determines the kind of symmetry that is present.

The *orbit* of a point p is the set of all points into which p can be moved by the action of the isometries of a space. Orbits are necessarily homogeneous (all physical quantities are the same at each point). An *invariant variety* is a set of points moved into itself by the group. This will be bigger than (or equal to) all orbits it contains. The orbits are necessarily invariant varieties; indeed they are sometimes called *minimum invariant varieties*, because they are the smallest subspaces that are always moved into themselves by all the isometries in the group.

Fixed points of groups of isometries are those points which are left invariant by the isometries (thus the orbit of such a point is just the point itself). These are the points where all Killing vectors vanish, so the dimension of the Lie algebra is zero here (however the derivatives of the Killing vectors there are non-zero; the Killing vectors generate isotropies about these points).

General points are those where the dimension of the space spanned by the Killing vectors (that is, the dimension of the orbit through the point) takes the value it has almost everywhere; *special points* are those where it has a lower dimension (e.g. fixed points). Consequently the dimension of the orbits through special points is lower than that of orbits through general points. The dimension of the algebra is the same at each point of an orbit, because of the equivalence of the group action at all points on each orbit.

The group is *transitive on a surface* S (of whatever dimension) if it can move any point of S into any other point of S. Orbits are the largest surfaces through each point on which the group is transitive; they are therefore sometimes referred to as *surfaces of transitivity*. We define their dimension as follows, and determine limits from the maximal possible initial data for Killing vectors:

dim surface of transitivity $= s =$ dim minimum invariant varieties, where in a space of dimension n, $s \leq n$.

At each point we can also consider the dimension of the isotropy group (the group of isometries leaving that point fixed), generated by all those Killing vectors that vanish at that point:

dim of isotropy group $= q$, where $q \leq 1/2n(n-1)$.

The *dimension* r *of the group of symmetries* of a space of dimension n is $r = s + q$ (translations plus rotations). From the above limits , $0 \leq r \leq n + (1/2)n(n-1) = (1/2)n(n+1)$ (the maximal number of translations and of rotations). This shows the Lie algebra of KVs is finite dimensional.

Maximal dimensions: If $r = 1/2n(n+1)$ we have a space(time) of constant curvature (maximal symmetry for a space of dimension n). In this case,

$$R_{ijkl} = K(g_{ik}g_{jl} - g_{il}g_{jk}) \tag{39}$$

with K constant; and K necessarily *is* constant if this equation is true and $n \geq 3$. One can't get $q = (1/2)n(n-1) - 1$ so $r \neq (1/2)n(n+1) - 1$.

A group is *simply transitive* if $r = s \Leftrightarrow q = 0$ (no redundancy: dimensionality of group of isometries is just sufficient to move each point in a surface of transitivity into each other point). There is no continuous isotropy group.

A group is *multiply transitive* if $r > s \Leftrightarrow q > 0$ (there is redundancy in that the dimension of the group of isometries is larger than is needed to move each point in an orbit into each other point). There exist non-trivial isotropies.

3.3 Classification of cosmological symmetries

For a cosmological model, because space-time is 4-dimensional, the possibilities for dimension of the surface of transitivity are $s = 0, 1, 2, 3, 4$. As to isotropy, we assume $(\mu + p) \neq 0$; then $q = 3,\ 1,$ or 0 because u^a is invariant and so the isotropy group at each point has to be a sub-group of the rotations acting orthogonally to u^a (and there is no 2-d subgroup of $O(3)$.) The dimension q of the isotropy group can vary over the space (but not over an orbit): it can be greater at special points (e.g. an axis centre of symmetry) where the dimension s of the orbit is less, but r (the dimension of the total symmetry group) must stay the same everywhere. Thus the possibilities for isotropy at a general point are,

a) Isotropic: $q = 3$, the Weyl tensor vanishes, kinematic quantities vanish except Θ. All observations (at every point) are isotropic. This is the RW family of geometries;

b) Local Rotational Symmetry ('LRS'): $q = 1$, the Weyl tensor is type D, kinematic quantities are rotationally symmetric about a preferred spatial direction. All observations at every general point are rotationally symmetric about this direction. All metrics are known in the case of dust (Ellis 1967) and a perfect fluid (Stewart & Ellis, 1968, see also van Elst and Ellis 1996).

c) Anisotropic: $q = 0$; there are no rotational symmetries. Observations in each direction are different from observations in each other direction.

Putting this together with the possibilities for the dimensions of the surfaces of transitivity, we have the following possibilities [See Table 1].

4 Bianchi Universes ($s = 3$)

These are the models in which there is a simply transitive group G_3 of isometries transitive on spacelike surfaces, so they are spatially homogeneous. There is only

Dim invariant variety

Dimension Isotropy Group	s=2	s=3	s=4
	inhomogeneous	spatially homogeneous	space-time homogeneous
q = 0 aniso-tropic	generic metric form known. Spatially self-similar, Abelian G_2 on 2-d spacelike surfaces, non-abelian G_2	Bianchi: orthogonal, tilted	Osvath/Kerr
q = 1 LRS	Bondi-Tolman family	Kantowski-Sachs, LRS Bianchi	Godel
q = 3 isotropic	none (can't happen)	Friedmann	Einstein static
	two non-ignorable coordinates no redshift	one non-ignorable coordinate	algebraic EFE

Dim invariant variety

s=0	s=1
Inhomogeneous,	No Isotropy Group
Szekeres-Szafron, Stephani-Barnes, Oleson type N The real universe!	General metric form independent of one coord; KV h.s.o., not h.s.o

Table 1: Classification of cosmological models ($\mu + p > 0$) by isotropy and homogeneity (see Ellis 1967).

one essential dynamical coordinate, and the EFE reduce to ordinary differential equations, because the inhomogeneous degrees of freedom have been 'frozen out'. They are thus quite special in geometric terms; nevertheless they form a rich set of models where one can study the exact dynamics of the full non-linear field equations. The solutions to the field equations will depend on the matter in the space-time. In the case of a fluid (with uniquely defined flow lines), we have two different kinds of models:

Orthogonal models, with the fluid flow lines orthogonal to the surfaces of homogeneity (Ellis and MacCallum 1969);

Tilted models, with the fluid flow lines not orthogonal to the surfaces of homogeneity; the fluid velocity vector components enter as further variables (King and Ellis 1973, see also Collins and Ellis 1979).

Rotating models must be tilted, and are much more complex than non-rotating models.

4.1 Constructing Bianchi universes

There are essentially three direct ways of constructing them, all based on properties of a triad of vectors $\mathbf{e}_\alpha$ that commute with the basis of Killing vectors ξ_β. Thus these approaches does not directly relate to the variables introduced in the previous section, although they will be important in understanding the Bianchi models.

The *first approach* (Taub 1951, Heckmann and Schücking 1962) puts all the time variation in the metric components:

$$ds^2 = -dt^2 + \gamma_{\alpha\beta}(t)(e^\alpha{}_i(x^\nu)dx^i)(e^\beta{}_j(x^\mu)dx^j) \tag{40}$$

where $e^\alpha{}_i(x^\nu)$ are 1-forms inverse to the spatial vector triad $e_\alpha{}^i(x^\mu)$, which have the same commutators $C^\alpha{}_{\beta\gamma}$ $(\alpha, \beta, \gamma, .. = 1, 2, 3)$ as the structure constants of the group of isometries and commute with the unit normal vector e_0 to the surfaces of homogeneity; that is, $e_\alpha = e_\alpha{}^i(\partial/\partial x^i)$, $e_0 = (\partial/\partial t)$ obey

$$[e_\alpha, e_\beta] = C^\gamma{}_{\alpha\beta}e_\gamma, \; [e_0, e_\alpha] = 0\,. \tag{41}$$

One can classify the Lie Algebra structure (following Schücking) by defining

$$C^\alpha{}_{\beta\gamma}\epsilon^{\beta\gamma\delta} = n^{\alpha\delta} + \epsilon^{\alpha\delta\kappa}a_\kappa \tag{42}$$

where $n^{\alpha\beta} = n^{(\alpha\beta)}$, $a_\gamma = C^\gamma{}_{\alpha\gamma}$. Then the Jacobi Identities (36) for these vectors are

$$n^{\alpha\beta} a_\beta = 0 \tag{43}$$

We define two major classes of structure constants (and so Lie Algebras):

Class A: $a_\alpha = 0$,
Class B: $a_\alpha \neq 0$.

One can diagonalise $n_{\alpha\beta}$ in both cases by suitable choice of basis, and choose a_α in the 1-direction. Most of the non-zero constants (represented as constant components of $n^{\alpha\beta}$ and a_α) can be normalised to ± 1 by change of basis (37), the structure constants transforming according to (38) (and so $n^{\alpha\beta}$ and a_α transforming as tensors). The EFE (1) become ordinary differential equations for $\gamma_{\alpha\beta}(t)$. We deal directly with these equations, without introducing the Weyl tensor components as additional variables (so we do not explicitly consider the full set of Bianchi identities in this approach; rather they are identities that will automatically be satisfied once the EFE are satisfied).

The *second approach* (Ellis and MacCallum 1969) uses an orthonormal tetrad, so the metric components g_{ab} are constants, putting all the time variation in the commutators of the basis vectors. In this case we have an orthonormal basis $\mathbf{e}_a$ $(a = 0, 1, 2, 3)$ such that

$$[e_a, e_b] = \gamma^c{}_{ab}(t) e_c \, . \tag{44}$$

The spatial commutator functions $\gamma^\alpha{}_{\beta\gamma}(t)$, which can be represented analogously to (47) above by a time-dependent matrix $n^{\alpha\beta}(t)$ and vector $a_\alpha(t)$, are equivalent to the structure constants $C^\alpha{}_{\beta\gamma}$ of the symmetry group at each point (i.e. they can be brought to the canonical forms of the $C^\alpha{}_{\beta\gamma}$ at that any by a suitable change of basis; however the transformation to do so is different at each point and at each time). The commutators $\gamma^a{}_{bc}(t)$, together with the matter variables, are then treated as the dynamical variables. The EFE (1) are first order equations for these quantities, supplemented by the Jacobi identities for the basis vectors which are also first order equations for the commutators.

The third approach is based on the automorphism group of the symmetry group. We will not consider it further here.

5 Dynamical systems approach

The most illuminating dynamical systems description of Bianchi models is based on the use of orthonormal tetrads, and is examined in detail in a forthcoming book (Wainwright and Ellis 1996). The variables used are essentially the commutator coefficients mentioned above, but rescaled by a common time dependent factor[1].

5.1 The reduced differential equations

The basic idea (Collins 1971, Wainwright 1988) is to write the Einstein field equations in a way that enables one to study the evolution of the various physical and geometrical quantities *relative to the overall rate of expansion of the universe,* as described by the rate of expansion scalar $\theta = u^a_{;a}$, or equivalently *the Hubble variable* H:

$$H = \tfrac{1}{3}\theta. \tag{45}$$

We consider here non-tilted fluids, where the 4-velocity $\mathbf{u}$ is orthogonal to the group orbits and t is a time variable which is constant on the group orbits, so that $\mathbf{u} = \frac{\partial}{\partial t}$. Let $\{\mathbf{e}_a\}$ be a group invariant orthonormal frame, with $\mathbf{e}_0 = \mathbf{u}$. We use the commutation functions γ^c_{ab} associated with the frame $\{\mathbf{e}_a\}$: as the basic gravitational field variables. The γ^c_{ab} are constant on the group orbits and can thus be regarded as a function of the time variable t: $\gamma^c_{ab} = \gamma^c_{ab}(t)$. Since $\mathbf{e}_0$ is normal to the group orbits, the non-zero commutation functions are

$$(\gamma^c{}_{ab}) = (H, \sigma_{\alpha\beta}, \Omega_\alpha, n_{\alpha\beta}, a_\alpha)\,, \tag{46}$$

where $H(t)$, $\sigma_{\alpha\beta}(t)$ are the expansion and shear of the normal flow lines, $\Omega_\alpha(t)$ is the rate of rotation of the spatial tetrad vectors relative to a parallel propagated basis along the fluid flow lines, and $n_{\alpha\beta}(t)$, $a_\alpha(t)$ represent the purely spatial commutators through the equation

$$\gamma^\alpha{}_{\beta\gamma}\epsilon^{\beta\gamma\delta} = n^{\alpha\delta}(t) + \epsilon^{\alpha\delta\kappa}a_\kappa(t) \tag{47}$$

(cf. (42)). At this stage the remaining freedom in the choice of orthonormal frame needs to be eliminated by specifying the variables Ω_α implicitly or explicitly (for example by specifying them as functions of the $\sigma_{\alpha\beta}$). This also simplifies the other quantities (for example choice of a shear eigenframe will result in the

[1] The following is adapted from notes by J. Wainwright

tensor $\sigma_{\alpha\beta}$ being represented by two diagonal terms). This leads to a reduced set of variables, consisting of H and the remaining commutation functions, which we denote symbolically by

$$\mathbf{x} = ({\gamma^c}_{ab}|_{reduced}). \tag{48}$$

The physical state of the model is thus described by the vector $(H, \mathbf{x})$. The details of this reduction differ for the class A and B models, and in the latter case there is an algebraic constraint of the form

$$g(\mathbf{x}) = 0\,, \tag{49}$$

where g is a homogeneous polynomial.

The idea is now to normalize $\mathbf{x}$ with the Hubble variable H. We denote the resulting variables by a vector $\mathbf{y} \in \mathbb{R}^n$, and write:

$$\mathbf{y} = \frac{\mathbf{x}}{H}. \tag{50}$$

These new variables are *dimensionless*, and will be referred to as *expansion-normalized variables*. It is clear that each dimensionless state $\mathbf{y}$ determines a 1-parameter family of physical states $(\mathbf{x}, H)$. The evolution equations for the ${\gamma^c}_{ab}$ lead to evolution equations for H and $\mathbf{x}$ and hence for $\mathbf{y}$. In deriving the evolution equations for $\mathbf{y}$ from those for $\mathbf{x}$, the *deceleration parameter* q plays an important role. The Hubble variable H can be used to define a scale factor ℓ, according to

$$H = \frac{\dot{\ell}}{\ell}, \tag{51}$$

where $\cdot$ denotes differentiation with respect to t. The deceleration parameter is then defined by

$$\dot{H} = -(1+q)H^2. \tag{52}$$

In order that the evolution equations define a flow, it is necessary, in conjunction with the rescaling (50) to introduce a *dimensionless time variable* τ according to

$$\ell = \ell_0 e^{\tau}, \tag{53}$$

where ℓ_0 is the value of the scale factor at some arbitrary reference time. Since ℓ assumes values $0 < \ell < +\infty$ in an ever-expanding model, τ assumes all real values, with $\tau \to -\infty$ at the initial singularity and $\tau \to +\infty$ at late times. It follows from equations (51) and (53) that

$$\frac{dt}{d\tau} = \frac{1}{H}, \tag{54}$$

and the evolution equation (52) for H can be written

$$\frac{dH}{d\tau} = -(1+q)H. \tag{55}$$

Since the right hand side of the evolution equations for the γ^c_{ab} are homogeneous of degree 2 in the $\gamma^c{}_{ab}$ the change (54) of the time variable results in H canceling out of the evolution equation for $\mathbf{y}$, yielding an autonomous DE:

$$\frac{d\mathbf{y}}{d\tau} = \mathbf{f}(\mathbf{y}), \quad \mathbf{y} \in \mathbb{R}^n. \tag{56}$$

The constraint $g(\mathbf{x}) = 0$ translates into a constraint

$$g(\mathbf{y}) = 0, \tag{57}$$

which is preserved by the DE. The functions $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}$ are polynomial functions in $\mathbf{y}$. An essential feature of this process is that the evolution equation for H, namely (55), decouples from the remaining equations (56) and (57). In other words, the DE (56) describes the evolution of the non-tilted Bianchi cosmologies, the transformation (5.5) essentially scaling away the effects of the overall expansion. An important consequence is that the new variables are bounded near the initial singularity.

5.2 Cosmological dynamical systems

5.2.1 Invariant sets and limit sets

The first step in the analysis is to formulate the field equations, using expansion-normalized variables, as a DE (56) in $\mathbb{R}^n$, possibly subject to a constraint (57). Since τ assumes all real values (for models which expand indefinitely), the solutions of (56) are defined for all τ and hence define a *flow* $\{\phi_\tau\}$ on $\mathbb{R}^n$. The evolution of the cosmological models can thus be analyzed by studying the orbits of this flow in the physical region of state space, which is a subset of $\mathbb{R}^n$ defined by the requirement that the energy density be non-negative, i.e.

$$\Omega(\mathbf{y}) = \frac{\mu}{3H^2} \geq 0 \tag{58}$$

where the density parameter Ω is a dimensionless measure of the matter density μ.

The *vacuum boundary*, defined by $\Omega(\mathbf{y}) = 0$, describes the evolution of vacuum Bianchi models, and is an invariant set which plays an important role in the qualitative analysis because vacuum models can be asymptotic states for perfect fluid models near the big-bang or at late times. There are other invariant sets which are also specified by simple restrictions on $\mathbf{y}$ which play a special role: the subsets representing each Bianchi type, and the subsets representing higher symmetry models, specifically the FL models and the LRS Bianchi models.

It is desirable that the dimensionless state space D in $\mathbb{R}^n$ is a compact set. In this case each orbit will have a non-empty α-limit set and ω-limit set, and hence there will exist a past attractor and a future attractor in state space. When using expansion-normalized variables, compactness of the state space has a direct physical meaning for ever-expanding models: if the state space is compact then at the big-bang no physical or geometrical quantity diverges more rapidly than the appropriate power of H, and at late times no such quantity tends to zero less rapidly than the appropriate power of H. This will happen for many models; however the state space for Bianchi VII$_0$ and VIII models is non-compact. This lack of compactness manifests itself in the behaviour of the Weyl tensor at late times.

5.2.2 Equilibrium points and self-similar cosmologies

Each ordinary orbit in the dimensionless state space corresponds to a one-parameter family of physical universes, which are conformally related by a constant rescaling of the metric. On the other hand, for an equilibrium point $\mathbf{y}^*$ of the DE (56) (which satisfies $\mathbf{f}(\mathbf{y}^*) = \mathbf{0}$), the deceleration parameter q is a constant, i.e. $q(\mathbf{y}^*) = q^*$, and we find

$$H(\tau) = H_0 e^{(1+q^*)\tau}.$$

In this case, however, the parameter H_0 is no longer essential, since it can be set to unity by a translation of τ, $\tau \to \tau +$ constant; then (54) implies that

$$Ht = \frac{1}{1+q^*}, \tag{59}$$

so that by (48) and (50) the commutation functions are of the form (constant) $\times\, t^{-1}$. It follows that the resulting cosmological model is self-similar. Thus, *to each equilibrium point of the DE (56) there corresponds a unique self-similar cosmological model.* In such a model the physical states at different times differ only by an overall change in the length scale. Such models are expanding, but in such a way that their dimensionless state does not change. They include the flat FL model ($\Omega = 1$) and the Milne model ($\Omega = 0$). All vacuum and non-tilted perfect fluid self-similar Bianchi solutions have been given by Hsu and Wainwright (1986).

The equilibrium points determine the asymptotic behaviour of other more general models. If the α-limit set of a point y is an equilibrium point y^*, then the orbit through y approaches y^* as $\tau \to -\infty$. The physical interpretation is that the self-similar model which corresponds to y^* approximates the dynamics of the model with initial state y, as $\tau \to -\infty$. This model is *asymptotically self-similar into the past*. A similar interpretation holds if the ω-limit set is an equilibrium point. The term *asymptotically selfsimilar* without a qualifier means that the model has this property into the past and into the future. In this case the orbit that describes the model will be *heteroclinic* (i.e. joins two equilibrium points).

Equilibrium points also influence the intermediate evolution by determining finite heteroclinic sequences which join the past attractor to the future attractor. The intermediate equilibrium points in the sequence determine quasi-equilibrium epochs, which may be important from an observational point of view. In this case an anisotropic model can spend an arbitrarily large time in an ϵ-neighbourhood arbitrarily close to a FL equilibrium point; and hence can for practical purposes by observationally indistinguishable from it, even though its very early and very late behaviour will both be completely different.

Many phase planes can be constructed explicitly. The reader is referred to Wainright and Ellis (1996) for a comprehensive presentation and survey of results attained so far.

5.3 Relation to full state space

The symmetric spaces, such as the Bianchi universes, form finite dimensional subsets of the full dynamical system, defining involutive subsets of the full state space of solutions. There are also involutive subspaces that are infinite dimensional, some of which are discussed in the next section. The challenge is to characterise them and to relate them to the finite dimensional subspaces, such as those associated with Bianchi models.

6 Other involutive subspaces of state space

We look at some of these infinite dimensional subspaces here, and then briefly comment on the relation to the finite dimensional subspaces in the following section.

6.1 *Pressure-free matter ('dust')*

A particularly useful dynamical restriction is

$$p = 0 = q_a = \pi_{ab}$$

so the matter (often described as 'baryonic') is represented only by its 4-velocity u^a and its density $\mu > 0$.

In this case momentum conservation shows that $\dot{u}_a = 0$: the matter moves geodesically (as expected from the equivalence principle), and all equations simplify considerably. This is the case of *pure gravitation*: it separates out the (non-linear) gravitational effects from all the fluid dynamic effects. It is a very large involutive subspace.

6.2 *Irrotational flow*

If we have a barotropic perfect fluid:

$$q_q = \pi_{ab} = 0, \;\; p = p(\mu) \;\; \Rightarrow \; \eta^{acd}\dot{u}_{c;d} = 0$$

then $\omega = 0$ is involutive, i.e.

$$\omega^a = 0 \; \Rightarrow \; \dot{\omega}^a = 0$$

follows from the vorticity conservation equations (and this is true also in the special case $p = 0$), see (Ehlers 1961, 1993; Ellis 1973). In such flows,

1. The fluid flow is hypersurface orthogonal, as there exists a cosmic time function t such that $u_a = -g(x^b)t_{,a}$,
2. The metric of the orthogonal 3-spaces is h_{ab},
3. The Ricci tensor of these 3-spaces is given by

$$\begin{aligned} {}^3R_{ab} \;&=\; h_a{}^f h_b{}^g \left[\dot{u}_{(f;g)} - \ell^{-3}(\ell^3\sigma_{fg})^{\cdot}\right] \;+\; \dot{u}_a\dot{u}_b \;+ \\ &\quad + \frac{2}{3}(-\frac{1}{3}\Theta^2 + \sigma^2 - \frac{1}{2}\dot{u}^c{}_{;c} + \Lambda + \kappa\mu) \;+\; \kappa\pi_{ab} \end{aligned} \qquad (60)$$

and their Ricci scalar by

$$^3R = 2\{\sigma^2 - \frac{1}{3}\Theta^2 + \Lambda + \kappa\mu\} \ , \tag{61}$$

which is a generalised Friedmann equation. These equations fully determine the curvature tensor $^3R_{abcd}$ of the orthogonal 3-spaces. Provided the matter is baryonic perfect fluid, this is an involutive subspace of large dimension.

6.3 Irrotational dust

Dust is a special case of a baryonic fluid, so the dust irrotational spaces form an involutive subspace which is the intersection of the two. Considering these solutions, $p = 0 \Rightarrow \dot{u}_a = 0$ and $\omega^a = 0$. Then the non-trivial (exact) evolution equations of Section 1.2 are,

$$\dot{\mu} + \mu\Theta = 0 \, , \tag{62}$$

$$\dot{\Theta} + \frac{1}{3}\Theta^2 + 2\sigma^2 + \frac{1}{2}\kappa\mu = 0 \, , \tag{63}$$

$$\dot{\sigma}_{ab} + \sigma_a{}^f\sigma_{fb} - h_{ab}\frac{2}{3}\sigma^2 + \frac{2}{3}\Theta\sigma_{ab} + E_{ab} = 0, \tag{64}$$

$$\dot{E}^{mt} + h^{mt}\sigma^{ab}E_{ab} + \Theta E^{mt} + J^{mt} - 3E_s{}^{(m}\sigma^{t)s} = -\frac{1}{2}\mu\sigma^{tm} \, , \tag{65}$$

$$\dot{H}^{mt} + h^{mt}\sigma^{ab}H_{ab} + \Theta H^{mt} - I^{mt} - 3H_s{}^{(m}\sigma^{t)s} = 0 \, , \tag{66}$$

where J^{mt} is 'curl H' and I^{mt} is 'curl E'.

The constraint equations are

$$h^{ab}(-\sigma_b{}^c{}_{;d}h^d_c + \frac{2}{3}\Theta_{,b}) = 0 \, , \tag{67}$$

$$H_{ad} = -h_a{}^t h_d{}^s \sigma_{(t}{}^{b;c})\eta_{s)fbc}u^f \ \ , \tag{68}$$

$$h^t{}_a E^{as}{}_{;d}h^d{}_s - \eta^{tbpq}u_b\sigma^d{}_p H_{qd} = \frac{1}{3}X^t \, , \tag{69}$$

$$h^t{}_a H^{as}{}_{;d}h^d{}_s + \eta^{tbpq}u_b\sigma^d{}_p E_{qd} = 0 \, . \tag{70}$$

In general these equations are consistent (Maartens et al. 1997).

6.4 FL universes (RW geometry)

A particularly important involutive subspace of the irrotational dust space-times is that of the Friedmann-Lemaître ('FL') universes, based on the everywhere-isotropic Robertson-Walker ('RW') geometry. It is characterized by a perfect fluid matter tensor and the conditions

$$\omega_{ab} = \sigma_{ab} = 0 = \dot{u}^a \;\Rightarrow\; E_{ab} = H_{ab} = 0,\; X_a = Y_a = Z_a = 0\,,$$

the first conditions stating these solutions are also shear-free and hence are locally isotropic, the second that they are conformally flat, and the third that they are spatially homogeneous. It follows then that:

1. ${}^3R_{ab}$ is isotropic, so the 3-spaces are 3-spaces of constant curvature;

2. The remaining non-trivial equations are the energy equation (26), the Raychaudhuri equation (20) which now takes the form

$$\dot{\Theta} + \frac{1}{3}\Theta^2 + \frac{1}{2}\kappa(\mu + 3p) = 0\,, \tag{71}$$

and the Friedmann equation that follows from (61):

$${}^3R = -\frac{2}{3}\Theta^2 + 2\kappa\mu = \frac{6k}{\ell^2}\,, \tag{72}$$

where k is a constant. Any two of these equations imply the third if $\theta \neq 0$ (the latter equation being a first integral of the other two).

3. From these equations, as well as finding simple exact solutions one can determine evolutionary phase planes for this family of models, see Refsdal and Stabell (1966), Madsen and Ellis (1988), and Ehlers and Rindler (1989).

6.5 The Shear-Free case

If $p = 0 \Rightarrow \dot{u}_a = 0$ and $\sigma_{ab} = 0$ in an open set $\mathcal{U}$ then all equation simplify in $\mathcal{U}$. In particular the vorticity equation becomes

$$h^a{}_b(\ell^2\Omega^b)^{\cdot} = 0 \;\Rightarrow\; \omega^a = \frac{\Omega^a}{\ell^2},\; (\Omega^a)^{\cdot} = 0 \tag{73}$$

and then (on using the energy conservation equation) we can integrate the Raychaudhuri equation to get a 'Friedmann equation'

$$3(\dot{\ell})^2 + \frac{2\Omega^2}{\ell^2} - \frac{M}{\ell} = E \tag{74}$$

where M, E are constants. This appears to allow the avoidance of an initial singularity, as the vorticity term can dominate at early times! BUT putting $\sigma_{ab} = 0$ converts the $\dot{\sigma}_{ab}$ equation (22) into a new constraint:

$$\omega_a\omega_b + h_{ab}(-\frac{1}{3}\omega^2) = -E_{ab}\,. \tag{75}$$

This has to be consistent with the time evolution of E_{ab}, which now takes the form

$$h^m{}_a h^t{}_c \dot{E}^{ac} + J^{mt} + \Theta E^{mt} - E_s{}^{(m}\omega^{t)s} = 0\,. \tag{76}$$

We must now systematically check consistency.

The *Procedure* is as follows: take the time derivatives of all new constraints that arise from our assumptions (here, (75)). If necessary, commute space and time derivatives in the resulting equations, using the Ricci identities to do so. Substitute for the time evolution terms from the evolution equations, and use Leibniz's rule to expand out the spatial derivatives. Collect terms, obtaining simplified equations without any time derivatives. The result is *either* a new constraint equation that must be satisfied if the original constraint is to be preserved in time, *or* an identity $(0 = 0)$. CONTINUE until all the constraints that arise in this way are identically conserved by the time evolution, *or* we get an inconsistency.

The result of this procedure (Ellis 1967) is that in order to be consistent, shear-free dust solutions cannot expand and rotate; in $\mathcal{U}$,

$$\omega\Theta = 0 \;\Rightarrow\; if\ \Theta \neq 0,\ then\ \omega = 0\,. \tag{77}$$

Thus the only expanding dust solutions with vanishing shear are the FL solutions. Hence this does not offer a route to singularity avoidance (for consistency, the constant Ω in equation (74) has to vanish, so the vorticity term cannot dominate the early expansion.) The involutive subspace of irrotational dust space-times defined by this condition is just the FL subspace.

6.6 Silent universes: $H_{ab} = 0$.

The evolution equations for irrotational dust, in general partial differential equations, become ordinary differential equations if $I^{mt} = 0 = J^{mt}$: with these restrictions, there are no spatial derivatives in these equations. Hence we then have what has been called a 'silent universe' —provided the constraints are satisfied initially, and are conserved by the evolution equations, each world line evolves independently of each other (this evolution being governed by o.d.e's). In this case the infinite dimensional dynamical system decomposes into the direct product of finite dimensional dynamical systems along each world line.

The simplest case is when $H_{ab} = 0$. Then the equation (66) becomes a new constraint:

$$I^{mt} = h_a{}^{(m}\eta^{t)rsd}u_r E^a{}_{s;d} = 0\,. \tag{78}$$

Is this constraint (and the other constraints) preserved along the flow lines? No they are not, as has been shown by Bonilla et al (1996) and by van Elst et al (1996)[2]. The proof is based on analysis using a tetrad that simultaneously diagonalizes σ_{ab} and E_{ab} (possible because of (70)). It is not known what the full set of consistent solutions is, that forms an involutive subset of the exact field equations; it includes Bianchi I universes and the Szekeres family of models. There may be no others.

6.7 $div\ H = 0$

Now consider the case of solutions with $div\ H = 0$. Equation (70) then shows

$$div\ H = 0 \;\Rightarrow\; \eta^{tbpq}u_b\sigma^d{}_p E_{qd} = 0\,, \tag{79}$$

so E_{ab} and σ_{ab} can be simultaneously diagonalised. This reduces the number of variables drastically. We now need to check the consistency of the new condition, that is, to examine the consequences of the equation $(div\ H)^{\cdot} = 0$, using the same procedure as before. A consistency analysis (Maartens et al 1997) [3] shows this is consistent, even if $H \neq 0$. This is an exact result following from the full field equations, and shows consistency of these equations with the usual results of linearised theory for gravitational waves. Hence this does form an involutive subset of the full space of solutions.

