Università di Napoli Federico II - Dipartimento di Scienze Fisiche
Dottorato di Ricerca in Fisica Fondamentale e Applicata - XX Ciclo
Anno Accademico 2007-2008

The Automorphic Universe

Coordinatore

Prof. G. Miele

Supervisore

prof. A. Sciarrino

Candidato

Luca Antonio Forte

On the cover: *Gravity*, M. C. Escher, Lithograph and watercolor (1952)
Property of the M. C. Escher Company B. V. - http://www.mcescher.com/

a mammà
a papà
a peppe

Contents

Introduction

Do I Dare disturb the universe?

The Love Song of J. Alfred Prufrock
T. S Eliot

Why the automorphic universe

The *automorphic universe* could be called in alternative ways the arithmetic universe or the chaotic universe [130]. We will try to explain both attributes in the following chapters. The word automorphic goes to add to an already established property of the universe, i.e. its chaotic behavior. The goal of this work is to explain how automorphic properties and chaotic ones are intimately related. We hope that the adjective automorphic with which we like describing our universe helps in unrevealing it.

Plan of the thesis

This work is divided in two main parts.

The first part deals mostly with mathematical aspects. The first chapter describes classical and quantum dynamical systems, in particular geodesic flows on the hyperbolic plane, the Selberg trace formula, and other topics like quantum chaos and quantum unique ergodicity. This chapter should be read in parallel with Appendices A and B. The second chapter contains an overview of Kac-Moody algebras and some results I derived about the hyperbolic Kac-Moody algebra $\mathrm{HA}_1^{(1)}$ and the primitive periodic orbits inside the fundamental domain of its Weyl group.

The second part deals mostly with applications to physics, in particular we review the known fact the dynamics of Einstein equations close to the cosmological singularity shows a chaotic behavior which can be studied in many similar ways (we describe some of them). We focus on the billiard representation for this dynamics, give its classical properties and carry on a quantum analysis using general arguments valid for quantum billiards. The result is that the wave function of the early universe is a certain automorphic L-function, specifically a Maass cusp form for the modular group. Some speculations are given together with the Conclusions.

Precise statements are given in each chapter; finally, some comments and a hopefully helpful bibliography are included at the end of each chapter. The last appendix explains the picture on the front cover.

Disclaimer: All the statements which sound like “this is new” implicitly contain the expression “modulo ever-present ignorance”.

Part of this work has been done during my visits at ULB under the supervisions of F. Englert, M. Henneaux, L. Houart, and at IHES under the supervision of T. Damour. I wish to express my gratitude to all of them and especially to my supervisor prof. A. Sciarrino for giving me the freedom to study many different topics, the freedom to think and to be wrong.

Part I

Mathematical Structures

Chapter 1

Chaotic Dynamical Systems

> One is struck by the complexity of this figure that I am not even attempting to draw. Nothing can give us a better idea of the complexity of the three-body problem and of all the problems of dynamics in general
> . . .
>
> *Collected Works*
> H. Poincaré

In this chapter we review standard facts about chaotic dynamical systems, focusing on the ones which have the highest degree of chaos (Anosov flows), especially geodesic and billiard flows. We deal also with the quantum version of them.

1.1 Ergodicity, Mixing, Hyperbolicity and All that

Here we briefly review basic notions of *chaos theory*, an important part of physics which for many years was just a prerogative of mathematicians under the less fancy name of *ergodic theory* (the branch of mathematics which studies transformations which preserve some measures). This theory strongly uses concepts from measure theory and probability theory. For more details

and bibliography see the last section of this chapter.
Let $(X, \mathcal{B})$ a measurable space. A transformation $T : X \to X$ is said to be measurable if $T^{-1}(B) \in \mathcal{B}$ for every $B \in \mathcal{B}$. A transformation $T : X \to X$ is called an *automorphism* if it is a bijection and both T, T^{-1} are measurable. Positive iterations $\{T^n\}$, $n \geq 0$, of a measurable transformation T make a semi-group.; all iterations $\{T^n\}$, $n \in \mathbb{Z}$, of an automorphism make a group. For any point $x \in X$ the sequence $T^n x$ is called the *trajectory* or the *orbit* of x. Measurable transformations with continuous time (*flows*) will be described later. Given a measurable space $(X, \mathcal{B})$, let us denote by $\mathcal{M}(X)$ the set of all probabilities (that is normalized to 1) measures on it. It is a convex set, as for any $\mu, \nu \in \mathcal{M}(X)$ and $0 < p < 1$ we have $p\mu + (1-p)\nu \in \mathcal{M}(X)$. A measurable transformation T induces a map (which we still denote by T) $T : \mathcal{M}(X) \to \mathcal{M}(X)$ defined by $(T\mu)(B) = \mu(T^{-1}B)$ (sometimes it is denoted by T_*). We say that T preserves a measure μ or that μ is $T-$*invariant* if $T\mu = \mu$. If, in addition, T is an automorphism, then $T\mu = \mu$ is equivalent to $T^{-1}\mu = \mu$, so that T and T^{-1} preserve the same measures (eventually more than one). Let us also denote by $\mathcal{M}_T(X)$ the set of all $T-$invariant probability measures; it is still a convex subset of $\mathcal{M}(X)$ [1].

[1] Let us only enunciate some properties of the set $\mathcal{M}(X)$. If μ_1 and μ_2 belong to $\mathcal{M}(X)$, then

- μ_1 is *absolutely continuous* with respect to μ_2 (and we write $\mu_1 << \mu_2$) if $\forall B \in \mathcal{B}$, $\mu_2(B) = 0 \Rightarrow \mu_1(B) = 0$

- μ_1 and μ_2 are *equivalent* ($\mu_1 \sim \mu_2$) if $\mu_1 << \mu_2$ and $\mu_2 << \mu_1$, that is if they have the same sets of zero measure

- μ_1 and μ_2 are *singular* ($\mu_1 \perp \mu_2$) if $\exists B \in \mathcal{B}$ such that $\mu_1(B) = 0$ and $\mu_2(B) = 1$

If $\mu_1 << \mu_2$, then the Radon-Nikodym theorem says that there exist $f \in L^1(X, \mathcal{B}, \mu_2)$ such that $\mu_1(B) = \int_B f d\mu_2 \; \forall B \in \mathcal{B}$. In this case one writes $f = \frac{d\mu_1}{d\mu_2}$.

Note also that $\mathcal{M}(X)$ is not empty, in fact it always contains the Dirac measure

$$\delta_x(B) = \begin{cases} 1 & x \in B \\ 0 & x \notin B \end{cases} \tag{1.1}$$

concentrated on a single point. If X is a compact topological space, we can define more structures for $\mathcal{M}(X)$. First, we define the *support* of $\mu \in \mathcal{M}(X)$ (denoted supp(μ)) to be the smallest closed set C with $\mu(C) = 1$.

For example, let us consider the unit interval with $\mathcal{B}$ the usual Borel $\sigma-$algebra. Let μ_1 the usual Lebesgue measure and $\mu_2 = \delta_0$, the Dirac measure concentrated on the point $x = 0$. These two measures are singular, in fact for $B = (0, 1]$ $\mu_1(B) = 1$, $\mu_2(B) = 0$. Moreover, supp$(\mu_1) = [0, 1]$ and supp$(\mu_2) = \{0\}$.

A measure μ is $T-$invariant iff for any measurable function $f : X \to \mathbb{R}$ we have

$$\int_X f \circ T \, d\mu = \int_X f \, d\mu \tag{1.2}$$

that is if one integral exists, so does the other and they are equal. More generally, for any $\mu \in \mathcal{M}(X)$ its image $\mu_1 = T\mu$ is characterized by

$$\int_X f \circ T \, d\mu = \int_X f \, d\mu_1 \tag{1.3}$$

A measurable transformation $T : X \to X$ induces a linear map U_T on the space of measurable functions $f : X \to \mathbb{R}$ defined by

$$(U_T f)(x) = (f \circ T)(x) = f(T(x)) \tag{1.4}$$

For any $T-$invariant measure μ and $p > 0$ the map $U_T : L^p(X, \mu) \to L^p(X, \mu)$ preserves the norm $|| \cdot ||_p$, and in the case $p = 2$ it preserves also the scalar product in $L^2(X, \mu)$. If T is an automorphism, then U_T is a bijection, thus a unitary operator on $L^2(X, \mu)$ (Koopman operator).

In the following T will always denote a measurable transformation T (sometimes an automorphism) preserving a measure $\mu \in \mathcal{M}(X)$. The quadruple $(X, \mathcal{B}, T, \mu)$ is called a *measure-preserving transformation* or a (time-discrete) dynamical system.

The fist result in chaos theory is perhaps *Poincaré's recurrence theorem* [2]. Let T preserve a measure $\mu \in \mathcal{M}_T(X)$ and $\mu(A) > 0$ for some measurable set $A \subset X$. Then for $\mu-$almost any point $x \in A$ we have

$$T^{n_i}(x) \in A \text{ for some sequence } n_1 < n_2 < \cdots \tag{1.5}$$

In this situation, the map

$$T_A(x) := T^{n_A(x)}(x)\,, \quad n_A(x) = \min\{n \geq 1 : T^n(x) \in A\} \tag{1.6}$$

We can define a topology on $\mathcal{M}(X)$, called the *weak* topology*, by $\mu_n \to \mu$ as $n \to \infty$ $\Leftrightarrow \int F d\mu_n \to \int F d\mu$ as $n \to \infty$, for some test functions, for example $\forall\, F \in C^0(X)$. With this topology, $\mathcal{M}(X)$ is a compact topological space. The weak* topology will be used in the section dedicated to the quantum unique ergodicity problem.

[2] The sentence at the beginning of this chapter alludes to Poincaré's theorem about eternal returns, and it expresses how complicated the evolution of a dynamical system may be.

is defined a.e. on A and is called the *Poincaré return map.* It preserves the conditional measure μ_A on A defined by $\mu_A(B) = \mu(A \cap B)/\mu(A)$.

A measurable set B is $T-$invariant if $T^{-1}B = B$; if in addition T preserves a measure μ, then a measurable set B is said to be $T - invariant$ (mod 0) if $T^{-1}B = B$ (mod 0). In this case, there exist a $T - invariant$ set $\widetilde{B}$ such that $\widetilde{B} = B$ (mod 0). A function $f : X \to \mathbb{R}$ is $T-$invariant if $U_T f = f$, i.e. $f \circ T = T$. In this case, f is constant on every trajectory of the map T. Again, if T preserves a measure μ, then we say that f is $T-$invariant (mod 0) if $f(x) = f(T(x))$ for $\mu-$a.e. point $x \in X$. Then there exist a $T-$invariant function $\widetilde{f}$ such that $\widetilde{f} = f$ (mod 0).

Let us come now the most important examples of measures. A measure μ is called *ergodic* if it is $T-$invariant ($\mu \in \mathcal{M}_T(X)$) and if for any $T-$invariant set $B \subset X$ we have $\mu(B) = 0$ or $\mu(B) = 1$. Equivalently, for any $T-$invariant (mod 0) set $B \subset X$ we have $\mu(B) = 0$ or $\mu(B) = 1$. A $T-$invariant measure μ is ergodic iff any $T - invariant$ function $f : X \to \mathbb{R}$ is a.e. constant, i.e. $\mu(x : f(c) = c) = 1$ for some $c \in \mathbb{R}$. Equivalently, μ is ergodic iff any $T-invariant$ (mod 0) function f is a.e. constant, i.e. $\mu(x : f(c) = c) = 1$ for some $c \in \mathbb{R}$. We usually say that T is ergodic if it is clear from the context which invariant measure is associated with T. A $T-$invariant measure is ergodic iff it is an extremal point in the convex set $\mathcal{M}_T(X)$. Any two distinct ergodic measures μ, ν are mutually singular (orthogonal). If a measurable transformation T has a unique invariant measure μ, this will be automatically ergodic. Then T is said to be *uniquely ergodic.*

Let us now introduce the notion of *isomorphism* between dynamical systems. Two measure-preserving transformations $(X_1, \mathcal{B}_1, T_1, \mu_1)$ and $(X_2, \mathcal{B}_2, T_2, \mu_2)$ are said to be isomorphic if for each $i = 1, 2$ there is a T_i-invariant set $B_i \subset X_i$ of full μ_i measure and a bijection $\phi : B_1 \to B_2$ such that (a) ϕ preserves measurable sets and measures, and (b) ϕ preserves the dynamics, i.e. $\phi \circ T_1 = T_2 \circ \phi$ on B_1. The map ϕ is called an isomorphism. As usual one does not distinguish between isomorphic dynamical systems, and we will describe many important properties invariant under isomorphism. For example, μ_1 is ergodic iff μ_2 is ergodic.

Let us now come to the first important result in ergodic theory, which dates back to Birkhoff, about the equality between spatial and time averages in certain cases of physical interest. Given a measurable function $f : X \to \mathbb{R}$, we can think of it as an observable (physical) quantity. For every $x \in X$, the sequence $\{f(T^n x)\}$ of values of f on the trajectory of x plays an important role, it is the value of f at time n. then $\{f(T^n x)\}$ can be regarded as a time series. Its partial sums

$$S_n(x) = f(x) + f(TX) + f(T^2 x) + \cdots + f(T^{n-1}x) \tag{1.7}$$

are called ergodic sums and the limit

$$f_+(x) = \lim_{n\to\infty} \frac{1}{n} S_n(x) \tag{1.8}$$

if it exists, is called the forward time average of the observable f along the orbit of x. If T is an automorphism, one can define also the backward time average

$$f_-(x) = \lim_{n\to\infty} \frac{1}{n} S_{-n}(x) \tag{1.9}$$

where $S_{-n}(x) = f(x) + f(T^{-1}x) + f(T^{-2}x) + \cdots + f(T^{-n+1}x)$. We have now the ingredients to state *Birkhoff ergodic theorem.* Let $(X, \mathcal{B}, T, \mu)$ be a measure-preserving transformation and $f \in L^1(X, \mu)$. Then

- for almost any point $x \in X$ the limit f_+ exists
- the function $f_+(x)$ is $T-$invariant, more precisely: if f_+ exists, then $f_+(T^n x)$ exists for all n and $f_+(T^n x) = f_+(x)$
- f_+ is integrable ($f_+(x) \in L^1(X, \mu)$) and $\int_X f_+ d\mu = \int_X f d\mu$
- if μ is ergodic, then $f_+(x)$ is a.e. constant and its value is $\int_X f d\mu$

If T is an automorphism, then the limit $f_-(x)$ exists as well and the two limits coincide a.e., $f_+(x) = f_-(x)$ (mod 0). The integral $\int_X f d\mu$ is the space average of the observable f. The last part of the theorem (that is when μ is ergodic) asserts *the time averages are equal to the space averages.* The theorem admits a generalization (*L^p Von Neumann ergodic theorem*): for every $p \geq 1$ and $f \in L^p(X, \mu)$ we have $||S_n/n - f_+||_p \to 0$ as $n \to \infty$. A first application of the ergodic theorem is the following. For any measurable set $A \subset X$ and $x \in X$, define the quantity

$$r_A(x) := \lim_{n\to\infty} \frac{\sharp\{0 \leq i \leq n-1 : \, T^i(x) \in A\}}{n} \tag{1.10}$$

called the asymptotic *frequency of visits* (returns) of the point x to the set A (when it exists). It immediately follows from the ergodic theorem that $r_A(x)$

exists for a.e. $x \in X$ (in this case, the function f is of the characteristic function of the set A, r_A). Moreover, $r_A(x) > 0$ for a.e. $x \in A$ by Poincaré recurrence theorem. If μ is ergodic, then $r_A(x) = \mu(A)$ for a.e. $x \in X$. Hence the orbit of a point $x \in X$ spends time in the set A proportional to its measure $\mu(A)$. In this sense, the ergodic measure μ describes the asymptotic distribution of almost every orbit $\{T^n x\}$, $n \geq 0$, in the space X.

Given two measurable sets A and B, ergodicity of a transformation T can be also reformulated as

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \mu(A \cap T^k B) = \mu(A)\mu(B) \tag{1.11}$$

which means that after a large number of applications of the mapping T moving B forwards in time, one approaches the statistical independence *on the average.* Ergodicity is not a very strong statistical property: it just indicates that a measurable set of a system is visited by a trajectory with a frequency proportional to its measure. Ergodic systems do not need to have sensitive dependence on initial conditions.