These examples show how examination of the integrability conditions of the exact field equations starts to delineate allowed subspaces in the space of cosmological space-times. There is much to be done here, for example extending the above analyses to the case where $\omega_{ab} \neq 0$, or to $p = p(\mu)$.

[2] Correcting previous incorrect claims by Lesame et al

[3] Correcting Lesame et al 1996, which is erroneous because of a sign error in the equations used.

7 Problems and Issues

A lot of progress has been made in recent times, but many issues remain outstanding. So far, the covariant approach has not been properly tied in to the exact solutions characterized by symmetries. That is, the two main sections above have not been related properly to each other. This is an unsolved problem at the present time - the Locally Rotationally Symmetric case has been solved (van Elst and Ellis 1996) and some partial results are known in other cases. But we do not have a simple characterization of the symmetric subspaces —for example the Bianchi universes— in terms of the covariant variables.

The broader aim is an understanding of the evolution of models in the space of space-times, characterizing invariant sets, fixed points, saddle points, attractors, etc. As seen above, we can find these features in some phase planes that are sections of the full space of space-times, corresponding to families of higher-symmetry solutions or to kinematic restrictions; they then determine the nature of the evolutionary curves in those families (Wainwright and Ellis 1996). The problem is to extend this understanding to broader classes of models, and the to relation between the covariant and symmetry approaches.

Other issues that have not yet been resolved are:

(1) finding a suitable measure of probability in the full space of space-times, and in its involutive subspaces. The requirement is a natural measure that is plausible. Progress has been made in the FL sub-case, but even here is not definitive.

(2) Relating descriptions of the same space-time on different scales of description. This leads to the issue of averaging and the resulting effective (polarization) contributions to the stress tensor, arising because averaging does not commute with calculating the field equations for a given metric.

(3) Related to this is the question of definition of entropy for gravitating systems in general, and cosmological models in particular. This may be expected to imply a coarse-graining in general, and so is strongly related to the averaging question. It is an important issue in terms of its relation to the spontaneous formation of structure in the early universe.

References

[1] M. A. G. Bonilla, M. Mars, J. M. M. Senovilla, C. F. Sopuerta and R. Vera (1996). *Phys. Rev.* D **54**, 6565.

[2] P. M. Cohn (1961). *Lie Groups*. Cambridge University Press, Cambridge.

[3] C. B. Collins (1985). *J. Math. Phys.* **26**, 2009.

[4] C. B. Collins and G. F. R. Ellis (1979). *Phys. Rep.* **56**, 63.

[5] J. Ehlers (1961). *Abh. Mainz Akad. Wiss. u. Litt.*, Mat-Nat. Kl., Nr. 11.

[6] J. Ehlers (1993). *Gen. Rel. Grav.* **25**, 1225.

[7] J. Ehlers and W. Rindler (1989). *Mon. Not. Roy. Ast. Soc.* **238**, 503.

[8] L. P. Eisenhart (1933). *Continuous groups of transformations*. Dover (reprinted).

[9] G. F. R. Ellis (1967). *J. Math. Phys.* **8**, 1171-1194 .

[10] G. F. R. Ellis (1971). In *General Relativity and Cosmology*, Proc. Int. School of Physics "Enrico Fermi" (Varenna), Course XLVII. Ed. R. K. Sachs (Academic Press), 104.

[11] G. F. R. Ellis (1973). In *Cargèse Lectures in Physics, Vol. VI*, Ed. E. Schatzman (Gordon and Breach), 1.

[12] G. F. R. Ellis and M. A. H. MacCallum (1969). *Comm. Math. Phys.* **12**, 108.

[13] O. Heckmann and E. Schücking (1962). In *Gravitation: An Introduction to current research*. Ed. L. Witten (Wiley), 438-469.

[14] A. R. King and G. F. R. Ellis (1973). *Comm. Math. Phys.* **31**, 209.

[15] L. Hsu and J. Wainwright (1986). *Class. Quant. Grav.* **3**, 1105.

[16] A. Krasinski (1993). *Physics in an Inhomogeneous Universe*. Preprint 1993/10, Applied Mathematics Department, University of Cape Town. To appear as a book, Cambridge University Press (1997).

[17] W. Lesame, G. F. R. Ellis and P. K. S. Dunsby (1996). *Phys. Rev.* D **53**, 738.

[18] R. Maartens, W. M. Lesame, and G. F. R. Ellis (1997). Consistency of dust solutions with div $H = 0$. To appear, *Phys. Rev.* D.

[19] M. S. Madsen and G. F. R. Ellis (1988). *Mon. Not. Roy. Ast. Soc.* **234**, 67.

[20] R. Stabell and S. Refsdal (1966). *Mon. Not. Roy. Ast. Soc.* **132**, 379.

[21] J. M. Stewart and G. F. R. Ellis (1968). *J. Math. Phys.* **9**, 1072-1082.

[22] A. Taub (1951). *Ann. Math.* **53**, 472.

[23] H. van Elst and G. F. R. Ellis (1996). *Class. Quant. Grav.* **13**, 1099-1127.

[24] H. van Elst, C. Uggla, W. M. Lesame, R. Maartens, and G. F. R. Ellis (1997). *Class. Quant. Grav.* (1997) **14**, 1151–1162.

[25] J. Wainwright (1988). In *Relativity Today: Proc. 2nd Hungarian relativity Workshop 1987*, Ed. Z. Perjes. (Singapore: World Scientific).

[26] J. Wainwright and G. F. R. Ellis (1996). *Dynamical Systems in Cosmology.*

The Geometry of Quasicrystals

Christian Janot
Institut Laue-Langevin
Grenoble, France

Abstract

Quasicrystals are a new form of solid state which differ from both crystal and amorphous compounds by possessing a new type of long-range translational order, quasiperiodicity, and a noncrystallographic orientational order. Several geometrical schemes can be used to described quasiperiodic structures, including cut and projection from an hyperspace periodic structure, space tiling with matching rules, selfsimilar packing of clusters or even simplistic growth procedure within some constraints.

Introduction

Quasicrystals are materials having a new type of long range order such that their diffraction patterns show Bragg reflections revealing symmetries which are incompatible with periodicity [1]. However, they are highly ordered systems [2] with correlation length of several tenths of a micrometer [3]. Large single (quasi)crystals have been grown [4] whose structural quality is such that dynamical diffraction has been observed [3]. Deciphering the atomic structure of quasicrystals via classical techniques of crystallography has been reasonably well achieved using the relation of a quasiperiodic function with its hyperspace periodic image [5], even if details about atom positions are still missing.

Aside from their peculiar structures, quasicrystals also exhibit very unexpected properties [6]. Their perhaps most intriguing feature is a very high electrical (and thermal as well) resistivity. Its value which is almost as large as that of insulators [7] is amazing indeed for a material containing about 70% of aluminium. Reduced surface wetting, low friction, high hardness, weak chemical reactivity are among other interesting properties of quasicrystals. It is a current consensus that such physico-chemical behaviours are rooted somewhere into the still unusual geometry of these structures.

Substitution rules for growing quasicrystals

Making a structure grow is always a tiling story: polyhedra are first selected, then decorated with atoms of one or several chemical species and finally the structure results from packing copies of these decorated polyhedra. The packing obtained is a tiling of the space if there are neither holes nor overlaps of polyhedra in the built structure. The classical (periodic) crystals, based on the observation of natural minerals, deals with the simplest tiling procedure you can think of: a single type of tile is added again and again by translation. But addition is not the only way to fill space in good order. Iterative substitution rules offer an interesting alternative. To illustrate the difference between geometrical addition (GA) and geometrical substitution (GS), consider linear (one-dimensional) chains built up with sequences of two segments one large (L) and one short (S). GA structures can be obtained by adding strips LS over and over again, resulting in the periodic chain $LSLSLSLS\ldots$ To obtain a GS chain, substitution rules must be used instead. There is of course an infinite variety of substitution rules. For instance, any given strip of L, S segments can be grown by substituting L by LS and S by L iteratively; this results in the following successive grown strips:

- initial strip: LS
- first substitution: LSL
- second substitution: $LSLLS$
- third substitution: $LSLLSLSL$
- fourth substitution: $LSLLSLSLLSLLS$
- etc.

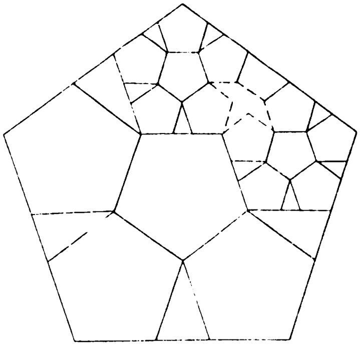


Figure 1: Substitution rules for planar tiling with a pentagonal symmetry.

The final chain is a perfectly ordered, deterministic sequence of L and S segments without any indication of periodicity. It is easy to see that the strip S_n obtained after n iterative steps is the simple addition of the strips S_{n-1} and S_{n-2} obtained after $n-1$ and $n-2$ steps respectively. Self similarity of the grown structure is then obvious.

Some other properties of the above quasiperiodic chain (Fibonacci chain) are of interest, inasmuch as they are easily generalised to two and three dimensional quasiperiodic structures.

First of all, let us count the number of L and S segments in the chain strips obtained after each substitution step.

This gives:

- start situation: one L, one S
- after one step: two L, one S
- after two steps: three L, two S
- after three steps: five L, three S
- after four steps: eight L, five S
- etc.

The number of $L(S)$ segments after n steps of substitution is equal to the sum of $L(S)$ segments found after $n-1$ and $n-2$ steps. The ratio of L segment numbers over S segment numbers takes successively, according to substitution steps, the values $1/1, 2/1, 3/2, 5/3, 8/5, 13/8$, etc. This is precisely the Fibonacci series whose limit is the golden mean $\tau = (1+\sqrt{5})/2 = 2\cos 36^o$ when the chain is grown ad infinitum.

The substitution rule as applied in the present case forces also the length ratio L/S to be equal to τ, indeed:

$$\begin{aligned} x &= \frac{L}{S} = \frac{L+S}{L} \\ \text{or: } x^2 &= x+1 \quad \text{with} \quad x > 1 \end{aligned}$$

which has the single mathematical solution $x = \tau$. Consequences will be that a quasicrystal must be made of at least two different chemical species mixed in strictly defined proportions and occupying well defined partial volumes. A growth rule may also be deduced in which density fluctuation would be bounded via the requisite that the numbers of L and S segment remain in a ratio close to τ.

Differences in properties for periodic GA and aperiodic GS chains can be easily anticipated. For instance, in a monatomic GA chain, such as a metal crystal, all atomic sites are strictly equivalent. If some electrons are loosely bonded to atoms they have no reason to locate on a particular site and can travel essentially freely through the bulk of the metal. This results in high conductivity and isotropy of the properties. Conversely, in GS structures strictly equivalent sites cannot be found if the fully extended surroundings of the sites is considered. The "free" electrons, if any, are forced to "locate" recurrently into hierarchies of sites according to an energy scale and within the constraints of Coulomb interactions. Actually, quasiperiodic structure, are such that identical site domains of any size can be found recurrently at distances apart of about twice the domain size. This is easily checked with the Fibonacci chain, if not too small domains are considered. Thus, delocalization via hopping between domains of a given class of local isomorphism can be reasonably expected.

For one-dimensional structures, it may be difficult to imagine why this awkward substitutional operation should be preferred instead of straightforward periodic packing. But in two and three dimensions, the latter may be just impossible. This is the situation, for instance, with pentagonal or icosahedral tiles whose fivefold symmetries cannot be accommodated by periodicity. Consequently aperiodic structures become the stable solution when chemical bonding favours such local "non-crystallographic" symmetries. This has been demonstrated by both numerical simulations [8] and experimental observations [5, 9].

Aperiodic tiling of the two-dimensional space

The substitutional growth is formally extended to two- and three-dimensional tiling without basic difficulties. This is illustrated in Figure 1 which shows how to grow a pentagonal tiling. Starting with a pentagonal area, six second generation pentagons and five triangles share the available space. Applying the same substitutional rules to the second generation pentagons introduces an additional "boat shape", as it is shown in Figure 2. In this particular example, R. Penrose [10] has proved that four and only four prototiles are needed to pursue the tiling ad infinitum: a pentagon, a triangle, a "boat shape", and a fivefold star. More precisely, holes which may form while growing the structure can always be filled in by one to these four tiles of the same generation. Simple geometrical derivations give the linear and surface deflation factors of the above procedure: τ^2 and τ^4 respectively. By multiplying the tiling size by these factors after each substitutional operation one get a growing tiling made of constant size elementary tiles.

Finally, it is worth pointing out that a simulated diffraction pattern with atoms sited on the vertices of such a tiling reproduces qualitatively experimental data (Figure 3) and clearly shows long range order and fivefold symmetries.

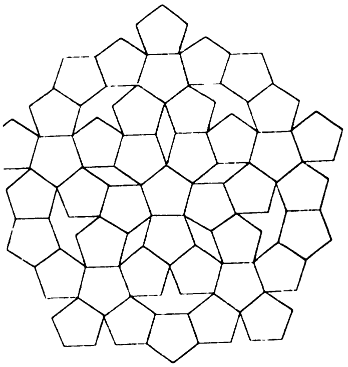


Figure 2: Further growth of the pentagonal tiling.

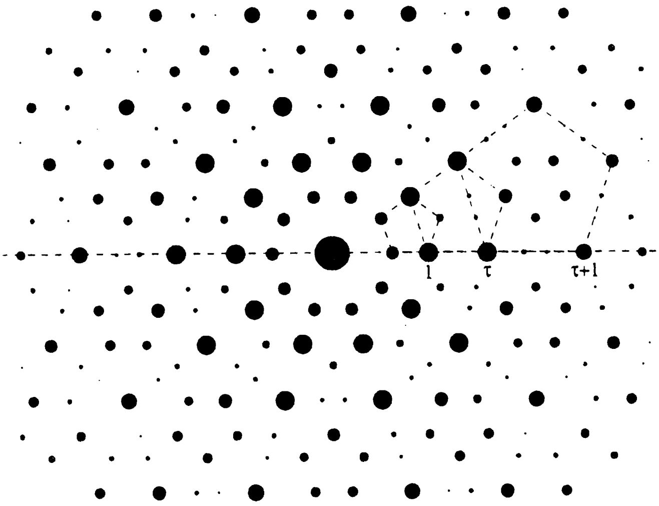


Figure 3: Numerical simulation of a diffraction pattern for the pentagonal tiling shown in Figure 2.

Again R. Penrose [10] has demonstrated that, in the case of most of the two-dimensional quasiperiodic tilings, the number of prototiles can be reduced to two. One example is given in Figure 4 for a fivefold tiling based on two triangles prototiles A and B with master angles $\pi/5$ and $2\pi/5$ respectively and area ratio equal to τ. The substitutional operation used to grow the structure is also illustrated in Figure 4 ($A_n = A_{n-1}B_{n-1}A_{n-1}$ and $B_n = A_{n-1}B_{n-1}$). It is very easy to verify that the sizes expand as $B_{n+1} = \tau^2 B_n$ and $A_{n+1} = \tau^2 A_n$. The grown structure is selfsimilar but not fractal ($d_f = 2$). Again the number of prototiles A divided by the number of prototiles B in the growing tiling is a figure which follows the Fibonacci series and has a limit equal to τ when the tiling is grown ad infinitum.

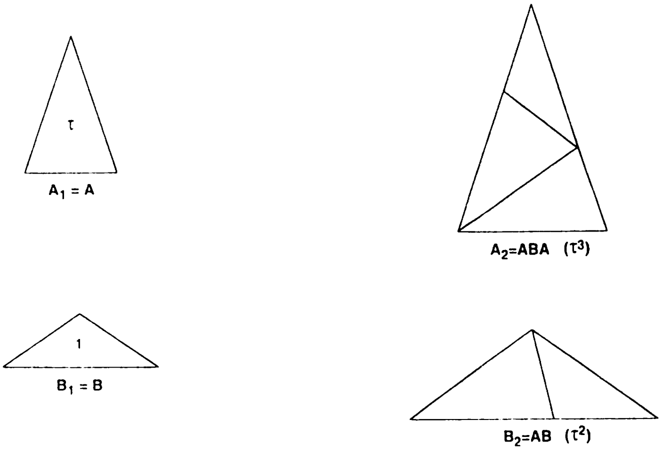


Figure 4: Triangular tiles and substitution rules for an alternative fivefold tiling of the plane.

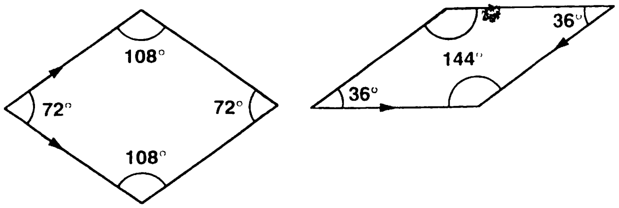


Figure 5: Rhombic prototiles with matching rules for the Penrose tiling.

The most famous modification of the fivefold two dimensional tiling is the so-called Penrose tiling which is based on two rhombic prototiles as shows in Figure 5, along with proper decorations which define matching rules in a growth procedure: when a tile is added, full dark or full white circles must be completed at the vertices, excluding mixed dark and white circles, and arrows on edges must match identically. Penrose tilings have fascinating properties. Despite being aperiodic, similar domains repeat in the structure over and over again, as clearly visible in Figure 6 (see for instance the "star" made of five "fat" rhombic units). One can also verify directly on Figure 6 that any type of domain, of any size, is reproduced ad infinitum at distances apart twice their size. Iterative substitution and matching rules are equivalent procedures to grow aperiodic structures (Fig. 7). A more tricky geometrical feature can be observed by looking carefully at Penrose tiling schemes: actually an infinite number of slightly different Penrose tilings can be obtained within proper respect of given matching rules; these various modifications cannot be globally superimposed to each other but, amazingly, any area selected in one of the tiling is also found in the other parent tilings. This curiosity is going to be explained via the cut/projection scheme later on in the paper.

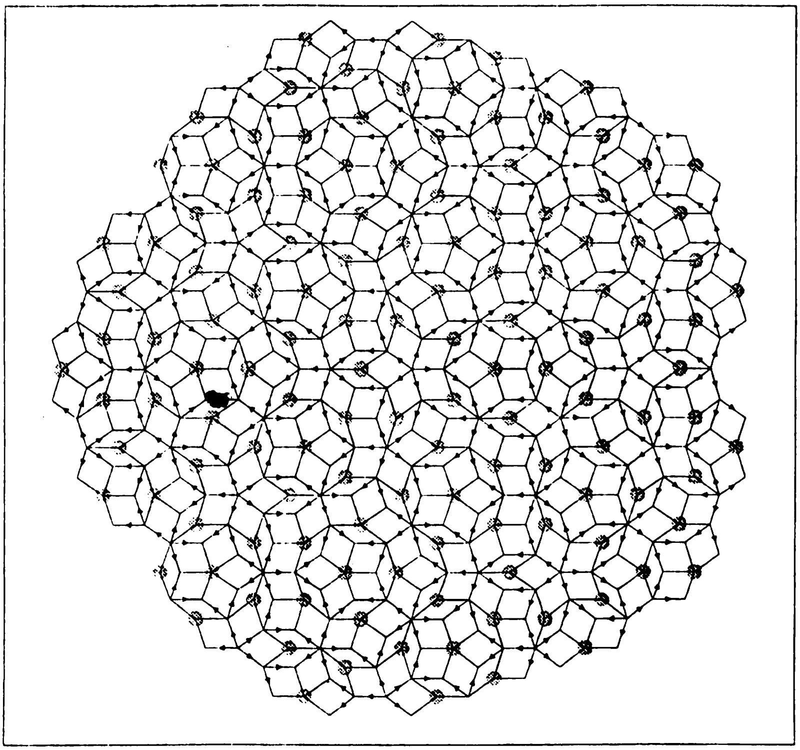


Figure 6: Piece of Penrose tiling as obtained with the prototiles and matching rules shown in Figure 5.

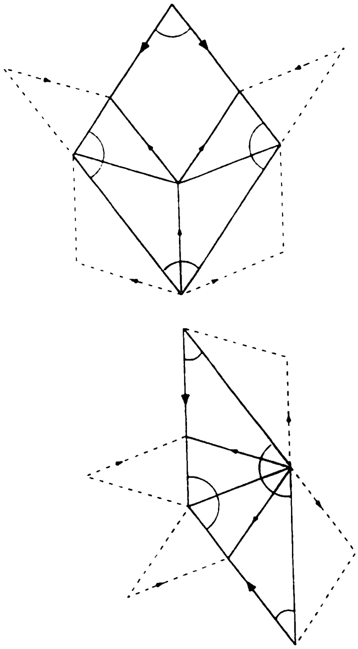


Figure 7: Iterative substitution rules for growing a Penrose tiling via self similar inflation.

The geometrical ingredient leading to Penrose tiling can be easily extended to planar tilings of any symmetries, except for two-, three-, four- and six fold rotations which allow periodic tilings. The example of a thirteen-fold tiling is shown in Figure 8.

Even the simplest iterative substitution rules seems to require more than one prototile to grow a planar quasiperiodic structure. But this has not been rigorously proved. On the other hand, one may accept to forget tiles and tiling and use motives instead. Conversely to a tile which is a bulky geometrical shape refusing overlaps, a motive is made of dots and can overlap. In the Penrose tiling of Figure 6, decagonal overlapping motives are clearly visible [11]. It has been suggested that such overlapping motives or equivalently, atomic clusters, are the pertinent basic units in the growth scheme of real quasicrystals [8, 12]. This is going to be advocated further in a forthcoming section of the paper.

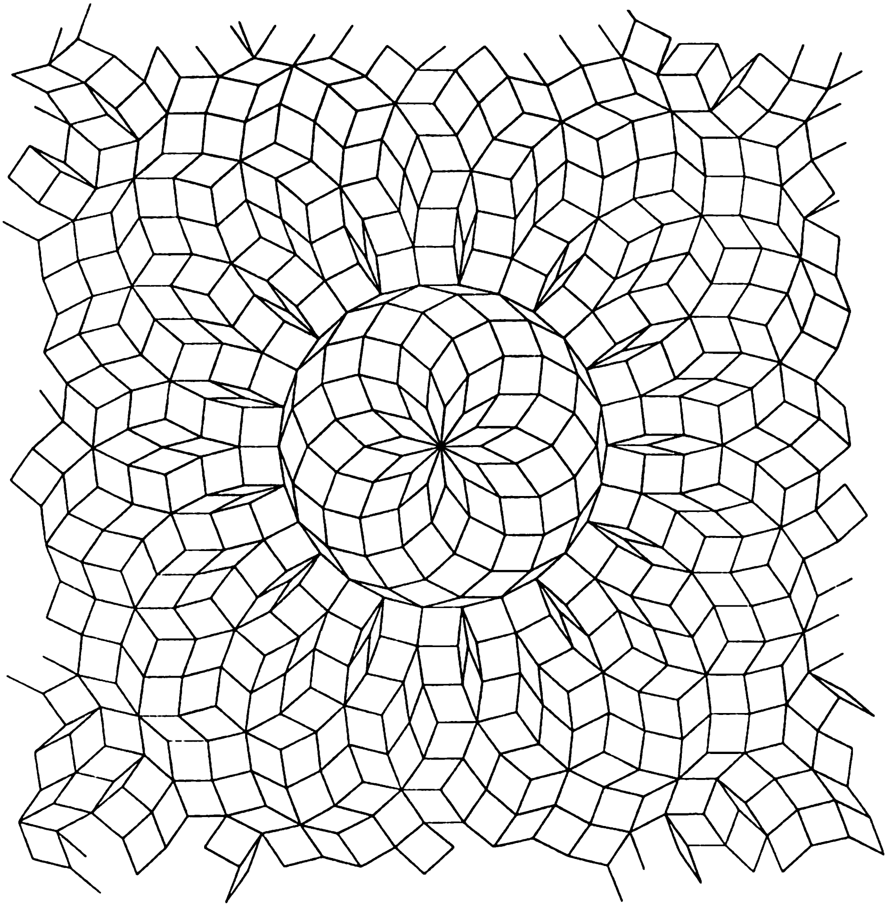


Figure 8: Example of a thirteen-fold planar tiling.

The three-dimensional situation

It is obviously possible to consider any aperiodic planar tiling of the sort described previously and to pile up them periodically in a direction perpendicular to the tiling plane. Real quasicrystals have indeed been obtained with such uniaxial symmetries but only five, eight, ten and twelve-fold rotations have been actually observed. Amusingly enough, the corresponding polygones are those which can be most easily drawn with only a ruler and a pair of compasses!

Now, if the three-dimensional space is to be tiled quasiperiodically in all directions, one must combine several rotations in such a way that the images of any point remains on a finite trajectory, inside a polyhedron which then can be used as a prototile. For instance, using the symmetry operations of a cube gives trajectories of 48 points. All other geometrical possibilities have been known for quite a long time. They include the 32 rotation groups which give rise to periodic crystal structures. Beyond them, there are only two more cases which correspond to symmetries of an icosahedral polyhedron: either the 60 rotations or 120 operations by adding mirror planes to the rotations . Fully three-dimensional quasicrystals can then be only of the icosahedral species. This is actually well consistent with physical reality.

Formal extension of the 2-dimensional Penrose tiling is straightforward. Instead of planar rhombic units, bulky rhombohedral tiles are used: they are designed in either an oblate or a prolate shape ; all edges are equal ; angles of their rhombic faces are 63.43^o or 116.57^o, precisely those found in between fivefold axes of an icosahedron. Assembling these rhombohedra to generate a 3-dimensional quasiperiodic order requires to select proper matching rules in the form of appropriate decoration of faces and vertices. The practical building of such a tiling suffers actual difficulties which make the procedure both effectively intractable and physically implausible. The hyperspace scheme offers a more acceptable alternative.

The periodic image of quasicrystals

Both crystal and quasicrystal structures can be analysed in terms of their Fourier components in that the space dependence of the density can be expressed as a sum of density waves, i.e:

$$\rho(r) = \frac{1}{V} \sum_G F(G) \exp(iG \cdot r) \tag{1}$$

For periodic crystals, the sum (1) is zero except for those G vectors which define a discrete reciprocal periodic lattice and can then be written as an integer linear combination of three basis vectors a_i^*, i.e:

$$G = ha_1^* + kA_2^* + la_3^* \tag{2}$$

in which the integers h, k, l, are the so-called Miller indices for the structure factor $F(G)$ appearing in Eq. (1).

The diffraction pattern of a quasicrystal, as the one shown in Figure 9, cannot obviously be interpreted with a lattice of G vector given by Eq. (2). But a careful investigation of the pattern suggests that actually only a few things must be modified. The density wave description by Eq. (1) is still valid; the G vectors still form a discrete set but Eq. (2) must be modified into:

$$G = n_1 a_1^* + n_2 a_2^* + n_3 a_3^+ n_4 a_4^* + n_5 a_5^* + n_6 a_6^* \qquad (3)$$

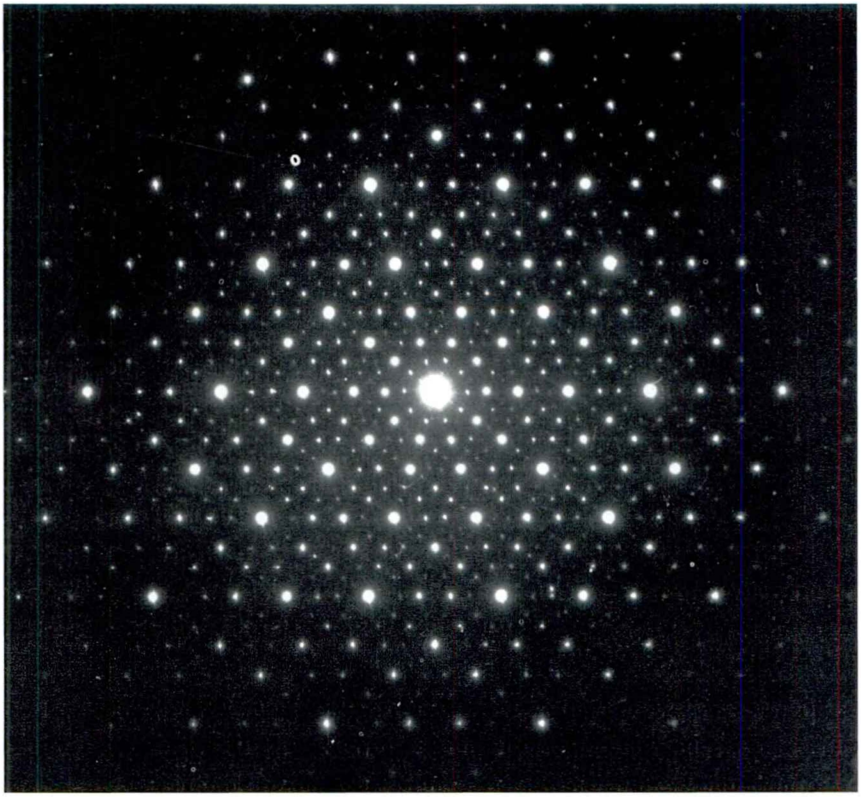


Figure 9: Electron diffraction pattern of an icosahedral quasicrystal of the AlFeCu system.

in which the n_i are integers and the a_i^* vectors are lying along the six fivefold axes of an icosahedron (Fig. 10). These a_i^*'s cannot be reduced to three members via any projection scheme on reference axes; the resulting "Miller indices" would be always fractional numbers due to irrationality of the cosine and sine functions for the angles between a_i^*'s $(63^o43or116^o57)$. Using the reference axes of the Figure 10, Eq. (3) can be given a equivalent expression in the form of three orthogonal components for the G vectors, i.e.:

$$G \begin{cases} h + \tau h' \\ k + \tau k' \\ l + \tau l' \end{cases} \tag{4}$$

$(h, h', k, k', l, l'$ are integers and τ is the golden mean).