Let us now state a stronger property than ergodicity. We say that a measure-preserving transformation $T : X \to X$ is *mixing* or *strongly mixing* if for all pairs of measurable sets $A, B \subset X$

$$\lim_{n\to\infty} \mu(T^{-n}A \cap B) = \mu(A)\,\mu(B) \tag{1.12}$$

i.e. if the events $T^{-n}A$ and B become asymptotically independent as $n \to \infty$. Note that $x \in T^{-n}A$ is equivalent to $T^n(x) \in A$, i.e. we are speaking about the events $x \in B$ (characterizing x at time 0) and $T^n(x) \in A$ (characterizing the image of x at time n). Thus mixing is commonly interpreted as asymptotic independence of the distant future from the present, but without averaging. Mixing says that if we fix B and let A evolve in time, then A will spread out and mix through the entire phase space , eventually intersecting the fixed set B. As the mixing become more thorough, any part of B will locally resemble the whole space and memory of the initial conditions will eventually be lost. Mixing is also equivalent to

$$\lim_{n\to\infty} \langle f \cdot (g \circ T^n) \rangle = \langle f \rangle \langle g \rangle \quad \forall f, g \in L^2(X, \mu) \tag{1.13}$$

where $\langle f \rangle = \int_X f d\mu$. Given two observable functions f and g, the quantity

$$\mathbf{C}_{f,g}(n) = \langle f \cdot (g \circ T^n) \rangle - \langle f \rangle \langle g \rangle \tag{1.14}$$

is called the *correlation* between f and g at time n (it the covariance of the random variables f and $g \circ T^n$). Mixing is equivalent to the convergence of correlations to zero, $\mathbf{C}_{f,g}(n) \to 0$, a property called the decay of correlations. A map T is *weak mixing* (with respect to an invariant measure μ) if for all pairs of measurable sets $A, B \subset X$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left|\mu(T^{-i}A) \cap B - \mu(A)\mu(B)\right| = 0 \tag{1.15}$$

In terms of correlations, this is equivalent to

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} |\mathbf{C}_{f,g}(n)| = 0 \tag{1.16}$$

There is also another notion of mixing (multiple mixing) which we do not really need here. Mixing properties are invariant under isomorphism. It is clear that strong mixing implies weak mixing (but not viceversa), weak mixing implies ergodicity (but not viceversa). A graphic illustration of the properties due originally to Gibbs envisages a fluid mixture of 10% rum and 90% cola (gin and martini in the book by Arnold and Avez [4], but I prefer rum and cola). If now one considers the proportion of rum in any fluid volume, then an *ergodic cocktail* ensures this proportion is 10% on the time average. A *weakly mixed cocktail* ensures that this proportion is eventually 10% except on occasional, infrequent moments, while a *strongly mixed cocktail* has the property that after some time the proportion of rum is always 10%. Mixing systems tend to an equilibrium as time goes to ∞.

Let us give now some examples of measure-preserving transformations. Perhaps the most popular one is the a *circle rotation*. Let $X = \mathbb{R}/\mathbb{Z}$ be the unit 1-torus, or a circle of length one, with a cyclic angular coordinate $x \in [0, 1]$ with the points 0 and 1 identified. The rotation through an angle α is defined by

$$T(x) = x + \alpha \pmod 1 \tag{1.17}$$

It preserves the standard Lebesgue measure $\mathbf{m}$ on X. If $\alpha = p/q$ is rational, then every point $x \in X$ is periodic with the same period q. If α is irrational, then the trajectory of any x is dense and uniformly distributed in X, i.e. for any $A \subset X$ $r_A(x) = \mathbf{m}(A)$. In this case the Lebesgue measure is ergodic but not mixing (not even weak mixing). Finally, $\mathbf{m}$ is the only invariant measure for T, hence T is uniquely ergodic. The higher-dimensional generalization

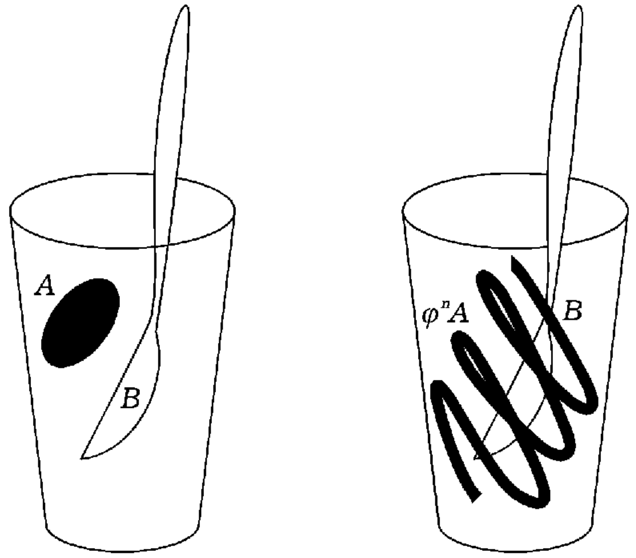


Figure 1.1: A mixed cocktail! Figure from [4], $\varphi^n A$ is what we call $T^n A$.

of the circle rotation is the *linear translations of tori.* Let $X = \mathbb{R}^d/\mathbb{Z}^d$ be the unit $d-$torus with angular coordinates $\mathbf{x} = (x_1, \ldots, x_d) \in [0,1]^d$ $(d \geq 2)$. The translation of X along a fixed vector $\mathbf{a} = (a_1, \ldots, a_d) \in \mathbb{R}^d$ is defined by

$$T_{\mathbf{a}}(\mathbf{x}) = \mathbf{x} + \mathbf{a} \pmod{\mathbf{1}} \tag{1.18}$$

The translation $T_{\mathbf{a}}$ is ergodic iff the components $(a_1, \ldots, a_d)$ of the vector $\mathbf{a}$ are rationally independent, i.e.

$$m_0 + m_1 a_1 + \cdots + m_d a_d \neq 0 \tag{1.19}$$

for any integers $m_0, m_1, \ldots, m_d \in \mathbb{Z}$ unless $m_0 = m_1 = \cdots = 0$. The map $T_{\mathbf{a}}$ is never weakly mixing. So in dimensions $d \geq 2$, the translations of the tori are less chaotic than the circle rotations in $d = 1$.

A still more powerful random property is the Bernoulli shift. This describes systems which are completely random. Roughly, their phase space can be partitioned into n sections each labelled by some k_i and having a probability p_i of rising during the evolution. If the system evolves at discrete intervals of time, then the dynamics are coded by a a random sequence of k_i. The simplest example would be a tossing coin with two possible outcomes

k_1, k_2 and $p_1 = p_2 = 0.5$. Let us formalize this concept and define an important object for our work, the so-called *symbolic space.* Let $S = \{1, \ldots, r\}$ a finite alphabet whit r letters. Let $\Sigma_+ = \Sigma_{+,r} = S^{\mathbb{Z}_+}$ denote the space of infinite sequences of letters; a point $\underline{\omega} \in \Sigma_+$ is a sequence $\underline{\omega} = \{\omega_n\}_{n=0}^{\infty}$ with $\omega_n \in S$ for each $n \geq 0$. Define also $\Sigma = \Sigma_r = S^{\mathbb{Z}}$, the space of double infinite sequences of letters, i.e. Σ consists of sequences $\underline{\omega} = \{\omega_n\}_{n=-\infty}^{+\infty}$ with $\omega_n \in S$ for any $n \in \mathbb{Z}$. The spaces Σ_+ and Σ are examples of symbolic spaces (the suffix r to remind the cardinality of our alphabet S is suppressed for brevity). We equip the set S with the discrete topology, where each subset of S is open, and the spaces Σ_+ and Σ with the product topology. The corresponding Borel $\sigma-$algebras are denoted by $\mathcal{B}_+$ and $\mathcal{B}$. The (left) shift homeomorphism $\sigma : \Sigma \longrightarrow \Sigma$ is defined by $\underline{\omega}' = \sigma(\underline{\omega})$ with $\omega'_i = \omega_{i+1}$. Similarly, the (left) shift $\sigma_+ : \Sigma_+ \longrightarrow \Sigma_+$ is defined by $\underline{\omega}' = \sigma_+(\underline{\omega})$ with $\omega'_i = \omega_{i+1}$ for all $i \geq 0$; it is a continuous $r-$to-1 map on Σ_+. Let us define the measures preserved by these transformations. Let μ_0 be a probability measure on the finite set S (different from a Dirac measure). Denote by μ_+ the corresponding product measure $\mu_0^{\mathbb{Z}_+}$ on Σ_+ and by μ the corresponding product measure $\mu_0^{\mathbb{Z}}$ on Σ. The measure space $(X, \mathcal{B}, \mu)$ corresponds to a sequence of independent identically distributed random variables each of which takes finitely many values, a classical object of study in probability theory. The shifts σ_+ and σ preserve respectively the measures μ_+ and μ. Both shifts are ergodic and mixing. The dynamical system $(X, \mathcal{B}, \sigma, \mu)$ is said to be a *Bernoulli shift.* This is completely characterized by the measure μ_0 on S, once we fix the number r of letters and the shift transformations.

Given an abstract measure-preserving transformation $(X, \mathcal{B}, T, \mu)$, one can associate to it a *symbolic representation.* Let $X = A_1 \cup \cdots \cup A_r$ a finite partition of X into disjoint measurable subsets. For every point $x \in X$, we can define its itinerary

$$\underline{\omega}(x) = \{\omega_n\}_{n=0}^{+\infty} \in \Sigma_+ : \quad T^n(x) \in A_{\omega_n} \ \forall n \geq 0 \tag{1.20}$$

$X = \bigcup_i A_i$ is said to be a generating partition if distinct points have distinct itineraries. Then the map $\phi : X \longrightarrow \Sigma_+$ defined by $\phi(x) = \underline{\omega}(x)$ is one-to-one; it induces a measure $\mu_X = \phi(\mu)$ on Σ_+ which is σ_+-invariant. The map ϕ is an isomorphism between the given system $(X, \mathcal{B}, T, \mu)$ and $(\Sigma_+, \mathcal{B}_+, \sigma_+, \mu_X)$, which is called the symbolic representation of the former. If T is an automorphism, then the itinerary of x is defined by

$$\underline{\omega}(x) = \{\omega_n\}_{n=-\infty}^{+\infty} \in \Sigma : T^n x \in A_{\omega_n} \ \forall n \in \mathbb{Z} \tag{1.21}$$

The same concept of generating partition applies, the map $\phi : X \to \Sigma$ defined by $\phi(x) = \underline{\omega}(x)$ is one-to-one and induces a measure $\mu_X = \phi(\mu)$ on Σ which is $\sigma-$invariant, and again we obtain an isomorphism between $(X, \mathcal{B}, T, \mu)$ and $(\Sigma, \mathcal{B}, \sigma, \mu_X)$. Finally, an automorphism $T : X \to X$ preserving a measure μ is said to be Bernoulli (or have Bernoulli property or B-property) if it is isomorphic to a Bernoulli shift. Equivalently, there is a finite generating partition $\xi = \{A_1, \ldots, A_r\}$ of X such that the corresponding symbolic representation of T is a Bernoulli shift, i.e. the induced measure μ_X on Σ is the product measure $\mu_0^{\mathbb{Z}}$ we defined above (this is the main point).

There exist also in literature the notion of *Kolmogorov automorphism* (or K-mixing transformation): this is a stronger notion of mixing but weaker than Bernoulli property. In particular, Bernoulli property implies (K-property which implies) mixing. In some known systems (in particular billiard flows) it is also true that the K-property implies the B-property, but this is not true in general. The K-property is invariant under isomorphisms.

To describe a system, it is useful to introduce some numerical quantities which characterize its chaotic behavior. The first important concept is *entropy*. Given a measure-preserving transformation $(X, \mathcal{B}, T, \mu)$, the entropy of a finite partition $\xi = \{A_1, \ldots, A_r\}$ of X is given by

$$H(\xi) = -\sum_{i=1}^{r} \mu(A_i) \ln \mu(A_i) \tag{1.22}$$

with the convention $0 \ln 0 = 0$. We have $0 \leq H(\xi) \leq \ln r$, with the minimum 0 attained on the trivial partition and the maximum $\ln r$ attained on equipartitions, which are characterized by $\mu(A_1) = \ldots = \mu(A_r) = 1/r$. Since the measure μ is $T-$invariant, the partition $T^{-n}\xi = \{T^{-n}A_1, \ldots, T^{-n}A_r\}$ has the same entropy, $H(T^{-n}\xi) = H(\xi)$ for every $n \geq 1$. It T is an automorphism, this is true for all $n \in \mathbb{Z}$. The entropy of T with respect to a finite partition ξ

$$h(T, \xi) = \lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) \tag{1.23}$$

this limit always exists and is non-negative, indeed the sequence on the right hand side decreases monotonically. Finally, the *metric entropy* of T is

$$h(T) = \sup_{\xi} h(T, \xi) \tag{1.24}$$

where the supremum is taken over all finite partitions ξ of X. Note that this entropy (also known as the *Kolmogorov-Sinai entropy*) has nothing to do with the dynamical entropy (the macroscopic one), which evolves in time; given the dynamical system, $h(T)$ is a fixed number, its range is $0 \leq h(T) \leq \infty$. The metric entropy is invariant under isomorphisms, i.e. two isomorphic dynamical systems have the same metric entropy (the converse is not true in general, but it is true for Bernoulli shifts where the K-S entropy is a complete invariant). We have $h(T^n) = nh(T)$ for any $n \geq 1$. If T is an automorphism, then $h(T^n) = |n|h(T)$ for every $n \in \mathbb{Z}$, in particular $h(T^{-1}) = h(T)$. An automorphism is $K-$mixing iff its entropy is positive, $h(T, \xi) > 0$ for any nontrivial finite partition ξ.

Let us now come to *dynamical systems with continuous time.* The single transformation T (which counts the discrete time) is replaced by a one-parameter group $\{S^t\}$, where each S^t is a measure-preserving transformation. Given a measurable space $(X, \mathcal{B})$, a dynamical system with continuous time or a *flow* is a one-parameter family $\{S^t\}_{t\in\mathbb{R}}$ of measurable transformations $S^t : X \to X$ that satisfies two condition: (a) $S^{t+s} = S^t \circ S^s$ (group property), S^0 is the identity, (b) the map $X \times \mathbb{R} \to X$ defined by $(x, t) \to S^t x$ is measurable. For every point $x \in X$ the set $\{S^t x\}$, $t \in \mathbb{R}$, is called the orbit of x. In most of applications, X is a topological space and $\{S^t x\}$ is a continuous curve for every $x \in X$. The flow preserves a measure $\mu \in \mathcal{M}(X)$ if $\mu(S^t(A)) = \mu(A)$ for all measurable subsets $A \subset X$ and all $t \in \mathbb{R}$. In other words, μ is a common invariant measure for all the automorphisms S^t included in the flow.

The previous properties of automorphisms extend to flows with some trivial modifications. A measurable set $B \subset X$ is invariant under a flow $\{S^t\}$ if $B = S^t B$ for every $t \in \mathbb{R}$. If the flow $\{S^t\}$ preserves a measure μ, then a measurable set B is said to be invariant (mod 0)under the flow if $B = S^t B$ (mod 0) for every $t \in \mathbb{R}$. If B is invariant (mod 0), then there exists an invariant set $\widetilde{B}$ such that $\widetilde{B} = B$ (mod 0). A function $f : X \to \mathbb{R}$ is invariant under $\{S^t\}$ if $f = f \circ S^t$ for all $t \in \mathbb{R}$. In this case, f is constant on every orbit of the flow $\{S^t\}$. If $\{S^t\}$ preserves a measure μ, then we say that a function $f : X \to \mathbb{R}$ is invariant (mod 0) under the flow if for every $t \in \mathbb{R}$ we have $f(x) = f(S^t x)$ for $\mu-$a.e. point $x \in X$. In that case there exists an invariant function $\widetilde{f}$ such that $\widetilde{f} = f$ (mod 0).

A flow $\{S^t\}$ is ergodic with respect to an invariant measure μ if any $\{S^t\}-$invariant (mod 0) set $A \subset X$ has measure 0 or 1. Equivalently, a flow $\{S^t\}$ is ergodic if any invariant (mod 0) function f is a.e. constant, i.e. $\mu(x : f(x) = c) = 1$ for some $c \in \mathbb{R}$. It turns out that if at least one automorphism S^t in the flow is ergodic, then the whole flow $\{S^t\}$ is ergodic. Conversely, if the flow is ergodic, then the automorphism S^t is ergodic for all but countably many $t \in \mathbb{R}$.

Let us give the version of Birkhoff ergodic theorem for flows. Given a measurable function $f : X \to \mathbb{R}$, its (forward and backward) time averages are defined by

$$f_{\pm}(x) = \lim_{T\to\pm\infty} \frac{1}{T} \int_0^T f(S^t(x))\, dt \tag{1.25}$$

Suppose as before that the flow preserves a measure μ and that $f \in L^1(X, \mu)$. Then

- for almost every point $x \in X$ the above limits exist and $f_+(x) = f_-(x)$
- the function $f_{\pm}(x)$ is $T-$invariant, more precisely if $f_{\pm}(x)$ exists, then $f_{\pm}(S^t x)$ exists for all $t \in \mathbb{R}$ and $f_{\pm}(S^t x) = f_{\pm}(x)$
- $f_{\pm}$ is integrable and $\int_X f_{\pm} d\mu = \int_X f d\mu$
- if $\{S^t\}$ is ergodic, then $f_{\pm}(x)$ is a.e. constant and its value is $\int_X f d\mu$

A flow $S^t : X \to X$ is mixing with respect to an invariant measure μ if for any $A, B \subset X$ we have

$$\lim_{t\to\pm\infty} \mu(A \cap S^t(B)) = \mu(A)\mu(B) \tag{1.26}$$

If a flow $\{S^t\}$ is mixing, then every map S^t,$t \neq 0$, is also mixing. The flow is a K-flow iff any map S^t, $t \neq 0$, is a K-automorphism. Finally, the flow is Bernoulli (B-flow) if at least one automorphism S^t, $t \neq 0$, is Bernoulli (in this case any other S^t is Bernoulli too). As in the discrete case, Bernoulli property implies K-property, which implies mixing which implies ergodicity (all of them are one-way implications).

The metric entropy of the map S^t of any flow $\{S^t\}$ is a linear function of time: $h(S^t) = |t|h(S^1)$. Thus *the entropy of the flow* is defined by $h(\{S^t\}) = h(S^1)$.

For an integrable Hamiltonian system $\{S^t\}$, $h(\{S^t\}) = 0$; but the converse is not true, i. e. a dynamical system with zero entropy is not necessarily integrable.

Generally, a dynamical system with positive metric entropy is chaotic, in the sense that nearby trajectories in phase space diverge at an exponential rate, contrary to what happens in integrable systems where the separation is a power of time.

1.1.1 The Gauss map

With the term *continued fraction* we mean the “infinite” fraction

$$a_0 + \cfrac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \frac{1}{a_4 + \cdots}}}} \tag{1.27}$$

where the a_i are positive integers (a_0 is allowed to be 0). Such a fraction is also denoted by $[a_0; a_1, a_2, a_3, \ldots]$. For the finite fraction we write $[a_0; a_1, a_2, \ldots, a_n]$, that is

$$a_0 + \cfrac{1}{a_1 + \frac{1}{a_2 + \cdots + \frac{1}{a_{n-1} + \frac{1}{a_n}}}} \tag{1.28}$$

Thus, for example

$$[a_0; a_1, a_2, \ldots, a_n] = a_0 + \frac{1}{[a_1; a_2, a_3, \ldots, a_n]} \tag{1.29}$$

A continued fraction is not just a formal object, in fact it converges to a real number. Namely,

$$u = [a_0; a_1, a_2, \ldots] = \lim_{n\to\infty} [a_0; a_1, \ldots, a_n] \tag{1.30}$$

$$= \lim_{n\to\infty} \frac{p_n}{q_n} = a_0 + \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{q_{n-1} q_n}$$

is absolutely convergent, because $q_0 = 1$, $q_1 = a_1$, $q_2 \geq 2$ and generally $p^k \geq 2^{(k-2)/2}$, $q^k \geq 2^{(k-2)/2}$ since $a_n \geq 1$ for all n. By construction, we have

$$[a_0; a_1, a_2, \ldots] = a_0 + \frac{1}{[a_1; a_2, \ldots]} \tag{1.31}$$

We say that $[a_0; a_1, \ldots,]$ is the *continued fraction expansion* for u, and u is irrational. Conversely, for any irrational number u, this expansion always exists and is unique (see below). The rational numbers

$$\frac{p_n}{q_n} = [a_0; a_1, \ldots, a_n] \tag{1.32}$$

with coprime numerator and denominator, are called the convergents of the continued fraction for u and provide very rapid rational approximations for u. A continued fraction in which some of the digits are allowed to be zero (but that is not allowed to end with infinitely many zeros) can always be rewritten with digits in $\mathbb{N}$.

Let X be the set of irrational numbers in the unit interval, $X = [0, 1]\backslash\mathbb{Q}$, and define a map $T : X \to X$ by

$$T(x) = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor \tag{1.33}$$

where $\lfloor t \rfloor$ denotes the greatest integer less than or equal to t. In other words, $T(x)$ is the fractional part $\left\{\frac{1}{x}\right\}$ of $\frac{1}{x}$. This map is called the *Gauss map* (see the picture for the graph).

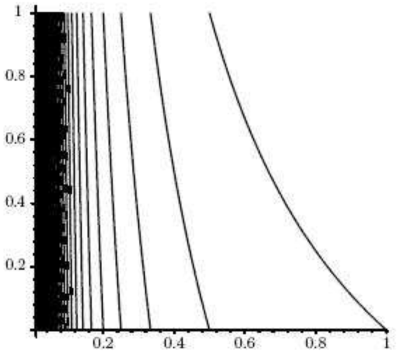


Figure 1.2: The Gauss map

Gauss observed that T preserves a probability measure (Gauss measure) on $[0,1]$ given by

$$\mu(A) = \frac{1}{\ln 2} \int_A \frac{1}{1+x}\, dx \tag{1.34}$$

for any measurable set $A \subseteq [0,1]$. The connection with continued fractions is the following. Fixed any $x \in X$ and $n \geq 1$, define the sequence of natural numbers $\{a_n\} = \{a_n(x)\}$ by

$$\frac{1}{1+a_n} < T^{n-1}(x) < \frac{1}{a_n} \tag{1.35}$$

or equivalently

$$a_n(x) = \left\lfloor \frac{1}{T^{n-1}x} \right\rfloor \in \mathbb{N} \tag{1.36}$$

Then for any irrational x in $[0,1]$, the sequence $\{a_n(x)\}$ gives the continued fraction expansion of x, i.e.

$$x = [a_0(x); a_1(x), a_2(x), \ldots] \tag{1.37}$$

It turns out that the Gauss measure is equivalent to the usual Lebesgue measure $\mathbf{m}$ on the unit interval (i.e. they have the same sets of zero measure), and moreover T is ergodic respect to μ. The Gauss map belongs to the class of so-called *expanding transformations* of the interval $[0,1]$, that is transformations $T : x \to Tx = f(x)$ with $|f'(x)| > 1$ on any interval between two discontinuities. For such expanding transformations, one can show that the metric entropy is given by

$$h(T) = \int_0^1 \ln |f'(x)|\, \rho(x)\, dx \tag{1.38}$$

where $\rho(x)$ is the density of the invariant measure, the one preserved by T (in our case, the Gauss measure). The Gauss map has a countable number of discontinuities (which form a set of zero measure), but the above formula is still valid, so its metric entropy is

$$h(T_{Gauss}) = \frac{2}{\ln 2} \int_0^1 \frac{|\ln x|}{1+x} = \frac{\pi^2}{6 \ln 2} \tag{1.39}$$

The Gauss map is isomorphic to a Bernoulli shift with the same metric entropy, and in general expanding transformations possess the property of exponential instability which leads to the appearance of strong stochastic properties.

A famous example is given by the golden ratio and its inverse

$$\begin{aligned} \frac{1+\sqrt{5}}{2} &= [1;1,1,1,1,\ldots] \\ \frac{-1+\sqrt{5}}{2} &= [0;1,1,1,1,\ldots] \end{aligned} \tag{1.40}$$

Note that these continued fraction expansions are periodic. This is a indeed a theorem. In fact it is possible to prove [98] that any irrational quadratic number (i.e. any number which satisfies a quadratic equation with integral coefficients) is represented by a periodic continued fraction and viceversa. A continued fraction

$$[a_0; a_1, a_2, \ldots] \tag{1.41}$$

is *periodic* if there exist positive integers k_0 and h such that for arbitrary $k \geq k_0$

$$a_{k+h} = a_k \tag{1.42}$$

We sill see that the fixed points of hyperbolic transformations on the hyperbolic plane lie on the real axis and are irrational quadratic. The continued fraction expansion for these points gives a code (see below) for all hyperbolic matrices in $\mathrm{SL}(2,\mathbb{Z})$ and we will use this fact to code the imaginary root system of the hyperbolic Kac-Moody algebra $\mathrm{HA}_1^{(1)}$.

1.1.2 Geodesic Flows and Billiards

Mathematical billiards describe the motion of a mass point in a domain with elastic reflections from the boundary. The theory of billiards is not a single one, but it is a mathematician's playground where various methods and approaches are tested. Indeed, very simple dynamical problems can be reduced to the investigation of billiards in polygons or polyhedrons. Following [144] consider the mechanical system of two point-masses m_1 and m_2 of coordinates x_1 and x_2 on the positive half-line $x \geq 0$. The collisions between the two masses and with the rigid wall at $x = 0$ are elastic. Then this mechanical system is isomorphic to the the billiard in the angle $\arctan\sqrt{m_1/m_2}$. Similarly, the configuration space of two (or more) points moving inside a segment is a simplex, and collisions between the particles and/or the two hard walls correspond to geometric reflections from the the boundary of this simplex according to the law "the angle on incidence equals the angle of reflections". It is clear that the theory of billiards has many relations with the geometrical optics too. We will show in this thesis that billiards appear in *general relativity* in a particular regime of the gravitational theory described by Einstein equations[3]. In order to define rigorously billiard games, we need first the notion of *geodesic flows*.

Let Q be a smooth compact $d-$dimensional Riemannian manifold. For each point $q \in Q$, we can define the tangent space T_qQ and the cotangent space T_q^*Q. The main object is the *unit tangent bundle* M on Q, $M = SQ = \{(q,v)|q \in Q, v \in T_qQ, ||v|| = 1\}$. If Q is a compact smooth manifold with piecewise smooth boundary, then M is also a manifold with the boundary $\partial M = \pi^{-1}(\partial Q)$ and dim $M = 2d - 1$. The geodesic flow on Q is a group $\{T^t\}$ of transformations of M such that a specific transformation T^t consists in moving an element (q, v) of M along the geodesic line which it determines by a distance t. If $d\sigma(q)$ is the element of the Riemannian volume and ω_q is the Lebesgue measure on the unit sphere S^{d-1} in T_qQ, the measure μ on M given by $d\mu = d\sigma(q)d\omega_q$ is invariant under $\{T^t\}$.

Geodesic flows belong to the class of the so-called *Hamiltonian dynamical systems*. In fact, an alternative way to introduce the geodesic flow is the following. The tangent bundle $TQ = \{(q,v)|q \in Q, v \in T_qQ\}$ can be naturally identified with the cotangent bundle $T^*Q = \{(q,p)|q \in Q, p \in T_q^*Q\}$. Each point $p \in T_q^*Q$ is uniquely determined by its components $(p_1, \dots, p_m)$. The non-degenerate canonical 2-form $\omega = \sum_{i=1}^d dq^i \wedge dp_i$ induces the symplectic structure on T^*Q and the geodesic flow $\{T^t\}$ which we have just introduced is naturally isomorphic to the restriction to the unit tangent bundle of the Hamiltonian dynamical system with Hamiltonian $H(p,q) = \frac{1}{2}\,||p||^2$.

Important properties of the geodesic flow on negatively curved Riemannian manifolds Q will be stated after we introduce the hyperbolic plane.

Generalizations of geodesic flows are *billiard flows*. Suppose Q is a closed $d-$dimensional manifold of class C^∞ and Q_0 is a subset given by the systems of inequalities of the form $f_i(q) \geq 0$, $q \in Q$, $f_i \in C^\infty(Q)$, $1 \leq i \leq r$. The

[3] Note that the relation between general relativity and geometric optics is well known, thus the relation between Einstein's theory and billiards is perhaps not completely surprising. The big question would be to understand if one can reformulate the *full* theory as a billiard problem in *any* regime, with different billiard tables of course depending on the specific symmetries (remember that Einstein's theory is theory with constraints). More will be said in the following.

phase space of the billiard in Q is the set M whose points are the pairs $x = (q, v)$, $q \in Int\, Q$, $v \in S^{d-1}$, as well as those $x = (q, v)$ for which $x \in \partial Q$, $v \in S^{d-1}$ and v is directed inside Q. The motion of a point $x = (q, v)$ under the billiard flow is the motion with unit speed along the trajectory of the geodesic flow until the boundary ∂Q is reached. At such moments, the point reflects from the boundary according to the "incidence angle equals reflection angle" rule and then continues its motion. As before, the measure $d\mu = d\sigma(q) d\omega_q$ is invariant under $\{T^t\}$. Thus, a billiard in a region Q_0 can also be defined as the Hamiltonian system with a potential $V(q) = 0$ inside Q_0 and $V(q) = \infty$ if $q \in \partial Q$.

If Q is not compact but of finite area, like in the case of the hyperbolic surfaces $\Gamma(N)\backslash\mathbb{H}$ (see below), one can define the geodesic flow in a similar way, but the structure of the fiber bundle is violated in a certain number of points.

Let $Q \subset \mathbb{R}^d$ a convex polyhedron, that is a closed bounded set $Q = \{q \in \mathbb{R}^d : f_i(q) \geq 0, i = 1, \ldots, r\}$ where the functions $f_i(q)$ are linear. The boundary of the billiard is the union of the faces $\Gamma_i, i = 1, \ldots, r$. Denote by n_i the unit vector orthogonal to each face Γ_i, directed inside Q. The trajectories of billiards in domains contained in Euclidean space are broken lines (segments). Let us consider the isometric mapping $\sigma_i : S^{d-1} \to S^{d-1}$ acting on every point $x = (q, v)$, $q \in \Gamma_i$, according to the mirror reflection

$$\sigma_i(v) = v - 2\,(n_i, v)\,n_i \tag{1.43}$$

where $(,)$ is the standard Euclidean scalar product and $(n_i, n_i) = 1$. We assume that there are trajectories in Q which have vertices in the faces with numbers $i_1, i_2, \ldots$. Then by means of successive reflections in the faces of Q, we can obtain a straight line instead of the broken one (*unfolding* of a billiard trajectory). The straight line intersects with the polyhedrons $Q, Q_{i_1}, Q_{i_1 i_2}, \ldots$, where $Q_{i_1 \cdots i_k}$ is the result of successive reflections of Q, relative to the faces $\Gamma_{i_1}, \ldots, \Gamma_{i_k}$, where Γ_{i_l} is a face of $Q_{i_1 \cdots i_{l-1}}$. Given a point $x_0 = (q_0, v_0)$, the vector $v_0 \in S^{d-1}$ defines the initial velocity of the billiard trajectory originating from the point $q_0 \in Q$. The velocity vector becomes $v_k = (\sigma_{i_k} \sigma_{i_{k-1}} \cdots \sigma_{i_1}) v_0$ between the $k-$th and the $(k+1)-$th reflections. Let us now consider the group G_Q generated by the reflections $\sigma_i, \ldots, \sigma_r$; it is a subgroup of all isometries of S^{d-1}. The ergodicity of the billiard depends on the group G_Q, precisely: if G_Q is a *finite* group, then the billiard flow inside Q is *not* ergodic. For $d = 2$, the finiteness of the group is equivalent to the commensurability of all angles on the polygon Q.

The situation for generic polyhedra is still open, in particular one knows that the entropy of a billiard inside an arbitrary, *not* necessarily convex, polyhedron is zero

$$h(\text{ inside a polyhedron }) = 0 \tag{1.44}$$

One can think that the every trajectory of a billiard in a convex polygon must be periodic or everywhere dense, but a result due to G. A. Galperin shows that this is not always the case: in his example there is a trajectory which is everywhere dense in some proper sub-domain of Q.