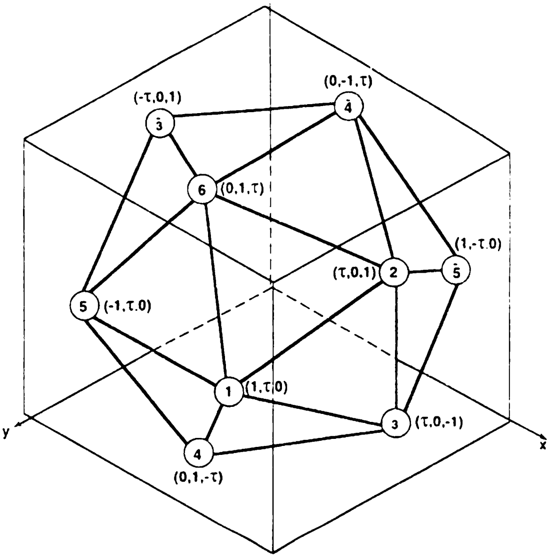


Figure 10: Icosahedron showing the fivefold axis vectors a*i with their components in an orthogonal frame:

$$\begin{array}{lll} a_1^* = (1, \tau, 0) & a_2^* = (\tau, 0, 1) & a_3^* = (\tau, 0, -1) \\ a_4^* = (0, 1, -\tau) & a_5^* = (-1, \tau, 0) & a_6^* = (0, 1, \tau) \end{array}$$

Consequences are manyfold:

– it is confirmed that the point symmetry is incompatible with periodic translational order in 3-dimension

– the G vectors do not define a reciprocal lattice but generate a set of points that fill the space densely

– the diffraction pattern is selfsimilar since tnG belongs to the set define by Eq. (3) or (4), given a vector G of this set

– and, last but not least, there is a periodic image of the quasiperiodic structure in a higher dimensional space. Indeed, Eq. (2) and (3) are formally equivalent. If Eq. (2) is used to define a 3-dim reciprocal lattice for a crystal structure, Eq. (3) can be used as well to define a 6-dim reciprocal lattice for a quasicrystal structure (or other high-dim image for other symmetries than icosahedral). Let us call $R^*_{3//}$ the space containing the 3-dim G vectors and R^*_6 such a 6-dim space containing a lattice of basic vectors e^*_i which project on a^*_i into $R^*_{3//}$. Then the vectors $\mathcal{G} = \sum_{i=1}^{6} n_i e^*_i$ span this 6-dim lattice when the $G = \sum_{i=1}^{6} n_i a_i$ span the $R^*_{3//}$; each $\mathcal{G}$ project into $R^*_{3//}$ on one and only one G vector.

To the dense distribution of G vectors in $R^*_{3//}$ corresponds a density distribution ρ_3 in a direct space $R_{3//}$ which is dual of $R^*_{3//}$, via the Eq. (1). $R_{3//}$ is our physical space and ρ_3 is the structure of the quasicrystal of interest. Similarly, to the periodic reciprocal lattice $\mathcal{G}$ in R^*_6 corresponds a direct periodic lattice bearing a density distribution ρ_6 in a direct space $R6$, which is dual of R^*_6. The density distribution ρ_6 can be dubbed as the periodic image of the quasicrystal structure ρ_3. Mathematics tell us that if distributions in two different spaces are related via projection, the Fourier transformed distributions in the dual associated spaces relate via a cut procedure. The correspondence scheme can then be summarised as follow:

$$\begin{tikzcd}
\rho_6(r) \arrow[r, leftrightarrow, "FT"] \arrow[d, leftrightarrow, "\text{cut of } R_6 \text{ by } R_{3//}"'] & F(\mathcal{G}) \arrow[d, leftrightarrow, "\text{Projection of } R^*_6 \text{ onto } R^*_{3//}"] \\
\rho_3(r_{//}) \arrow[r, leftrightarrow, "FT"'] & f(G)
\end{tikzcd}$$

in which FT means of course Fourier transform and $r_{//}$ is the components in R_3 of the 6-dim vector r.

Using the high-dimensional image is a very efficient and economical way to describe a quasicrystal. We are thus back to normal crystallography in which one needs only to know a unit cell and a metric to design the whole structure. Moreover, this gives the easy way to operate diffraction experiment for structure determination: the diffraction peaks are indexed with six Miller indices according to Eq. (3) or (4) and then "lifted" into R^*_6 formally to produce $F(\mathcal{G})$ whose Fourier transform gives $\rho_6(r)$; a final cut of $\rho_6(r)$ by our 3-dim physical space generate the quasiperiodic structure $\rho_3(r_{//})$.

It is, however, useful to describe in somewhat more details both the "Bravais" lattice and the unit cell motive of the periodic image $\rho_6(r)$. It is first of common use to refer to the physical space $R_{3//}$ as the parallel space or the internal space; the 3-dim space that must be added to $R_{3//}$ in order to complete R_6 is dubbed complementary space, or perpendicular space (hence $R_{3\perp}$ with its dual $R^*_{3\perp}$)

or external space. Each basis vector e_i^* of the reciprocal lattice in R_6^* projects on one $e_{i//}^* = a_i^*$ and on one $e_{i\perp}^*$ into $R_{3//}^*$ and $R_{3\perp}^*$ respectively. The scheme which relates the quasiperiodic structure to its periodic image imposes that at any symmetry operation in $R_{3//}^*$ correspond associated symmetry operations in $R_{3\perp}^*$ and R_6^*; in other words the reciprocal lattice in R_6^* is invariant in any operation which preserve $e_{i//}^*$ and $e_{i\perp}^*$. Thus the six fivefold planes $(e_{i//}^*, e_{i\perp}^*)$ are mirror planes of the 6-dim reciprocal lattice which, hence, is cubic and so is the direct lattice in R_6 (an N-cube has N mirror planes perpendicular to the rotational axes of the highest order). There are also ten threefold and fifteen twofold mirror planes. It is said that the point group of the lattice in R_6 is isomorphic to the icosahedral point group. The subspaces $R_{3//}$ and $R_{3\perp}$ have the same symmetries.

Now what does the density distribution $\rho_6(r)$ in this cubic lattice look like? First of all, the cut of $\rho_6(r)$ by $R_{3//}$ must generate a set of points that will accept atom positions. Thus, $\rho_6(r)$ must have no thickness in $R_{3//}$, i.e., must be a distribution of objects being "flat" in $R_{3//}$ or, in other words, completely located into $R_{3\perp}$. Let us call $A_{3\perp}$ these 3-dim objects which have been commonly named **Atomic Surfaces** (AS). The main requisite to design the $A_{3\perp}$ are the following:

- they must be 3-dim polyhedra having symmetries of an icosahedron.

- they must obey a so-called hard core condition which constrains their size and shape so that cutting by $R_{3//}$ does not generate unphysically too short atom pair distances.

- they must allow energy translational invariance of the quasiperiodic structure parallel to both $R_{3//}$ and $R_{3\perp}$ spaces. Flatness in $R_{3//}$ guarantees translation invariance in this subspace. Translation invariance in $R_{3\perp}$ means that the $A_{3\perp}$ must form subset in which piecewise connection prevents annihilation/ creation of atoms under any $R_{3\perp}$ translation, while structures with differences into their detailed geometry may be generated. This is called the closeness condition [13].

- density and composition of the quasicrystal also operate on size, shape and partitioning of the atomic surfaces.

The simplest shape that may be attributed to the $A_{3\perp}$ volumes in spherical. But reducing the $A_{3\perp}$ objects of the high-dim image to their spherical approximation is obviously accepting a low resolution description of the structure. Here, the expression "low resolution" means that in the Fourier transform of the $A_{3\perp}$ atomic surfaces the high-order Fourier components are not really accounted for. Sphere sizes are mostly fixed by density and composition constraints.

One possible method of introducing the high-order Fourier components is to parametrize the atomic surfaces in terms of linear combinations of symmetry-adapted functions associated with their point group symmetry [14]. In the case of an icosahedral quasicrystal, the perpendicular space is three-dimensional. The *spherical harmonics* are then a natural choice for expressing the boundaries of any radial functions $r(\theta, \phi)$. Hence the set of symmetry-adapted orthonormalized functions, invariant for icosahedral point group symmetry, can be chosen according to the decomposition

$$r(\theta, \phi) = \sum_{li} a_{li} Z_{li}(\theta, \phi) \tag{5}$$

with

$$Z_{li}(\theta, \phi) = \sum_{m} Z_{lm}(i) Y_{lm}(\theta, \phi)$$

in which Y_{lm} are the classical spherical harmonics, Z_{lm} are determined by the point group symmetry of the $A_{3\perp}$ plus the normalization conditions of Z_{li}, and a_{li} are continuous parameters to be fitted in structural diffraction analysis: the index i allows for the possible existence of several orthogonal invariant functions within the same subspace of functions having a given value of l.

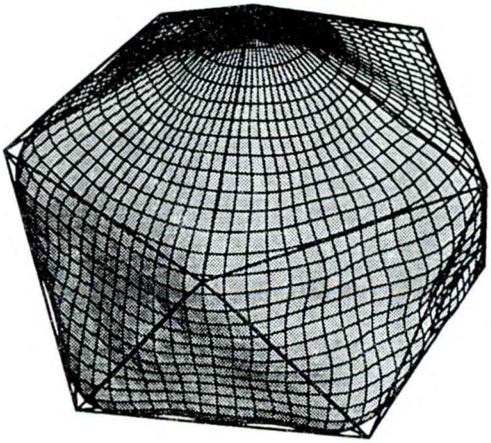


Figure 11: Approximation of an icosahedron by four spherical harmonics.

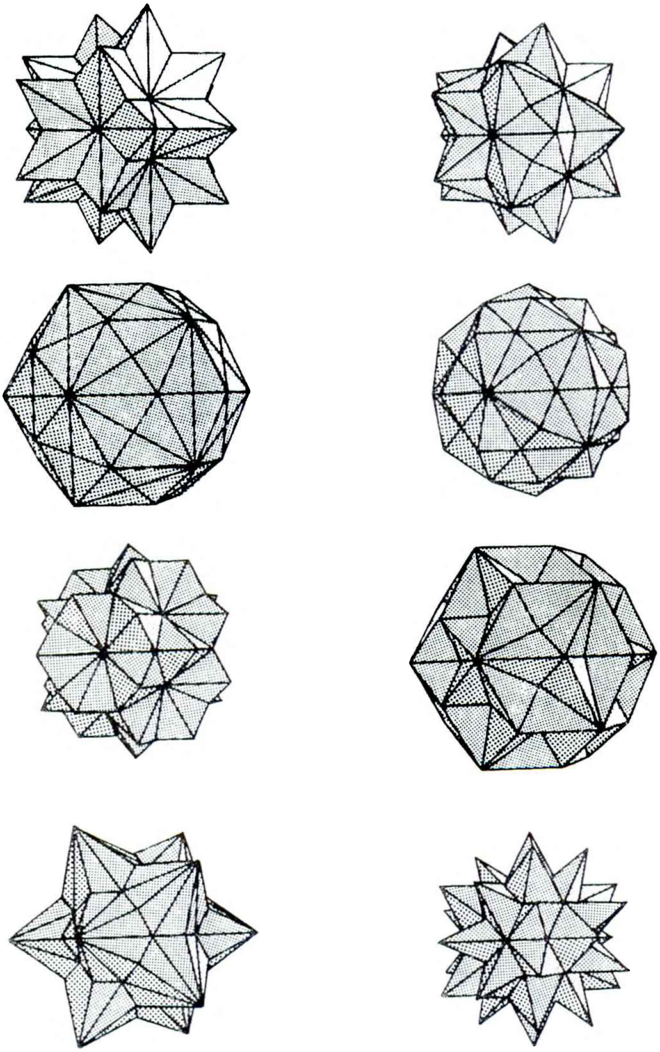


Figure 12: The eight basic polyhedra bounded by 2-fold planes in $R_{3\perp}$ for the 6-dimensional image structure of icosahedral quasicrystals [13].

If the point group is large enough, there will be many empty subspaces. For instance, with the icosahedral point groups there is a single invariant function (for l up to 15) only for l values of 0, 6, 10 and 12. Beyond $l = 15$, contributions to Eq. (5) are expected to be very weak. As an illustration, Figure 10 shows that these four components are sufficient for the reconstruction of an icosahedron certainly beyond experimental resolution. Any physical constraint on the $A_{3\perp}$, such as those induced by realistic atomic distances, density, composition, etc., can be introduced in the refinement via penalty functions. This is probably a good basis for allowing successful least-squares refinement processes to obtain realistic faceted $A_{3\perp}$ objects.

The hard-core and closeness conditions mentioned above are satisfied if the $A_{3\perp}$ objects are bounded by piecewise connected surfaces, mostly parallel to the complementary space, without overlapping in this space, and globally invariant under point group symmetries. These conditions are satisfied for surface boundaries which are mirror planes of the structures. As a consequence, possible faceted $A_{3\perp}$ volumes for the 6-dim images of icosahedral quasicrystals would have 2-fold, 3-fold, or 5-fold plane boundaries. This point has been demonstrated in detail for 2-fold plane boundaries, [13] and the shapes of the eight corresponding polyhedra are presented in Figure 12. The acceptable volumes for decorating the 6-dim cube must be one of these polyhedra, or any τ-scaling and/or intersection of them. Obviously, this leaves a number of alternative solutions and the formal faceting conditions, as they stand, have to be considered mostly as a negative test to reject improper solutions.

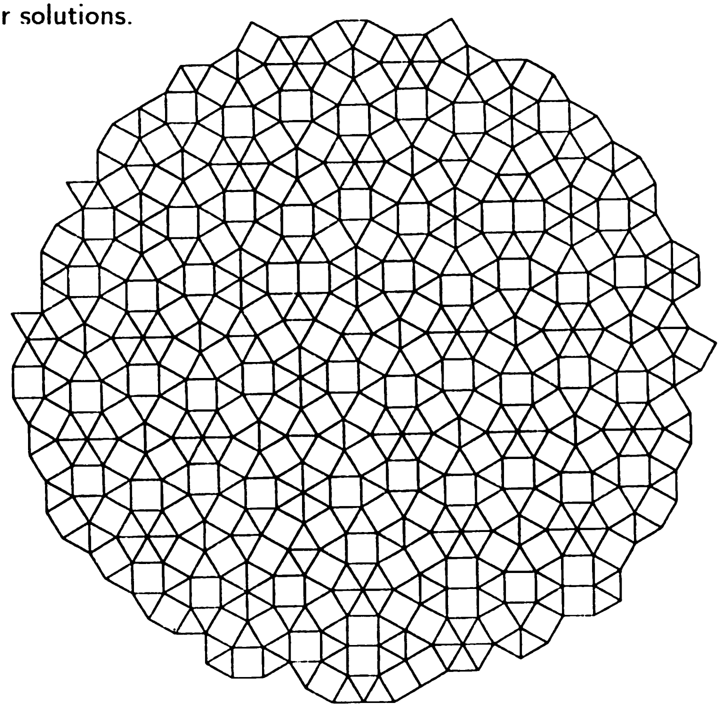


Figure 13: Finite portion of a dodecagonal planar quasicrystal with square- triangle prototiles [15].

So far we have assumed that the $A_{3\perp}$ atomic objects of the high-dim image are (faceted) polyhedra. This has induced conditions for these atomic objects. It may be of interest to consider whether the polyhedral solution is imposed in every case. There is no general answer to this question, and the point has received very little investigation, with restriction to 1-dim and some 2-dim quasiperiodic structures. One example has been reported by Baake et al. 15. They generated a quasiperiodic dodecagonal tiling of the plane using squares and regular triangles arranged with simple deflation-inflation symmetries (Fig. 13). This 2-dim structure has been "lifted" (embedded) into a 4-dim periodic lattice and the acceptance domain (or $A_\perp$ objects) has been iteratively constructed to generate the vertex set of the square-triangle tiling. The result is shown in Figure 14. The procedure leads to a fractally bounded $A_\perp$. It can be shown that there is no polyhedral alternative solution if the square-triangle tiling is to be obtained with a single type of $A_\perp$.

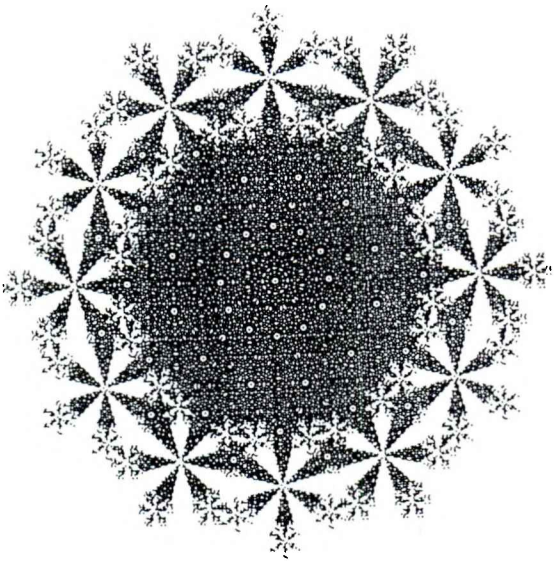


Figure 14: Acceptance domain in $R_\perp$ filled by lifting 32000 vertices of the tiling shown in Figure 13 [15].

The hyperspace image of the Fibonacci chain

As a quasiperiodic 1-dim structure requires at least two different segments for avoiding periodicity, the corresponding periodic image is at least two- dimensional. In the absence of 1-dim point group, this 2-dim Bravais lattice may be any of the five existing ones. A square lattice may be the best choice for the sake of geometrical simplicity and also for mimicking at best the 6-dim cubic lattices that correspond to icosahedral real quasicrystals. The atomic surfaces $A_\perp$ must be "flat" in the direction $R_\perp$ which is perpendicular to the direction $R_{//}$ of the chain. Hence, they are simple straight line pieces with the length Δ, as shown in Figure 15. The position of $R_{//}$ (and then $R_\perp$) in R_2 is fixed by the angle α of $R_{//}$ with respect to the horizontal raw of the square lattice. If $\tan\alpha$ is an irrational number, the structure of the chain is aperiodic, with two different tiles $L = a\cos\alpha$ and $S = a\sin\alpha$. The closeness condition is fulfilled provided that

$$\Delta = a(\cos\alpha + \sin\alpha)$$

The average density of the chain must be transferred to its image and, hence, is equal to $\Delta/a^2 = (\cos\alpha+\sin\alpha)/a$. Finally, L/S being equal to τ in a Fibonacci chain fixes the angle α and there is no free parameter left for the periodic image.

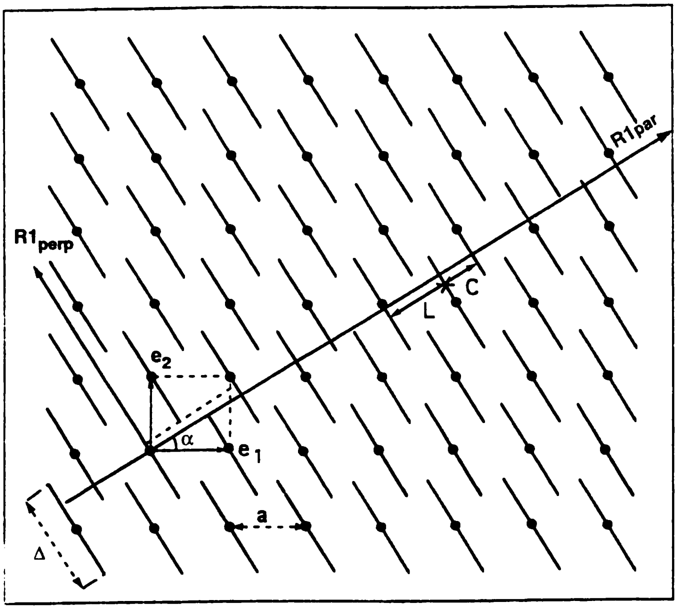


Figure 15: Toy-model of the hyperspace image for a quasicrystal. Here is shown a 1-dimensional Fibonacci chain and its 2-dimensional periodic image as a decorated square lattice.

Moving the $R_{//}$ direction across the decorated square lattice generates all the equiprobable structures with the same energy but differing locally in their geometry features. All these isomorphic structures relate to each other via "atomic jumps", so-called phason-jump, due to flipping in $L - S$ sequences (see Fig. 15).

The physical generation of quasiperiodic patterns may be very simple

Detailed description of the atomic surfaces using diffraction data with real quasicrystals is very difficult, may be impossible to be achieved and, so far, only low resolution structures have been obtained. But the main drawbacks of the high dimensional scheme are twofold: (i) the crude resulting physical structure in the three-dimensional space is concealed in a list of atomic positions without any clear guides on how to design straightforwardly space occupation and, even more disturbing, (ii) there is a total lack of how to grow the whole structure by adding atomic positions one by one or, to the least, cluster by cluster.

Actually, growing a piece of matter within certain rules for short and long range order is not an easy task. For regular periodic crystals, the sequence of atoms that exists in a seed cluster repeats exactly again and again; so it appears that the atom to be added must interact only with a small number of atoms at some places on the cluster surface. Moreover, there is a single ground state structure for a given space group which means that the structure is energetically stabilised and can be grown perfectly. The various mathematical procedures that have been used so far to generate quasiperiodic lattices are somewhat suggestive that growing a perfect quasicrystal would be a daunting task. The sequence of atoms never exactly repeats, so that atoms added to the surface of a cluster must interact with each atom in the seed cluster to ensure that is sticks at a site consistent with perfect quasiperiodic order. As the cluster grows, this requirement imposes arbitrary long-range interaction, which is physically implausible. Matching rules, particularly well exemplified with the Penrose tiling, would then seem to offer a potential mitigating factor to these growth problems. The classical edge-matching rules are typically indicated by placing different arrows on the edges of tiles that constrain the way two tiles must match edge-to-edge. Penrose clearly showed that the only plane-filling tiling consistent with the matching rules is a perfect Penrose tiling. Do these edge-matching rules also represent viable local rules for growing a tiling by adding one tile at a time to a random chosen edge? Unfortunately not. Mistakes are made which are not revealed at once and the catastrophe can be appreciated only after many further building steps. Removing tiles for another or other tries is obviously a dismal failure of the edge-matching rules as a growth procedure [16]. Replacing edge-matching rules by "forced vertex-matching rules" has certainly relaxed part of the difficulties but the basic drawbacks remain the same [17]. Recently, Moody and Patera [18] have described a mathematical procedure to grow quasiperiodic structures via strictly local rules in which a point is added to the growing patch if and only if (i) an ideal configuration is not violated and (ii) the point phase in the physical space remains within a chosen range of values. But, still, local phases are correlated to each other and are not exactly direct space parameters. However, it is possible to keep the spirit of the method and derive a purely local growth procedure that, moreover, is consistent with structure and properties of real quasicrystals.

Among the many properties of quasicrystals observed so far, two of them deserve to be selected for the present purpose: (i) their structure appears as basically built from packing of very rigid atomic clusters with "forbidden" symmetries and (ii) their shear modulus [19] is as large as those obtained with semiconductors revealing strong directional atomic bonding. One very simple way to preserve what can be preserved of that while growing the structure is to proceed as follows [20]:

(i) A "star" of atomic bonding is deduced from a given cluster of atoms. In the two-dimensional example of a decagonal centred cluster (Fig. 16(a)), the "star" is made of the ten radial vectors linking centre to vertices and dispatched at $2\pi/10$ angles from each other (cluster requisite).

(ii) The above "star" of vectors defines the only possible translations, originating from an existing site at the surface of the growing structure, to create new sites (directional bonding requisite).

(iii) New sites that would introduce too short pair distances, with respect to the already existing sites, are rejected (finite density requisite). The point is illustrated with Figures 16(b) and 16(c) which show a second growing step from a decagonal cluster and what is preserved if too close positions are erased; in this example, the threshold pair distance has been fixed to the length of the decagonal edges. Figure 17 shows a piece of one structure that can be iteratively generated by pursuing again and again the above addition procedure with the same decagonal star.

That this procedure is easily feasible and allows structure growth via purely single local rules is then now obvious. It remains that at least two basic questions must be carefully addressed: (i) does this mechanism generate a single state structure or, conversely, a variety of "energetically" equivalent structures and (ii) does it truly result into quasiperiodicity.

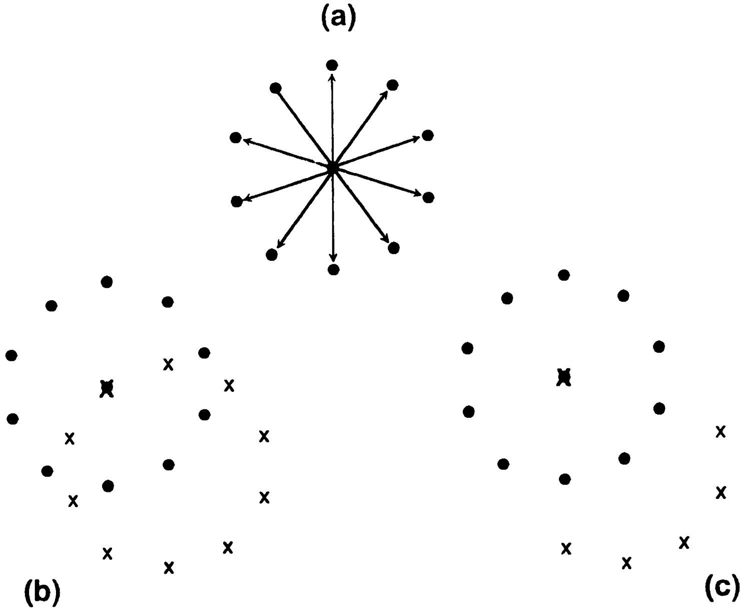


Figure 16: (a) Decagonal cluster of sites defining a ten-fold star of vectors; (b) A second decagon of sites (•) has been added to the one shown in (a), with its centre on a vertex site; (c) same as in (b) except that sites overly close to those of (a) are removed.

If combinations of two-, three-, four and six-fold stars of vectors are used in the growing sequence, one gets trivially lattices of sites which are periodic crystal lattices and it has been well known for almost a century that one given star generates one and only one lattice. With pentagonal, decagonal, icosahedral ... stars it is not possible any more to generate lattices and fully dense set of sites are obtained instead, it there is no restriction made on pair distances. The physical constraint on density via the rejection rule of too close atoms is then usefully applied. But, clearly, which particular new sites have to be rejected strongly depends on which sites are already there, which in turn depends on the order chosen to explore the surface sites of the growing structure. This exploration can be made at random which gives prospects for an infinite number of very slightly different structures (Fig. 18); or else one may decide to circle the seed always in, say, a step by step clockwise exploration of the surface sites, which generate the more regular structure of the family with an overall symmetry axis in its centre. Conclusively, the procedure does not generate a single state structure but, rather, a farnily of very similar packing of sites.

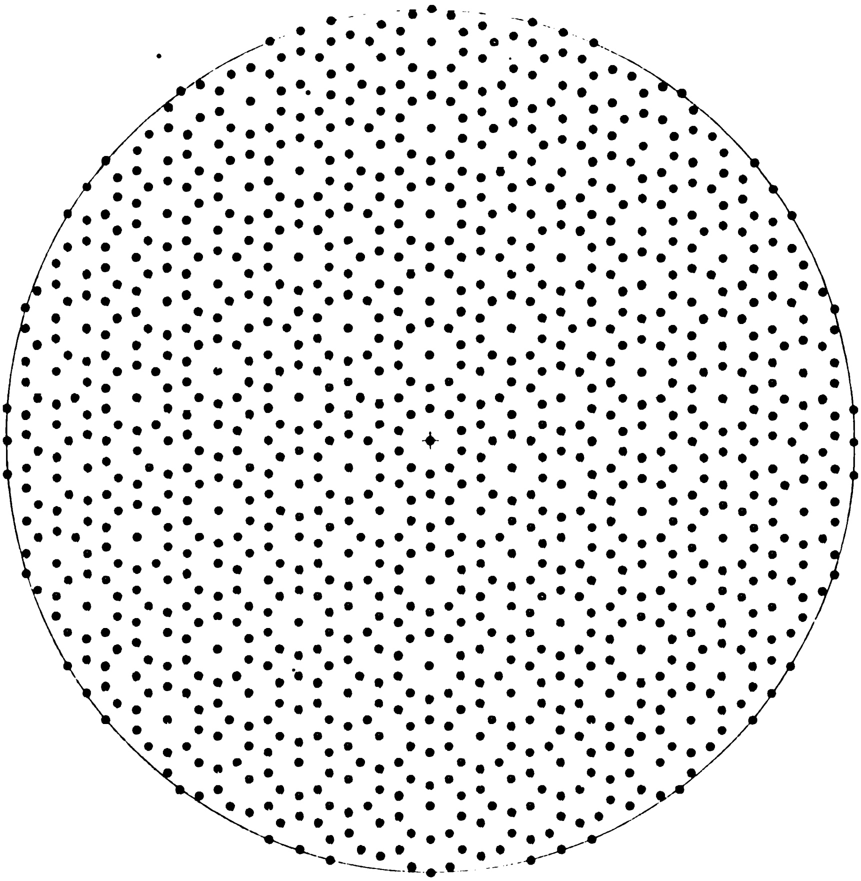


Figure 17: A piece of structure grown using a tenfold star and shortest distances slightly below the edge length of the decagon.

The question of true quasiperiodicity of the obtained structures is less easily overcome even if encouraging insight can be proposed. Of course, the trivial requisite is that vector stars that genèrate crystal lattices must be strictly avoided. This being said, the geometrical scheme described here is more related to the hyperspace description than it would seem at first sight. Indeed, specifying the high dimensional Bravais lattice by its translation vectors corresponds to the selection of a star of vectors in the three-dimensional physical space. Adding to this hyperlattice some atomic surfaces is equivalent to define acceptance domains for permitted atomic pair distances and directions. Thus, our growing scheme exhibits the two main ingredients that should allow to generate quasiperiodic structure. As a positive illustration of this statement, Figure 19 presents an impressive comparison between a structure obtained via the "star-short distance" scheme ($SSDS$), the one already shown in Figure 2 actually and, on the other hand, a five-fold planar cut of an atomic arrangement deduced from diffraction data with an AlPdMn real quasicrystal via the hyperspace method and within spherical approximations for the atomic surfaces. This is a strong support to the double suitability of the procedure to generate structures that can be quasiperiodically ordered and even describe real quasicrystal quite nicely. But it seems that the conclusion is not universal whatsoever. A contradictory example is given in Figure 20 which shows a structure grown with a pentagonal star and short-distance threshold equal to the pentagon radius; this is obviously nothing else than a pentatwinned regular crystal. This is actually not deeply surprising and may be related to the well known possibility to generate either quasicrystals or some sorts of twinned crystals via the hyperspace scheme. At this stage we cannot refrain from thinking that, finally, quasicrystals are less difficult to describe than suggested so far and, also, that they can be grown via the simplest mechanism, i.e. adding one atom at a time.