Let us give some more examples in dimension 2. Let $\Omega \subset \mathbb{R}^2$ be a compact domain with boundary $\partial\Omega$. If one imagines hard walls at the boundary $\partial\Omega$, we obtain a planar billiard. The trajectories of the particle consist of segments of straight lines with elastic reflections at $\partial\Omega$. The Hamiltonian of such a planar billiard is not smooth, but rather discontinuous

$$H(\mathbf{p},\mathbf{q}) = \begin{cases} \mathbf{p}^2/2m & \mathbf{q} \in \Omega \\ 0 & \mathbf{q} \notin \Omega \end{cases} \tag{1.45}$$

It turns out that the billiard dynamics depends very sensitively on the shape of the boundary $\partial\Omega$. In fact, if the boundary is a *circle*, an *ellipse* or a *square*, the system is integrable while a boundary of the shape of a stadium leads to a strongly chaotic system, the well known *Bunimovich billiard.* Note that after the introduction of the Sinai billiard, it was believed that convex billiards were too focusing to be chaotic, contrary to the dispersive behavior

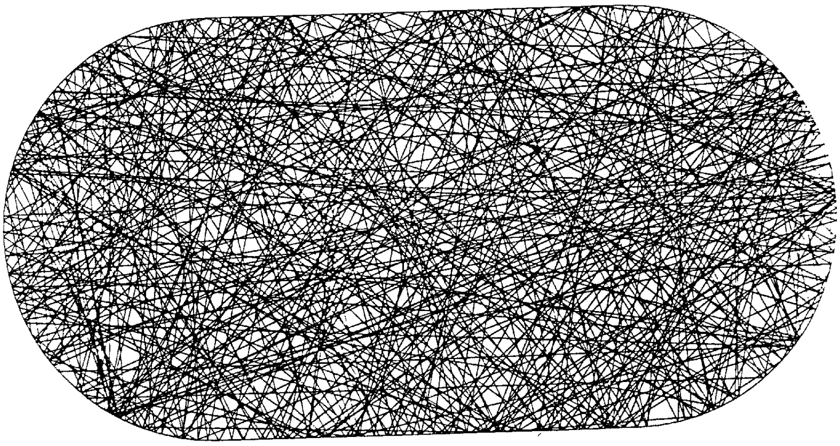


Figure 1.3: The Bunimovich stadium

of Sinai-like billiards (see the figure at the end of this chapter). But the Bunimovich stadium is an example of focusing billiard which is also chaotic; if the two horizontal lines collapse to points, the stadium becomes a circle, and we have a transition from a chaotic to an integrable billiard. Note also the boundary of the stadium is not smooth (we mean C^∞). Usually, one assumes that the each arc in the boundary of a billiard is of class C^3, i.e. the curvature is continuously differentiable. This a technical assumption which ensures that there are no trajectories having an infinite number of collisions in a finite time interval.

For *smooth* strictly convex domains, there is an important result due to V. F. Lazutkin. First remember that by *caustic* for a billiard Ω, we mean a smooth closed curve $\gamma \subset \Omega$ such that if one link of the billiard trajectory is tangent to γ, then every other link of this trajectory is also tangent to γ. For a circle, there is a unique family of caustics, namely the concentric circles. For ellipses, one has two different families, confocal ellipses and hyperbolas. It is not known if only ellipses have this property. Lazutkin proved that there exist an uncountable set of caustics, of positive measure in Ω, if the boundary $\partial\Omega$ is convex and sufficiently smooth. A billiard inside a sufficiently smooth convex figure is not ergodic. Lazutkin also constructed quasi-eigenfunctions (quasi-modes) and quasi-eigenvalues for the Dirichlet problem in Ω. The support of any such eigenfunction is localized in a neighborhood of one of the invariant sets of a billiard defined by caustics. For ergodic flows and billiards, the situation is different, see the last sections of this chapter for the quantum ergodicity theorem.

1.2 Quantum chaology, not quantum chaos

The most striking property of deterministic chaos is the sensitive dependence on initial conditions such that neighboring trajectories in phase space separate at an exponential rate. As a result, the long-time behavior of a strongly chaotic system is *unpredictable.* There arises the basic question whether this well established phenomenon of classical chaos manifests itself in the quantum world in an analogous phenomenon which could be called *quantum chaos.* By this we mean the following: given a classical dynamical system which is strongly chaotic, is there any manifestation in the corresponding quantum system which betrays its chaotic behavior? The first place where one should seek for a possible chaotic behavior in quantum mechanics seems to be the long-time behavior in analogy to the classical case. It turns out, however, that the large-time limit in quantum mechanics is well under control due to the fundamental fact tat the time-evolution operator $e^{-i\widehat{H}t/\hbar}$ is unitary and thus its spectrum lies on the unit circle. This is in contrast to classical systems whose time-evolution is ruled by the Liouville operator. If the classical system is mixing and chaotic, the spectrum of the Liouville operator has a continuous part on the unit circle and thus the time-evolution in unpredictable for large times. This fundamental difference is the main reason of the absence of chaos in quantum mechanics (together with the linearity of the Schrödinger equation), in the sense of exponential sensitivity to initial conditions. The study of semiclassical, not classical, limit of systems which exhibit classical chaos has been called *quantum chaology* by M. Berry [21]. Semiclassical means as Planck's constant $\hbar$ tends to zero. This limit is non trivial because quantum mechanics, considered as depending on a complex parameter $\hbar$, is essentially singular at the origin $\hbar = 0$, in ways that differ from system to system. Because of the essential singularity at $\hbar = 0$, the semiclassical limit of quantum mechanics (and also the geometrical-optics limit of electromagnetism) is complicated and conceals a rich variety of phenomena. Quantum theory is a non-perturbative extension of classical mechanics, unlike, say, special relativity, which grows out of Newtonian mechanics by a convergent perturbation expansion in velocity v/c.

Let us give some more remarks about quantum chaos. Chaos is problematic because the way a quantum wave develops in time is determined by the associated energy levels. A mathematical consequence of the existence of energy levels is that quantum time development contains only periodic motions with definite frequencies - the opposite of chaos. Therefore there is no chaos in quantum mechanics, only regularity. How then, can there be chaos in the world? There are two answers. One is that as the semiclassical limit is approached - as objects get bigger and heavier - the time taken for chaos to be suppressed by quantum mechanics gets ever longer, and would be infinite in the strict limit. However, this explanation fails because the chaos suppression time is often surprisingly short: just a few decades even for Hyperion (a satellite of the planet Saturn), which has an erratic rotation.

The true reason for the prevalence of chaos is that large quantum systems are hard to isolate from their surroundings. Even the patter of photons from the Sun (whose re-emission gives the light by which we see Hyperion) destroys the delicate interference underlying the quantum regularity. This effect, of large quantum systems being dramatically sensitive to uncontrolled external influences, is called *decoherence.* In the semiclassical limit, the quantum suppression of chaos is itself suppressed by decoherence, allowing chaos to re-emerge as a familiar feature of the large scale world. Smaller quantum systems, such as atoms in strong magnetic fields, molecules vibrating strongly, or electrons confined in quantum dots with unsymmetrical boundaries, can be effectively isolated. Therefore decoherence is irrelevant and there is no quantum chaos, even though the corresponding classical systems are chaotic. Nevertheless, these quantum systems reflect classical chaos in several ways, whose systematic study is quantum chaology. With this premise, we also adopt the term quantum chaos as usual in the literature.

For integrable systems with N degrees of freedom, one has the so-called EBK quantization rules. Each orbit of the dynamical system lies on a $N-$dimensional sub-manifold which has the topology of a torus. In this case it is possible to introduce new coordinates, the so called action-angle variables $(\mathbf{I}, \mathbf{w})$, through a canonical transformation. The angles w_k vary from 0 to 2π and are interpreted as new coordinates, the actions I_k play the role of new conjugate momenta. If w_k runs from 0 to 2π, it defines a loop L_k in the original $(\mathbf{p}, \mathbf{q})$ phase space, where L_k is the $k-$th irreducible homotopy circuit of the torus. The I_k's are the new constants of motion. Then the EBK quantization condition reads

$$I_k = (n_k + \beta_k/4)\,\hbar \tag{1.46}$$

where the $n_k \geq 0$ are integer quantum numbers and the integers $\beta_k \geq 0$ are the Maslov indices (the motion takes place on a so-called Lagrangian manifold, and the Maslov index, which can be understood as the number of conjugate points of the Morse index of a trajectory, is determined by the topology of the Lagrangian manifold in phase space with respect to configuration space). These quantization rules are contained in a paper by A. Einstein in 1917 [46], without the integers β_k. In fact, it was in the fifties that the mathematician J. Keller rediscovered Einstein's paper (forgotten for almost 40 years) and found that the most general semiclassical quantization rules turned out to be exactly Einstein's torus quantization rules plus corrections coming from Maslov indices.

Contrary to what is commonly believed, in this paper Einstein did *not* consider ergodic systems ([62] contains an Italian translation of this important work).

Anyhow, for ergodic systems, the EBK quantization rules can not be applied, because there are no invariant tori in phase space. In fact, for a chaotic system, the phase space carries two mutually transverse foliations, each leave of dimension N. Every trajectory is the intersection of two manifolds, one from each foliation. The distance between two neighboring trajectories increases exponentially along the unstable manifold and decreases exponentially along the stable one (see below for the definition of an Anosov system). Thus, there remains the task to find a semiclassical quantization rule for generally chaotic systems. The formulas one can build in these cases are trace formulas which typically relate the level density of a quantum system to classically periodic orbits.

The first answer in this direction came from the work by M. Gutzwiller [67] with the introduction of the *Gutzwiller trace formula.* This is a formal formula, since it is divergent, we discuss it in the next section.

1.3 The Gutzwiller Trace Formula

The general framework is Feynman's formulation of quantum mechanics in terms of his sum over histories or path integrals. In the semiclassical limit when $\hbar$ tends to zero, it is well known the leading contribution to the path integral comes from the classical orbits. Taking the trace of the time-evolution operator, the contribution comes from those classical orbits which are *closed* in coordinate space. Gutzwiller made the important observation that the trace of the energy-dependent Green's function (which is the Fourier transform of the time-evolution operator) is given by formal sum over all classical orbits which are closed in phase space, i.e. all *periodic orbits.* The sum has only a formal meaning because there are infinitely many periodic orbits whose growth in number as a function of the period is exponential for chaotic systems (see Margulis asymptotics for Anosov systems below), and thus the sum is in general not even conditionally convergent for physical energies.

As an illustration of the semiclassical theory for chaotic systems, let us consider (Euclidean) planar billiards. For the quantum Hamiltonian $\widehat{H}$ we get $\widehat{H} = -(\hbar^2/2m)\,\nabla$ where $\nabla = \partial^2/\partial q_1^2 + \partial^2/\partial q_2^2$ is the Euclidean Laplacian. The hard walls at the billiard boundary $\partial\Omega$ are incorporated by demanding that the quantum wave functions $\psi_n(\mathbf{q})$ should vanish at $\partial\Omega$. Then the Schrödinger equation for the given quantum billiard is equivalent to the following eigenvalue problem of the Dirichlet Laplacian

$$\begin{aligned} -\frac{\hbar^2}{2m}\,\nabla\psi_n(\mathbf{q}) &= E_n\psi_n(\mathbf{q}) \quad \mathbf{q}\in\Omega \\ \psi_n(\mathbf{q}) &= 0 \quad \mathbf{q}\in\partial\Omega \\ \int_\Omega \psi_m(\mathbf{q})\,\psi_n(\mathbf{q})\,d^2q &= \delta_{mn} \end{aligned} \tag{1.47}$$

The following properties of this eigenvalue problem are standard: there exist a discrete spectrum corresponding to an infinite number of bound states whose energy levels $\{E_n\}$ are strictly positive, $0 < E_1 \leq E_2 \leq \ldots$, and $E_n \to \infty$. The eigenvalues scale in $\hbar, m, R$ in the form $E_n = -\frac{\hbar^2}{2mR^2}\,\epsilon_n$, where ϵ_n is dimensionless and independent of $\hbar, m, R$ (R is an arbitrary but fixed length scale). This implies that the semiclassical limit corresponds to the limit $E_n \to \infty$ and thus requires a study of the highly excited states, i.e. of the high energy behavior of the quantum billiard. Notice that the semiclassical limit is identical to the macroscopic limit $m \to \infty$ where the mass of the atomic bouncing ball is becoming so heavy that one is dealing with a macroscopic point particle.

The Dirichlet problem for compact domains is an old one. It described a vibrating membrane with clamped edges (Helmholtz). The cases in which one can solve exactly this problem correspond to the integrable billiards inside a rectangle, an equilateral triangle and a circle (and other domains corresponding to affine Weyl chambers, see the book by M. Berger [18]). The problem turns out to be highly non trivial in cases when the billiard table is chaotic; indeed, in these cases, no explicit formula is known for the energy levels or for the the wave functions.

Thus, let us assume that the billiard domain Ω has been chosen in such a way that the corresponding classical systems is strongly chaotic, i.e. with positive metric entropy. All periodic orbits are unstable and isolated. The periodic orbits are characterized by their primitive length spectrum $\{l_\gamma\}$ where l_γ denotes the Euclidean length of the primitive periodic orbit (ppo) γ. Multiple traversals of γ have lengths kl_γ, where $k = 1, 2, \ldots$ counts the number of repetitions of the ppo γ. Let $\mathbf{M}_\gamma$ be the monodromy matrix of the p.p.o. γ, where $|\ \mathrm{Tr}\ \mathbf{M}_\gamma| > 2$, since all orbits are (direct or inverse) hyperbolic (this implies that all Lyapunov exponents are strictly positive, see the book by Gutzwiller [67] for more details). Moreover, let us attach to each ppo γ a character $\chi_\gamma \in \{\pm 1\}$ depending on the Maslov index of γ. Then the

Gutzwiller trace formula for the trace of the resolvent of $\widehat{H}$ (i.e. the trace of the Green's function) reads

$$\text{Tr}\ (\widehat{H}-E)^{-1} = \sum_{n=1}^{\infty} \frac{1}{E-E_n} \sim \overline{g}(E) + g_{osc}(E) \quad (\hbar \to 0) \tag{1.48}$$

where $\overline{g}(E)$ denotes the so-called zero length contribution which comes from direct trajectories going from $\mathbf{q}''$ to $\mathbf{q}'$ whose length tends to zero if $\mathbf{q}'' \to \mathbf{q}'$. The contribution from the periodic orbits is given by the formal sum

$$g_{osc}(E) = \frac{i}{2\,\hbar\,\sqrt{E}} \sum_{\gamma} \sum_{k=1}^{\infty} \frac{l_\gamma\, \chi_\gamma^k\, e^{i\,k\,\sqrt{E}\,l_\gamma/\hbar}}{|2-\text{Tr}\ \mathbf{M}_\gamma^k|} \tag{1.49}$$

The first problem with this trace formula comes from the fact that the resolvent operator $(\widehat{H}-E)^{-1}$ is not of trace class. This follows directly from *Weyl's asymptotic formula* which reads for two-dimensional planar billiards with area A

$$\lim_{n\to\infty} \frac{E_n}{n} = \frac{4\pi}{A}\,\hbar^2 \tag{1.50}$$

Thus $E_n = O(n)$ for $n \to \infty$ and the sum over n in (1.48) diverges. In order to cure this problem, one could simply consider the trace of a regularized resolvent, for example the trace of $[(\widehat{H}-E)^{-1} - (\widehat{H}-E')^{-1}]$ where E' is an arbitrary but fixed subtraction point. The real problems with the original trace formula arise, however, from the sum over the periodic orbits. Due to the exponential increase

$$N(l) \sim \frac{e^{\tau\, l}}{\tau\, l} \quad l \to \infty \tag{1.51}$$

of the number $N(l)$ of ppo γ whose lengths l_γ are smaller or equal to l, the infinite sum over γ is in general divergent. Since the divergence problems are a consequence of the exponential law and thus of the existence of a *topological entropy* $\tau > 0$, they are not just of a formal mathematical nature but rather a direct signature of classical chaos in quantum mechanics. A positive entropy τ is the most important global property of a strongly chaotic system which expresses the fact that the information about the system is lost exponentially fast. We therefore see that the periodic-orbit expression has only a formal meaning. One can calculate corrections in $\hbar$ (as in the paper by P. Gaspard [58]) to the Gutzwiller trace formula. We do insist on this.

In fact, as noted by the same Gutzwiller, if one considers the free geodesic motion on constant negative curvature manifolds (which is a strongly chaotic motion being Bernoullian and Anosov, see below), then the Gutzwiller trace formula becomes *exact* and corresponds to the Selberg trace formula, which is absolutely convergent (if the curvature is negative but not constant the motion is still chaotic, but there is no analog of the Selberg trace formula). Note that there exist an improved version of the Gutzwiller trace formula due to F. Steiner et al, which is convergent; in fact, the test functions satisfy the same conditions as in the Selberg trace formula (see below). This general trace formula establishes a striking duality between the quantum energy spectrum $\{E_n\}$ and the length spectrum $\{l_\gamma\}$ of the classical periodic orbits. The class of test functions satisfying the conditions in order to make the trace formula convergent is rather large, thus the trace formula represents an infinite number of periodic-orbits sums rules. That is, an infinite number of semi-classical quantization rules, which, at the moment, provide the only substitute for quantum systems whose classical limit is strongly chaotic. This will be more transparent when we deal with the Selberg trace formula, but before we need some notions of hyperbolic geometry in 2 dimensions.