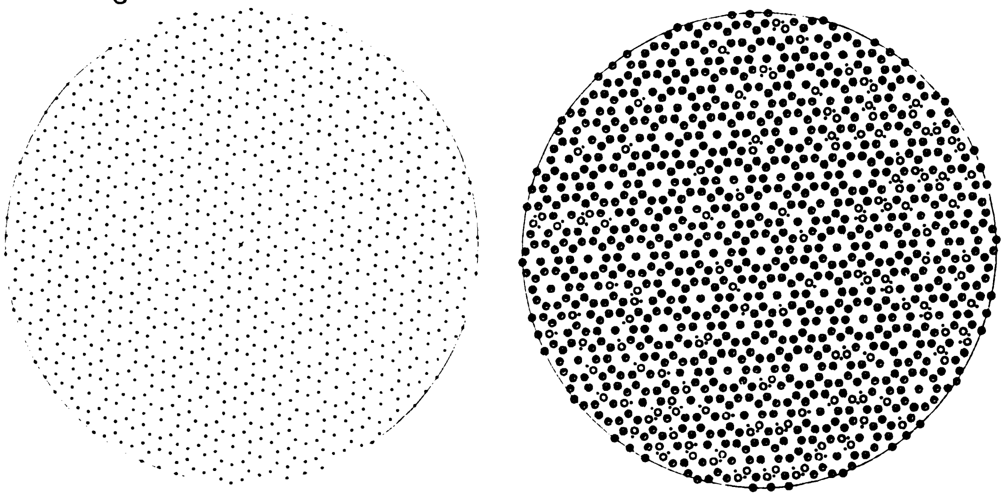


Figure 18: (a) Same conditions as in Figure 2 but with another random exploration of the surface sites; (b) Figure 17 and 18(a) have been superimposed. Solid dots appear whenever the two structures coincide.

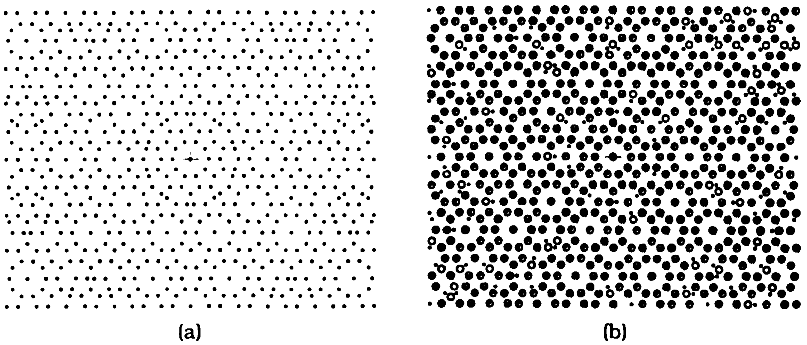


Figure 19: (a) One planar cut of the structure of a real quasicrystal; (b) Figure 17 and 19(a) have been superimposed. The overlap is quite impressive [20].

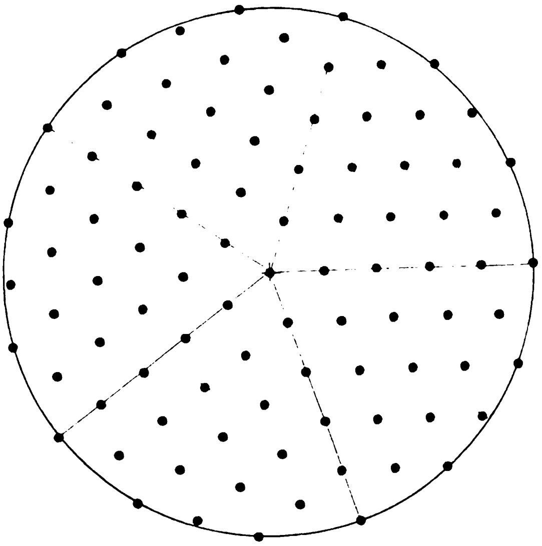


Figure 20: The structure grown with a five-fold star and shortest distance equal to the radius of the pentagon is a pentatwinned crystals.

A few more words about the geometry of real quasicrystals

The structure of real quasicrystals is generally obtained from diffraction data via the hyperspace approach to their periodic images. As already said, one must unfortunately be contented with rather low resolution structure, due to the poor definition of the atomic surfaces actually reached. However, clear building rules are revealed, with evidences of both geometrical and chemical order. The point is going to be exemplified with the structure of AlPdMn quasicrystals which offer the rare privilege to be growable as centimetre size single grain [21].

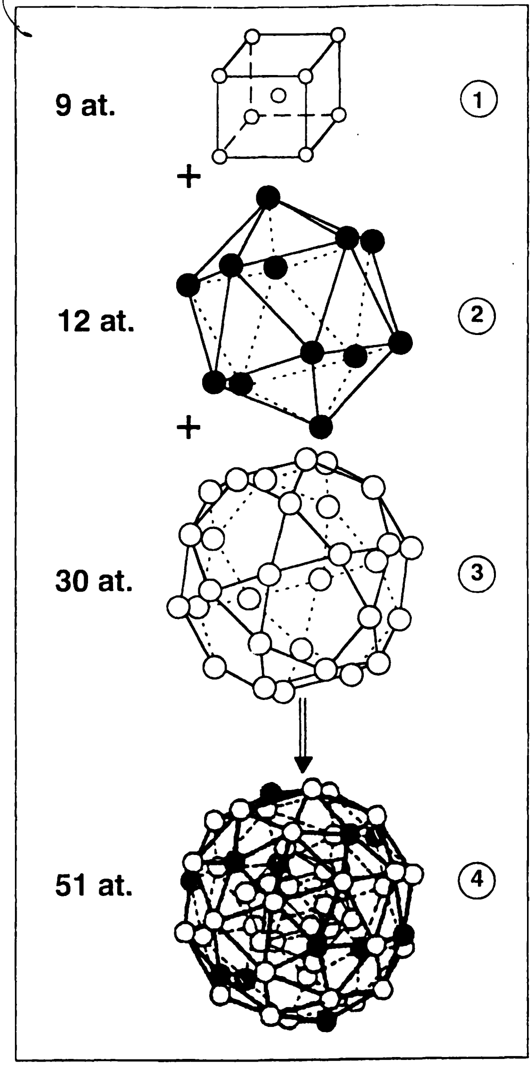


Figure 21: Successive atomic shells of a pseudo-Mackay icosahedron (PMI).

First of all, everything in the structure is based on atomic units containing 51 atoms in total, named pseudo-Mackay icosahedra (PMI) hereafter, and made of three centrosymmetrical shells as shown in Figure 21: an inner small centred cubic core of 9 atoms, an intermediate icosahedron of 12 atoms, and an external icosidodecahedron of 30 atoms. The last two shells have practically equal radii and constitute altogether the boundary of the PMI whose diameter is slightly less that 10 Å. Apart from this well-defined geometry, the PMI's show three different chemical compositions: one family (PMI-A) has 6 manganese plus 7 palladium atoms on the icosahedron and centre sites and 38 aluminium atoms elsewhere while the two other families (PMI-T) exhibits 20 or 21 palladium atoms among the 30 of the icosidodecahedron, the rest (30 or 31 atoms) being aluminium atoms. The calculated atomic density of an individual PMI is 0.064 atoms/Å^3, which compares quite well with the measured density of the bulk material, within experimental accuracy. It is, however, fair to say that several ingredients in the description of the PMI's do not show up directly from diffraction data. The Patterson analysis strongly suggests that the PMI cores are made of about 8-9 atoms distributed into pieces of dodecahedra; it is indeed a speculation to state that these pieces are arranged in centred cubic geometry. It is probably better to consider that we have a dodecahedral core whose 20 sites are only partially occupied.

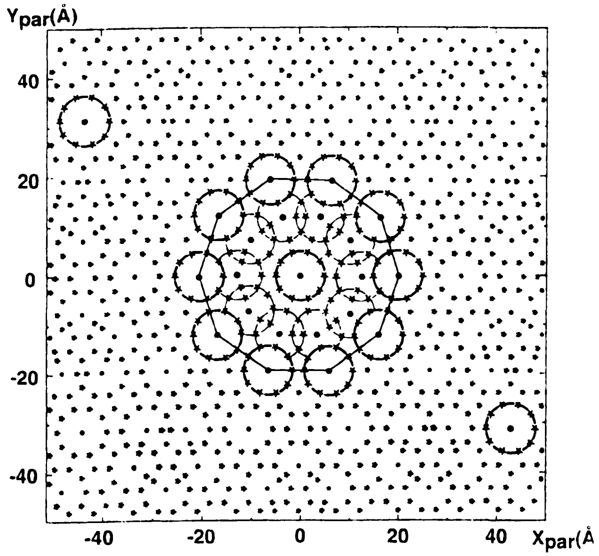


Figure 22: Fivefold planar cut of the structure of the AlPdMn quasicrystal. Rings of ten atoms are equatorial section of a PMI. The τ^3 and τ^2-inflated rings are visible [21].

Then, these PMI units combine to reproduce a selfsimilar geometry within inflation by a scale factor close to τ^3. This is shown in Figure 22 which presents the cut of a piece of the structure by a plane perpendicular to a fivefold axis. In the figure centre, the equatorial section of a PMI shows up. Around this central PMI, there are 42 PMI's whose centres are distributed on the combined sites of the icosahedron plus the icosidodecahedron of a big PMI with a radius τ^3 as large as that of the base unit (about 42 Å, namely). An intermediate shell, with τ^2 inflated radius, is also visible in Figure 22. This shell is made of overlapping PMI's and is the inflated modification of the partially occupied dodecahedral core of the basis PMI unit. The overlapping is such that preservation of the density is ensured, within very low residual fractality.

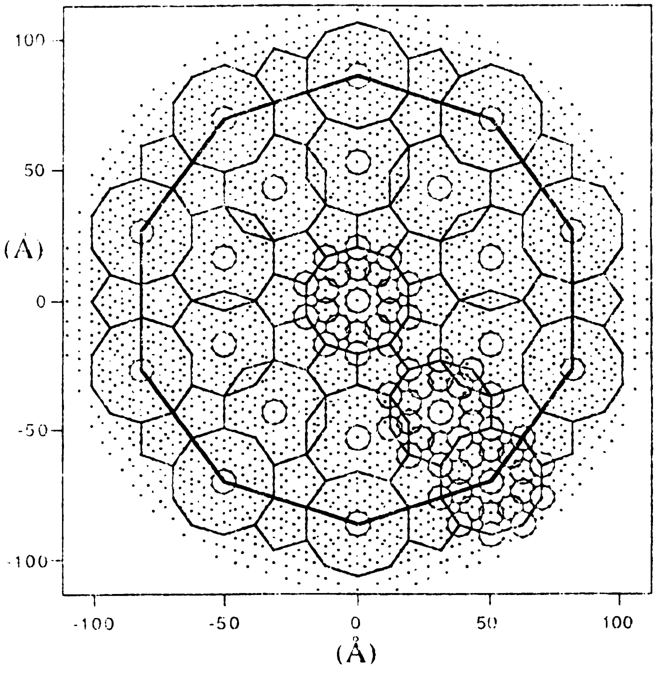


Figure 23: Same as in Figure 22 but with an additional inflation step of the structure.

The structure subsequently develops via successive steps of τ^3 inflation operations. Figure 23 shows a planar projection of a layer of atoms presenting the result of a $\tau^3 \times \tau^3$ inflation, with a $\underline{o}tau^3$-PMI in the centre, a shell of 42 τ^3-PMI on a $\tau^3 \times \tau^3$ radius "sphere" and the $\tau^3 \times \tau^2$ intermediate shell of overlapping truncated τ^3-PMI. Pentagonal "tiles" at various scales are also visible in the figure; they come from PMI and inflated PMI whose equatorial plane is not in the figure. In conclusion, at any inflation stage, we have a cluster of PMI clusters.

The Figures 19(a) and 23 are actually slightly different representations of the same data which, at two different levels, demonstrate the leading aspect of the basic cluster stability into natural growth of quasiperiodicity. It has been also observed that the selfsimilarity rules that describe the geometry also apply to the chemistry of quasicrystals [21].

References

[1] Senechal M., *The geometry of quasicrystals*, (Cambridge University Press) 1995; Janot C., *Quasicrystals: a primer* (Oxford University Press, Cambridge) 2nd edition 1994.

[2] Tsai A. P., Inoue A. and Masumoto T., *Phil. Mag. Lett.* **62** (2) (1990) 95.

[3] Kycia S. W., Goldman A.I., Lograsso T.A., Delaney D.W., Sutton M., Dufresne E., Bryning R. and Rodricks B., *Phys. Rev.* B **48** (1993) 3544.

[4] Boudard M., Bourgeat-Lami E., de Boissieu M., Janot C., Durrand-Charre M., Klein H., Audier M. and Hennion B., *Phil. Mag. Lett.* **71** (1995) 11; Audier M. and Hennion B., Durand-Charre M. and de Boissieu M., *Phil. Mag.* B **68** (1993) 607.

[5] Boudard M., de Boissieu M., Janot C., Heger G., Beeli C., Niessen H. U., Vincent H., Ibberson R., Audier M. and Dubois J. M., *J. Phys. Condens. Matter* **4** (1992) 10149.

[6] Janot C., *Europhysics Letter* **27** (1996) 60.

[7] Pierce F. S., Guo Q. and Poon S.J., *Phys. Rev. Lett.* **73** (1994) 2220 ; Guo Q. and Poon S.J., *Phys. Rev. Lett.* (in press).

[8] Dzugutov M., *Phys. Rev.* A **46** (1992) R2984; Phys. Rev. Lett. **70** (1993) 2924; Quemerais P., *J. Phys. I* (France) 4 (1994) 1669.

[9] Zurkirsh M., Atrei A., Erbudak M. and Hochstraser M., *Phil. Mag. Lett.* **73** (1996) 107; Ebert P., Feuerbacher M., Tamura N., Wollengarten M. and Urban K., *Phys. Rev. Lett.* **77** (1996) 3827.

[10] Penrose R., *Bull. Inst. Math. Appl.* **10** (1974) 266.

[11] Burkov S. E., J. Phys. I (France) 2 (1992) 695; Gummelt P., in *Quasicrystals* (Ed. C. Janot and R. Mosseri) World Scientific, Singapore (1995) p. 84; *Geometriae Dedicata* (in press).

[12] Jeong H. C. and Steinhardt P. J., *Phys. Rev. Lett.* **73** (1994) 1943; Khanna S. N. and Jena P., Phys. Rev. B **51** (1995) 13705; Janot C. and de Boissieu M., *Phys. Rev. Lett.* 72 (1994) 1674.

[13] Katz A. and Gratias D., *J. Non-Cryst. Solids* **153–154** (1993) 187; Katz A., in *Number Theory and Physics* (Ed. J.M. Luck, P. Moussa, W. Waldschmidt and C. Itzykson) Springer, Berlin (1990) p. 100; Katz A. and Gratias D., in *Quasicrystals* (Eds. C. Janot and R. Mosseri) *World Scientific*, Singapore (1995) p. 164.

[14] El Coro L., Perez-Mato J.M. and Madariaga G., *J. Non-Cryst. Solids* **153–154** (1993) 155.

[15] Baake M., Klitzing R. and Schlottman M., *Physica A* **1991** (1992) 554.

[16] Penrose R., in *Introduction to the Mathematics of Quasicrystals* (Ed. M. Jaric) Academic Press, New York, 1989, p. 53; Gardner M., *Sci. Am.* **236** (1977) 110.

[17] Onoda G. Y., Steinhardt P. J., Di Vincenzo D. P. and Socolar J. E. S., *Phys. Rev. Lett.* **60** (1988) 2653.

[18] Moody R. V. and Patera J., *Lett. Math. Phys.* **36** (1996) 291.

[19] Tanaka K., Mitarai Y. and Koiwa M., *Phil. Mag.* A **73** (1996) 1715.

[20] Janot C. and Patera J., *Phys. Rev. Lett.* (submitted).

[21] Janot C., Phys. Rev. B **53** (1996) 181; *J. Phys.: Condens. Matter* (submitted).

Topological Quantum Field Theory: A Prosperous Link Between Physics and Mathematics

José M. F. Labastida
Universidade de Santiago
Santiago de Compostela, Spain

Abstract

Quantum Field Theory has played a fundamental role in our understanding of the behavior of elementary particles. In the eighties it was discovered that quantum field theory could also be a very useful tool to study some aspects of low-dimensional topology, and the concept of Topological Quantum Field Theory was introduced. The richness of quantum field theory encoded in its different methods of study has been applied to this new concept, and new unexpected results have been obtained. The introduction of Seiberg-Witten invariants and of their relation to Donaldson invariants on four-manifolds, as well as the construction of integral representations of Vassiliev invariants for knots and links on three-manifolds, are two of the most salient accomplishments of topological quantum field theory. These have been achieved by a combination of some of the perturbative and nonperturbative methods of quantum field theory. From these results there emerges a new picture for some sets of topological invariants in which these are classified in terms of universality classes.

1 Introduction

During the last decade we have witnessed the emergence of a remarkable new relation between physics and mathematics. The most advanced elements of theoretical physics have become tools to create new mathematics. This type of relation is unprecedented in this century. It is also different than the usual relations in previous centuries in which often new mathematics were created because they were needed to describe physical situations. In the present case there is no such a need: physical theories are now used because they are able to provide new insights in mathematics whose present relevance comes entirely from the mathematical side. The field of theoretical physics which takes part in this relation is quantum field theory, and the special quantum field theories which are involved are called topological quantum field theories (TQFTs).

Quantum field theories are physical theories which are both quantum and relativistic. This means that they implement consistently two of the main physical principles discovered in this century: quantum mechanics and special relativity. These theories are therefore used to describe physical situations in which quantum and relativistic effects are important. They have been very successful in the description of the behavior of elementary particles at high energies. The Standard Model, which is based in quantum field theory, has been confronted with experiments to a high degree of accuracy. However, quantum field theory and the Standard Model itself have many problems and leave many questions unanswered. For example, quantum field theory is based on functional integrals, which are in general not well defined, and the Standard Model leaves aside gravity, one of the four fundamental interactions.

From a theoretical point of view the situation is rather unsatisfactory. This has led theoretical physicists to develop a variety of methods to study quantum field theory, and to consider a new kind of quantum theory which could accommodate gravity consistently. The methods are classified mainly in two types: perturbative and non-perturbative. On the other hand, with regard to the new kind of quantum theory, there exists at the moment a very promising theory, string theory, which certainly incorporates gravity and, furthermore, it might provide a unified theory involving all the fundamental interactions. The problem is that we do not know yet how to correctly formulate it.

A series of important events occurred in the eighties which made us turn into the new decade with a very promising tool to develop. In 1982, S. Donaldson discovered that the study of instantons, objects which appear in quantum field theories when they are analyzed from a non-perturbative point of view, provides very important information to study compact oriented smooth four-manifolds. Also in 1982, E. Witten, trying to unravel the structure of two-dimensional sigma models, generalized Morse theory to what is now known as Morse-Witten theory, an ancestor of TQFT. This theory was later rigorously reformulated by A. Floer who applied similar ideas to compact three-manifolds, constructing in this way new important objects from a topological point of view. In 1988, E. Witten, inspired in part by the work by Floer, formulated the first TQFT which in fact contains the topological invariants first studied by Donaldson at the beginning of the decade. The resulting TQFT is known as Donaldson-Witten theory.

In this brief history of the eighties there are two other important protagonists who played fundamental roles. One is M. Atiyah who soon was convinced that Donaldson theory could be formulated in terms of quantum field theory. His efforts to construct such a theory and to attract Witten to think on the problem were crucial. The second is string theory. String theory had a vertiginous development after 1985. Many theoretical physicists jumped in those days to heavily work on this theory. This development was strongly influenced by topological and geometrical ideas, creating a fruitful atmosphere for TQFT. A scenario where a quantum field theory of topological type could fit was found in 1987. It was then discovered that at high temperature strings could be described in terms of a theory with no degrees of freedom. The formulation of a theory with a feature like this, known then as new phase of gravity, was a goal whose achievement influenced Witten to construct his first TQFT in 1988. This first relation between string theory and TQFT did not have important consequences. However, it is very likely that string theory will provide a very useful tool to study the topology of low-dimensional manifolds and perhaps this will be the new breakthrough that we will witness in the second half of the present decade.

The eighties were completed by the formulation by Witten of two other fundamental TQFTs: topological sigma models for two-dimensional manifolds, which contain the Gromov invariants, and Chern-Simons gauge theory for three-manifolds which contains knot invariants as the Jones polynomial and its generalizations.

The present decade started with the work by Atiyah and Jeffrey on the formulation of TQFT using the Mathai-Quillen formalism. That work provided a general framework to understand the meaning of certain type of TQFTs from a mathematical point of view. However, it was not very useful to solve these theories. The first half of the present decade is characterized in fact by the opposite. The application of physical methods to certain class of TQFTs has led to their solution and to obtain a entirely new point of view from a mathematical perspective. The main physical concept which has been involved in this remarkable development is duality. Its use by Seiberg and Witten has originated a revolution in the program on four-manifolds started by Donaldson. They provided a framework which contains simpler but somehow equivalent topological invariants. These invariants are known as Seiberg-Witten invariants. It is very likely that this new framework will open new scopes not only in four dimensions but in other low-dimensional manifolds. Though these results are very astonishing, it is not unplausible that string theory is behind all this and that we have discovered only a small fraction of what will be found once string theory is understood.

In this talk, after introducing quantum field theory and TQFT I will describe, using the Aharonov-Bohm effect, why topology is relevant in quantum mechanics. This will allow us to get into Chern-Simons gauge theory and its knot invariants. Its various studies will permit us to understand the usefulness of both, perturbative and non-perturbative methods, and will allow us to discuss Vassiliev invariants for knots. Then we will leave these theories and will start with supersymmetric gauge theories and the TQFTs which are derived from them. Duality properties of supersymmetric gauge theories will be then applied obtaining the new framework which contains Seiberg-Witten invariants. I will end describing some generalizations which induce the idea of universality classes of topological invariants which is in part already present in Chern-Simons gauge theory.

2 Quantum field theory and TQFT

As was already mentioned in the introduction, quantum field theory is a theory which reconciles quantum physics with special relativity providing a helpful framework to describe the behavior of elementary particles. We cannot go here into details but we can give a general picture on how this theory is used and what are the mathematics involved.

As in any other theory, in quantum field theory one begins considering a set of input data and then computes some quantities which are of interest because in principle they could be measured in laboratories. These quantities are the predictions of the theory. The standard experimental setting which is behind quantum field theory consists of a collision in which incoming and outgoing particles participate, the input data being the classical properties of these particles, their masses, their momenta, their spins, etc.. Given a specific situation, quantum field theory is the tool to be used to compute quantities which could be measured such as cross sections, decay rates, etc. These quantities are basically probabilities for a given event characterized by the input data to happen.

Once we have a picture of what is involved in quantum field theory let us describe the type of mathematics which one has to confront in doing the calculations needed to obtain the probability for an event to occur. The basic ingredient is the generalization to the case of fields of the Feynman path integral. In quantum field theory one first associates a field $\Phi_{m_i,p_i,s_i,\ldots}(A)$ to each particle of the input data. This field contains the information which characterizes the state of particle i, namely, its mass, m_i, its momentum, p_i, its spin, s_i, etc., and is expressed in terms of the basic fields of the theory which are collectively denoted by A. The quantity which one computes and is associated to the probability for the event to happen is called vacuum expectation value of the product of fields $\Phi_{m_i,p_i,s_i,\ldots}(A)$, $i = 1, \ldots, n$, and it is basically the average value of this product weighted by a function which contains the most fundamental ingredient of the theory: the action or integral over space-time of the lagrangian density. Vacuum expectation values are denoted by open brackets and have the following form,

$$\begin{aligned} &\langle \Phi_{m_1,p_1,s_1,\ldots} \Phi_{m_2,p_2,s_2,\ldots} \cdots \Phi_{m_n,p_n,s_n,\ldots} \rangle \\ = \; &\frac{1}{Z} \int [DA] \, \Phi_{m_1,p_1,s_1,\ldots}(A) \, \Phi_{m_2,p_2,s_2,\ldots}(A) \cdots \Phi_{m_n,p_n,s_n,\ldots}(A) \\ &\exp\big(iS(A)\big), \end{aligned} \qquad (2.1)$$

where $[DA]$ denotes some integration measure over the space of configurations of the basic fields, $S(A)$ denotes the action, and Z is the partition function of the theory:

$$Z = \int [DA] \exp\big(iS(A)\big). \qquad (2.2)$$

Out of the three ingredients of a quantum field theory, the one on which we have more control is on the action $S(A)$. The form of the action is in general very much constrained by the symmetries of the theory. For example, in the case of the Standard Model, the presence of a gauge symmetry based on the gauge group $SU(3) \times SU(2) \times U(1)$ severely constraints its form. The action in this case is known except for a small fraction of it and the part which is widely accepted has been tested experimentally to a high degree of accuracy. The fields $\Phi_{m_i,p_i,s_i,\ldots}(A)$, $i = 1, \ldots, n$, are harder to control, specially in theories like quantum chromodynamics, which is part of the standard model, in which the property of confinement takes place. But the really unsurmountable problem is to define a measure for the functional integration involved in the computation of vacuum expectation values. It is not known in general how to do it. This has led theoretical physicists to develop a variety of methods to circunvent the problem. In fact, the richness of quantum field theory resides in the existence of this variety of methods which in practice turn out to be complementary since each of them provides partial information on the structure of the quantum field theory involved. As mentioned in the introduction, these methods usually fall into two categories: perturbative and non-perturbative. One of the main goals of this talk is to explain precisely how the application of these methods to TQFT has led to the recent successful results which have changed our way of looking at certain sets of invariants of low-dimensional manifolds.

It is now the turn of TQFT. These theories are special cases of quantum field theories. One of the properties which singularizes these theories is that now the space-time in which they are defined is a general smooth manifold and that the input data are not labels of particles but labels of topological or geometrical origin related to that manifold. These labels might be, for example, homology cycles, loops, etc. Another property which characterizes TQFTs is that their actions are such that the resulting vacuum expectation values do not depend on the metric on the manifold. The result of the computation of a vacuum expectation value in TQFT does not have an interpretation as a probability for an event to happen. These quantities turn out to be topological invariants. The reason for this is that they correspond to quantities which do not vary under deformations of the metric.

As in ordinary quantum field theory, the hard problem in TQFT is to define properly the functional integration measure. The problem of finding the equivalent of the fields $\Phi_{m_i,p_i,s_i,\ldots}(A)$ is much simpler in this case. Due to the problem with the measure one cannot think of the results so obtained with TQFT as rigorous from a mathematical point of view. Perturbative and non-perturbative methods are used to obtain those results and these methods contain a part based on the intuition that physicists had acquired through their work during many years trying to make sense of quantum field theory and confronting their results with experiments. The rigorous mathematical work that definitely describes the invariants predicted by TQFT is carried out using different methods. This work is certainly necessary and completes the formulation, making this new relation between physics and mathematics very fruitful. It is likely that in the future the arrow will turn backwards and physics will profit having at its disposal an elaborate theory on functional integration. This would be a very rewarding outcome of this relation.

3 Topology and quantum mechanics: the Aharanov-Bohm effect

The two branches of physics and mathematics which are particularly involved in TQFT are quantum mechanics and topology. At first sight, one would not anticipate a relation between the two. However, there is a simple qualitative argument to expect a link between them: both, topology and quantum mechanics, lead to discrete quantities out of continuous data. One could think for example of the Euler number for smooth manifolds in the case of topology, or the spectrum of the hydrogen atom in the case of quantum mechanics.

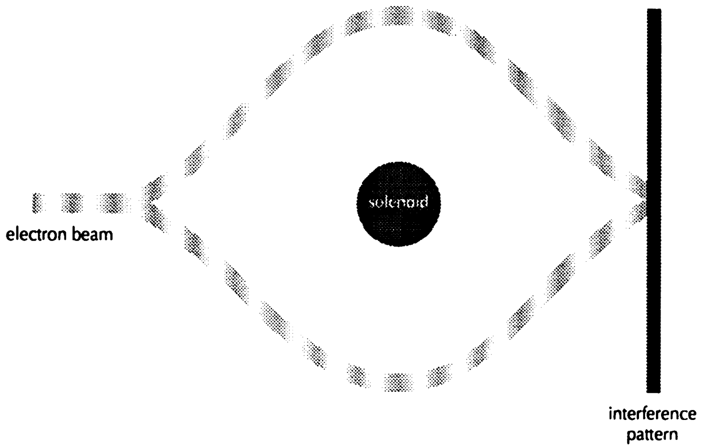


Figure 1: Sketch of the experiment proposed by Y. Aharanov and D. Bohm.

The deep relation between topology and quantum mechanics became manifest after the Aharanov-Bohm effect was understood in the late fifties. In 1959, Yakir Aharanov and David Bohm proposed an experiment which showed that the global properties of space-time were important in the description of quantum processes. The experiment was carried out in 1960 by R. G. Chambers and since then the physical process which takes place is known as the Aharanov-Bohm effect. In order to understand the role played by topology in this effect let us briefly describe the experimental setting in which it is observed and its theoretical explanation in terms of quantum mechanics.

The experimental arrangement consists of a very thin and long solenoid, which creates a magnetic field, and an electron beam which is split into two partial beams, each traveling along one side of the solenoid. The two partial beams are then recombined so that they interfere. A transversal section of the experimental situation is schematically depicted in Figure 1. For a thin and long enough solenoid the setting is such that the magnetic field vanishes along the paths travelled by the two partial beams. This means that at least classically one would expect that the interference pattern would be the same whether or not an electric current goes through the solenoid. This is not what is observed experimentally. Chambers found in 1960 that the interference pattern gets shifted when the current going through the solenoid is increased. This effect does not have an explanation classically. According to the classical equations of electrodynamics, Maxwell's equations, if the magnetic and the electric fields vanish, charged particles do not feel the interaction. In quantum mechanics, however, the interaction between electromagnetic fields and particles is described making use of the electromagnetic potential. If one could argue that the electromagnetic potential is different in each region travelled by the partial beams, and that such a difference depends on the current going through the solenoid, one could explain the shift in the interference pattern which is observed. This is in fact the way to understand the Aharanov-Bohm effect.

The first question which we must address to analyze the experiment is what is the value of the electromagnetic potential in each region. This does not have a unique response due to the existence of a gauge symmetry. The presence of a gauge symmetry implies that there are several descriptions in terms of electromagnetic potentials. Each description is related to the others by a gauge transformation. To choose one specific description is called to choose a gauge. One obvious question to ask is if in a situation in which the magnetic field is zero one could always choose a gauge in which the vector potential (part of the electromagnetic potential associated to the magnetic field) vanishes. If the answer were positive, one would not be able to explain the Aharanov-Bohm effect the way we intend to do. But if that were the case one would enter also in contradiction with Maxwell's equations. According to these equations, the integral of the vector potential $\mathbf{A}$ along the loop C pictured in Figure 1 should equal the magnetic flux Φ through the solenoid:

$$\oint_C \mathbf{A} d\mathbf{l} = \Phi. \tag{3.1}$$

When the current through the solenoid is turned on, the magnetic flux Φ is not zero and Maxwell's equations would be inconsistent for a null vector potential. If one thinks instead that $\mathbf{A}$ is pure gauge one is still in trouble, because then $\mathbf{A} = \nabla\phi$ for some scalar function ϕ, and then the left hand side of (3.1) would vanish. The puzzle is solved in the theory of electromagnetism allowing multivalued functions ϕ or, equivalently, vector potentials which are defined only locally. If ϕ were multivalued the left hand side of the last equation would not vanish in general. In such a situation the difference in the value of ϕ when one goes along the loop C should be just the magnetic flux. Notice that this picture does not lead to any singularity due to the fact that there is a region excluded: the region containing the solenoid. In other words, the space where the description is valid is not simply connected. A different but equivalent point of view consists of splitting space in regions which overlap (patches) and assume that the vector potential is different in each region while differing in the overlapping regions by gauge transformations. For a non-vanishing magnetic flux Φ two regions are enough to obtain a satisfactory description consistent with equation (3.1). Again, this framework does not lead to singularities due to the fact that space is not simply connected.