1.4 Hyperbolic Geometry and Fuchsian Groups

In this section we review hyperbolic geometry and Fuchsian groups. As it is well known, Euclid's fifth postulate was noticeably more complicated than the other axioms, looking more like a theorem than a self-evident proposition. For centuries, starting with Archimedes, mathematicians tried to prove it from the other axioms. Hyperbolic geometry was discovered by C F. Gauss, who never published his results because at the time it was not clear if non-Euclidean geometries were consistent. Finally, in 1868 the Italian mathematician E. Beltrami established its independence by finding models for the hyperbolic plane, proving the conjecture of Gauss, Boylai and Lobachevski as to the existence (i.e. internal consistency) of this non-Euclidean geometry. Today we know that in 2 and 3 dimensions, hyperbolic geometry is far more important than Euclidean geometry. We will see that Fuchsian groups are similar to lattices in $\mathbb{R}^n$ which are discrete groups of orientation-preserving Euclidean isometries. However, for $n = 2$, while the quotients of the latter are always compact surfaces homeomorphic to the torus, the quotient of Fuchsian group acting on the hyperbolic plane $\mathbb{H}$ may not be a torus. Indeed, all orientable surfaces (compact or not) other than the sphere, torus, plane or punctured plane, are quotients of Fuchsian groups acting on $\mathbb{H}$ without fixed points (in other words, for any integer $g > 1$, there exists a Fuchsian group Γ acting on $\mathbb{H}$ without fixed points such that $\Gamma \backslash \mathbb{H}$ has genus g). In 3 dimensions, the situation is much more complicated as shown by W. Thurston; his geometrization conjecture roughly states that 3-dimensional manifolds allow for 8 different geometric structures.

The reader can consult the books by S. Katok [91] and by J. Ratcliffe [127] for more details. We will mainly use the upper-half plane as a model for hyperbolic geometry, see the previous books for formulas on the Poincaré disk and [9].

The best way to introduce hyperbolic geometry is to think of it as the differential geometry on a Riemannian manifold. In particular, let us introduce the *upper-half plane* or *Poincaré plane*

$$\mathbb{H} = \{z = x + iy \in \mathbb{C} \,|\, \text{Im } z = y > 0\} \tag{1.52}$$

endowed with the metric

$$ds^2_{\mathbb{H}} = \frac{dx^2 + dy^2}{y^2} \tag{1.53}$$

which is conformally flat, $ds^2_{\mathbb{H}} = ds^2_{Eucl}/y^2$ (thus hyperbolic angles on $\mathbb{H}$ are the same as the Euclidean ones). More generally, one can consider

$$ds^2_{\mathbb{H}} = R^2 ds^2_{Eucl}/y^2 \tag{1.54}$$

which has Gaussian curvature $\mathcal{K} = -1/R^2$. We will always put $R = 1$, that is $\mathbb{H}$ is the unique connected, simple connected hyperbolic surface with negative constant Gaussian curvature $\mathcal{K} = -1$. It is a non-compact Riemannian manifold (of infinite volume) of dimension 2. But it is *also* a Riemann surface.

The *hyperbolic distance* between two points $z, w \in \mathbb{H}$ is defined by

$$\rho(z, w) = \inf l_H(\gamma) \tag{1.55}$$

where the infimum is taken over all γ joining z and w. $l_H(\gamma)$ is the hyperbolic length of the curve γ

$$l_H(\gamma) = \int_0^1 \frac{\sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2}}{y(t)} = \int_0^1 \frac{\left|\frac{dz}{dt}\right|}{y(t)} \tag{1.56}$$

A useful expression for the distance between two points is the following

$$\cosh \rho(z, w) = 1 + \frac{|z - w|^2}{2 \operatorname{Im} z \operatorname{Im} w} \tag{1.57}$$

The *geodesics* in $\mathbb{H}$ are semi-circles and straight (vertical) lines orthogonal to the real axis $\mathbb{R}$. Observe also that every *hyperbolic circle* $\{z \in \mathbb{H} | \rho(z, z_0) = r^2\}$ is a Euclidean circle (with different center of course) and viceversa. This implies the topology on $\mathbb{H}$ induced by the hyperbolic metric is the same as the topology induced by the Euclidean metric.

Let us consider the group of linear fractional transformations of $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ (the Riemann sphere) given by

$$g\, z = \frac{az + b}{cz + d} \quad a, b, c, d \in \mathbb{R}, ad - bc > 0 \tag{1.58}$$

and denote by $\mathrm{GL}^+(2, \mathbb{R})$ the group of 2×2 real matrices of positive determinant. A linear fractional transformation g determines the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}^+(2, \mathbb{R})$ up to a scalar because the matrices $\begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$ with $\alpha \neq 0$ give the identity transformation. Dividing by a scalar, we can always represent g by a matrix of determinant 1. We can thus identify the factor group $\mathrm{PSL}(2, \mathbb{R}) = \mathrm{SL}(2, \mathbb{R})/\{\pm \mathbf{1}\}$ with the linear fractional transformations. This is called the *Mobius group* [4] and it is isomorphic to the group of the positive isometries (i.e. the transformations of $\mathbb{H}$ which preserve the hyperbolic distance) of the hyperbolic plane

$$\mathrm{Isom}^+(\mathbb{H}) = \mathrm{PSL}(2, \mathbb{R}) \tag{1.59}$$

All these positive (i.e. orientation-preserving) isometries are analytical automorphisms of the upper-half plane. The negative isometries (which do not form a group of course) are generated by the *reflections* $z \to -\overline{z}$, which are orientation-reversing, not analytic maps

$$g \in \mathrm{Isom}^-(\mathbb{H}) = \langle z \to -\overline{z} \rangle \Leftrightarrow g\, z = \frac{a\overline{z} + b}{c\overline{z} + d}, \quad ad - bc = -1 \tag{1.60}$$

Thus we have the disjoint union

$$\mathrm{Isom}(\mathbb{H}) = \mathrm{Isom}^+(\mathbb{H}) \cup \mathrm{Isom}^-(\mathbb{H}) = \mathrm{PSL}(2, \mathbb{R}) \cup \langle z \to -\overline{z} \rangle \tag{1.61}$$

[4] In the following we do not usually distinguish between the matrices and the linear transformation that they define.

and $\mathrm{PSL}(2,\mathbb{R})$ is a subgroup of $\mathrm{Isom}(\mathbb{H})$ of index 2. Positive isometries are conformal, while negative ones are anti-conformal, i.e. they preserve the absolute values of angles but change the signs.

The Mobius transformations transform a circle into a circle subject to the convention that a straight line is a circle passing through ∞. Of course, the center of a circle may not be mapped onto the center, save for the $g = \pm \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix}$ which is a translation.

For a subset $A \subset \mathbb{H}$, we define by $\mu(A)$ the *hyperbolic area* of A

$$\mu(A) = \int_A \frac{dx\,dy}{y^2} \tag{1.62}$$

when the integral exists. It is clear that this notion of area (when it exists) is invariant under $\mathrm{PSL}(2,\mathbb{R})$, that is $\mu(gA) = \mu(A)$ for any $g \in \mathrm{PSL}(2,\mathbb{R})$. As in elementary Euclidean geometry, one can define *hyperbolic n-sided polygons*, which are closed subsets of $\mathbb{H} \cup \mathbb{R} \cup \{\infty\}$ bounded by hyperbolic geodesic segments. A *vertex* is a point where two sides meet; we allow vertices on $\widehat{\mathbb{R}}$, but no segment of the real axis can belong to a hyperbolic polygon. The simplest polygons are the hyperbolic triangles, whose area is given through the Gauss-Bonnet theorem only in terms of the angles

$$\mu(triangle) = \pi - \alpha - \beta - \gamma \tag{1.63}$$

thus in hyperbolic geometry the angles of a triangle sum up to a quantity less than π (greater than π in spherical geometry). For a polygon with n sides and n angles θ_i

$$\mu(n - gon) = (n-2)\pi - \sum_{i=1}^{n} \theta_i \tag{1.64}$$

This formula also shows that in hyperbolic geometry rectangles do ont exist. In all the previous expressions an overall factor R^2 is implicit, if one considers general hyperbolic metrics as described above. Finally, given three numbers α, β, γ whose sum is less than π, then there exist a unique (up to isometries) hyperbolic triangle with angles α, β, γ.

The linear fractional transformations are rigid motions of the hyperbolic plane and they move points in distinct ways. Given $g \in \mathrm{PSL}(2,\mathbb{R})$ we denote its conjugacy classes by

$$\{g\} = \{h\,g\,h^{-1}\,|\,h \in \mathrm{PSL}(2,\mathbb{R})\} \tag{1.65}$$

Conjugate motions act on $\mathbb{H}$ similarly, so the classification will be invariant under conjugation. The identity motion forms a class by itself, since every $z \in \mathbb{H}$ is a fixed point. Any other motion $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ has one or two fixed points in $\widehat{\mathbb{C}}$. Three cases are possible

1. g has one fixed point on $\widehat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$
2. g has two fixed points on $\widehat{\mathbb{R}}$
3. g has one fixed points in $\mathbb{H}$ and the complex conjugate one in $\overline{\mathbb{H}} = \{z \in \mathbb{C} | \text{Im } z < 0\}$

Accordingly g is called *parabolic*, *hyperbolic* or *elliptic*. By conjugating, we can bring g to one of the following types

1. $z \to z + t$ (translation, fixed point ∞)
2. $z \to pz$ (dilation, fixed points $0, \infty$)
3. $z \to k(\theta)z$ (rotation, fixed point i)

where $t \in \mathbb{R}$, $p \in \mathbb{R}^+$ and $k(\theta)$ is the usual rotation matrix $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ The number of fixed points of a rigid motion is invariant under conjugation, therefore the above classification applies naturally to the conjugacy classes. The same classification can be also described in terms of the trace (which is an algebraic invariant under conjugation), namely if $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \neq \pm\mathbf{1}$, then we can classify the positive isometries in the following way

1. g is parabolic iff $|a + d| = 2$
2. g is hyperbolic iff $|a + d| > 2$
3. g is elliptic iff $|a + d| < 2$

A parabolic motion moves points along *horocycles* (circles in $\mathbb{H}$ tangent to $\widehat{\mathbb{R}}$). An elliptic motion moves points along circles centered at its fixed point in $\mathbb{H}$. The geodesic in $\mathbb{H}$ joining the two fixed points of a hyperbolic transformation g is called the *axis* of g (*hypercycles*); such a geodesic is globally invariant under the action of g, but not pointwise (except of course for the two fixed points of g). A hyperbolic motion moves points along its axis. Of the two fixed points u, w, one, say u, is repelling, the other, w, is attracting, $g'(u) > 1$ and $g'(w) < 1$. These two fixed points are the roots of the equation

$$cz^2 - (d - a)z - b = 0 \quad \text{(hyperbolic fixed points)} \tag{1.66}$$

For Fuchsian groups (which we define below), a, b, c, d are integer, so the two fixed points are irrational quadratic. We will see that the invariant axis of a hyperbolic g belonging to a Fuchsian group Γ, oriented from the repelling to the attracting point, becomes a closed geodesic in the quotient space $\Gamma\backslash\mathbb{H}$. Moreover, if g_1 and g_2 are conjugate in Γ, i.e. $g_1 = gg_2g^{-1}$ for some $g \in \Gamma$, then g maps the axis of g_2 to the axis of g_1, hence they represent the same oriented closed geodesic in $\Gamma\backslash\mathbb{H}$. Conversely, every oriented closed geodesic in $\Gamma\backslash\mathbb{H}$ represents the conjugacy class of a primitive hyperbolic transformation in Γ.

As concerns negative isometries, they are the product of a positive isometry with a pure symmetry across a geodesic. The latter is a *hyperbolic reflection* in a geodesic γ, that is a negative isometry which fixes pointwise γ (unlike a positive hyperbolic transformation, which fixes its axis globally). Every hyperbolic reflection R has order 2, $R^2 = I$. In order to classify negative isometries A, it is convenient to consider the square of these A^2. Each matrix cancels its characteristic polynomial

$$A^2 - (\text{Tr } A)A + (\text{det A})I = 0 \tag{1.67}$$

A^2 is a positive isometry, whose trace will be Tr $A^2 = (\text{Tr } A)^2 + 2$. First, A^2 can never be elliptic. If Tr $A \neq 0$, then A^2 is hyperbolic and A corresponds to the product of a positive hyperbolic isometry with a hyperbolic reflection. If Tr $A = 0$, then $A^2 = I$, which means that A is a pure hyperbolic reflection in a geodesic pointwise fixed by the action of A. This classifies negative isometries.

Given a Mobius transformation $T(z) = \frac{az+b}{cz+d}$, we can define a norm $|| \cdot ||$ on PSL$(2, \mathbb{R})$

$$||T|| = (a^2 + b^2 + c^2 + d^2)^{1/2} \tag{1.68}$$

which makes PSL$(2, \mathbb{R})$ into a topological group with respect to the metric $||T - S||$. The full group of Isom$(\mathbb{H})$ is topologized similarly. A subgroup Γ of Isom$(\mathbb{H})$ is called *discrete* if the induced topology on Γ is a discrete topology, that is if Γ is a discrete set in the topological space $\mathrm{Isom}(\mathbb{H})$. Γ is discrete iff for any sequence $\{T_n\}$ of elements of Γ converging to the identity I, then $T_n = I$ for n sufficiently large.

Let us now come the main object of our applications. A *Fuchsian group* is a discrete subgroup of $\mathrm{PSL}(2,\mathbb{R})$. Discrete subgroups of Lie groups are are sometimes called *lattices* by analogy with lattices in $\mathbb{R}^n$ which are discrete groups of isometries of $\mathbb{R}^n$. The latter have the following important property: their action on $\mathbb{R}^n$ is discontinuous in the sense that every point of $\mathbb{R}^n$ has a neighborhood which is carried outside itself by all elements of the lattice except for the identity. In general, discrete groups of isometries do not have such discontinuous behavior, for if some elements have fixed points these points cannot have such a neighborhood. However, they satisfy a slightly weaker discontinuity condition called a *properly discontinuously* action (for the precise definition see [91]). It turns out that a subgroup Γ of $\mathrm{PSL}(2,\mathbb{R})$ is a Fuchsian group iff its action on $\mathbb{H}$ is properly discontinuous.

Remember the notion of fundamental region. If X is a metric space and G a group of homeomorphisms acting properly discontinuously on X, a closed region $\mathcal{F} \subset X$ is said to be a *fundamental region* (or domain) for the action of G on X if

- $\bigcup_{T\in G} T(\mathcal{F}) = X$
- $\mathrm{Int}\ \mathcal{F} \cap T(\mathrm{Int}\ \mathcal{F}) = \emptyset$

The set $\partial\mathcal{F} = \mathcal{F} - \mathrm{Int}\ \mathcal{F}$ is called the boundary of the fundamental region. The family $\{T(\mathcal{F}) | T \in G\}$ is called the *tessellation* (or tiling) of X under the action of X. Each copy $T(\mathcal{F})$ of the fundamental region is a *tile*.

When the area of a fundamental region is finite, then it is a numerical invariant of the group, that is $\mu(\mathcal{F}_1) = \mu(\mathcal{F}_2)$ for two fundamental regions $\mathcal{F}_1$ and $\mathcal{F}_2$, when $\mu(\mathcal{F}_i)$ exists. Note also that a fundamental region is not uniquely determined by the group. A Fuchsian group with a fundamental region of infinite area is the group generated by $z \longrightarrow z+1$ (each vertical strip of length 1 is a fundamental region). Indeed, one usually classifies Fuchsian groups according to the properties of their fundamental regions. A Fuchsian group is said to be *of the first kind* or *co-finite* if a fundamental region has finite hyperbolic area, and *of the second kind* if a fundamental region has infinite hyperbolic area. Moreover, the fundamental region of a Fuchsian group of the first kind can be compact (*co-compact group*) or not compact. The compactness of the fundamental region implies that the corresponding

Fuchsian group does not contain parabolic elements. Moreover, a co-compact group is said to be a *strictly hyperbolic Fuchsian group* if it contains only hyperbolic elements. These are the ones which behave in the best possible way (see below).