The mathematics behind the description based in a vector potential only defined locally is the theory of principal fiber bundles. The vector potential plays the role of a connection. Thus the mathematical description is intrinsically related to geometry and topology. In either description the vector potential is different in the region travelled by each partial beam and therefore, since in quantum mechanics charged particles couple to the vector potential, one expects a shift in the interference pattern as the flux Φ or, equivalently, the electric current through the solenoid, is increased. The detailed mathematical analysis leads precisely to the prediction of a shift in the interference pattern which is in full agreement with the one which is observed experimentally.

The Aharanov-Bohm effect was the starting point of a continuous presence of geometry and topology in quantum physics. The crucial point is that in quantum physics interactions are described by potential fields and these objects have a fundamental meaning from the point of view of geometry and topology. Since then many objects of geometrical or topological origin have played an important role. The most important case is non-abelian gauge theory, which is a generalization of electromagnetism in which the potential is a connection associated to a non-abelian group, in contrast to the case of electromagnetism in which it is abelian. The roots of many developments in theoretical physics during the last decades are based on objects of geometrical or topological origin. Examples of this are magnetic monopoles, solitons, instantons, strings, etc. The Standard Model itself is a non-abelian gauge theory whose gauge group is $SU(3) \times SU(2) \times U(1)$.

4 Chern-Simons gauge theory and link invariants

The effect described in the previous section revealed the importance of geometry and topology in quantum mechanics. In fact, the quantity that is computed integrating along the path C is a topological quantity. If one slightly deforms the path C, the value of the integral remain unchanged or, if one goes around the solenoid one more time one gets twice the magnetic flux. The path integral of the vector potential is proportional to the number of times that the path winds around the region of space which is excluded. This winding number is clearly topological.

To obtain more interesting topological quantities one can think of replacing the electromagnetic field of the Aharanov-Bohm effect by a non-abelian gauge field. In fact, this could lead to an interesting theory from a geometrical or topological point of view without the solenoid because non-abelian gauge theories, contrary to electromagnetism, are self-interacting. However, the situation is not so simple for two reasons: first, the path integral of equation (3.1) is not gauge invariant and one has to consider its gauge invariant generalization, the Wilson loop; second, for this quantity small deformations of the path C imply a change in its value.

The two problems plus the self-interaction property get nicely combined if one lowers the dimension of space-time and chooses a special action for the corresponding gauge theory: the Chern-Simons action. This action is based on a geometrical object known as the Chern-Simons form. The value of a Wilson loop remains invariant under deformations of the integration path which do not lead to crossings. Thus, in this theory, to each loop or set of loops embedded in three-dimensional space one gets a quantity which is invariant under small deformations which do not imply crossing lines. One seems to be dealing with topological quantities. Furthermore, these quantities are not trivial because Chern-Simons gauge theory is self-interacting. This theory possesses a contact interaction which modifies the value of the Wilson loop when lines cross to each other or to themselves. Indeed, the quantities associated to these sets of embedded loops in three-dimensional space are knot invariants.

A knot is a one-dimensional curve traced in three-dimensional space in such a way that it begins and ends at the same point and does not intersect itself. A link is a set of one-dimensional curves of the same type which do not intersect to each other. Knottedness and linkedness are not properties of the curves but of the way they are embedded in three space. Knots and links are specific of three dimensions, precisely the dimension for which Chern-Simons gauge theory exists. In Figure 2 some simple knots and links are shown: the first three are knots (or links of one single component) and the fourth is the simplest among two-component links: the Hopf link.

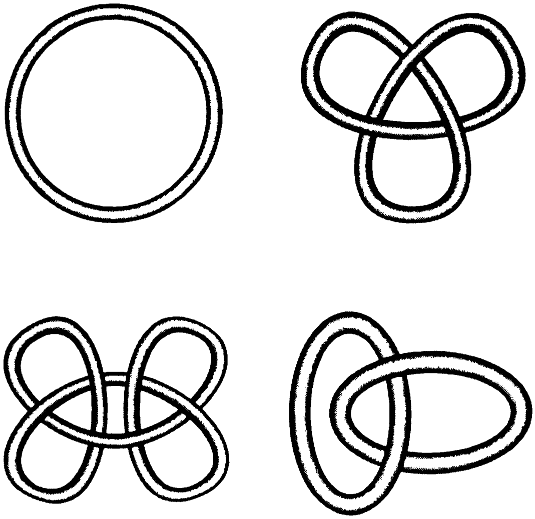


Figure 2: Some examples of knots and links: unknot, trefoil, squarea knot and Hopf link.

Interest in knot theory started in the 19th century when William Thompson (Lord Kelvin) proposed a model for atoms based on knots. Though this idea was soon discarded to describe atoms, it aroused interest in the problem of classifying knots. In 1900 Peter G. Tait published the first table of knots and links, and formulated a series of conjectures that in some cases waited eighty years for a proof. Since then knot theory has been a field of interest in mathematics. It has been very fruitful in its application to the study of the topology of three-manifolds.

One of the goals of knot theory is to classify knots and links. Two links (often in this paper knots will be treated as links of one component) are topologically equivalent if one can be obtained from the other by a continuous deformation, in other words, when no intersections occur in the deformation. Thinking of links as a series of knotted and linked strings with their loose ends attached, two links are equivalent if one can be deformed into the other without breaking any of the strings. In Figure 2 no pair of links contains two which are topologically equivalent. This statement, though it does not have a simple proof, seems plausible due to the simplicity of the links involved. However, for two complicated links it may be extremely difficult to decide whether they are topologically equivalent.

Mathematicians have developed techniques to discriminate between links. One of these techniques is based on the construction of link invariants or quantities associated to links which are invariant under continuous deformations. Two links having different link invariants are topologically inequivalent. However, two links having the same invariant might or might not be topologically equivalent. The more discrimination is achieved by a link invariant the better, but as yet there is not a complete classification of links.

In 1923 W. Alexander introduced a polynomial link invariant which had a good discrimination power as compared to previous invariants. However, it was soon realized that many topologically inequivalent links had the same Alexander polynomial. For example, knots which are not topologically equivalent to their mirror image knots (as the trefoil knot) have the same Alexander polynomial. Fundamental progress in knot theory was achieved by V. F. R. Jones in 1984 after the discovery of a new polynomial link invariant. This invariant is much more powerful than the Alexander polynomial; for example, in general, it distinguishes knots from their mirror images when they are not topologically equivalent. Nevertheless, soon it was discovered that there are non-equivalent knots with the same Jones polynomial. Invariants with more discrimination power were needed. After Jones' discovery, other polynomial invariants as the HOMFLY polynomial were constructed. Many of these new invariants, like the Jones polynomial itself, were formulated from mathematical structures whose study was in part motivated by statistical mechanics.

Chern-Simons gauge theory was formulated in 1988 providing an entirely new point of view in knot theory. This gauge theory is three-dimensional and so it provides an intrinsically three-dimensional formulation of polynomial link invariants. All previous formulations of these invariants were basically two-dimensional, defined on plane projections. This feature allows to obtain link invariants for arbitrary smooth oriented three-manifolds, and not only for flat space or for the three-sphere as was the case in previous formulations. In Chern-Simons gauge theory there exists a polynomial invariant for each representation of each simple Lie group. All previous polynomial invariants correspond to some specific choice of group and representation, or a special limit of some of them. It is not known yet if this huge amount of link invariants discriminates all topologically inequivalent links.

Chern-Simons gauge theory possesses the general problems of any quantum field theory, in particular, its integration measure is not well defined. However, being topological, it is simpler than the ordinary ones. Non-perturbative methods have been applied to this theory leading to its exact solution, at least for the case of simple three-manifolds. Chern-Simons gauge theory is one of the few quantum field theories whose exact solution is known. The solution consists in a series of rules which allow to compute vacuum expectation values of any product of Wilson loops. These rules are particularly simple for special cases. For example, for the case of the gauge group $SU(2)$ and Wilson loops in its fundamental representation the rule is shown in Figure 3. This rule has to be understood in the following way: project the link on a plane labeling overcrossings and undercrossings. Then, for three links which differ only in a part as depicted in Figure 3 the relation between the corresponding vacuum expectation values is:

$$\frac{1}{t}W_{L_+} - tW_{L_-} = (\sqrt{t} - \frac{1}{\sqrt{t}})W_{L_0}, \tag{4.1}$$

where t is a function of the coupling constant, $g = 1/\sqrt{k}$, of the theory:

$$t = \exp\left(\frac{2\pi i}{k+2}\right). \tag{4.2}$$

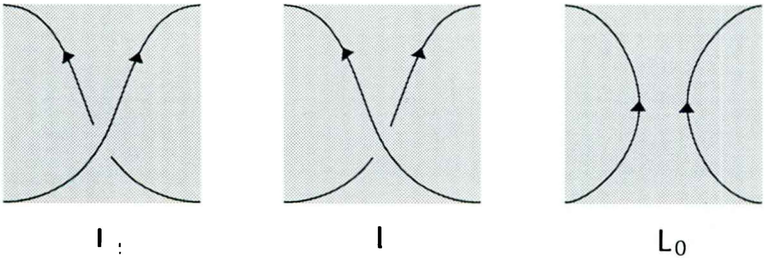


Figure 3: Skein rules for the Jones polynomial.

The rule so obtained (called skein rule) is precisely the rule which defines the Jones polynomial. The normalization is taken usually in such a way that for the unknot the polynomial invariant is 1. It is clear from this rule that the resulting invariant is a polynomial in $\sqrt{t}$ with positive and negative powers.

Chern-Simons gauge theory leads to a link invariant for each irreducible representation of each simple group. In general one does not find a skein rule as simple as the one for the Jones polynomial, but with the help of other non-perturbative methods one can complement the skein rule to design a calculation procedure. One important property of the solution found is that the vacuum expectation values are analytic in the coupling constant, $g = 1/\sqrt{k}$. This implies that the power series that results from a perturbative approach has to match the power series in $g = 1/\sqrt{k}$ streaming from the exact solution. When a situation like this occurs one says that there are not non-perturbative effects in the theory. But why worry about the power series if one knows the exact sum? There is an important reason for this. Perturbation theory provides path and space integral expressions for the coefficients of the power series expansion. If the whole series is a link invariant, each coefficient is also a link invariant since a continuous deformation of the link changes the expressions for the integrals of the coefficients but not the expansion parameters $g = 1/\sqrt{k}$.

Let us consider the trefoil and its Jones polynomial as an example. This polynomial is:

$$W_T = t + t^3 - t^4, \tag{4.3}$$

which, after expanding in powers of the coupling constant, results in:

$$W_T = 1 - 12\left(\frac{\pi i}{k}\right)^2 - 48\left(\frac{\pi i}{k}\right)^3 + \dots \tag{4.4}$$

In doing this calculation one removes first the shift by 2 of k in the denominator of the exponential in (4.2). This shift is controlled in perturbation theory by loop insertions related to finite renormalizations and can be ignored if one discards contributions from Feynman diagrams corresponding to those insertions. The integral expression which is provided by perturbation theory for the -12 appearing in the expansion (4.4) is the following:

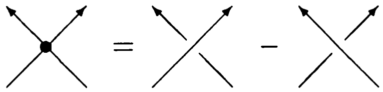
$$\begin{aligned} -12 \quad &= \frac{1}{2} - \frac{3}{4\pi^2} \oint_T dx_\mu \int^x dy_\nu \int^y dz_\rho \int^z dw_\tau \Delta^{\mu\rho}(x-z)\Delta^{\nu\tau}(y-w) \\ &+ \frac{3}{16\pi^3} \oint_T dx_\mu \int^x dy_\nu \int^y dz_\rho \int d^3\omega \epsilon^{\alpha\beta\gamma} \Delta^{\mu\alpha}(x-w) \\ &\quad \Delta^{\nu\beta}(y-w)\Delta^{\rho\gamma}(z-w) \end{aligned} \tag{4.5}$$

where,

$$\Delta^{\mu\nu}(x) = \epsilon^{\mu\nu\sigma}\frac{x_\sigma}{|x|^3} \tag{4.6}$$

Notice that in this expression there is a path integral and a space integral. Though its invariance under continuous deformations of the path T is a consequence of Chern-Simons gauge theory, it is worth proving that it is so. This has been in fact achieved.

Perturbation theory provides an infinite series of numerical invariants as the one shown. These invariants can be identified as Vassiliev invariants or numerical invariants of finite type. V. A. Vassiliev introduced his invariants in 1989 studying the cohomology of the space of all knots. These invariants have the property that if one defines from them invariants for singular knots using the equation:



there exists a finite value n such that for knots with $n+1$ singular points it vanishes. These values of n determine their orders or degrees. It turns out that the coefficient of the power series expansion of a Wilson loop which multiplies $1/k^n$ is a numerical knot invariant of order n. The invariant shown in (4.5) is of degree two.

Different representations of different groups provide different polynomial invariants and therefore different integral expressions for Vassiliev invariants. One could ask if at a given order in perturbation theory one could extract from the power series coefficient the contribution from the representation and group chosen. The answer to this question is positive due to the property of factorization intrinsic to the Feynman rules of Chern-Simons gauge theory. The power series expansion can be written as:

$$W_C = \sum_{i=0}^{\infty} \sum_{j=1}^{d_i} \alpha_{ij}(C) r_{ij}(G,R) \frac{1}{k^i}. \tag{4.7}$$

The factor $r_{ij}(G,R)$ (group factor) contains all the dependence on the group and representation chosen, while the factor $\alpha_{ij}(C)$ (geometrical factor) contains all the dependence on the path C. The quantity d_i denotes the number of independent group factors. Let us explain what is meant by this. The Feynman rules generate many more group factors than d_i. However, if one considers their possible values in the space of all representations of all semi-simple groups one observes that all of them can be written in terms of just a few. A minimum set of these few is selected as the set of independent group factors. In fact, these factors build a vector space and what one is doing in (4.7) is just to choose a basis. Its dimension is the number d_i in (4.7). The dimensions d_i, which are known only up to order 9, are shown in Table 1. The geometrical factors $\alpha_{ij}(C)$ constitute a basis of Vassiliev invariants of order i.

i	1	2	3	4	5	6	7	8	9
d_i	0	1	1	3	4	9	14	27	44

Table 1: Numbers of independent group factors

As we have described, Chern-Simons gauge theory provides integral expressions for Vassiliev invariants. But the description presented is not the only one available to obtain representations of Vassiliev invariants from Chern-Simons gauge theory. There are many other ways to do perturbation theory, each providing a different representation. Chern-Simons gauge theory is a gauge theory and therefore has a gauge symmetry. Gauge invariant quantities like the Wilson loop can be computed in different gauges all leading to the same result. The expression presented in (4.5) is obtained in a specific gauge. Other gauge would lead to a different expression providing an alternative representation.

Given the space of all representations of all semi-simple Lie groups one obtains from Chern-Simons gauge theory an infinite sequence of sets of Vassiliev invariants. One could ask if the Vassiliev invariants so obtained are a complete set. In other words, that there is not a finite type invariant of a given order which cannot be expressed in terms of the ones originated from Chern-Simons gauge theory. The answer to this question seems to be negative. Possibly a structure bigger than semi-simple Lie groups is needed to accommodate all Vassiliev invariants.

Another important subject related to Chern-Simons gauge theory is the study of its partition function. This quantity leads to very interesting three-manifold invariants.

Chern-Simons gauge theory has opened a variety of new points of view in knot theory and on the topology of three-manifolds. Its field-theoretical study using non-perturbative and perturbative methods has provided a rich framework to analyze its topological invariants. A consequence of this analysis is that polynomial invariants based on representations of semi-simple Lie groups are in the same universality class of invariants as a subset of Vassiliev invariants. By being in the same universality class of invariants we mean that all the topological information which can also be obtained from one set of invariants can be obtained from the other. If two non-equivalent knots have the same polynomial invariant for all representations of all semi-simple Lie groups, they will clearly have the same Vassiliev invariants, at least of the mentioned subset. It is known that Chern-Simons gauge theory for semi-simple Lie groups do not detect non-invertible knots and mutant knots. The question of whether Vassiliev invariants ever discriminate among these types of knots is open.

5 Supersymmetry and Donaldson-Witten theory

In our previous attempt to find a non-abelian version of the Aharanov-Bohm effect we had to lower the dimension of space-time to construct a TQFT. The result was Chern-Simons gauge theory. There exist, however, TQFTs in four dimensions which are non-abelian gauge theories. In fact, there are two types: theories which are related to supersymmetry and theories which are not. The last set contains theories which share some common features with Chern-Simons gauge theory and are called BF theories. We will not discuss them here. Among the theories related to supersymmetry the first TQFT formulated by Witten in 1988 stands out. This theory deals with Donaldson invariants for oriented smooth four-manifolds. Theories in this second set present a series of special featUres which characterize them. But before going into detail let us make a short detour to introduce supersymmetry.

Supersymmetric quantum field theories possess a symmetry which consists of the invariance of the theory under a transformation which interchanges bosons and fermions. There are theories in which this interchanging can be done in different ways and then one has theories with $N = 2, 3, 4, \ldots$ supersymmetries. For gauge theories in four dimensions $N = 4$ is the maximum number of supersymmetries if one excludes particles with spin two and higher. Particles in supersymmetric theories appear grouped into multiplets. For $N = 4$ there is only one multiplet, the gauge multiplet. For $N = 2$ there are two types of multiplets: gauge or vector multiplets, and matter multiplets or hypermultiplets.

Theories with a higher number of supersymmetries are simpler to solve but much more restricted. For example, $N = 4$ supersymmetric gauge theory does not renormalize and is conformal invariant. However, this theory is very restrictive and the only freedom is the choice of gauge group. $N = 4$ supersymmetric gauge theory has some common features with Chern-Simons gauge theory in three dimensions and one expects that soon it could be solved exactly. In fact, fundamental progress has been made in this direction during the last three years.

As it will be discussed below, $N = 4$ supersymmetric gauge theories possess a symmetry called duality which is extremely helpful towards the search of its exact solution. Though, strictly speaking, duality is not a symmetry for theories with a lower number of supersymmetries, it constitutes a helpful tool to analyze these theories and, indeed, its use has led to substantial progress towards the search for their exact solution. The use of the resulting information about this solution has led to the discovery of a new point of view in the theory of Donaldson invariants.

TQFTs of the type under consideration can be regarded as originated from supersymmetric theories with a least two supersymmetries. Theories with $N = 2$ supersymmetry are less restrictive than theories with $N = 4$ and can be labeled by a Lie group and a finite number of representations. Starting with an $N = 2$ supersymmetric theory one obtains a TQFT through the process of twisting. On flat space, a twisting consists of a rewriting of the theory in such a way that some fields are relabeled so that they have exotic new labels. Recall when we introduced quantum field theory that we attached labels to particles denoting their mass, their spin, etc. For the case of fields one also possesses a set of labels to characterize them. Of particular importance are the labels denoting their representation respect to the space-time group: the Lorentz group. In this sense one talks about scalar fields, spinor fields, vector fields, etc. Furthermore, $N = 2$ supersymmetric theories have an extra symmetry together with the space-time symmetry. This symmetry is called internal and its group is $SU(2)$. Fields also carry labels indicating how they transform under the internal symmetry group. A twisting consists of choosing an exotic relabeling of the fields or a particular mixing between the space-time symmetry group and the internal symmetry group. For theories with only two supersymmetries this can be done only in one non-trivial way while for theories with $N = 4$ there are three non-equivalent ways.

The theory resulting after the twisting is the same as the original one on flat space-time. However, it is different when considered on curved space. The reason is that the coupling of the fields to the background Riemannian metric is dictated by their spin, which has been changed in the twisting. Twisted theories have three important properties. First, they have a scalar symmetry even when they are considered on an arbitrary smooth four-manifold. Second, due to the presence of this symmetry these theories are such that the vacuum expectation values of quantities which are invariant under this symmetry are invariant under deformations of the Riemannian metric. Third, again due to the presence of the scalar symmetry, the vacuum expectation values are independent of the coupling constant of the theory. This is not exactly true for twisted theories originated from $N = 4$ supersymmetric gauge theories where there remains a dependence which, however, is simple to control. We will exclude these theories in our discussion. These properties indicate that vacuum expectation values are just numbers (not functions of the coupling constant as in Chern-Simons gauge theory) which are topological invariants of the four-manifold where the theory is defined. The input data which characterize these vacuum expectation values are labeled by the homology of the four-manifold. To a specific selection of homology cycles correspond a number which is a topological invariant.

The topological invariants which are obtained after the twisting of an $N = 2$ supersymmetric gauge theory with no representation labels and gauge group $SU(2)$ are Donaldson invariants. This was shown by Witten in his seminal paper of 1988. He proved this connection using perturbative methods. The basic idea is the following. Twisted theories are TQFTs whose vacuum expectation values are independent of the coupling constant of the theory. This means that the calculation of these quantities in the $g \to 0$ limit is exact. But the $g \to 0$ limit is rather simple: one has just to keep the first term of the perturbative series expansion. This was done by Witten in 1988 showing that the resulting expression were the same as the ones proposed by Donaldson to define his invariants for four-manifolds. This was rather satisfactory because, finally, Atiyah's proposal of giving a quantum field theory interpretation to Donaldson theory was implemented. However, Witten's formulation did not lead to further progress towards the computation of these invariants.

Let us briefly discuss what kind of invariants one is dealing with in Donaldson-Witten theory. The perturbative analysis of the theory leads to the conclusion that one has to compute certain quantities on the space of solutions of a set of equations which are very familiar in physics, the instanton equations,

$$F^+_{\mu\nu} = 0, \tag{5.1}$$

where $F_{\mu\nu}$ is the field strength or curvature associated to the gauge connection A_μ, and the symbol plus indicates that one is equating to zero only the self-dual part. The solutions of this equation are called instantons and the space formed by those solutions is the moduli space of instantons, which will be denoted by $\mathcal{M}$. In this space two instanton solutions which are related by a gauge transformation are considered equivalent. From the input data, which, as indicated, were labeled by homology cycles, $\gamma_1, \gamma_2, \ldots$, the perturbative analysis leads to a well defined prescription to map to each set of labels a cohomology cocycle $\Omega_{\gamma_1,\gamma_2,\dots}$ on the moduli space of instantons $\mathcal{M}$. The integrals of these forms over the moduli space,

$$\int_{\mathcal{M}} \Omega_{\gamma_1,\gamma_2,\dots}, \tag{5.2}$$

are the numbers which correspond to Donaldson invariants. The problem related to the compactification of this moduli space (in general it is not compact) is the same one as in Donaldson theory. From the perturbative point of view TQFT does not bring anything new to this problem, it just shows that the theory we are dealing with is in fact the TQFT of Donaldson invariants. Insight from quantum field theory could come if one were able to carry out the analysis of the theory for a different value of the coupling constant g, for example, $g \to \infty$, or strong coupling limit. However, the corresponding analysis required non-perturbative information which was not available until recently. Before getting into the non-perturbative analysis we need to discuss some aspects of duality.

6 Duality and Seiberg-Witten invariants

Electromagnetic duality is a symmetry of Maxwell's equations without matter which allows to interchange the electric and magnetic fields. If one writes Maxwell's equations in terms of the complex field $\mathbf{E} + i\mathbf{B}$, where $\mathbf{E}$ and $\mathbf{B}$ are the electric and magnetic fields respectively,

$$\begin{aligned} \nabla \cdot (\mathbf{E} + i\mathbf{B}) &= 0, \\ \nabla \wedge (\mathbf{E} + i\mathbf{B}) &= i\frac{\partial}{\partial t}(\mathbf{E} + i\mathbf{B}), \end{aligned} \tag{6.1}$$

duality is the invariance of these equations under the transformation:

$$\mathbf{E} + i\mathbf{B} \to e^{i\phi}(\mathbf{E} + i\mathbf{B}). \tag{6.2}$$

When matter is included in Maxwell's equation, duality is only maintained if one assumes that matter is composed of classical point particles carrying electric and magnetic charges. If these charges are q and g respectively, duality is kept if these transform as:

$$q + ig \to e^{i\phi}(q + ig). \tag{6.3}$$

The price one has to pay to preserve duality is the inclusion of unobserved magnetic charge.

As was discussed before, the quantum description of the coupling of charged particles to electromagnetic fields is made using the electromagnetic potential. In the presence of magnetic charges the coupling is consistent only if some constraints are satisfied. In 1931 Dirac proved that a magnetic charge g_1 carrying no electric charge could occur in the presence of an electric charge q_2 carrying no magnetic charge provided the following condition is satisfied:

$$q_2 g_1 = 2\pi n\hbar, \qquad n = 0, \pm 1, \pm 2, \ldots \tag{6.4}$$

being $\hbar$ the Plank's constant. This is known as the Dirac quantization condition and it implies that if a magnetic charge g_1 exists, electric charge is quantized. Quantization of electric charge is a feature of nature and this explanation is perhaps the best yet found. Particles carrying only magnetic charge are called monopoles.

One of the problems with Dirac's quantization condition is that it is not invariant under duality. It took some time to realize how this condition has to be generalized to accommodate duality. The new input is to assume that there are particles carrying electric and magnetic charges. These particles are called dyons. Applying Dirac's argument to dyons carrying, respectively, charges (q_1, g_1) and (q_2, g_2) one finds:

$$q_1 g_2 - q_2 g_1 = 2\pi n\hbar, \qquad n = 0, \pm 1, \pm 2, \ldots \tag{6.5}$$

This is known as Schwinger quantization condition and it is invariant under the duality transformation (6.3). One of the consequences of Schwinger quantization condition is that the set of possible electric and magnetic charges form a two-dimensional lattice. This is a property that must be satisfied by the electric and magnetic charges of the particle spectrum of any quantum field theory having duality as a symmetry.

During the last years, evidence has accumulated to make plausible that $N = 4$ supersymmetric $SU(2)$ gauge theory is a theory where duality is realized exactly. In this theory there is a part of the spectra obtained by spontaneous symmetry breaking. The rest of the spectra is realized through monopole and dyon solitons. A consequence of duality is that this theory possesses many equivalent descriptions. For example, one could choose to describe it via a Higgs mechanism applied to some other part of the spectra, realizing now the original ones as solitons. This indeed can be done provided one changes properly the coupling constant of the theory. To choose a particular description is basically to make a choice of basis in the lattice of allowed electric and magnetic charges. Depending on the choice one has a different coupling constant. All these choices are related by a duality group of transformations. For example, there exist dual descriptions in which the coupling constant g is interchanged by $1/g$, in other words, the interchange of weak and strong couplings.

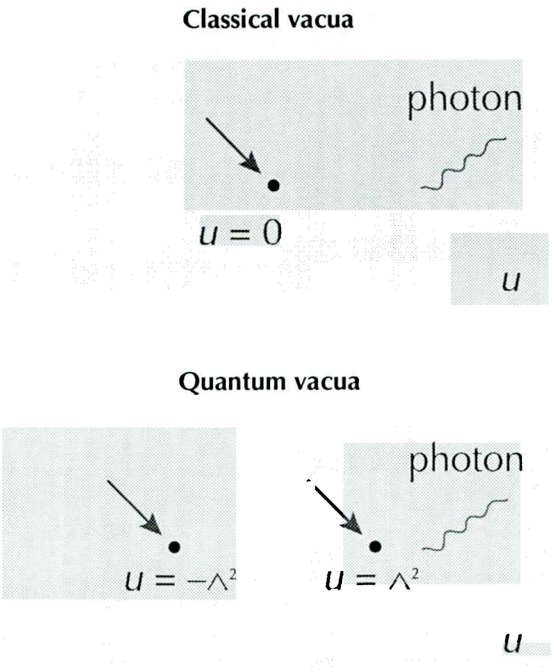
Although there is not a proof yet that in $N = 4$ supersymmetric $SU(2)$ gauge theory duality is exactly realized, this has been verified in one of its twisted versions. The partition function of this twisted theory has been computed for some four-manifolds obtaining a result which is invariant under the full duality group. This question should be addressed for other gauge groups and for the other two non-equivalent twistings of $N = 4$ supersymmetric gauge theory.

$N = 2$ supersymmetric gauge theories are rather different than their $N = 4$ counterpart. The first important difference is that in general these theories are not conformal invariant and therefore the coupling constant gets renormalized. The second difference is that in $N = 2$ supersymmetry there exist two kinds of supersymmetric multiplets, the gauge multiplet and the matter multiplet or hypermultiplet. In this theories one does not expect duality to be realized exactly. However, there is a variant of this symmetry which plays a fundamental role.

$N = 2$ supersymmetric gauge theory is asymptotically free. This means that at high energy (ultraviolet regime) the theory is weakly coupled, the effective coupling constant becomes small. At low energies (infrared regime) the theory is strongly coupled becoming its effective coupling constant big. Seiberg and Witten discovered that for $N = 2$ supersymmetric gauge theories duality become the statement that the strongly coupled limit is equivalent to the weak coupling limit of some other system. They found that system for the case under consideration. Notice that the statement is consistent with what we found for $N = 4$ supersymmetric gauge theories. What distinguishes $N = 4$ is that the 'other system' is again $N = 4$ supersymmetric $SU(2)$ gauge theory. In $N = 4$ supersymmetry there is only one multiplet and therefore the 'other system' has to be of the same type. Only the gauge group could be modified. In fact, that seems to be the case when considering more complicated groups. In $N = 2$ supersymmetry there are two multiplets and therefore there are many more possibilities for the 'other system'. Seiberg and Witten found that the strongly coupled limit of $N = 2$ supersymmetric $SU(2)$ gauge theory is equivalent to a weakly coupled $N = 2$

supersymmetric abelian gauge theory coupled to matter hypermultiplets.

In the weak coupling limit, or perturbative regime, one deals with the space of classical vacua of the theory. For $N = 2$ supersymmetric $SU(2)$ gauge theory this space is parametrized by a complex parameter u. Of particular importance, specially in its application to TQFT, is the massless spectra for each value of u. It turns out that for $u \neq 0$, since the gauge symmetry is spontaneously broken, there is only one massless particle: a photon described by an abelian gauge field. At $u = 0$ the full gauge symmetry is restored and there are three massless particles corresponding to the three gauge bosons. This point is called singular.


Classical vacua

Quantum vacua

Figure 4: Classical and quantum vacua.

In an asymptotically free theory, as the one under consideration, the strong coupling limit correspond to the quantum vacua of the theory. Seiberg and Witten showed that this space of vacua is parametrized again by a complex parameter u. For values of $u \neq \pm\Lambda^2$, where Λ is certain mass scale, the only massless particle corresponds to an abelian gauge field. For $u = \pm\Lambda^2$ there are additional massless particles: among them a magnetic monopole for $u = \Lambda^2$ and a dyon for $u = -\Lambda^2$. Full $SU(2)$ symmetry is never restored. At the quantum vacuum $u = \Lambda^2$ the weakly coupled theory is an $N = 2$ supersymmetric abelian gauge multiplet coupled to a massless hypermultiplet. The points $u = \pm\Lambda^2$ are called singular points. The classical and quantum moduli spaces of vacua are represented in Figure 4.

In the previous section we carried out the perturbative analysis of Donaldson-Witten theory. This analysis was done for $g \to 0$ and therefore it corresponds to the ultraviolet regime or weakly coupled limit. Since this TQFT is independent of g the exact result is just a sum over classical vacua. We actually did not do this in our analysis of the previous section. There, of all the values of u we just took the contribution from $u = 0$. We ignored the rest. This is justified if the manifold satisfies the topological property $b_2^+ > 1$. We succinctly assumed this to hold in the previous perturbative analysis. The condition $b_2^+ > 1$ means that the number of self-dual cohomology 2-cocycles is bigger than 1. Precisely smooth manifolds satisfying this condition are the most studied in Donaldson theory. The condition $b_2^+ > 1$ implies that the contributions from $u \neq 0$ vanish.

The analysis of Donaldson-Witten theory in the strong coupling limit should provide a new point of view on Donaldson invariants. This seemed hard to achieve before 1994 but after Seiberg and Witten's work on the strong coupling limit of $N = 2$ supersymmetric $SU(2)$ gauge theory, this goal was reached. The main, piece of the argument is that in the strong coupling limit the contributions come only from the space of quantum vacua. Again, the condition $b_2^+ > 1$ notably simplifies the analysis because in this case only contributions from the points $u = \pm\Lambda^2$ survive. Actually, it is enough to work out the contribution from $u = \Lambda^2$ since there is a symmetry which relates both points. At $u = \Lambda^2$ the weakly coupled theory is known and one has just to work out its twist. The result is obtained using the previous perturbative methods in this weakly coupled theory and one finds that the contributions come from the solutions of a different set of equations:

$$\begin{aligned} &F^+_{\mu\nu} + \overline{M}\gamma^+_{\mu\nu}M = 0, \\ &\gamma^\mu\nabla_\mu M = 0. \end{aligned} \tag{6.6}$$

In these equations M is a commuting chiral spinor, and γ_μ, $\gamma^+_{\mu\nu}$ are Dirac matrices. These equations are known as the monopole equations or Seiberg-Witten equations. They are simpler than the instanton equations because the field strength $F_{\mu\nu}$ corresponds to an abelian gauge field. The second equation in (6.6) is just the Dirac equation for the chiral spinor M. The vacuum expectation values analyzed in the perturbative approach can be rewritten now as a sum over solutions of the Seiberg-Witten equations. Actually, these sums have a very simple form. They turn out to be:

$$\sum_{x\in\Gamma} n_x f_{\gamma_1,\gamma_2,\ldots}(x) \tag{6.7}$$

where the n_x are the Seiberg-Witten invariants. In this equation Γ is a set of cohomology classes x which satisfy certain constraint (also known as basic classes) and $f_{\gamma_1,\gamma_2,\ldots}(x)$ is a function of x which involves the input data. Recall that this input data consisted of a finite set of homology cycles $\gamma_1, \gamma_2, \ldots$. The Seiberg-Witten invariants n_x involve a sum over solutions of the Seiberg-Witten equations for an abelian gauge field in the class x. In fact, n_x is just the partition function of a TQFT for fixed x.

Equation (6.7) presents some similarities with the one found for Chern-Simons gauge theory involving Vassiliev invariants. There, the Vassiliev invariants contained all the topological information on the knot. Here, the Seiberg Witten invariants contain all the topological information on the the smooth four-manifold.

Seiberg-Witten invariants were totally unexpected in mathematics. They certainly have opened a new door. Conjectures about four-manifolds which were waiting for a proof based on Donaldson theory were quickly proved using Seiberg-Witten invariants. At the moment we are lacking a proof of equation (6.7), but this is very hard with today's knowledge, and perhaps not the most interesting thing to do. Seiberg-Witten invariants stand out by themselves and can be used disregarding their origin from Donaldson theory. It is very likely that the first proof of (6.7) will come from string theory. Nevertheless, further developments of this theory are necessary before having a glimpse on how this could be achieved.

7 Non-abelian monopoles

We have limited our discussion to Donaldson-Witten theory with gauge group $SU(2)$. Certainly, we could ask about why not to consider other groups and couplings to twsited $N=2$ hypermultiplets. TQFTs of these types are in general labeled by a group and a finite number of representations which denote the ones chosen for the twisted hypermultiplets. No much has been explored in this direction. Only the case of $SU(2)$ with a hypermultiplet in its fundamental representation has been studied. We will briefly describe it in what follows.

The perturbative analysis is similar to the one in Donaldson-Witten theory. The input data are labeled in the same way and one ends with an integration over the moduli space of the non-abelian version of the monopole equations (6.6):

$$\begin{aligned} &F^{a+}_{\mu\nu} + \overline{M}\gamma^{+}_{\mu\nu}T^a M = 0, \\ &\gamma^\mu \nabla_\mu M = 0, \end{aligned} \tag{7.1}$$

where T^a is an $SU(2)$ generator. These equations are called non-abelian monopole equations. The resulting moduli space contains the moduli space of instantons as a subset. It has basically the same type of problems as that moduli space. The non-perturbative analysis of the physical theory has been done by Seiberg and Witten. As in the previous case, the quantum vacua possess a massless abelian gauge field, but now it contains three singular points related by a symmetry. One of these points corresponds again to a massless magnetic monopole and the contributions come again from solutions of the abelian monopole equations (6.6). The final expression in the strong coupling limit can be written as:

$$\sum_{x\in\Gamma} n_x \tilde{f}_{\gamma_1,\gamma_2,\ldots}(x) \tag{7.2}$$

where n_x are the same Seiberg-Witten invariants as before. However, the function multiplying them, $\tilde{f}_{\gamma_1,\gamma_2,\ldots}(x)$, is different. Comparing the results (6.7) and (7.2) we can assure that no new topological information is obtained analyzing the moduli space of non-abelian monopoles. All that information is already contained in the moduli space of instantons.

These observations bring again the idea of universality classes of topological invariants. It seems that Seiberg-Witten invariants represent a class in the sense that topological invariants associated to several moduli spaces can be written in terms of them. This is certainly true for the two cases studied but presumably it holds for other groups. It is very likely that Seiberg-Witten invariants are the first set of a series of invariants, each defining a universality class. TQFTs originated from the twist of $N = 2$ supersymmetric gauge theory constitute a big set of theories labeled by the group and a finite number of representations. Only two elements of this set have been studied. Presumably, many more new invariants and many more new relations among invariants of different moduli spaces are waiting there to be discovered.

8 Final remarks

In this talk I have described several examples which show how ideas from quantum field theory and, in particular, from TQFT have been very successful in the discovery of new results in the topology and geometry of smooth low-dimensional manifolds. We have analyzed situations in which the physical approach occurred first leading to new mathematics, and situations in which, though mathematics came first, physics provided an important generalization. The examples described, and many other which we have not treated, show that TQFT makes correct predictions in mathematics. These days, quantum field theorists, though working with a tool which is not rigorous, are being encouraged not only by the excellent experimental agreement achieved by physical theories, but also by the success accumulated with this other type of predictions. Physicists and mathematicians should join efforts to construct a rigorous and sound foundation for quantum field theory.

TQFTs are simpler than ordinary quantum field theories and presumably it is easier to make them rigorous. The difficulties in their rigorous definition by analytic methods could be overcome by axiomatizing them. In fact some TQFTs can be constructed using combinatorial and algebraic methods. However, it is likely that the richness inherent to the methods developed by quantum field theorists is a much more powerful tool to obtain unexpected relations between different sets of invariants, or a variety of representations for each of them. In this talk the success of the these methods has been described for theories in three and four dimensions. The results are summarized in Table 2, where the invariants and the methods used in their analysis are presented.

	$d=3$	$d=4$
perturbative	Vassiliev	Donaldson
non-perturbative	Jones	Seiberg-Witten

Table 2: Topological invariants in the perturbative and the non-perturbative regimes for $d=3$ and $d=4$.

Physicists have started to accumulate a big amount of knowledge on the behavior of $N=2$ supersymmetric gauge theories. The application of these results to TQFT has led to the prediction of Seiberg-Witten invariants. This should be regarded as a first result of possibly a series of important relations between different sets of topological invariants. Duality would be at the heart of these developments. There is some evidence that duality has its roots in string theory and that the evolution of this theory will provide new insights in supersymmetric physical theories and in their topological counterparts. From this point of view duality might relate also different sets of invariants for manifolds with dimension different than four. Some results in this direction have been recently obtained in three dimensions. String theory itself could provide new unexpected results in geometry and topology. However, though a considerable amount of progress has been made in the last years, we are still far from the fundamental formulation of string theory. What is becoming firmly accepted is that in such a formulation duality will play an important role. This is a very encouraging feature towards future developments of TQFT. The best is yet to come.

Acknowledgements

I would like to thank D. Jou and the organizers of the "Jornades sobre Física i Geometria" for inviting me to give this talk. I would like to thank also A.V. Ramallo and my collaborators M. Alvarez, C. Lozano, M. Mariño and E. Pérez for very helpful discussions on many aspects of the subjects treated in this talk.

References

[1] V.F.R. Jones, *Bull. AMS* **12** (1985) 103; *Ann. of Math.* **126** (1987) 335.

[2] M. F. Atiyah, "New invariants of three and four dimensional manifolds", in *The Mathematical Heritage of Herman Weyl, Proc. Symp. Pure Math.* **48**, *American Math. Soc.* (1988) 285-299.

[3] E. Witten, *Comm. Math. Phys.* **117** (1988) 353.

[4] E. Witten, *Comm. Math. Phys.* **121** (1989) 351.

[5] S. K. Donaldson, *Topology* **29** (1990) 257.

[6] V.A. Vassiliev, "Cohomology of knot spaces" in *Theory of Singularities and its applications*, (V.I. Arnold, ed.), *Amer. Math. Soc.*, Providence, RI, 1990, 23.

[7] N. Seiberg and E. Witten, *Nucl. Phys.* **B426** (1994) 19; Erratum, *Nucl. Phys.* **B430** (1994) 485; *Nucl. Phys.* **B431** (1994) 484.

[8] E. Witten, *Math. Res. Lett.* **1** (1994) 769.

Unanswered Mathematical Questions Raised by Fractal Geometry

Benoît B. Mandelbrot
Yale University
New Haven, USA

Abstract

This text sketches diverse questions of pure mathematics that fractal geometry raised over the years. Some are broadly-based challenges. Others are fully-fledged conjectures that resist repeated efforts to answer them. Some can be understood by a good secondary-school student, while others are delicate or technical. Their perceived importance ranges from high to low, but they are alike in three ways. First, they did not arise from earlier mathematics, but in the course of practical investigations into diverse fields of science and engineering, some of them old and well-established, others newly revived, and a few of them altogether new. Second, they originate in careful inspection of actual pictures that were generated by computer. Third, they built upon the century-old mathematical "monster shapes" that were for a long time guaranteed to lack any contact with the real world.

The scope of this paper is necessarily limited. Many other fractal challenges and/or conjectures remain unanswered. Still others have been met and/or confirmed (especially in the context of multifractals). Among fields of research, fractal geometry seems to exemplify the shortest distance and the greatest contrast between a straightforward core, which is by now known to children and adult amateurs, and multiple frontiers filled with every kind of major difficulty, some of them linked with practical problems and other of purely mathematical interest.

1 Introduction

For three reasons listed in the abstract, the unanswered questions raised in this paper bear on an issue of great consequence. Does pure (or purified) mathematics exist as an autonomous discipline, one that can —and ideally should— adhere to a Platonic ideal and develop in total isolation from "sensations" and the "material" world? Or, to the contrary, is the existence of totally pure mathematics a myth?

In my work, the role of "sensations" is dominated by the role of fully-fledged pictures that are as detailed as possible and go well beyond sketches and diagrams. Their original goal was to help already formulated ideas and theories become accepted, by bridging cultural gaps between fields of science and mathematics. Then they went on to help me and many others generate new ideas and theories.

Many of these shapes strike everyone as being of exceptional and totally unexpected beauty. Some have the beauty of the mountains and clouds that they mean to represent; others seem wild and unexpected at first, but after brief inspection appear totally familiar. In front of our eyes, the visual geometric intuition built on the practice of Euclid and of calculus is being retrained with the help of new technology.

Pondering these pictures proves central to a different philosophical issue. What is beauty, and how does the beauty of these mathematical pictures relate to the beauty that a mathematician sees in his trade after long and strenuous practice? My lectures often underlined these questions, by showing what certain mathematical shapes really look like. By now, those pictures have become ubiquitous.

Next, consider the relation between pure mathematics and the "material" world. Everyone agrees that an awareness of physics, numerical experimentation and geometric intuition are very beneficial in some branches of mathematics, but elsewhere physics is reputed to be irrelevant, computation powerless, and intuition misleading. The irony is that history consistently proves that, as branches or branchlets of mathematics develop, they suddenly either lose or acquire deep but unforeseen connections with the sciences —old and new. As to numerical experimentation— which Gauss found invaluable but whose practice was waning until yesterday —it has seen its power multiply a thousandfold thanks to computers, and later, to computer graphics.

In no case that I know is this irony nearly as intense as in fractal geometry, a branch of learning that I conceived, developed and described in my book [FGN]. I put it to use in models and theories relative to diverse sciences, and it has become widely practiced. A "Polish school" of mathematics that had viewed itself as devoted exclusively to *Fundamenta*, added mightily to the list of the monster shapes, and greatly contributed to the creation of a chasm between mathematics and physics. Specifically ironical, therefore, is the fact that my work, that of my colleagues, and now that of many scholars, made those monster shapes, and new ones that are even more "pathological", into everyday tools of science.

This article uses freely the term *fractal*, which I coined in 1975 from the Latin word for "rough and broken up", namely *fractus*, and which is now generally accepted. Loosely, a "fractal set" is one whose detailed structure is a reduced-scale image of its overall shape. Among linear reductions, when the reduction ratio is the same in all directions, a fractal is "self-similar"; when those ratios differ, the fractal is self-affine. "Dust" will denote a totally disconnected set.

2 Complex Brownian bridge; Brownian cluster and its boundary; the self-avoiding plane Brownian motion

We begin with the open conjecture that is easiest to state and to understand.

Definitions. *The Wiener Brownian motion* $B(t)$ *is self-affine. Setting* $B(0)=0$, *recall that a* Brownian bridge $B_{bridge}(t)$ *is a periodic function of* t, *of period* 2π, *given for* $0 \leq t \leq 2\pi$ *by*

$$B_{bridge}(t) = B(t) - (t/2\pi)B(2\pi).$$

In distribution, $B_{\text{bridge}}(t)$ is identical to a sample of $B(t)$ conditioned to return to $B(0)=0$ for $t=2\pi$. It is the sum of Wiener's trigonometric series, whose n-th coefficient is G_n/n, where the G_n are independent reduced Gaussian random variables.

Take $B_{\text{bridge}}(t)$ to be complex of the form $B_r(t)+iB_i(t)$ and define a *Brownian plane cluster* Q as the set of values of $B_{\text{bridge}}(t)$. This non-traditional concept is the map of the time axis by the complex function $B_{\text{bridge}}(t)$. The classical map of the time axis by the complex $B(t)$ is everywhere dense in the plane, and the map of a time interval by the complex $B(t)$ is an inhomogeneous set. In contrast to the preceding example, when the origin Ω of the frame of reference belongs to Q, all the probability distributions concerning Q are independent of Ω; therefore Q is a *conditionally homogeneous* set.

The *self-avoiding planar Brownian motion* $\tilde{Q}$ is defined in [FGN] as being the closed set of points in Q accessible from infinity by a path that fails to intersect Q.

The unanswered "4/3 conjecture." The set $\tilde{Q}$ has a fractal dimension of $4/3$, in some suitable sense: Hausdorff-Besicovitch, or perhaps Bouligand ("Minkowski"), Tricot ("packing"), and/or other.

Comment. The original illustration of Q in Plate 243 of [FGN] is reproduced as Figure 1. It looked to me like an island with an especially wiggly coastline, hence visual intuition nourished by experience in the sciences suggested $D \sim 4/3$. This value was confirmed by direct numerical tests I commissioned and by further indirect numerical tests.

Literature: It is extensive and endowed with its own web site: math.duke.edu/ faculty/lawler. Major contributions include C. Burdzy, G.F. Lawler, W. Werner, E.E. Puckette, and C. Bishop, P. Jones, R. Permantle & Y. Peres, (*J. Functional Analysis* in press).

Comments on the dimension 4/3, self-avoidance and squigs. There are two reasons for the term "self-avoiding Brownian motion": by definition, $\tilde{Q}$ does not self-intersect and its conjectured dimension $4/3$ is the value found in the self-avoiding random walk on a lattice. The latter $4/3$, which is unquestioned, was obtained by analytic arguments that are geometrically opaque, and the interpretation of $4/3$ as a dimension implies yet another unproven conjecture.

Squids and a wide open issue that combines fractals and topology. To avoid those difficulties, [FGN] (Chapter 24) introduced a class of recursive constructions, *squigs*, that create self-avoidance by recursive interpolation. The simplest is of dimension $\log 2.5/\log 2 \sim 1.3219...$; my original heuristic argument was confirmed by J. Peyrière (*C.R. Acad. Sc.* (Paris), **286**, 1978, 937 ; *Ann. Institut Fourier*: **31**, 1981, 187.) I suspect that the discrepancy between $4/3$ and 1.3219... follows from the fact that squigs involve a recursive subdivision or "triangulation" of the plane. Viewing this discrepancy as of secondary importance, I suspect that self-avoidance is linked in a profound and intrinsic way to the dimension $4/3$. The nature of this link is a mystery and a challenge.

3 Tools of fractal analysis: new, or old but obscure

Need to elaborate on the concept of fractal dimension. The preceeding section hinges on a single concept: a fractal dimension. This concept deserves several distinct comments. For the simplest self-similar fractals, D has a unique value, and to evaluate it is not much of a challenge, even for bright high-school students (if not younger). Those simplest examples contribute to the popularity of fractal geometry and to its pedagogical usefulness, but they happen to be utterly exceptional, hence very misleading. As to the Hausdorff measure, its definition is very short, and Hausdorff gave its value for the Cantor dusts. But it is known numerically in only a few other cases.

The multiplicity of dimensions of self-affine sets. After self-affine sets began to appear in concrete problems, several distinct fractal dimensions turned out to be needed. For example, the Hausdorff-Besicovitch (H.B.) dimension is a *local* concept, but *global* dimensions are also needed; they have been studied for very few cases, in references that are inaccessible but will be included in [SH]. Furthermore, the evaluation of the H-B dimension often proves extremely difficult (C. McMullen, Y. Peres, K. Falconer, to mention only a few) and conjectures abound. Even the graph of the Weierstrass function is of unknown H-B dimension.

The infinity of dimensions for a multifractal measure. While fractal sets call for a finite number of fractal dimensions, fractal measures call for an infinite number, in fact for a real function $f(\alpha)$ of a real variable α. This is one of the several reasons for calling such measures *multifractal.* Starting with the application to turbulence discussed in [SN], Multifractal measures occur in several areas of physics, and [SE] shows how they recently spread into finance. Besides, they are the topic of many studies of purely mathematical character. The literature is extremely large and there is no way of summarizing it here. But to assist the reader unacquainted with the topic and help introduce negative dimension, later in this section, it is good to briefly describe of the original random cascade multifractal.

Construction of a cascade multifractal. Given an integer base b, form the following array of independent and identically distributed (i.i.d.) random variables (r.v.): b r.v. $W(g)$, then b^2 r.v. $W(g,h)$, then b^3 r.v. $W(g,h,k)$ etc... Given a point $t \in [0,1]$, write it in base b as $t = 0, t_1 t_2, \ldots t_n, \ldots$ Define $X'_n(t) = W(t_1)W(t_1,t_2)W(t_1,t_2,t_3)\ldots W(t_1,t_2,\ldots,t_n)$, and

$$X_n(t) = \int_0^t X'_n(s)ds.$$

In a paper reproduced in [SN] (in particular, *J. Fluid Mech.*, **62**, 1974, 331–358 and *C.R. Acad. Sc.* (Paris), **278A**, 1974, 289–292 & 355–358) I posed and solved in part many problems that are relative to a variety of classes of W's, and concern the weak or strong convergence of $X_n(t)$ to a non-vanishing limit $X(t)$, the numbers of finite moments of $X(t)$ and the dimension of the set of t's on which $X(t)$ varies. *Partial answers:* J.P. Kahane and J. Peyrière (*Adv. in Math.* **22**, 1976, 131–145 –translated in [SN]) confirmed and/or extended these conjectures and theorems. Here is one example: when $C = -EW \log_b W < 1$, the measure is non vanishing and can be said to be supported by a set of codimension C.

The fixed points of related smoothing transformations of probability distributions. Take b i.i.d.r.v. W_g, and b i.i.d.r.v. X_g having the same distribution as the $X(1)$ in the preceding paragraph. The weighted average $(1/b)\Sigma W_g X_{\hat{g}}$ (with the sum from 0 to $b-1$) has the same distribution as each X_g, meaning that $X(1)$ is a fixed point of the weighted averaging operation.

Negative dimensions and corresponding challenges and conjectures. Suppose now that the above-defined codimension C is > 1. If so, the measure almost surely vanishes and no further question about it was raised by mathematical analysis. Concrete needs, to the contrary, forced me to distinguished between several distinct levels of emptiness, and the dimension-like quantity $1-C$, which is negative, is an excellent way of fulfilling this need. There is a discussion in *J. Fourier Analysis and Appl.* **Special issue**, 1995, 409–432.

Beyond all fractal dimensions. From the 1960s to the 1980s, the H-B dimension played an important role in helping fractal geometry be started and accepted. Today, the H-B dimension subsists as one of the many alternatives, at best a *primus inter pares.*

More important even is the fact that a careful analysis of both mathematical and observed fractals (in particular in the two sections that follow) showed the need for in many additional old or new tools. We now proceed to two examples.

Distinguishing between the Sierpinski curves on the basis of Urysohn-Menger ramification. Two ancient decorative designs occur in Sierpinski's investigations in the 1910's: one became known as the "carpet", and the second I called "gasket". Sierpinski used the carpet to show that a plane curve can be "topologically universal", that is, contain a homeomorphic transform of every other plane curve. The construction starts with a square, divides it into nine equal subsquares and erases the middle one, which I call a "trema" ($\tau\rho\eta\mu\alpha$ is the Greek term for "hole"). One proceeds in the same fashion with each remaining subsquare, and so on ad infinitum. As to the "gasket", Sierpinski used it to show that a curve can have branching points everywhere. The construction starts with an equilateral triangle, divides it into four equal subtriangles and erases the middle one as trema. One proceeds in the same fashion with each remaining subtriangle, and so on ad infinitum.

During the 1920's, the distinction between the carpet and the gasket became essential to the theory of curves. Piotr Urysohn and Karl Menger took them as prime examples of curves having, respectively, an infinite and a finite "order of ramification."

[FGN] quotes influential mathematicians who took the "gasket" as prime evidence that geometric intuition is powerless, because it can only conceive of branch points as being isolated, not everywhere dense. In fact, Gustave Eiffel himself wrote (as I interpret him) that he would have made his Tower even lighter, with no loss of strength, had the availability and cost of finer materials allowed him to increase the density of double points. From the Eiffel Tower to the Sierpinski gasket is an intellectual step that intuition can be trained to take.

The theory of curves that studies carpets, gaskets and the order of ramification became a stagnant corner of mathematics. Where can one find the latest facts about these notions? The surprising answer is that these notions came to be viewed as "unavoidable", once they were introduced in the statistical physics of condensed matter. Once ridden of the cobwebs of abstraction, they prove to be very practical and enlightening geometric tools to work with (e.g, Gefen, Mandelbrot & Aharony *Phys. Rev. Lett* **45**, 1980, 855–858). Physicists make them the object of scores of articles, and invent scores of generalizations that were not needed in 1915.

A new fractal tool: lacunarity. As is well-known, the most standard construction of a Cantor dust proceeds recursively as follow. The "initiator" is the interval $[0,1]$. Its first stage ends with a generator made of N subintervals, each of length r. In the second stage, each generator interval is replaced by N intervals of length r^2, etc... The resulting limit set arose in the study of trigonometric series, but first attracted wider interest because of its topological and measure-theoretical properties. From those viewpoints, all Cantor dusts are equivalent. Much later, Hausdorff introduced his generalized dimension; this and every other definition of dimension yield the similarity dimension $D = \log N/\log 1/r$, an expression that has become widely known: the value of dimension splits the topological Cantor dusts into finer classes of equivalence parametrized by D.

Fractal geometry began by showing those classes of equivalence to be of great concrete significance. In due time, it went further, because the needs of science rather than mathematics required an even finer subdivision. To pose a problem, consider the Cantor–like constructions stacked in Figure 2. In the middle line, $N = 2$ and $r = 4^{-1}$; k steps below the middle line, $N = 2^k$, $r = 4^{-k}$ and the generator intervals are uniformely spaced ; *k* steps above the middle line, $N = 2^k$, $r = 4^{-k}$, but the generator intervals are crowded close to the endpoints of $[0,1]$. The Cantor dusts in this stack share the common value $D = 1/2$, but look totally different. The Latin word for hole being *lacunar,* motion down (or up) the stack is said to correspond to decreasing (or increasing) *lacunarity*.

Challenge. As $k \to \infty$, the bottom line becomes "increasingly dense" on $[0, 1]$, and the top line "increasingly close to two dots". Provide a mathematical characterization of this "singular" passage to the limit.

Second challenge. [FGN], Chapters 33 to 35, describes and illustrates several constructions that allow a control of lacunarity. However, for the needs of both mathematics and science, the differences between the resulting constructs remained to be quantified. The existing studies of this quantification show that it is not easy and also not unique. Special complications occur when all the reduction ratios are identical, like in Figure 2. Of the alternative methods investigated in the literature, one is based on the prefactor of the relation $M(R) = FR^D$ that yields the mass $M(R)$ contained in a ball of radius R.

Another method is based on the prefactor in the Minkowski content. I studied it in *Fractal Geometry and Stockasties* (ed. C. Bandt et al) Birkhauser, 1995, pp. 12–38.

A third method has the advantage that defines a neutral level of lacunarity that separates positive and negative levels. On the line, this level is achieved by any randomized Cantor dust S with the following property. Granted that any choice of origin Ω in S divides the line into a right and a left half lines, lacunarity is said to be neutral when the intersections of S by those half lines are statistically independent. Increasingly positive (resp. negative) correlations are used to express and measure increasingly low (resp. high) levels of lacunarity. These notions will be used in the two sections that follow.

4 Major fractal clusters in statistical physics

While Brownian motion is fundamental in physics as well as in mathematics, the Brownian clusters in the first section are a mathematical curiosity. However, the property of fractality is shared by all the major real clusters (turbulence, galaxies, percolation, Ising, Potts) and all the major real interfaces (turbulent jets and wakes; metal and glass fractures; diffusion fronts). Each of these categories raises numerous open mathematical questions, of which a few will be listed.

Percolation clusters at criticality. (D. Stauffer & A. Aharony. *Introduction to Percolation Theory*. Second edition. London: Taylor & Francis.) Take an extremely large lattice of tiles. Each tile is chosen at random: with the probability p, it is made of vinyl and with the probability $1 - p$, of copper. Allow electric current to flow between two tiles if they have a side in common. A "cluster" is defined as a collection of copper tiles such that electricity can flow between two arbitrary points in the cluster. For an alternative, but equivalent, construction, define at the center of every tile, a random "relief function" $R(P)$ whose values are independent random variables uniformly distributed from 0 to 1. If this relief is flooded up to level p, each cluster stands out as a connected "island." Physicists conjectured, and mathematicians eventually proved, that there exists a "critical probability" p_C, such that a connected infinite island, i.e., a connected infinite conducting cluster, almost surely exist for $p < p_C$ but not for $p > p_C$.

The geometric complication of percolation clusters at criticality is extreme, and many of the basic conjectures arise not from pure thought, but careful examination of graphics.

Open conjecture A. Take an increasingly large lattice and resize it to be a square of unit side. At p_C, the infinite cluster converges weakly to a "limit cluster" that is a fractal curve.

Open conjecture B. The fractal dimension of this limit cluster is 91/48. This is the value obtained from a partly heuristic "field theoretical" argument that yields characteristic exponents.

Open conjecture C. The limit cluster is a finitely ramified curve in the sense of Urysohn-Menger.

Open conjecture D. Almost every linear cross-section of the limit cluster is a Lévy dust, as defined in [FGN]. Experimental evidence is found in Mandelbrot & Stauffer, *J. Physics* A 28, 1995, L 213 and Hovi, Aharony, Stauffer and Mandelbrot *Phys. Rev Lett.* **77**, 1996, 877–890.

The Ising model of magnets at the critical temperature. At each node of a regular lattice, the Ising model places a spin that can face up or down. The spins interact via forces between neighbors. By themselves, these forces create an equilibrium (minimum potential) situation in which all spins are either up or down. In addition, the system is in contact with a heat reservoir, and heat tends to invert the spins. When the temperature T exceeds a critical value T_C, heat overwhelms the interaction between neighbors. For $T < T_C$, local interactions between neighbors create global structures of greatest interest.

My work touched upon several issues in the shape of the up (or down) clusters at criticality.

Long open implicit question: Beginning with Onsager, it is known that in Euclidean space $\mathbb{R}^E$ the necessary and sufficient condition for magnets to exist is that $E > 1$. There are the innumerable mathematical differences between the $\mathbb{R}^E$ for $E = 1$ and $E > 1$. Identify differences that matter for the existence of magnets.

Partial answer: The specific examples of the Sierpinski curves and of related fractal lattices suggest that magnets can exist when and only when the order of ramification is infinite ([FGN], p. 139; Gefen, Mandelbrot and Aharony, *Phys. Rev. Lett* **45**, 1980, 855).

Conjecture: The above answer is of general validity.

Unanswered challenge. Rephrase the criterion of existence of magnets from the present indirect and highly computational form, to a direct form that would give a chance of proving or disproving the preceding conjecture.

Actual geometric implementation of the fractional-dimensional spaces of physics. Physicists are very successful with a procedure that is mathematically very dubious. They deal with spaces whose properties are obtained from those of Euclidean spaces by interpolation to "noninteger Euclidean dimensions." The dimension may be $4 - \varepsilon$ or $1 + \varepsilon$, where ε is in principle infinitesimal but is occasionally set to $\varepsilon = 1$. Calculations are carried out, in particular, expansions are performed in ε, and at the final stage, the "infinitesimal" ε is set to be the integer. Mathematically, these spaces remain unspecified, yet the procedure turns out to be extremely useful.