The following theorem is useful in applications. Let Γ be a discrete subgroup of $\mathrm{Isom}(\mathbb{H})$ (thus Γ can be a Fuchsian group or also contain negative isometries) and Λ be a subgroup of Γ of index n. If

$$\Gamma = \Lambda T_1 \cup \Lambda T_2 \cup \cdots \cup \Lambda T_n \tag{1.69}$$

is a coset decomposition of Γ into $\Lambda-$cosets and if $\mathcal{F}$ is a fundamental region for Γ, then

- $\mathcal{F}_\Lambda = T_1(\mathcal{F}) \cup T_2(\mathcal{F}) \cup \cdots \cup T_n(\mathcal{F})$ is a fundamental region for Λ
- if $\mu(\mathcal{F})$ is finite and $\mu(\partial\mathcal{F}) = 0$, then $\mu(\mathcal{F}_\Lambda) = n\,\mu(\mathcal{F})$

We show now that each Fuchsian group Γ possesses a nice (connected and convex) fundamental region. Let us define the *Dirichlet region* for Γ centered at p

$$D_p(\Gamma) = \{z \in \mathbb{H} |\, \rho(z,p) \leq \rho(z, T(p))\,\forall\, T \in \Gamma\} \tag{1.70}$$

where p is not fixed by any element of $\Gamma - \{I\}$ (such elements always exist). Since ρ is invariant under $\mathrm{PSL}(2,\mathbb{R})$, we can also write

$$D_p(\Gamma) = \{z \in \mathbb{H} |\, \rho(z,p) \leq \rho(T(z), p)\} \tag{1.71}$$

Fix an element $T_1 \in \mathrm{PSL}(2,\mathbb{R})$ and consider the geodesic segment joining p and $T_1 p$. The line given by the equation

$$\rho(z,p) = \rho(z, T_1 p) \tag{1.72}$$

is the geodesic orthogonal to the middle-point of the geodesic segment joining p and $T_1 p$ (perpendicular bi-sector); let us call it $L_p(T_1)$. Consider now the hyperbolic half-plane $H_p(T_1)$ bounded by $L_p(T_1)$ and containing the point p. It is not difficult to show that

$$D_p(\Gamma) = \bigcap_{T \in \Gamma - \{I\}} H_p(T) \tag{1.73}$$

that is the Dirichlet region is an intersection of closed half-planes, hence it is closed and convex. Moreover, it is path-connected, hence connected.

The shape of a Dirichlet region can be quite complicated, since it is bounded by geodesics in $\mathbb{H}$ and possibly by segments of the real axis. It two geodesics intersect in $\mathbb{H}$, their point of intersection is called a *vertex*. It can be shown that vertices are isolated, thus a Dirichlet region is bounded by a union of (possibly infinitely many) geodesics and possibly segments of the real axis.

In general, two points $u, v \in \mathbb{H}$ are *congruent* if they belong to the same $\Gamma-$orbit; in a fundamental region $\mathcal{F}$ this means that the two points belong to the boundary $\partial\mathcal{F}$. Let us choose for $\mathcal{F}$ a Dirichlet region and consider congruent vertices of $\mathcal{F}$. The congruence is an equivalence relation on the vertices of $\mathcal{F}$ and the equivalence classes are called *cycles*. If one vertex of the cycle is fixed by an elliptic element, then all the vertices of that cycle are fixed by conjugate elliptic elements. Such a cycle is called an *elliptic cycle* and the vertices are called *elliptic vertices*. The number of elliptic cycles is equal to the number of non-congruent elliptic points in $\mathcal{F}$.

It is clear that every point $w \in \mathbb{H}$ fixed by an elliptic element S' of Γ lies on the boundary of $T(\mathcal{F})$ for some T. Hence $u = T^{-1}(w)$ lies on the boundary of $\mathcal{F}$ and is fixed by the elliptic element $S = T^{-1}S'T$. This element has finite order k (remember that elliptic elements are conjugate to rotations). If $k \geq 3$, then as S is an isometry fixing u which maps geodesics to geodesics, u must be a vertex whose angle θ is at most $2\pi/k$. The hyperbolically convex region $\mathcal{F}$ is bounded by a union of geodesics. The intersection of $\mathcal{F}$ with these geodesics is either a single point or a segment of a geodesic. These segments are called sides of $\mathcal{F}$. If S has order 2 ($k = 2$), then its fixed point u might lie on the interior of a side of $\mathcal{F}$. In this case, S interchanges the two segments of this side separated by the fixed point. We will include such elliptic fixed points as vertices of $\mathcal{F}$, the angle at such a vertex being π. Thus a *vertex* of $\mathcal{F}$ is a point where two bounding geodesics meet or a fixed point of an elliptic element of order 2.

A parabolic element can be considered as an elliptic element of infinite order, it has a unique fixed point on $\widehat{\mathbb{R}}$. If Γ contains parabolic elements (then $\mathcal{F}$ is not compact) and $\mu(\mathcal{F}) < \infty$, then $\mathcal{F}$ has at least one *vertex at infinity*; such a vertex is a parabolic fixed point.

Let us now consider the congruence of sides of $\mathcal{F}$, a Dirichlet region for a Fuchsian group Γ. If s is a side and $T(s)$ is also a side of $\mathcal{F}$ ($T \in \Gamma - \{I\}$), then s and $T(s)$ are called *congruent sides*. But $T(s)$ is also a side of $T(\mathcal{F})$, a copy of the Dirichlet region under T, so that $T(s) \subseteq \mathcal{F} \cap T(\mathcal{F})$. If a side of $\mathcal{F}$ has a fixed point of an elliptic element S of order 2 on it, then S interchanges the two segments of this side. It is convenient to regard these two segments as distinct sides separated by a vertex. With this convention, for each side of $\mathcal{F}$ there exist another side of $\mathcal{F}$ congruent to it. Thus the sides of $\mathcal{F}$ fall into congruent pairs. Hence, if the numbers of sides of a Dirichlet region is finite, it is always even. These considerations allow to show that if $\{T_i\}$ is the subset of Γ consisting of those elements which pair the sides of some fixed Dirichlet region $\mathcal{F}$, then $\{T_i\}$ is a set of generators for Γ.

The most important class of groups containing negative isometries are the so-called *hyperbolic reflection groups.* Let m_i $(i = 1, 2, 3)$ be positive integer or ∞ such that $\frac{1}{m_1} + \frac{1}{m_2} + \frac{1}{m_3} < 1$ and let r be a hyperbolic triangle with vertices v_1, V_2, v_3, angles $\pi/m_1, \pi/m_2, \pi/m_3$ at these vertices and sides M_1, M_2, M_3 opposite to these vertices. Such a triangle always exist and is unique up to isometry [5]. Moreover, a generic hyperbolic triangle tiles the hyperbolic plane iff its angles are of the form $\pi/n, \pi/m, \pi/l$ with n, m, l positive integers (one of them is allowed to be ∞). The triangle groups we are going to define are often indicated with (n, m, l).

Let R_i the hyperbolic reflection in the geodesic containing the side M_i and let Γ^* be the group generated by the reflections $\{R_i\}$. Since hyperbolic reflections are negative isometries, $R_i \notin \mathrm{PSL}(2, \mathbb{R})$, Γ^* is not a Fuchsian group; it is called a *triangle reflection group.* Let us consider the intersection $\Gamma = \Gamma^* \cap \mathrm{PSL}(2, \mathbb{R})$. It is clear that $\Gamma^* = \Gamma \cup \Gamma R_1$, since the composition of two orientation-reversing isometries is orientation-preserving. If we denote by τ the region inside the triangle, then $\{T(\tau) |\, T \in \Gamma^*\}$ forms a tessellation of $\mathbb{H}$, that is every point of $\mathbb{H}$ belongs to some Γ^*-image of τ and any two images of τ may overlap only on the boundary. It follows that τ is a fundamental region for Γ^*. For any point p inside τ, the Γ^*-images of p are points of other triangles of the tessellation, hence they form a discrete set. As the $\Gamma-$orbit of each point of $\mathbb{H}$ is a discrete set, Γ is a Fuchsian group called a *triangle group* (it does not contain reflections). From what we said above, $\tau \cup R_1(\tau)$ is a fundamental region for Γ (see the picture). The sides v_2v_1 and v_2v_1' are paired by R_1R_3 and the sides v_3v_1 and v_3v_1' are paired by R_1R_2. Finally, $\{v_1, v_1'\}$ is an elliptic cycle and both vertices are stabilized by cyclic groups of order m_1, $\{v_2\}$ and $\{v_3\}$ are elliptic cycles whose vertices v_2 and v_3 are stabilized by cyclic groups of order m_2 and m_3 respectively.

[5] This is different from Euclidean geometry, where there exist similar, but not isometric triangles with the same angles. The reason is that on the hyperbolic plane the Gaussian curvature $\mathcal{K}$ is strictly negative, in particular different from zero, thus in hyperbolic geometry there exists a preferred length scale.

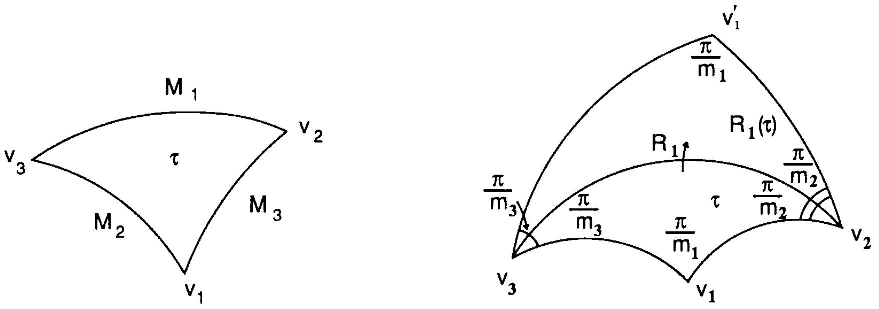


Figure 1.4: A triangle group generated by hyperbolic reflections (figure from [91]).

Before giving some examples, let us mention an important sub-class of Fuchsian groups, the so-called *arithmetic Fuchsian groups.* In full generality, the notion of an arithmetic subgroup of a semisimple Lie group uses tools from the theory of linear algebraic groups. We are not so ambitious, for precise definitions see the last chapter of [91]. Here we will be happy with an intuitive idea. A first example of arithmetic group is $\mathrm{PSL}(2,\mathbb{Z})$ (see below), obtained by $\mathrm{PSL}(2,\mathbb{R})$ just restricting the field $\mathbb{R}$ to the ring $\mathbb{Z}$. The same construction, restriction to integers, applies equally well to obtain arithmetic subgroups of larger matrix groups, e.g. $\mathrm{SL}(n,\mathbb{Z})$ in $\mathrm{SL}(n,\mathbb{R})$, $\mathrm{Sp}(2n,\mathbb{Z})$ in $\mathrm{Sp}(2n,\mathbb{R})$ etc. In order to have an arithmetic Fuchsian group, let $g \to T(g)$ be a finite-dimensional representation of $\mathrm{PSL}(2,\mathbb{R})$. The elements of $\mathrm{PSL}(2,\mathbb{R})$ which correspond to matrices $T(g)$ with integer coefficients form a discrete subgroup of $\mathrm{PSL}(2,\mathbb{R})$. All subgroups thus obtained and also their subgroups of finite index are called arithmetic Fuchsian groups. This is not easy to check. A result of A. Weil states that the list of all arithmetic subgroups of $\mathrm{SL}(2,\mathbb{R})$ is exhausted up to commensurability by Fuchsian groups derived from quaternion algebras over totally real number fields. The list of all arithmetic Fuchsian groups can be found in a paper by K. Takeuchi and it is reproduced in [25].

To finish this section, let us introduce the concept of *hyperbolic surface.* Let Γ be a Fuchsian group and $\mathcal{F}$ a fundamental region. The group Γ induces a natural projection (continuous and open) $\pi : \mathbb{H} \to \Gamma\backslash\mathbb{H}$ and the points of $\Gamma\backslash\mathbb{H}$ are the $\Gamma-$orbits. The restriction of π to $\mathcal{F}$ makes $\Gamma\backslash\mathbb{H}$ into a oriented surface (Γ does not contain reflections) with possibly some *marked points* (which correspond to elliptic cycles) and *cusps* (which correspond to non-congruent vertices at infinity of $\mathcal{F}$). Such a surface is known as an *orbifold.* If we take for $\mathcal{F}$ a Dirichlet region, $\Gamma\backslash\mathbb{H}$ is homeomorphic to $\Gamma\backslash\mathcal{F}$. It is clear that if $\mu(\mathcal{F})$ is the area of a fundamental region $\mathcal{F}$, this induces an area on the hyperbolic surface $\Gamma\backslash\mathbb{H}$, and one has $\mu(\Gamma\backslash\mathbb{H}) = \mu(\mathcal{F})$. Moreover, $\Gamma\backslash\mathbb{H}$ is compact iff $\mathcal{F}$ is compact, i.e. if Γ is a co-compact Fuchsian group. If, in addition, Γ acts on $\mathbb{H}$ without fixed points, then $\Gamma\backslash\mathbb{H}$ is a *compact Riemann surface.* Thus for any strictly hyperbolic Fuchsian group Γ , $\Gamma\backslash\mathbb{H}$ is a compact Riemann surface whose genus g is ≥ 2 and its fundamental group is isomorphic to Γ. If Γ is co-compact with elliptic elements, then $\Gamma\backslash\mathbb{H}$ is a compact Riemann surface with punctures. If Γ is co-finite with parabolic elements, then $\Gamma\backslash\mathbb{H}$ is a Riemann surface with punctures and cusps.

If Γ contains elliptic elements, the structure of the fiber bundle is violated in a finite number of points, the ones we have called marked points. Finally, $\Gamma\backslash\mathbb{H}$ is compact iff $S(\Gamma\backslash\mathbb{H})$ is compact.

1.4.1 The regular octagon

An example of chaotic billiard studied in [9] is a free particle moving in a particular domain of the Poincaré disc, the *regular octagon.* This is a fundamental domain for the discrete group generated by the hyperbolic transformations which pair the opposite sides (see the figure). The octagon is compact, thus it is associated to a compact Riemann surface of genus 2, i.e. the double torus. This discrete group is an example of a co-compact group and moreover it contains *only* hyperbolic elements, i.e. it is a strictly hyperbolic Fuchsian group.

1.4.2 The modular group and some of its distinguished subgroups

Let Γ the triangle group $(2, m, \infty)$. According to the construction described above, we first generate a reflection group Γ^* by hyperbolic reflections in the sides of the triangle with vertices v_1, $v_2 = i$ and $v_3 = \infty$ and angles π/m, $\pi/2$ and 0 respectively. Explicitly, $R_1(z) = -\overline{z}$, $R_2(z) = -\overline{z} + 2\cos\frac{\pi}{m}$, $R_3(z) = \frac{1}{\overline{z}}$ (see the picture). The triangle group Γ (the Fuchsian group) is generated by $R_1R_3 = -\frac{1}{z}$ and $R_2R_1 = z + 2\cos\frac{\pi}{m}$ which identify the sides $v_1'v_2$ with v_1v_2, and $v_1'v_3$ with v_1v_3 respectively. The corresponding group is called a *Hecke triangle group* and is denoted by $\Gamma(2\cos\frac{\pi}{m})$. [91] proves that $\Gamma(2\cos\frac{\pi}{m})$ is

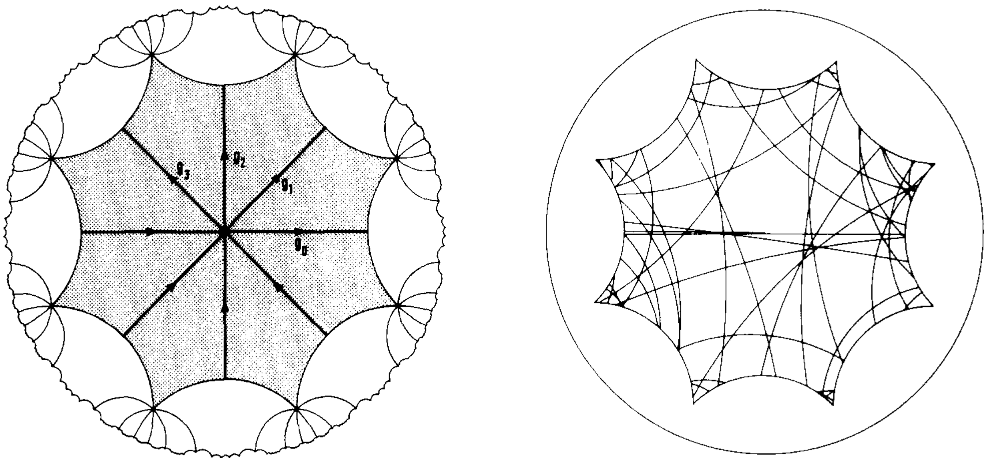


Figure 1.5: A regular octagon on the Poincaré disc and the exponential divergence of two nearby trajectories (figure from [9]). The figure on the right shows the trajectories of a particular point starting from the origin of the Poincaré disc with an angular deviation of 10^{-3} from a periodic trajectory.

arithmetic only for $m = 3, 4, 6$.

The most important Hecke triangle group is the *modular group* $\mathrm{PSL}(2,\mathbb{Z})$ or in the previous notation $\Gamma(2\cos\frac{\pi}{3})$. Its fundamental domain (see the picture) is

$$\mathcal{F}(\mathrm{PSL}(2,\mathbb{Z})) = \left\{ z \in \mathbb{H} : |z| \geq 1,\ |x| \leq \frac{1}{2} \right\} \tag{1.74}$$

The modular group derives from the reflection group $\Gamma^* \cong \mathrm{PGL}(2,\mathbb{Z})$ (sometimes known as the *extended modular group*). Its domain is the halved-modular domain, in agreement with the general construction for triangle groups. $\mathrm{PGL}(2,\mathbb{Z})$ contains $\mathrm{PSL}(2,\mathbb{Z})$ as a subgroup of index 2. In fact the standard generators for $\mathrm{PSL}(2,\mathbb{Z})$

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{1.75}$$

are given in terms of the hyperbolic reflections of $\mathrm{PGL}(2,\mathbb{Z})$ by the general formulae above with $m = 3$

$$T(z) = z + 1 = R_2 R_1 \tag{1.76}$$

$$S(z) = -\frac{1}{z} = R_1 R_3 \tag{1.77}$$

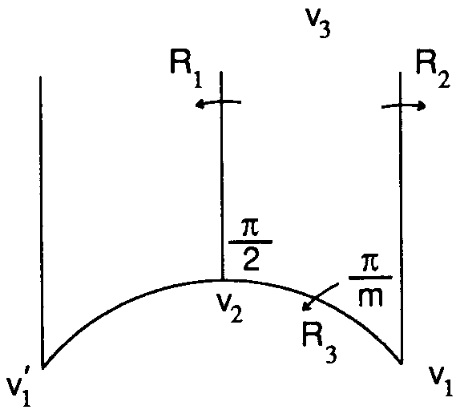


Figure 1.6: The typical fundamental domain of a Hecke triangle group (from [91])

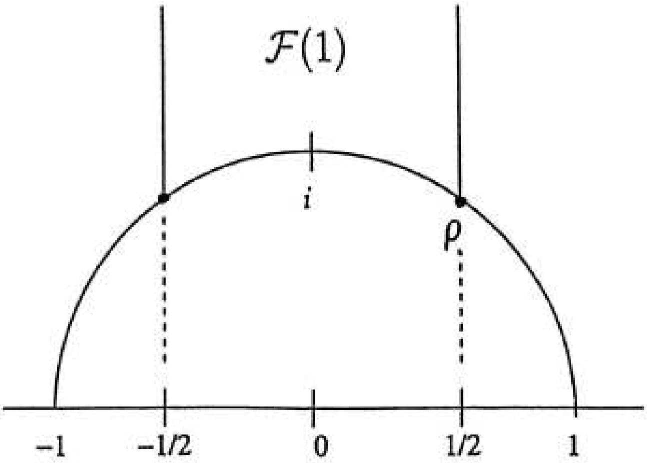


Figure 1.7: The standard modular domain (from [137]).

It is easy to show that the modular domain is the Dirichlet region centered at $p = ki$, $k > 1$ (the point p is not fixed by any element in $\mathrm{PSL}(2,\mathbb{Z})$). This region has 4 vertices, including the point i, which is an elliptic point of order 2 fixed by S . The other elliptic fixed points are ρ and $\rho+1$, of order 3. With this convention, the modular domain has 4 sides: T pairs the two vertical sides, S pairs the two semi-arcs, thus T and S generate $\mathrm{PSL}(2,\mathbb{Z})$. The area of the fundamental domain is easily calculated by the Gauss-Bonnet theorem

$$\mu(\mathcal{F}(\mathrm{PSL}(2,\mathbb{Z})) = \frac{\pi}{3} = 2\,\mu(\mathcal{F}(\mathrm{PGL}(2,\mathbb{Z})) = 2\,\frac{\pi}{6} \tag{1.78}$$

The most important subgroups of $\mathrm{SL}(2,\mathbb{Z})$ are its congruence subgroups. For any $N \geq 1$, the *principal congruence subgroup of level* N is

$$\Gamma(N) = \{\gamma \in \mathrm{SL}(2,\mathbb{Z})|\, \gamma \equiv I \,(\mathrm{mod}\ N)\} \tag{1.79}$$

where I is the 2×2 identity matrix; $\mathrm{SL}(2,\mathbb{Z})$ is identified with $\Gamma(1)$. Finally a *congruence group* Γ is a subgroup of $\mathrm{SL}(2,\mathbb{Z})$ for which there exist an integer M such that Γ contains the principal congruence group of level M, i.e. $\Gamma(M) \subset \Gamma \subset \Gamma(1)$. Some importance also have the groups $\Gamma_0(N)$ defined as

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}) : c \equiv 0 \,(\mathrm{mod}\ N) \right\} \tag{1.80}$$

It is clear that

$$\Gamma(N) \subset \Gamma_0(N) \subset \mathrm{SL}(2,\mathbb{Z}) \tag{1.81}$$

and that $\Gamma(1) = \Gamma_0(1) = \mathrm{SL}(2,\mathbb{Z})$. The principal congruence groups are the kernel of the applications

$$\Gamma(N) = \ker\left(\mathrm{PSL}(2,\mathbb{Z}) \to \mathrm{PSL}(2,\mathbb{Z}/N\mathbb{Z})\right) \tag{1.82}$$

which explains the term *congruence*, and they are normal subgroups of finite index in $\mathrm{SL}(2,\mathbb{Z})$ (note that *not* all subgroups of $\mathrm{SL}(2,\mathbb{Z})$ can be described by congruence relations). Thus $\Gamma(N)$ is arithmetic for any N. For N prime, $\Gamma_0(N)$ is also of finite index in $\mathrm{SL}(2,\mathbb{Z})$ (and thus arithmetic). In fact, for N prime and for every $V \in \mathrm{SL}(2,\mathbb{Z})$ such that $V \notin \Gamma_0(N)$, there exist an element $P \in \Gamma_0(N)$ and an integer $0 \leq k < N$ such that

$$V = PST^k \tag{1.83}$$

where S and T are the standard generators for $\mathrm{SL}(2,\mathbb{Z})$. Moreover, if $\mathcal{F}(1)$ is a fundamental region for $\mathrm{SL}(2,\mathbb{Z})$ and N is a prime, then

$$\mathcal{F}(1) \cup \bigcup_{k=0}^{N-1} ST^k(\mathcal{F}(1)) \tag{1.84}$$

is a fundamental region for $\Gamma_0(N)$.

In this thesis, we will be mainly concerned with $\mathrm{SL}(2,\mathbb{Z})$ and $\mathrm{GL}(2,\mathbb{Z})$.

As we mentioned, a hyperbolic surface is the quotient of the hyperbolic plane by a Fuchsian group [6]. We are mainly interested in the hyperbolic surfaces $X(N) = \Gamma(N)\backslash\mathbb{H}$, especially the *modular surface* $X(1) = \mathrm{PSL}(2,\mathbb{Z})\backslash\mathbb{H}$. Each $X(N)$ is a finite area, non-compact surface. It is also a Riemann surface (with the complex structure inherited by $\mathbb{H}$) whose genus grows like N^3 when N gets large. $X(1)$ has a cusp at infinity, which corresponds to the fixed point $i\infty$ of the parabolic transformation $T : z \rightarrow z + 1$. In the next section, we describe the spectral problem for the surfaces $X(N)$.

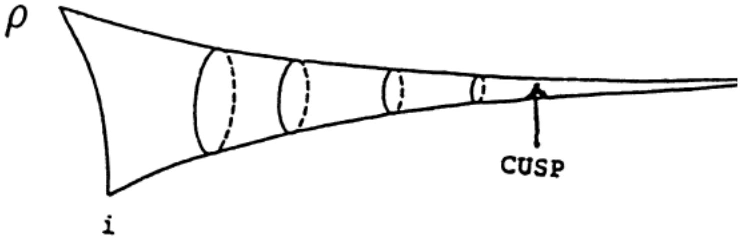


Figure 1.8: The modular surface (from [137]).

[6] To be honest, the Fuchsian group must be *torsion-free*, that is there are no non-trivial elements of finite order. $\mathrm{PSL}(2,\mathbb{Z})$ is not torsion-free, because $S^2 = 1$ and $S \neq 1$, while the principal congruence groups are torsion-free and each $X(N)$ with $N \geq 2$ is a well-defined hyperbolic 2-manifold.

1.5 Maass automorphic forms and the Selberg Trace Formula

The *fundamental spectral problem of quantum chaos* is the following

$$\begin{cases} \Delta\phi + \lambda\phi = 0 \\ \phi(\gamma z) = \phi(z) \quad \forall\, \gamma \in \Gamma(N) \\ \int_{X(N)} |\phi(z)|^2 d\mu(z) < +\infty \end{cases} \tag{1.85}$$

Here, Δ is the hyperbolic Laplacian, $\Delta = y^2(\partial_x^2 + \partial_y^2)$. The numbers $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \ldots$ for which the spectral problem has solutions form a discrete set (eigenvalues), the discrete spectrum of $X(N)$. The only eigenvalue known is $\lambda_0 = 0$, the corresponding eigenfunction $\phi_0(z)$ is a constant (the surface has finite hyperbolic area). We call a solution to (1.85) a *Maass form*, after the mathematician H. Maass who first introduced them. Their existence is not obvious at all, since the surface is not compact. In particular, no explicit eigenvalues are known or expected for $X(1)$, although, for example, for $\Gamma(4)$ Maass himself produced an explicit subsequence of eigenvalues [137].

Regarding the existence of Maass waveforms, let us first remember a classical result of H. Weyl. Let Ω be a compact domain in $\mathbb{R}^2$, with smooth boundary $\partial\Omega$. The Dirichlet problem for the Euclidean Laplacian $\Delta = \partial_x^2 + \partial_y^2$ is

$$\begin{aligned} \Delta\phi(z) + \lambda\phi(z) &= 0 \qquad \text{for } z \in \Omega \\ \phi|_{\partial\Omega} &= 0 \end{aligned} \tag{1.86}$$

If we denote by $N_\Omega(R)$ the number of eigenvalues λ counted with their multiplicity such that $\lambda \leq R$, then *Weyl's law* says that

$$N_\Omega(R) \sim \frac{Area(\Omega)}{4\pi} R \quad \text{as } R \to \infty \tag{1.87}$$

This result has been generalized to compact Riemannian manifolds of any dimension. Usually the favorite way to prove it is by analyzing the small time asymptotics of the heat kernel on $\mathbb{R} \times \Omega$; the propagation of singularities for the wave kernel allows of get some reminder terms for such Weyl asymptotics.

Let us now come back to finite area hyperbolic surfaces. As we said, since these surfaces are not compact, it is not clear that there exist solutions to the problem (1.85) for $\lambda > 0$. In fact, the discrete spectrum we are looking for is *embedded* in the continuous spectrum, and this makes difficult to isolate the eigenvalues analytically. A. Selberg found that the continuous spectrum is the whole interval $[\frac{1}{4}, \infty)$, with multiplicity equal to the number of cusps of $X_\Gamma = \Gamma\backslash\mathbb{H}$. Here we can consider a general hyperbolic surface X_Γ (not necessarily some $X(N)$). The corresponding (not-normalizable) eigenfunctions are given by the *Eisenstein series.* For $X(1)$ they read as follows [135]

$$E(z,s) = \sum_{g\,\in\,\Gamma_\infty\backslash\Gamma(1)} \frac{y^s}{|cz+d|^s} \quad \text{for Re } s > 1 \tag{1.88}$$

they extend meromorphically to $\mathbb{C}$ and are analytic on Re $s = 1/2$ ($\Gamma(\infty)$ is the stabilizer in $\Gamma(1)$ of ∞, which is the only cusp). Thus, the continuous spectrum is furnished by these generalized eigenfunctions $E\left(z, \frac{1}{2} + it\right)$, $t \geq 0$,

$$\Delta E\left(z, \frac{1}{2} + it\right) + \left(\frac{1}{4} + t^2\right) E\left(z, \frac{1}{2} + it\right) = 0 \tag{1.89}$$

and of course they are $\Gamma(1)-$periodic

$$E(gz,s) = E(z,s) \text{ for any } g \in \Gamma(1) \tag{1.90}$$

The constant term in the Fourier expansion of the Eisenstein series, $\phi_\Gamma(s)$, is meromorphic in $\mathbb{C}$ and is called the determinant of the scattering matrix in the Lax-Phillips scattering theory for automorphic functions [101]. The pole of $\phi_\Gamma(s)$ in Re $s \geq \frac{1}{2}$ are in $(\frac{1}{2}, 1]$ and the residues at these poles furnish solutions to (1.85), called the *residual spectrum* of X_Γ. The poles of $\phi_\Gamma(s)$ in Re $s < \frac{1}{2}$ instead give resonances for problem (1.85).

If we now take the orthogonal complement in $L^2(X_\Gamma, \mu)$ of the continuous and residual spectrum, we obtain the *cuspidal space* $L^2_{cusp}(X_\Gamma)$. It is invariant under the (hyperbolic) Laplacian and the resolvent $(\Delta - \lambda)^{-1}$ is compact when restricted to $L^2_{cusp}(X_\Gamma)$. A Maass form, i.e. a solution to (1.85), which also lies in L^2_{cusp} is called a *Maass cusp form.* These particular cusp forms are the building blocks of the theory of automorphic forms. Their existence is tied to the size of L^2_{cusp}, in fact $L^2_{cusp}(X) \neq \{0\}$ is not obvious at all for a general hyperbolic surface X.

For the modular surfaces $X(N)$, Selberg was able to show using his trace formula that there is an abundance of Maass cusp forms. For these surfaces, $\phi_{\Gamma(N)}(s)$ can be expressed through Dirichlet $L-$functions; for example for $\Gamma(1)$

$$\phi_{\Gamma(1)}(s) = \frac{\zeta^*(2s-1)}{\zeta^*(2s)} \tag{1.91}$$

where $\zeta^*(s)$ is the completed Riemann zeta-function (see Appendix A). For any N, $\phi_{\Gamma(N)}$ has no poles in $(\frac{1}{2}, 1)$, which means that there is no residual spectrum (besides $\lambda = 0$) and any Maass form is automatically is a cusp form. Selberg proved that for the modular surfaces using the expression of $\phi_{\Gamma(N)}(s)$ in terms of Dirichlet $L-$functions the contribution of the continuous spectrum to the Weyl law is negligible, that is

$$N^{cusp}_{\Gamma(N)}(R) := \sum_{0<\lambda_j\leq R} 1 \sim \frac{\mu(X(N))}{4\pi} R \quad (R \to \infty) \tag{1.92}$$

Thus, solutions to (1.85) exist and in abundance, at least for the modular surfaces $X(N)$.

It is of course of interest to understand when solutions exist for more general hyperbolic surface. We do not discuss that here, but we may ask if there is a characterization of those Γ which have many Maass cusp forms. This question was addressed by Phillips and Sarnak and the answer lies in the *arithmeticity* of the group. In fact, it is believed that there are infinitely many solutions to the problem $\Delta u + \lambda u = 0$, $u \in L^2(\mathcal{F}_q)$, $\partial_n u|_{\partial\mathcal{F}_q} = 0$ (Neumann boundary conditions) if and only if $q = 3, 4, 6$, that is iff one considers the arithmetic Hecke triangle groups $\Gamma(2\cos\frac{\pi}{q})$ with fundamental domain $\mathcal{F}_q$. $q = 3$ corresponds of course to even Maass cusp forms for $\Gamma(1)$, $q = 4, 6$ to other congruence subgroups of $\mathrm{SL}(2, \mathbb{Z})$. All other integer values of q give, via reflections in the sides of the triangle, non-arithmetic subgroups of $\mathrm{SL}(2, \mathbb{R})$. One can also consider the case $q \notin \mathbb{Z}$, then the reflections in the sides of $\mathcal{F}_q$ do not generate a discrete group any more, but the eigenvalue problem still makes sense. The numerical evidence supports the absence of eigenvalues (and thus of Maass forms) in these cases. In conclusion, the Maass forms for $X(N)$ are very fragile objects and their existence is tied to the arithmeticity of $\Gamma(N)$.

Regarding the low-energy spectrum of the $X(N)$, the lowest eigenvalue is $\lambda_0 = 0$. Let $\lambda_1(X(N))$ be the next eigenvalue. A deep conjecture (still open) due to Selberg state that

$$\lambda_1(X(N)) \geq \frac{1}{4} \tag{1.93}$$

for any $N \geq 1$. Since for $X(N)$ there is no residual spectrum, we could take for $\lambda_1(X(N))$ the smallest eigenvalue of a Maass cusp form on $X(N)$. Remember that the continuous spectrum is the interval $\left[\frac{1}{4}, \infty\right)$ and that a result of H. McKean shows that the spectrum of Δ the universal covering

$L^2(\mathbb{H})$ is $\left[\frac{1}{4}, \infty\right)$, thus $\lambda_0(\mathbb{H}) \geq \frac{1}{4}$. Besides, the cuspidal spectrum of Δ becomes dense in $\left[\frac{1}{4}, \infty\right)$ as $N \to \infty$. Finally, the assumption that Γ is a congruence subgroup of $\mathrm{SL}(2, \mathbb{Z})$ can not be dropped: in fact there exist finite index subgroups Γ' of $\mathrm{SL}(2, \mathbb{Z})$ for which $\lambda_1(X_{\Gamma'}) < \epsilon$ given any $\epsilon > 0$.