Mathematical challenge: Show that the properties postulated for those spaces are mutually compatible, show that they do (or do not) have a unique implementation; describe their implementation constructively.

Very partial solution: A very special example of such space has been implemented indirectly ([FGN], second printing, p. 462; Gefen, Meir, Mandelbrot & Aharony, *Phys. Rev. Lett.* **50**, 1983, 145). We showed that the postulated properties of certain physical problems in this space are identical to the *limits* of the properties of corresponding problems in a Sierpinski carpet whose "lacunarity" is made to converge to 0, in the sense that it tends to 0 as one moves down the stack on Figure 2.

5 The origin of fractality in partial differential equations

To establish that many features of nature (and, as shown in [SE], also of the Stock Market!) are fractal was the daunting task to which a large proportion of [FGN] is devoted. Important new examples keep being discovered, but the hardest present challenge is to discover the causes of fractality. Some cases remain obscure, but others are reasonably clear.

Thus, in the case of the physical clusters discussed in the preceeding section, fractality is the geometric counterpart of the techniques of statistical physics called scaling and renormalization, which show that the analytic properties of those objects follow a wealth of "power-law relations". Many mathematical issues, some of them already mentioned, remain open, but the overall renormalization framework is very firmly rooted.

Similarly, the study of dynamical systems features renormalization and resulting fractality in arguments that involve attractors, repellers and boundaries of basin of attraction. The fractal dimension of those boundaries directly affects the degree of sensitive dependence on initial conditions that characterizes chaotic dynamics. Renormalization also led to the Feigenbaum-Coullet-Tresser theory on bifurcations, and plays an important role in the study of complex quadratic maps (to be considered in a later section).

Unfortunately, additional examples of fractality proved to be beyond the usual renormalization. A notorious case concerns the diffusion-limited aggregates (DLA). Yet another source that covers many very important occurences of fractality led me to a very broad challenge-conjecture which was stated in [FGN], Chapter 11, and which will now be discussed.

Are smoothness and fractality doomed to coexist? *A quandary.* It is universally granted that physics is ruled by diverse partial differential equations, such as those of Laplace, Poisson, and Navier-Stokes. All differential equations imply a great degree of local smoothness, even though closer examination shows isolated singularities or "catastrophes". To the contrary, fractality implies everywhere dense roughness and/or fragmentation. This is one of the several reasons why fractal models in diverse fields were initially perceived as being "anomalies" that stand in direct contradiction with one of the firmest foundations of science.

A conjecture. There is no contradiction at all, in fact, fractals arise unavoidably in the long time behaviour of the solution of very familiar and "innocuous"–looking equations. In particular, many concrete situations where fractals are observed involve equations having free and moving boundaries, and/or interfaces, and/or singularities. As a suggestive "principle", [FGN] (Chapter 11) described the possibility that, under broad conditions that largely remain to be specified, those free boundaries, interfaces and singularities converge to suitable fractals.

Many equations were examined from this viewpoint, but two are of critical importance.

The large scale distribution of galaxies ; Newton's law and fractality. *Background.* Among astronomers, the near universally held view is that the distribution of galaxies is homogenous, except for local deviations. However ([FGN], Chapter 9), philosophers or science fiction writers played with the notion that the distribution is hierarchical, in a way unknowingly patterned along a spatial Cantor set. Hierarchical models were dismissed as unrealistic, in fact, largely forgotten. They are excessively regular, for example the reduction ratio must be (positive or negative) power of a basic ratio r_0. They necessarily imply that the Universe has a center and the model and reality can not only be matched by introducing a host of ad-hoc "patches". Last but not least, the hierarchical models predict nothing, that is, have no property that was not put in beforehand, and raise no new question.

Conjecture that the distribution of galaxies is properly fractal. ([FGN], Chapters 9 and 33 to 35). This conjecture results from a search for invariants that was central to every aspect of my construction of a fractal geometry. Granted that the distribution of galaxies certainly deviates in some ways from homogeneity, two broad approaches were tried. One consists in correcting for local inhomogeneity by incorporating local "patches". The next simplest global assumption is that the distribution is non-homogenous but scale-invariant. I chose to follow up this assumption, while excluding the strict hierarchies.

A surprising and noteworthy finding rewarded a detailed mathematical *and* visual investigation of sample sites generated by two concrete constructions of random fractal sets. As they are random, their self-similarity can only be statistical, which may be viewed as a drawback. But a more than counter-acting strong asset is that the self-similarity ratio can be chosen freely. It is not restricted to powers of a prescribed r_0, that is, the hierarchical structure is not a deliberate and largely arbitrary input. Quite to the contrary, the existences of clear-cut clusters are an unanticipated property of the construction. The details are given in [FGN]. The first construction is *The Seeded Universe,* based on a Lévy flight. Its Hausdorff-dimensional properties were known. Its correlation properties (Mandelbrot *C.R. Acad. Sc.* (Paris), **280A**, 1975, 1075) are nearly identical to those of actual galaxy maps. The second construction is *The Parted Universe*, which is obtained by subtracting from space a random collection of overlapping sets, already described as being called "tremas". Here, the tremas are allowed to overlap. Either construction yields sets that are highly irregular and involve no special center, yet exhibit a clear-cut clustering that was not deliberately inputted. They also exhibit "filaments" and "walls", which could not possibly have been inputted, because I did not know that they have been observed.

Conjecture that the observed "clusters", "filaments" and "walls" need not be explained separately, but necessarily follow from "scale free" fractality. This subtitle consists in conjecturing that the properties that it lists do not result from unidentified specific features of the models that have actually been studied, but follow as consequences from a variety of unconstrained forms of random fractality.

In the preceding title and the sentence that elaborates it, the word "conjecture" cannot be given its strict mathematical meaning, until a mathematical meaning is advanced for the remaining terms.

Lacunarity. A problem arose when careful simulations of the Seeded Universe proved to be visually far more "lacunar" than the real world. This notion, which was already mentioned, means that the simulations show the holes larger than in reality. The Parted Universe model fared better, since its lacunarity can be adjusted at will and fitted to the actual distribution, as shown in Mandelbrot, *C.R. Acad. Sc.* (Paris), **288**, 1979, 81–83.

A lowered lacunarity is expressed by a positive correlation between masses in antipodal directions. Testing this specific conjecture is a challenge for those who analyze the data.

Conjectured mathematical explanation of why one should expect the distribution of galaxies to be fractal. Consider a large array of point masses in a cubic box in which opposite sides are identified to form a 3 dimension' torus. How this array evolves under the action of inverse square attraction is a problem that obeys the Laplace equation, with the novelty that the singularities of the solution are the positions of the points, therefore, movable. All simulations I know of (beginning with those performed by IBM colleagues around 1960) suggest the following. Even when the pattern of the singularities begins by being uniform or Poisson, it gradually creates clusters and a semblance of hierarchy, and appears to tend toward fractality. It is against the preceeding background that I conjectured that the limit distribution of galaxies is fractal, and that the origin of fractality lies in Newton's equations.

The Navier Stokes and Euler equation of fluid motion and fractality of their singularities. *Background.* It is worth mentioning that the first concrete use of a Cantor dust in real spaces is found in a 1963 paper on noise records by Berger & Mandelbrot (reprinted in [SN]). This work was near simultaneous with Kolmogorov's work on the intermittence of turbulence. After numerous experimental tests, designed to create an intuitive feeling for this phenomenon (e.g., listening to turbulent velocity records that were made audible), I extended the fractal viewpoint to turbulence, and was led circa 1964 to the following conjecture.

Conjecture concerning facts. The property of being "turbulently dissipative" should *not* be viewed as attached to domains in a fluid with significant interior points, but as attached to fractal sets. In a first approximation, those sets' intersection with a straight line is a Cantor-like fractal dust having a dimension in the range from 0.5 to 0.6. The corresponding full sets in space should therefore be expected to be fractals with a Hausdorff dimension in the range from 2.5 to 2.6.

Actually, Cantor dust and Hausdorff dimension are not the proper notions in the context of viscous fluids, because viscosity necessarily erases the fine detail that is essential to Cantor fractals. Hence the following.

Conjecture: ([FGN], Chapter 11 and Mandelbrot, *C.R. Acad. Sc.* (Paris), **282A**, 1976, 119, translated as Chapter 19 of [SN]). The dissipation in a viscous fluid occurs in the neighborhood of singularity of a nonviscous approximation following Euler's equations, and the motion of a nonviscous fluid acquires singularities that are sets of dimension about 2.5 to 2.6. *Open mathematical problem:* To prove or disprove this conjecture, under suitable conditions.

Comment A. Several numerical tests agree with this conjecture (e.g. Chorin, *Comm. Pure and Appl. Math.*, **34**, 1981 853–866).

Comment B. I also conjectured that the Navier-Stokes equations have fractal singularities, of much smaller dimension. A technical inequality equivalent to this conjecture turned out to be present in a 1934 paper by J. Leray (*Acta Mathematica.*) Once revied and provided with the appropriate geometric interpretation, it led to extensive work by V. Scheffer, and then many others.

Comment C. As is well-known to students of chaos, a few years after my work, fractals in phase space entered in the study of the transition from laminar to turbulent flow, through the work of Ruelle & Takens and their followers. The task of unifying the real and phase-space roles of fractals is not yet completed.

6 Iterates of the complex map $z^2 + c$. Julia and Mandelbrot sets

The study of the iterates of rational functions of a complex variable reached a peak circa 1918. Fatou and Julia succeeded so well that – apart from the proof of the existence of Siegel discs – their theory remained largely unchanged for sixty years.

The J set or Julia set. This set, defined as the repeller of rational iteration, is typically a fractal: a nonanalytic curve or a "Cantor-like" dust. Julia called those repellers "very irregular and complicated." The computer –which I was the first to use systematically– reveals they are beautiful. To associate forever the name of Fatou and Julia, the *complement* of the Julia set is best called *Fatou set* and its maximal open components, *Fatou domains*. The wildly colorful displays that represent them must now be familiar to every reader.

Starting with the quadratic map $z \to z^2 + c$, I explored numerically how the value of c affects the nature of quadratic dynamics, and in particular, the shape of the Julia set.

The M_0 set. Of greatest interest from the viewpoint of dynamics, hence, of physics, is the set M_0 of those values of c for which $z^2 + c$ has a finite stable limit cycle.

The M_0 set having proved to be hard to investigate directly, I moved onto the computer-assisted investigation of a set that is easier to study, and seemed closely related.

The M set or Mandelbrot set. Douady & Hubbard gave this name to the set of those parameter values c in the complex plane, for which the Julia set is connected.

M, called μ-map in [FGN] (Chapter 19), proved to be a most worthy object of study, first for "experimental mathematics" and then for mathematics, and many facts are known about it. It even created a new form of art! It is so well and so widely known, that no further reference is needed. This discussion will limit itself to one major unsolved conjecture.

Conjecture that M is the closure of M_0. A computer approximation can only yield a set smaller than M_0, and a set larger than M. Extending the duration of the computation seemed to make the two representations converge to each other. Furthermore, when c is an interior point of M, not too close to the boundary, it was easily checked that a finite limit cycle exists. Those observations led to the conjecture that M is identical to M_0 together with its limit points.

In terms of its being simple and understandable without any special preparation, this conjecture comes close to where this paper starts: the "dimension $4/3$" conjecture about Brownian motion. Again, I could think of no proof, even of a heuristic one. More significantly, after eighteen-odd years, the conjecture remains unanswered.

The MLC conjecture. Many equivalent statements were identified, the best known being that the Mandelbrot set is locally connected. This statement acquired a "nickname", MLC; it has the great advantage of being local. (J.C. Yoccoz received high praise for proving it for a very large subset of the boundary). But, compared to the original form, it has the great drawback of being incomparably more sophisticated and, for most people, far from intuitive.

References

[FGN] Mandelbrot, B.B. 1982, *The Fractal Geometry of Nature,* W.H. Freeman and Co., New York and Oxford. The second and later printings include an Update and additional references. Earlier versions were *Les objets fractals: forme, hasard et dimension,* Flammarion, Paris, 1975 (fourth edition, 1995) and *Fractals: Form, Chance and Dimension,* Freeman, 1977.

[SE] Mandelbrot, B.B. 1997E, *Fractal and Scaling in Finance: Discontinuity, Concentration,Risk* (Selecta, Volume E) Springer-Verlag, New York.

[SN] Mandelbrot, B.B. 1998N, *Multifractals and 1/f Noise: Wild Self-Affinity in Physics.* (Selecta, Volume N). Springer-Verlag, New York (expected early in 1998).

[SH] Mandelbrot, B.B. 1998H, *Gaussian Fractals and Beyond* (Selecta, volume H). Springer-Verlag, New York (expected late in 1998).

Note: Specific references are added in the text.

Fractal Geometry and Physical Phenomena

Luciano Pietronero
Università di Roma "La Sapienza"
Roma, Italy

Abstract

In the last years there has been a growing interest in the understanding of a vast variety of scale invariant and critical phenomena occurring in nature. Experiments and observations indeed suggest that many physical systems develop spontaneously power law behavior both in space and time. Pattern formation, aggregation phenomena, biological and geological systems, disordered materials, clustering of matter in the universe are just some of the fields in which scale invariance has been observed as a common basic feature. In this respect fractal geometry has changed the way we look at nature and it has expanded the frontiers of physical sciences to include a wide variety of strongly irregular systems and complex phenomena. The value and impact of fractals, however, is still rather controversial. In this lecture we discuss the real advancements as well as the present limitations of this field by presenting it along three distinct lines, which constitute evolutionary stages: (i) Fractal geometry as a *mathematical framework* that allows us to identify and characterize scale invariant properties in natural phenomena. (ii) The development of *physical models* for the spontaneous development of fractal structures in well defined physical phenomena. (iii) The attempts to construct *physical theories* that should provide a full understanding for the self-organized origin of fractal structures in various systems. The style of the present discussion will be colloquial but the references can give a clue for a more technical level.

1 Introduction

Statistical physics is undergoing a profound transformation. The introduction of new ideas, inspired by fractal geometry and scaling, irreversible and non-ergodic dynamical systems leading to self-organization and stochastic processes of various types, leads to a considerable enrichment of the traditional framework and provides efficient methods for characterising and understanding complex systems.

The physics of scale-invariant and complex systems is a novel field which is including topics from several disciplines ranging from condensed matter physics to geology, biology, astrophysics and economics [1]. This widespread interdisciplinarity corresponds to the fact that these ideas allow us to look at natural phenomena in a radically new and original way, eventually leading to unifying concepts independently of the detailed structure of systems.

In scale invariant phenomena, events and information spread over a wide range of length and time scales, so that no matter what is the size of the scale considered one always observes surprisingly rich structures. These systems, with very many degrees of freedom, are usually so complex that their large scale behaviour cannot be predicted from the microscopic dynamics. New types of collective behaviour arise and their understanding represents one of the most challenging areas in modern statistical physics.

The interest in this field has been largely due to two factors. First the emerging availability of high powered computers over the past decade has enabled to readily simulate complex and disordered systems. Second the cross disciplinary mathematical language for describing these phenomena evolving under conditions far from equilibrium has only become available in the past years. The study of critical phenomena in second order transitions introduced the concepts of scaling and power law behavior [2]. Fractal geometry [3] provided the mathematical framework for the extension of these concepts to a vast variety of natural phenomena.

The physics of complex systems, however, turned out to be effectively new with respect to critical phenomena. The theory of equilibrium statistical physics is strongly based on the ergodic hypothesis and scale invariance develops at the critical equilibrium between order and disorder. Reaching this equilibrium requires the fine tuning of various parameters. On the contrary, most of the scale-free phenomena observed in nature are *self-organized*, in the sense that they spontaneously develop from the generating dynamical process. One is then forced to seek the origin of the scale invariance in nature in the rich domain of nonequilibrium systems and this requires the development of new ideas and methods.

The realization that certain structures exhibit fractal properties does not tell us why this happens but it is crucial to formulate the right questions. The impact of fractals in physics can be assessed along three different lines of increasing complexity:

(a) Fractal geometry merely as a *mathematical framework* which leads to a re-analysis of known data that results in a revamping of long-standing points of view. This permits to include into the scientific areas many phenomena characterised by intrinsic irregularities which have been previously neglected because of the lack of an appropriate mathematical. The main examples of this type can be found in the geophysical and astrophysical data and in Section 3 we consider one example in detail. The possibility of extending these methods also to biological evolution in terms of complex adaptive systems is also an active field of research.

(b) The development of *physical models* for systems that exhibit fractal and Self-Organized Critical (SOC) behaviour. From a mathematical point of view the problems explored are paticularly difficult. Often they consist of iterative systems with many degrees of freedom and irreversible dynamics. Very little can be predicted a priori for systems of this complexity, even though sometimes they can be very easy to fomulate. In this respect computer simulations represent an essential method in the physics of complex and scale invariant systems. While the great majority of the theoretical activity is based upon "toy models" which barely resemble real nature, it is important to build a bridge between theory and real experiments and this another basic task of computer simulations. This implies the development of models with the properties of a greater realism and large scale simulations which can be used also in material characterization. A byproduct of this approach is the application of fractal concepts to the solution of particular experimental problems (oil industry, disordered materials, phase nucleation, crystal growth etc.)

(c) The construction of complete *physical theories* that allow us to understand the self-organized origin of fractal structures as well as all the other relevant properties in various physical systems and phenomena. At a phenomenological level, scaling theory, inspired to usual critical phenomena, has been successfully used. This is essential for the rationalization of the results of computer simulations and experiments. This method allows us to identify the relations between different properties and exponents and to focus on the essential ones. The situation is completely different in relation to the formulation of a microscopic fundamental theory. The theoretical approach is particularly difficult because the statistical physics of systems far from equilibrium lies far beyond the usual equilibrium theory. This implies that the time development is intrinsically irreversible and that it cannot be eliminated by some form of the ergodic hypothesis. In equilibrium statistical mechanics it is in fact possible to eliminate the specific dynamical evolution and to assign directly a Boltzmann weight to a given configuration. In the case of self-organized fractal structures this is usually not possible and a full knowledge of the dynamical history is necessary. This implies the development of theoretical concepts of novel type.

2 Scale invariance and intrinsic irregularity

Most of theoretical physics is based on analytical functions and differential equations. This implies that structures should be essentially smooth and irregularities are treated as single fluctuations or isolated singularities. The study of critical phenomena and the development of the Renormalization Group (RG) theory in the seventies was therefore a major breakthrough [1, 4]. One could observe and describe phenomena in which intrinsic self-similar irregularities develop at all scales and fluctuations cannot be described in terms of analytical functions. The theoretical methods to describe this situation could not be based on ordinary differential equations because self-similarity implies the absence of analyticity and the familiar mathematical physics becomes inapplicable. In some sense the RG corresponds to the search of a space in which the problem becomes again analytical. This is the space of scale transformations but not the real space in which fluctuations are extremely irregular. For a while this peculiar situation seemed to be restricted to the specific critical point corresponding to the competition between order and disorder. In the past years instead, the development of fractal geometry [3], has allowed us to realize that a large variety of structures in nature are intrinsically irregular and self-similar.

Mathematically this situation corresponds to the fact that these structures are singular in every point. This property can be now characterized in a quantitative mathematical way by the fractal dimension and other suitable concepts. However, given these subtle properties, it is clear that making a theory for the physical origin of these structures is going to be a rather challenging task. This is actually the objective of the present activity in the field [5].

The main difference between the popular fractals like coastlines, mountains, trees, clouds, lightning, etc. and the self-similarity of critical phenomena is that criticality at phase transitions occurs only with an extremely accurate fine-tuning of the critical parameters involved. In the more familiar structures observed in nature instead the fractal properties are self-organized, they develop spontaneously from the dynamical process. It is probably in view of this important difference that the two fields of critical phenomena and fractal geometry have proceeded

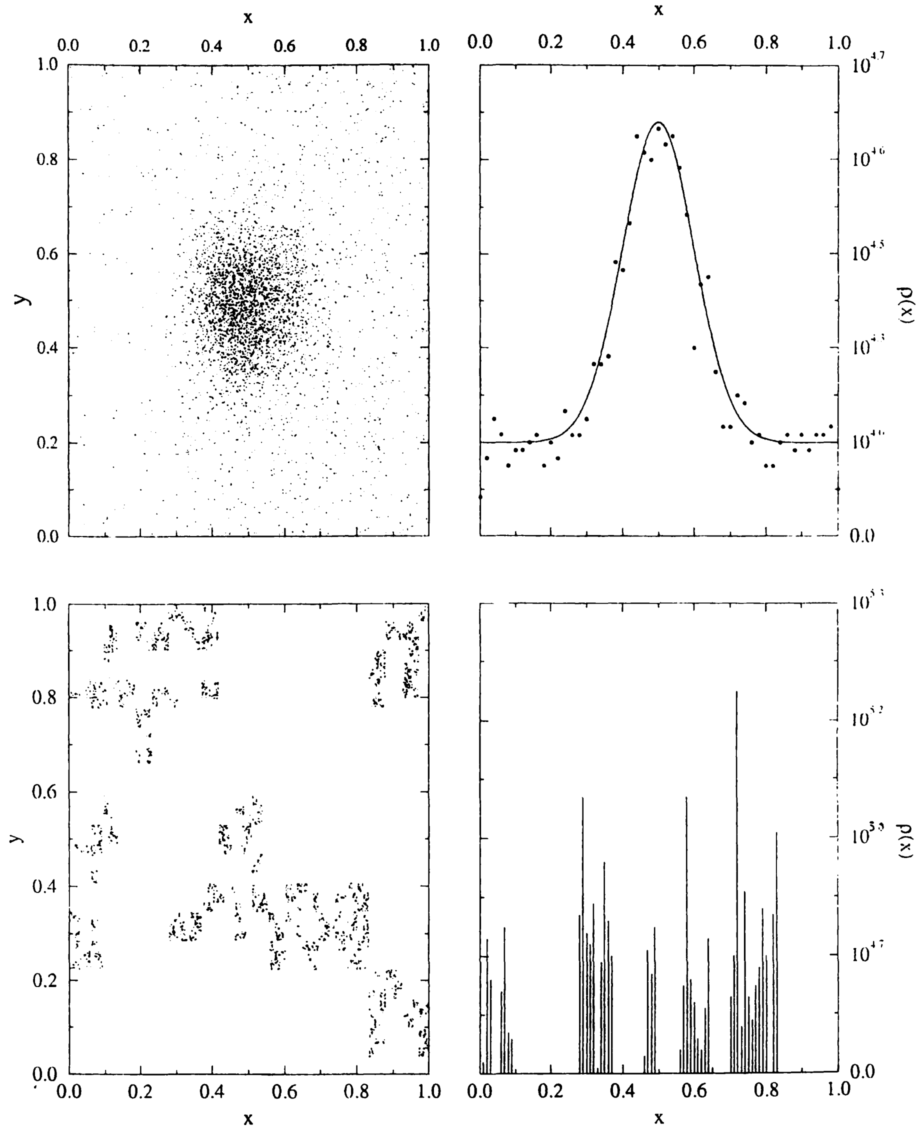


Figure 1: Example of analytical and nonanalytic structures. Top panels: (Left) A cluster in a homogenous distribution. (Right) Density profile. In this case the fluctuation corresponds to an enhancement the two-dimensional Euclidean space. (Right) Density profile. In this case the fluctuations are non-analytical and there is no reference value, i.e. the average density. The average density scales as a power law from any occupied point of the structure.

somewhat independently, at least at the beginning. The fact that we are traditionally accustomed to think in terms of analytical structures has crucial consequences on the type of questions we ask and on the methods we use to answer them. If one has never been exposed to the subtleness on nonanalytic structures, it is natural that analyticity is not even questioned. It is only after the above developments that we could realize that the property of analyticity can be tested experimentally and that it may or may not be present in a given physical system.

3 Fractal properties of the large-scale universe

In this section we discuss an example of the first category mentioned in the introduction in which the concept of Fractal Geometry, used as a mathematical tool, discloses new properties for the large-scale strucure of the universe and leads to fascinating and controversial perspectives.

The three-dimensional distribution of galaxies appears quite irregular and it consists of large structures and large voids. In the example shown in Figure 2 our galaxy is at the center and the empty slice corresponds to the galactic plane in which observations are difficult. Note that the picture is a projection (orthogonal) and this gives a smoothing effect to the eye. If one could rotate this picture as in a video the large structures and large voids would be better defined. Despite these structures the universe is believed to be homogeneous at large scale and this property is supposed to be in agreement with the data of Figure 2.

Some years ago we proposed a new approach for the analysis of galaxy and cluster correlations based on the concepts and methods of modern statistical physics. This led to the surprising result that galaxy correlations are fractal and not homogeneous up to the limits of the available catalogues. In the meantime many more red shifts have been measured and we have extended our methods also to the analysis of various other properties [6, 8].

The usual statistical methods, based on the assumption of homogeneity [9], appear therefore to be inconsistent for all the length scales probed until now. A new, more general, conceptual framework is necessary to identify the real physical properties of these structures, and theories should shift from "amplitudes" to "exponents" in the sense discussed in the previous section.

The new analysis shows that all the available data are consistent with each other and show fractal correlations (with dimension $D = 2$) up to the deepest scales probed until now (1000Mpc) [7, 8]. In these units, megaparsecs, the radius

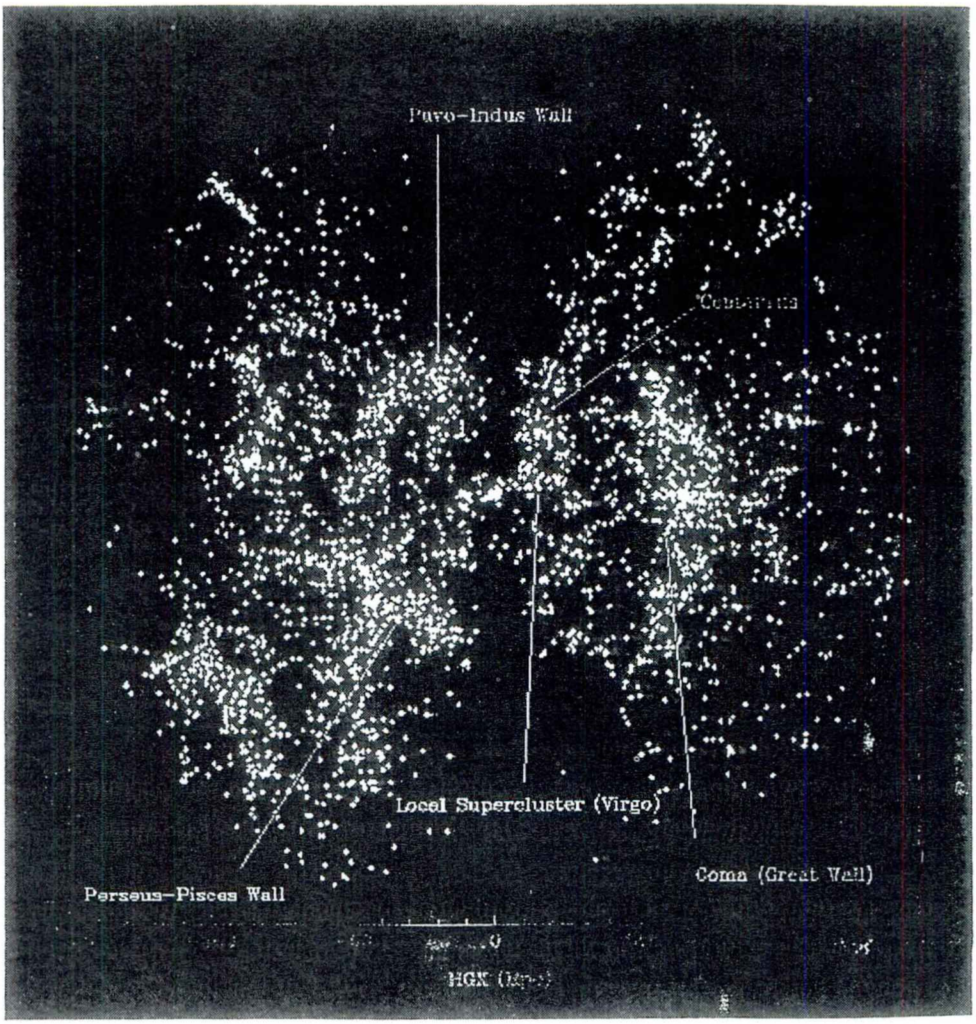


Figure 2: Three-dimensional distribution of galaxies around our galaxy (central point). The zone represented corresponds to about one tenth of the size of the entire universe.

of the entire universe is about 4000Mpc, while the size of a single galaxy (a point in our analysis) is about 0.01-0.1Mpc. The distribution of visible matter in the universe is therefore fractal and not homogeneous. In addition, the luminosity distribution is correlated with the space distribution in a specific way characterized by multifractal properties. These facts lead to fascinating conceptual implications about our knowledge of the universe and to a new scenario for the theoretical challenge.

This result has caused a large debate in the field [6] because it is in contrast with the usual assumption of large-scale homogeneity which is at the basis of most theories. Actually homogeneity represents much more than a working hypothesis for theory, it is often considered as a paradigm or principle and for some authors it is conceptually absurd even to question it [9].

For other authors instead, homogeneity is just the simplest working hypothesis and the idea that nature might actually be more complex is considered as extremely interesting [10]. These two points of view are not so different after all because, if something considered absurd becomes real, then it is indeed very exciting.

The problem is that these concepts touch directly the so-called Cosmological Principle (CP), which represents one of the landmarks of the field of cosmology. It is quite reasonable to assume that we are not in a very special point of the universe and to consider this as a principle, the CP. The usual mathematical implication of this principle is that the universe must be homogeneous [9]. This reasoning implies the hidden assumption of analyticity that often is not even mentioned. In fact the above reasonable requirement only leads to local isotropy. For an analytical structure this also implies homogeneity [10]. However, if the structure is not analytical, the above reasoning does not hold. For example a fractal structure has local isotropy but not homogeneity. In simple terms this means that all galaxies live in similar enviroments made of structures and voids (statistical isotropy). Therefore a fractal structure satisfies the CP in the sense that all the points are essentially equivalent (no center or special points) but this does not imply that these points are distributed uniformly.

The usual correlation analysis is performed by estimating at which distance (r_0) the density fluctuations are comparable to the average density in the sample. In practice this is done by considering the function $\xi(r) =< n(0)n(r) > / < n >^2 -1$, and by defining the characteristic length (r_0) as the point at which $\xi(r_0) = 1$. Now everybody agrees that there are fractal correlations at least at small scales. The important physical question is therefore to identify the distance λ_0 at which, possibly, the fractal distribution has a crossover into a homogeneous one. This would be the real correlation length beyond which the distribution can be approximated by an average density. The problem is therefore to understand the relation between r_0 and λ_0: are they the same or closely related or do they correspond to different properties?

This is actually a subtle point with respect to the concepts discussed in Section 2. In fact, if the galaxy distribution becomes really homogeneous at a scale λ_0 within the sample in question, then the value of r_0 is proportional to λ_0 and is related to the real correlation properties of the system.

If, on the other hand, the fractal correlations extend up to the sample limits, then the resulting value of r_0 has nothing to do with the real properties of the galaxy distribution but it is fixed just by the size of the sample [6].

Given this situation of ambiguity with respect to the real meaning of r_0, it is clear that the usual correlation study in terms of the function $\xi(r)$ is not the appropriate method to clarify these basic questions. The essential problem is that, by using the function $\xi(r)$, one defines the amplitude of the density fluctuations by normalizing them to the average density of the sample in question. This implies that the observed density should be the real one and it should not depend on the given sample or on its size apart from Poisson fluctuations. However, if the distribution shows long-range (fractal) correlations, this approach becomes meaningless. For example if one studies a fractal distribution with $\xi(r)$ a characteristic length r_0 will be identified, but this is clearly an artifact because the structure is characterized exactly by the absence of any defined length [6].

The appropriate analysis of pair correlations should therefore be performed using methods that can check homogeneity or fractal properties without assuming a priori either one. The simplest method to do this is to consider directly the conditional density $\Gamma(r) =< n(0)n(r) >$ without comparing it to the average density. This is not all however because one has also to be careful not to make hidden assumptions of homogeneity in the treatment of the boundary conditions [6, 8]. For these reasons the statistical validity of a sample is limited to the radius (Rs) of the largest sphere that can be contained in the sample.

The main data of our correlation analysis are collected in Figure 3 in which we report the conditional density as a function of distance for various galaxy catalogues. The properties derived from different catalogues are compatible with each other and show a power law decay (fractal correlations) for the conditional density from 1Mpc to 150Mpc without any tendency towards homogenization (flattening). Using also other data for which only a limited analysis is possible, one can see that the fractal behavior continues up to about 1000Mpc. (For a detailed discussion see [8]). This implies necessarily that the value of r_0 (derived from the $\xi(r)$ approach) is actually spurious and it will scale with the sample size Rs as discussed in detail in [8]. The behaviour observed corresponds to a fractal structure with dimension $D = 2$. A homogeneous distribution would correspond to a flattening of the conditional density which is never observed.

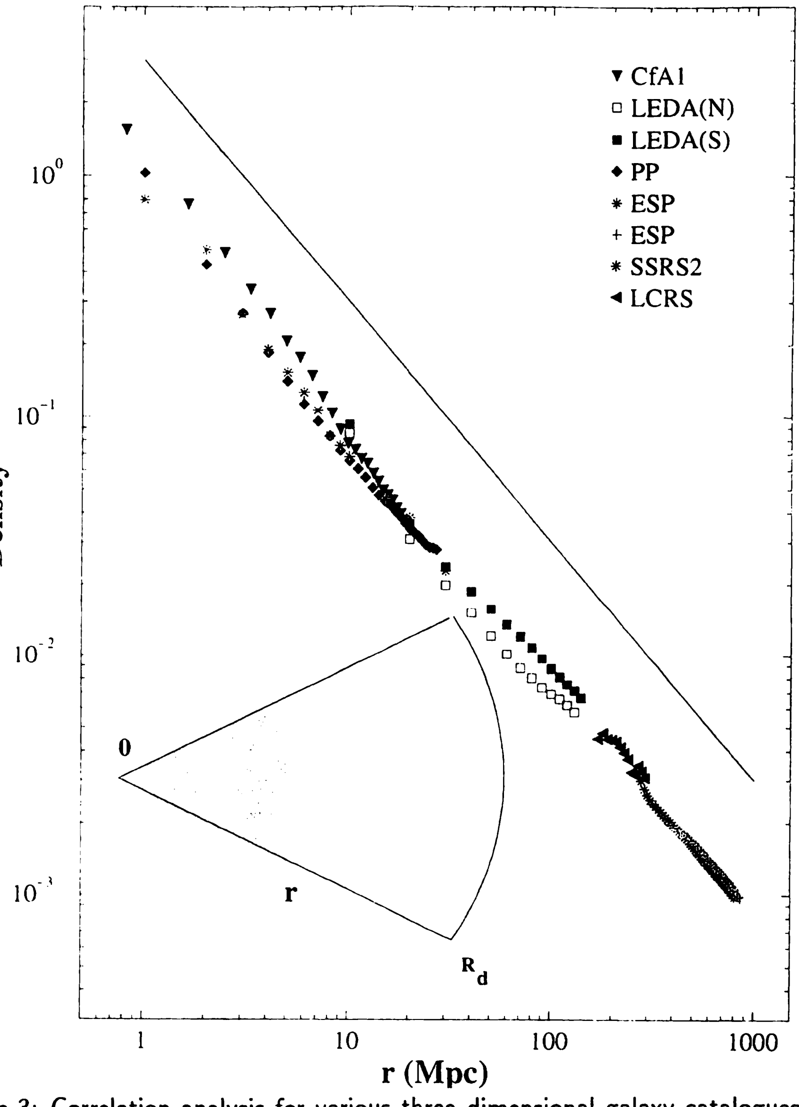


Figure 3: Correlation analysis for various three dimensional galaxy catalogues in the range 0.1 - 1000Mpc. The plot refers to the behavior of the conditional density as a function of distance. A reference line with slope -1 is also shown (i.e. fractal dimension $D = 2$). A constant density would correspond to a flat behavior.

It is important to remark that the usual correlation analyses have had a profound influence in the field in various ways [9]: first the various catalogues appear in conflict with each other. This has generated a strong mutual criticism about the validity of the data between different groups. In other cases the discrepancies observed have been considered as real physical problems for which various theoretical approaches have been proposed. These problems are, for example, the galaxy-cluster mismatch, luminosity segregation, the richness clustering relation and the linear and non-linear evolution of the perturbations corresponding to the "small" or "it large" amplitudes of fluctuations. We can now see that all this problematic is not real and it arises only from a statistical analysis based on inappropriate assumptions that do not find a correspondence in physical reality. It is also important to note that, even if the galaxy distribution would eventually become homogeneous at some large scale, the use of the above statistical concepts is anyhow inappropriate for the range of scales in which the system shows fractal correlations as those shown in Figure 3.

Up to now we have discussed galaxy correlations only in terms of the set of points corresponding to their position in space. Galaxies can be also characterized by their luminosity (related to their mass) and the luminosity distribution is then a full distribution and not a simple set. It is natural then to consider the possible scale invariant properties of this distribution. This requires a generalization of the fractal dimension and the use of the concept of multifractality [8]. A multifractal analysis shows that also the full distribution is scale invariant and this leads to a new and important relation between the Schechter luminosity distribution and the space correlation properties. This allows us to understand various morphological features (like the fact that large elliptic galaxies are typically located in large clusters) in terms of multifractal exponents. This leads also to a new interpretation of what has been called the luminosity segregation effect [8].

In summary our main points are:

(a) The highly irregular galaxy distributions with large structures and voids strongly point to a new statistical approach in which the existence of a well defined average density is not assumed a priori and the possibility of non-analytical properties should be addressed specifically.

(b) The new approach for the study of galaxy correlations in all the available catalogues shows that their properties are actually compatible with each other and they are statistically valid samples. The severe discrepancies between different catalogues that have led various authors to consider these catalogues as not fair, were due to the inappropriate methods of analysis.

(c) The correct two-point correlation analysis shows well-defined fractal correlations up to the present observational limits, from 1 to 1000Mpc. with fractal dimension $D = 2$.

(d) The inclusion of the galaxy luminosity (mass) leads to a distribution which is shown to have well-defined multifractal properties. This leads to a new, important relation between the luminosity function and that galaxy correlations in space.

From the theoretical point of view, the fact that we have a situation characterized by self-similar structures implies that we should not use concepts which make reference to the average density or related properties. One cannot talk about "small" or "large" amplitudes for a self-similar structure because of the lack of a reference value like the average density. Physics should shift from "amplitudes" towards "exponents" and the methods of modern statistical Physics should be adopted. This leads to a new, fascinating situation, that has been uncovered by the introduction of the concepts of self-similarity and fractal geometry.

4 Fractal physical models

The key question is *how does nature produce fractal structures.* The first physical model that shed light on this question was the *Diffusion Limited Aggregation* (DLA) model of Witten and Sander [11] introduced in 1981. The model was inspired by the observation of growing aggregates that were found to exhibit fractal structures. One starts with a seed particle and introduces a new particle at some (large enough) distance R that executes a random walk on a lattice. When the particle reaches a site adjacent to the seed, it is frozen in that position and extends the seed. A new particle is then introduced until it touches the new seed and so on. The iteration of this simple algorithm produces structures of great complexity with a fractal dimension $D = 1.7$ (for planar growth). An interesting variant of DLA is the *Cluster-Cluster* aggregation model [12] where one starts with many particles executing random walks that are allowed to aggregate into clusters. Clusters of all sizes continue to execute random walks forming cluster aggregates and so on. Each cluster turns out to be fractal with a dimension that is lower than in the DLA model. In addition the distribution of cluster sizes exhibits power-law behavior. The Cluster-Cluster model captures the physics of dust or smoke clouds and colloids [13] as shown in Figure 4.

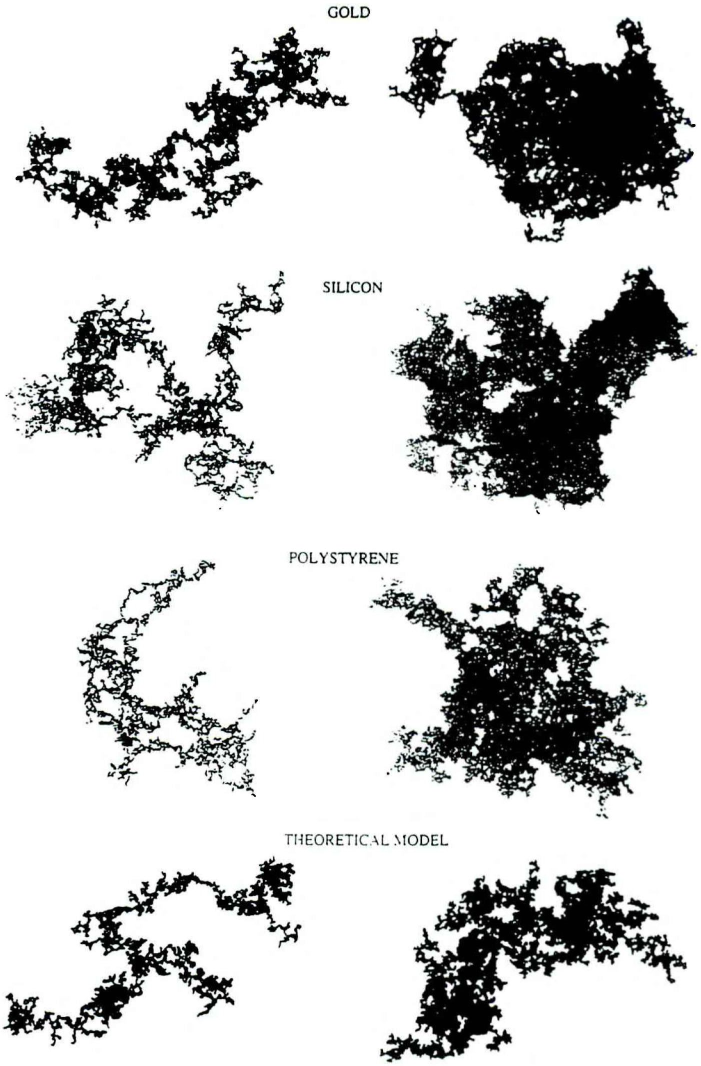


Figure 4: Real aggregation processes compared with the theoretical models. On the left we report the case of growth conditions in which the structure cannot rearrange after aggregation (low temperature). On the right, instead, structures may rearrange (high temperature). The first three cases correspond to real materials while the figures at the bottom correspond to the result of the theoretical model for the two growth regimes. (Courtesy of D. Weitz [12].)

In 1984 Niemeyer et al. introduced the Dielectric Breakdown Model (DBM) [14] inspired by discharges in gases (e.g. lightning). The discharge pattern is assumed to be composed of discrete points connected by bonds (see Figure 5) and

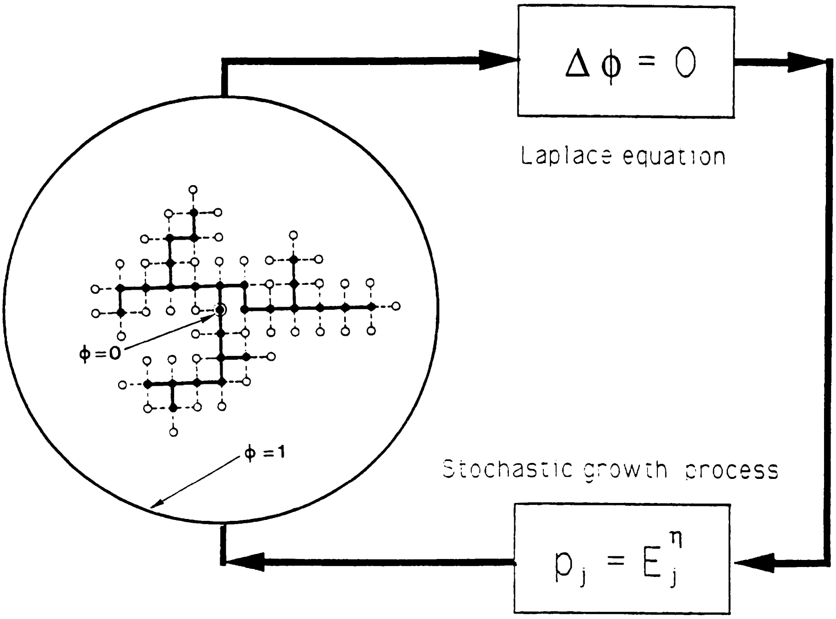


Figure 5: Schematic picture of DBM growth process. The grown structure is assumed to be equipotential. From the Laplace equation one can compute the local field for the bonds around the structure. The growth probability is related to the local field. A bond is selected and added to the structure and the process is then iterated.

the entire pattern at a given time is considered equipotential. At each perimeter point, a growth probability is assigned to be proportional to the local electric field E or to a power E^η. The electric field is determined from Laplace's equation for the electrical potential. The stochastic iteration of the model produces fractal structures with fractal dimensions that depend on h. In the case $h = 1$ one recovers the DLA structure. Apart from generalizing the DLA growth process the DBM illustrates the underlying mathematical properties in relation to partial differential equations like Laplace equation. This connection is quite surprising because usually Laplace equation produces smooth solutions: the potential at a given point is the average of the potential of the neighburing points. Here we see instead that a stochastic growth scheme in which the probabilities are defined by Laplace equation drives *spontaneously* the growth boundaries into a highly irregular fractal structure. In Figure 6 we show a fractal structure of the DLA/DBM in which the black and white stripes provide a visual impression of the variation of the electrical potential around the structure. A pair of black and white stripes corresponds to a change of a decade in the potential.

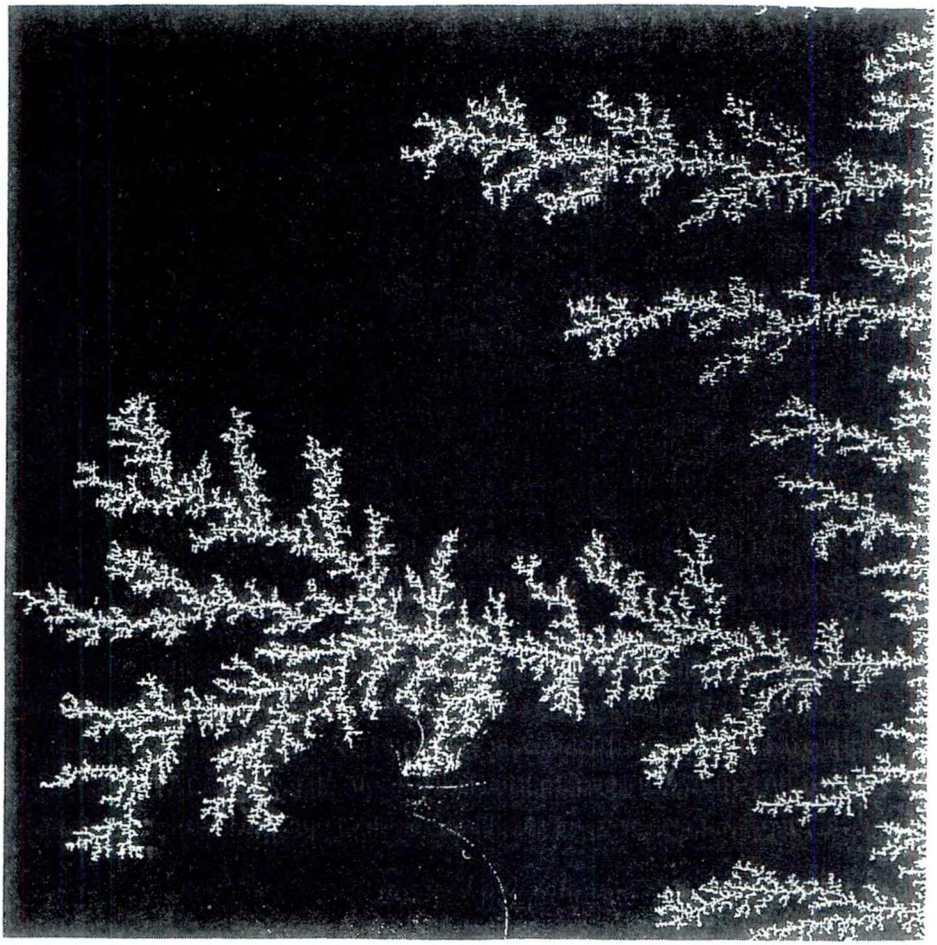


Figure 6: DLA/DBM cluster with potential stripes.

These "Laplacian" fractals are believed to capture the essential fractal properties of a variety of phenomena such as electrochemical deposition, dielectric breakdown, viscous fingering in fluids, the propagation of fractures and various properties of colloids [15].

The essential properties of these growth models are (for a detailed discussion see [5]):

— The growth process is *irreversible*. There is a growing interface and a "frozen" zone that will not be modified by further growth. The asymptotic properties and the fractal dimension refer to the frozen asymptotic structure.

— In order to assign a statistical weight to a structure it is necessary to know its entire *growth history*.

— The dynamics of these models evolves *spontaneously* into a fractal structure without the fine-tuning of any critical parameter, as is instead the case in ordinary critical phenomena.

— The degree of *universality* appears to be reduced with respect to equilibrium critical phenomena. For example in DBM the fractal dimension is a continuous function of the parameter h, but even for the standard DLA model, radial or cylindrical boundary conditions produce non-trivial differences.

The concept of spontaneous generation of complex or critical structures, also called *Self-Organized Criticality* (SOC), has been recently emphasized and investigated in the *sandpile models* introduced by Bak and coworkers [16]. To illustrate the basic ideas of SOC, they introduced a cellular automaton model of sandpiles. The random addition of sand grains drives the system towards a stationary state with a scale-free distribution of avalanches. As in the previous fractal growth models, also in this model criticality seems to emerge automatically without the fine-tuning of parameters. Because of the enormous conceptual power, SOC ideas have invaded rapidly throughout the sciences, from physics and geophysics to biology and economics, as a prototype mechanism to understand the manifestation of scale invariance and complexity in natural phenomena. It is interesting to compare in Table I the properties of these new models of fractal growth and SOC with those of standard critical phenomena represented by the Ising model.

Another model that was developed to simulate the displacement of a fluid in a porous medium is Invasion Percolation (IP) [15]. The porous medium is represented by a lattice where each bond has an assigned (quenched) value for its conductance. The dynamics of the fluid is to invade the bond with highest conductance within all its perimeter bonds. This model leads spontaneously to a fractal structure that is essentially identical to the percolating cluster of standard percolation [16]. The IP model, characterised by an extremal statistics, has recently inspired simple SOC models aimed at the description of the propagation of

SELFSIMILARITY: PHYSICAL MODELS		
Ising-Type (70's)	**DLA/DBM (81)**	**Sandpile (87)**
Equilibrium Statistical Mechanics Ergodicity	NON LINEAR, IRREVERSIBLE DYNAMICAL EVOLUTION. Assigning the statistical weight of a structure requires the knowledge of its complete growth history.	
Boltzmann Weight Standard Critical behaviour Fine Tuning: $T = T_c$	CRITICAL BEHAVIOR IS SELF-ORGANIZED ATTRACTIVE FIXED POINT	
Repulsive Fixed Point $\xi = (T - T_c)^{-\nu}$ Approach to the critical point	Asymptotically frozen fractal structure Long range interactions (Laplacian)	Dynamical driven stationary state with avalanches of all sizes
$\Gamma(r) = \frac{1}{r^{(d-2+\eta)}}$	Complex continuum limit: Lattice regularization seems to be essential	
Anomalous dimension exactly at $T = T_c$	Problem: understand and compute the fractal dimension D	Problem: distribution of avalanche sizes $P(s) = s^{-\tau}$
Theory: Renormalization Group	Theory: NEW CONCEPTS ARE NEEDED	

Table 1: Comparison between the Ising model and two of the most popular models that generate fractal or scale invariant structures in a self-organized way.

irregular surfaces or interfaces in a disordered medium and of scale-free events in biological evolution. Extremal dynamics in a quenched medium is also the essential theoretical ingredient of the Bak and Sneppen model of biological evolution [17].

Another principal subject where fractals play an essential role is the study of interface growth in disordered systems e.g. Kardar Parisi Zhang (KPZ) equation [18]. If we consider the DBM model and eliminate the effect of the Laplace equation by setting $h = 0$, all the perimeter bonds have the same growth probability. This is the Eden model [14] that leads to compact structures with an irregular surface characterized by a critical exponent. These models of surface growth are meant to describe the deposition of particles, the propagation of chemical reaction or fire fronts, the interface between fluids or a fluid in a porous medium under appropriate conditions [19].

5 New theoretical concepts and self-organization

The physical models discussed in the previous sections illustrate a number of physical situations that can lead to the generation of fractal structures. Comparison with experimental data suggests that these models capture the essential physics of various phenomena that produce fractal structures in nature. Such models however do not constitute a physical theory, and this is the next step of our discussion.

From the theoretical point of view the idea of many authors is that DLA/DBM and the other SOC models pose questions of a new type for which it would be desirable to have a common theoretical scheme [20]. The attempts to use the theoretical concepts developed for critical phenomena like field theory and the RG have been quite problematic for these new phenomena. The basic differences with respect to equilibrium phase transition is that the dynamics is irreversible and self-organized. There is no ergodic principle and it is not possible to assign a Boltzmann weight to a configuration without knowing its entire growth history.

The theoretical effort in this field can be separated into phenomenological or scaling theories and microscopic theories. The first approach has been extensively developed in the past years and it consists in defining consistency relations between the assumed scaling properties of the system. This phenomenological approach is essential in the analysis of computer simulations to identify and extract the relevant essential information. The microscopic approach consists in a comprehensive understanding of all the SOC and fractal properties of the system directly from the knowledge of the microscopic dynamics. In some specific cases exact results can also be obtained. The development of a microscopic theory is an extremely difficoult task in which some interesting progress has been made but many fundamental questions are still open.

It was natural however to expect that some of the theoretical concepts developed for critical phenomena should also work for fractal models. A notable attempt in this direction was made by Kardar, Parisi and Zhang (KPZ) [18] who showed that the dynamics of the growing profile of the Eden model surface growth can be described by a stochastic differential equation for which field theory and RG methods can be succesfully applied. This approach corresponds to mapping the irreversible dynamics of the problem onto an equilibrium problem for the statistics of the profile. Various experiments of surface growth show however surface fluctuations with exponents that are higher than those predicted by the KPZ equation. This is probably due to quenched disorder that cannot be described in terms of an effective equilibrium problem [19].

This brings us to the crucial problem of fractal growth. We have seen that most fractal growth models like DLA, DBM, Cluter-Cluster aggregation, Invasion Percolation and the sandpile models are characterized by an intrinsically irreversible dynamics. As a result the statistical weight of a configuration can be defined only with the knowledge of its entire history. In other words the temporal evolution is just as important as the spatial correlations, which is not at all the case in equilibrium phase transitions. In the latter, the ergodic principle allows one to eliminate the temporal dynamics and assign a statistical weight for each configuration in terms of the Boltzmann factor. Another important difference is that most fractal structures are *self-organized*. For these and other more technical reasons like the absence of an upper critical dimension in some of these models the usual methods of field theory and RG did not lead very far for this class of models.

One attempt of constructing a physical theory for the self-organized fractals with irreversible dynamics is the Fixed Scale Transformation (FST) [5]. This approach combines a technique of lattice path integrals to take into account the irreversible dynamics with the study of the scale invariant dynamics inspired by the RG theory. This combination allows us to compute the pair correlations induced by the irreversible dynamics between block variables of arbitrary size. In this way it is possible to understand the origin of self-organization in fractal growth in terms of an attractive fixed point for the scale invariant dynamics and to compute analytically the fractal dimension. At the moment the FST framework seems to be the only general approach for the broad class of self-organized fractals and related phenomena. This method has been succesfully applied to DLA/DBM, to Cluster-Cluster aggregation, to fracture models, to Invasion Percolation and related models [5] and finally also to the sandpile models [21]. This situation supports therefore the conjecture [20] that DLA and the sandpile models pose questions of a new type for which it would be desirable to define a common theoretical scheme.

There are several other approaches that address similar issues for specific problems, e.g. the work of Nagatani and others [22] and of Halsey [23] for DLA, the elegant algebraic methods of Dhar et al [24], the field theory approaches of Kardar et al [25] and of Bak and coworkers [26] for certain properties of SOC models and the Run Time Statistics [27] to deal with problems with quenched disorder like IP and the Bak and Sneppen model.

6 Open problems and further developments

As we have mentioned there has been some relevant progress on the theoretical side with the introduction of new ideas and methods. However, many important questions remain open. The objective would be to develop these ideas into a general and systematic theoretical framework with microscopic predictive power in relation to fractal growth and SOC properties. It would also be important to clarity the relations between these new models and usual critical phenomena especially in relation to the properties of self-organization and the concept of universality. For example a crucial issue is the role of universality in fractal and SOC phenomena. In usual critical phenomena the same exponents that define the onset of magnetisation also describe the liquid vapour transition in water. This strong universality appears to be a characteristic of equilibrium systems. Self-organized systems, on the other hand, do not seem to exhibit the same degree of universality as the fractal dimension can be easily altered by relatively simple changes in the growth process. This reduced universality is sometimes viewed as a negative element because one is forced to describe specific systems instead of a single universal model. The truth is probably the opposite. Some theoretical concepts can be considered as general or universal, but the inherent diversity of the various models that have been studied adds another fascinating dimension in the intellectual search. After all, the SOC fractal structures we observe in nature are quite various and different from each other. The preliminary knowledge we have at the moment suggests that there are some universal principles but the specific properties depend on the specific process. It is possible that this has to do with the fact that the domain of irreversible phenomena is much broader than that of equilibrium statistics. The definition of the classes and laws for this broader area is certainly one of the main tasks of the theoretical effort in this field.

References

[1] Evertsz, C.J.G., Peitgen, H.O. and Voss, R.F. Eds., (1996). *Fractal Geometry and Analysis,* (World Scientific, Singapore).

[2] Amit D., (1978). *Field theory, the Renormalization Group and Critical Phenomena* (Mc Graw-Hill, New York).

[3] Mandelbrot B., (1982). *The Fractal Geometry of Nature,* Freeman, New York.

[4] Wilson K.G., (1974). *Phys. Rep.* **12**, 75.

[5] Erzan A, Pietronero L., Vespignani A., (1995). *Rev. Mod. Phys.* 67, 554.

[6] Coleman, P.H. and Pietronero, L., (1992). *Phys. Rep.* **231**, 311.

[7] Pietronero, L., Montuori, M. and Sylos Labini, F., in *Critical Dialogues in Cosmology,* Princeton (June 1996). Ed. by Turok, N. et al., (World Scientific, Singapore).

[8] Sylos Labini, F., Montuori, M. and Pietronero, L., (1997). *Scaling in Galaxy Clustering, Physics Reports,* in print.

[9] Peebles, P.E.J., (1980). *The Large Scale Structure of The Universe* (Princeton Univ.Press.); Peebles, P.E.J. (1993). *Principles of physical Cosmology* (Princeton Univ. Press.).

[10] Weinberg, S. E., (1972). *Gravitation and Cosmology,* Wiley, New York.

[11] Witten, T.A. and Sander, L.M., (1981). *Phys. Rev. Lett.* **47**, 1400.

[12] Meakin, P., (1983). *Phys, Rev. Lett.* **51**, 1119; Kolb, M., Botet, R. and Jullien, R. (1983). *Phys. Rev. Lett.* **51**, 1123.

[13] Weitz, D., (1984). *Phys. Rev. Lett.* **52**, 1433.

[14] Niemeyer, L., Pietronero, L. and Wiesmann, H.J., (1984). *Phys. Rev. Lett.* **52**, 1033.

[15] Vicsek, T., (1992). *Fractal Growth Phenomena*, (World Scientific, Singapore).

[16] Bak, P., Tang, C. and Wiesenfeld, K., (1987). *Phys. Rev. Lett.* **59**, 381.

[17] Bak, P. and Sneppen, K., (1993). *Phys. Rev. Lett.* **71**, 4083.

[18] Kardar, M., Parisi, G. and Zhang, Y.C., (1986). *Phys. Rev. Lett.* **56**, 889.

[19] Family, F. and Vicsek, T., Eds., (1991). *Dynamics of Fractal Surfaces* (World Scientific, Singapore).

[20] Kadanoff, L.P., (1990). *Physica* A **163**, 1. See also *Physics Today*, (March 1991), p. 9.

[21] Pietronero, L., Vespignani, A. and Zapperi, S., (1994). *Phys. Rev. Lett.* **72**, 1690.

[22] Nagatani, T., (1987). *Phys. Rev.* A **36**, 5812; Wang, X. R., Shapir, Y. and Rubenstein, M., (1989). *Phys. Rev.* A **39**, 5974.

[23] Halsey, T.C., (1994). *Phys. Rev. Lett.* **72**, 1228.

[24] Dhar, D., (1990). *Phys. Rev. Lett.* **64**, 1613.

[25] Hwa, T. and Kardar, M., (1989). *Physica* D **38**, 198.

[26] Paczuski, M., Maslov, S. and Bak, P., (1994) *Europhys. Lett.* **27**, 97.

[27] Marsili, M., (1994) *Europhys. Lett.* **28**, 385.

INSTITUT
D'ESTUDIS
CATALANS