Finally, it is expected that the cuspidal spectrum of Δ on $X(1)$ is *simple*, a conjecture first stated by P. Cartier. The numerical computations support this, the situation for the other $X(N)$, $N \geq 2$, may be different. Let us mention also that all the numerical analysis supports the fact the high-energy spectrum follows a Poissonian distribution (see Appendix B), because the geodesic flow on $X(1)$ is chaotic, but arithmetic as well.

In order to motivate the Selberg trace formula, it is useful to briefly recall the spectral theory on the torus $\mathbb{T}^n$, which is the quotient of $\mathbb{R}^n$ by the translation group $\mathbb{Z}^n$ viewed as a discrete and torsion-free subgroup of the group $G = \mathbb{R}^n$ acting on the Euclidean space by translations. The group G makes the Euclidean space a homogeneous space. The Euclidean space can be endowed with a Riemannian metric

$$ds^2 = dx_1^2 + \cdots + dx_n^2 \tag{1.94}$$

coming from the scalar product $x \cdot y = x_1 y_1 + \cdots + x_n y_n$. This metric has null curvature and induces a flat Riemannian metric on the torus $\mathbb{T}^n$. All the elements of G are isometries for the Euclidean metric.

Let $\Delta = -\partial_{x_1}^2 - \cdots - \partial_{x_n}^2$ be the Laplacian on $\mathbb{R}^n$; it commutes with the action of G on the Euclidean space and defines a second-order differential operator on the torus $\mathbb{T}^n$. It is obvious that the exponential functions

$$\phi(x) = \exp\left[2\pi i\left(x_1\xi_1 + \cdots + x_n\xi_n\right)\right] \equiv e(x \cdot \xi) \tag{1.95}$$

are eigenfunctions of Δ on the torus

$$\Delta\,\phi = \lambda\,\phi\,, \qquad \lambda = 4\pi^2 ||\xi||^2 \tag{1.96}$$

The spectral resolution of the Laplacian is given through the classical Fourier inversion

$$\widehat{f}(\xi) = \int_{\mathbb{R}^n} f(x)\, e(-x \cdot \xi)\, dx \tag{1.97}$$

$$f(x) = \int_{\mathbb{R}^n} \widehat{f}(x)\, e(-\xi \cdot x)\, d\xi \tag{1.98}$$

where the functions obey some regularity conditions (rapid decay at infinity etc).

The Selberg trace formula can be understood as a non-Abelian generalization of the classical *Poisson summation formula*

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{m\in\mathbb{Z}} \widehat{f}(m) \tag{1.99}$$

with $f \in \mathcal{S}(\mathbb{R})$. This formula is proven by considering the periodic function $F(x) = \sum_n f(x+n)$, taking its Fourier series and putting $x = 0$. For $n = 1$, harmonic analysis on $\mathbb{R}/\mathbb{Z}$ is an important step in understanding the Riemann zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s} = \prod_p (1-p^{-s})^{-1} \tag{1.100}$$

in fact the Poisson summation formula allows to define the completed zeta-function, for more details see Appendix A.

The modern theory on automorphic functions is concerned in part with spectral problems associated with quotients of more general (non-abelian) groups, their homogeneous and symmetric spaces and the formation of related zeta functions.

Selberg discovered that Poisson summation formula of classical analysis had a non-commutative generalization (now referred to as the Selberg trace formula) with important applications to number theory and the theory of automorphic functions. The similarity between this trace formula and an explicit formula due to A. Weil let Selberg to introduce the *Selberg zeta function.* Let us briefly discuss these objects.

As we said, the eigenvalue problems studied by Maass turns out to be a problem of upmost complexity. Precise results about the individual eigenvalues of the fundamental spectral problem are lacking, but asymptotic results can be derived from the Selberg trace formula. This formula is as simple as possible for co-compact groups which moreover do not have elliptic elements (think of the regular octagon). Let Γ be such a group, with $\lambda_0 = 0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$, $\lambda_n = 1/4 + r_n^2$, the eigenvalues of Δ. Let $h : \mathbb{C} \to \mathbb{C}$ be an even function, which is holomorphic and satisfies the growth condition $h(r) = O((1+|r|^2)^{-1-\delta})$ for $|r| \to \infty$ uniformly in the strip $|\mathrm{Im}\, r| < \frac{1}{2} + \delta$ ($\delta > 0$). Let $g(u)$ be the Fourier transform of $h(r)$

$$g(u) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} h(r)\, e^{-iru}\, dr \tag{1.101}$$

Then *the Selberg trace formula for Fuchsian groups having only hyperbolic elements* is

$$\begin{aligned}\sum_{n=0}^{+\infty} h(r_n) &= \frac{\mu(\mathcal{F})}{4\pi}\int_{-\infty}^{+\infty} r\, h(r)\, \tanh(\pi r)\, dr \\ &+ \sum_{\{P\}} \frac{\ln N(P_0)}{N(P)^{1/2}-N(P)^{-1/2}}\, g(\ln N(P)) \qquad (1.102)\end{aligned}$$

where the sum on the right-hand side extends over all $\Gamma-$conjugacy classes $\{P\}$ of hyperbolic elements $P \in \Gamma\backslash\{I\}$. $N(P)$ denotes the norm of P, that is $N(P)$ is equal to the square of the eigenvalue of the matrix which defines P with larger absolute value. P_0 is the primitive hyperbolic element associated with P, that is there exist no $P \in \Gamma$, no integer $m \geq 1$ such that $P_0 = P^m$. All the sums and the integrals in the above trace formula are absolutely convergent.

For Fuchsian groups which do have parabolic and elliptic elements too, one has to compute also the contributions of these elements. Besides, the right-hand side must contain also the continuous spectrum. The trace formula for $X(1)$ reads as follows. Let $g \in C_0^\infty(\mathbb{R})$ be an even smooth function of compact support and let $h(\xi) = \widehat{g}(\xi/2\pi)$ (h is an entire function). Then *the Selberg trace formula for* $PSL(2,\mathbb{Z})$ is

$$\sum_{t_\phi} h(t_\phi) - \frac{1}{2\pi}\int_{-\infty}^{\infty} h(t)\,\frac{\phi'_{\Gamma(1)}}{\phi_{\Gamma(1)}}\left(\frac{1}{2}+it\right) dt$$

$$\begin{aligned}&= \frac{\mu(X(1))}{2\pi}, \int_{-\infty}^{+\infty} \tanh(\pi t)\, t h(t)\, dt - \frac{1}{\pi}\int_{-\infty}^{+\infty} h(t)\frac{\Gamma'}{\Gamma}(1+it)\, dt \\ &- \ 2\ln 2 g(0) + h(0) \\ &+ \sum_{\{R\}}\sum_{1\leq\nu\leq m-1} \frac{2}{m\,\sin\frac{\pi\nu}{m}}\int_{-\infty}^{+\infty} \frac{h(r)e^{-\frac{\pi\nu}{m}}}{1+e^{-2\pi r}}\, dr \\ &+ \ 2\sum_{\{P\}}\sum_{k=1}^{+\infty} \frac{\ln N(P)}{N(P)^{k/2}-N(P)^{-k/2}}\, g(k\,\ln N(P)) \qquad (1.103)\end{aligned}$$

The t_ϕ's run through the discrete spectrum of $X(1)$ (as usual $\lambda_\phi = \frac{1}{4}+t_\phi^2$). $\phi_{\Gamma(1)}$ is the constant term in the Eisenstein series as before given in terms of

$\zeta^*(s)$; Γ is the Euler gamma function. The sum $\{R\}$ is over elliptic conjugacy classes, for $\mathrm{SL}(2,\mathbb{Z})$ there are two of them, one of order $m=2$ and one of order $m=3$. Again, the sum $\{P\}$ is over primitive hyperbolic conjugacy classes of $\Gamma(1)$. Remember that a hyperbolic $A\in$SL$(2,\mathbb{R})$ can be conjugated into the form $\pm\begin{pmatrix} (N(A))^{1/2} & 0 \\ 0 & (N(A))^{-1/2}\end{pmatrix}$ with $N(A)>1$. A fixes a unique geodesic γ in $\mathbb{H}$, whose length (on $X(1)$) is $\ln N(A)$. Thus, as we already said, the set $\{P\}$ corresponds to the set of primitive closed geodesics on $X(1)$ [7].

The left-hand side of this formula is the spectral one, it contains a sum over the discrete and the continuous spectrum. It is the *quantum-mechanical* side. The right-hand side is geometrical, because it contains a sum over all closed geodesics. Thus it the *classical* side. The equality, somehow, sanctions an equivalence between classical mechanics and quantum mechanics

$$\text{quantum mechanics} = \text{classical mechanics}$$

in the semiclassical limit. As we already mentioned, unlike the divergent Gutzwiller trace formula, the Selberg trace formula is absolutely convergent, but its contents depend on the test function h. The class of test functions is pretty large, so the Selberg trace formula gives an infinite number of semiclassical quantization rules, which, at the moment, are the only tools for quantum systems whose semi-classical limit is chaotic.

Note two more things. First, the Selberg trace formula is valid for manifolds whose curvature is *constant*. There is no generalization of this result to surfaces of non-constant negative curvature. Second, the Poisson summation formula is a one-dimensional formula, in the sense that in higher dimensions one simply considers the product of one-dimensional Poisson formulas to get the spectral resolution of the Laplacian on, say, the torus $\mathbb{T}^n$. For the Selberg trace formula, things are more complicated. Yet, there exist a higher-dimensional generalization of it, due to J. Arthur [7], known as the *Arthur trace formula*. This formula uses the *adelic* language, unfortunately explaining that would lead us too far.

[7] To be more precise, for $X(1)$, the lengths of primitive closed geodesics are the numbers $2\,\ln\epsilon_d$ where $0<d\equiv 0$ or $1 \bmod 4$ is square-free and ϵ_d is the fundamental solution $\frac{t_0+\sqrt{d}u_0}{2}$ to the Pell equation $t^2-du^2=4$, with multiplicity the class number $h(d)$ of integral binary quadratic forms of discriminant d.

1.6 The geodesic flow on the hyperbolic plane

In this section, we state some important results about the geodesic flow on a $d-$dimensional Riemannian manifold Q of negative curvature. We assume that Q is compact (like the double torus associated with the regular octagon) or of finite volume (like any $X(N)$). The curvature does not need to be constant. Then the geodesic flow on Q is a so-called *Anosov flow* [8] in any dimension d. Anosov flows are examples of the most chaotic dynamical systems known. In particular, let $\{g^t\}$ the geodesic flow on Q. Then the following holds

- The flow $\{g^t\}$ is isomorphic to a Bernoulli flow; in particular $\{g^t\}$ is ergodic, mixing, has positive entropy and K-property;
- the flow $\{g^t\}$ is topologically mixing, in particular topologically transitive;
- periodic orbits of $\{g^t\}$ are dense in $M = SQ$; the number $P(T)$ of

[8] A dynamical system is called an Anosov system if every trajectory is uniformly completely hyperbolic and the constants C and λ below can be chosen independently of the point. A trajectory $\{S^t\}$ is completely uniformly hyperbolic if there exist subspaces $E^s(S^t(x))$ and $E^u(S^t(x))$ and constants $C > 0, \lambda, \mu$ such that

$$0 < \lambda < 1 < \mu$$

and for all t and $\tau \geq 0$, one has

$$T_{S^t(x)} = E^s(S^t(x)) \oplus E^u(S^t(x)) \oplus X(S^t(x))$$

$$dS^t E^s(x) = E^s(S^t x) \quad dS^t E^u(x) = E^u(S^t(x))$$

$$||dS^\tau v|| \geq C\lambda^\tau ||v|| \quad v \in E^s(S^t(x))$$

$$||dS^\tau v|| \leq C^{-1}\mu^\tau ||v|| \quad v \in E^u(S^t x)$$

$$\gamma(S^t(x)) \geq const.$$

where γ here is the angle between the subspaces $E^s(S^t(x))$ and $E^u(S^t(x))$, respectively called stable and unstable subspaces. $X(S^t(x))$ is the subspace generated by the flow (and invariant under it). In other words, the notion of (uniform) *hyperbolicity* means that the tangent space is split into a direct sum of three subspaces invariant under dS^t, where $dS^t|E^s$ is a contraction and $dS^t|E^s$ is an expansion. The presence of the unstable subspace is at the origin of the exponential divergence of nearby trajectories in phase space. The remaining subspace is neutral, in the sense that vectors lying in it may contract and expand but not too fast.

periodic orbits of period $\leq T$ is finite and

$$P(T) \sim \frac{e^{hT}}{hT}, \quad T \to \infty \tag{1.104}$$

where h is the *topological entropy* of the flow.

The last statement is due to G. A. Margulis [106]; it turns out that for a $d-$dimensional manifold Q of *constant* negative curvature $\mathcal{K}$, the topological entropy is given by

$$h = (d-1)\sqrt{-\mathcal{K}} \tag{1.105}$$

In this thesis, we are interested in the case $d = 2$, especially the modular surface $X(1) = \mathrm{PSL}(2, \mathbb{Z})\backslash\mathbb{H}$ for which $h = 1$ (we always put $R = 1$ in the hyperbolic metric, so $\mathcal{K} = -1/R^2 = -1$). Note that the geodesic flow on the *whole* hyperbolic plane is not ergodic, indeed it is integrable.

Note that the same asymptotic relation is valid considering *only* primitive periodic orbits (ppo) γ_0, that is

$$\sharp\{\gamma_0 |\, T(\gamma_0) \leq T\} \sim \frac{e^{hT}}{hT} \tag{1.106}$$

Now we can use the *Pesin formula*, which for strongly chaotic systems (in particular Anosov flows) gives the entropy as the sum of the positive Lypaunov exponents. Since on $\mathbb{H}$ there is only one Lyapunov exponent λ, $h = \lambda = v/R$ [9], where v is the speed and in our units $R = 1$. Thus $hT = vT/R = l/R$ with l the hyperbolic length of the orbit, and we can write down an asymptotic formula for the counting function of ppo γ_0 labelled by their lengths $l(\gamma_0)$

$$\mathcal{R}_0(l) = \sharp\{\gamma_0 |\, l(\gamma_0) \leq l\} \sim \frac{e^{l/R}}{l/R} \tag{1.107}$$

as $l \to \infty$. This asymptotic behavior can also be derived from the Selberg trace formula. Again, the counting function for the lengths of all orbits, primitive or repeated, has exactly the same exponential growth

$$\mathcal{R}(l) = \sharp\{\gamma |\, l(\gamma) \leq l\} \sim \frac{e^{l/R}}{l/R} \tag{1.108}$$

Indeed, an orbit of length l is either primitive or twice a primitive orbit of length $l/2$, etc, hence $\mathcal{R}(l) = \sum_{n=1}^{\infty} \mathcal{R}_0(l/n)$ and in this sum the first term $\mathcal{R}_0(l)$ exponentially dominates the others. This means that the exponential proliferation of longer primitive orbits overwhelms by far the increase brought by the iteration of shorter orbits.

1.6.1 Artin modular billiard

E. Artin [8] studied the case of a billiard flow inside the halved modular domain, that is the geodesic flow on $Q = \mathrm{PGL}(2,\mathbb{Z})\backslash\mathbb{H}$. Let us state Artin's main theorem. Let γ be a geodesic in Q and $\widetilde{\gamma}$ one of its liftings to $\mathbb{H}$. Denote by $\widetilde{x} = \widetilde{\gamma}(-\infty)$ and $\widetilde{y} = \widetilde{\gamma}(+\infty)$. Suppose that $\widetilde{x} > 0$ and $\widetilde{y} < 0$ and let $\widetilde{x} = [\widetilde{n}_1, \widetilde{n}_2, \ldots]$ and $-\widetilde{y} = [\widetilde{m}_1, \widetilde{m}_2, \ldots]$ be the continued fraction expansions of $\widetilde{x}$ and $\widetilde{y}$ with $n_i, m_i > 0$. Now let $\widehat{\gamma}$ another lifting of γ in $\mathbb{H}$, $\widehat{x} = \widehat{\gamma}(-\infty)$ and $\widehat{y} = \widehat{\gamma}(+\infty)$. As before, suppose that $\widehat{x} > 0$, $\widehat{y} < 0$ and $\widehat{x} = [\widehat{n}_1, \widehat{n}_2, \ldots]$, $-\widehat{y} = [\widehat{m}_1, \widehat{m}_2, \ldots]$. Then $\widehat{\gamma} = g\widetilde{\gamma}$ for some $g \in \mathrm{PGL}(2,\mathbb{Z})$. Artin theorem says that $(\widetilde{x}, \widetilde{y})$ and $(\widehat{x}, \widehat{y})$ define the same geodesic in $M = SQ$ if and only if

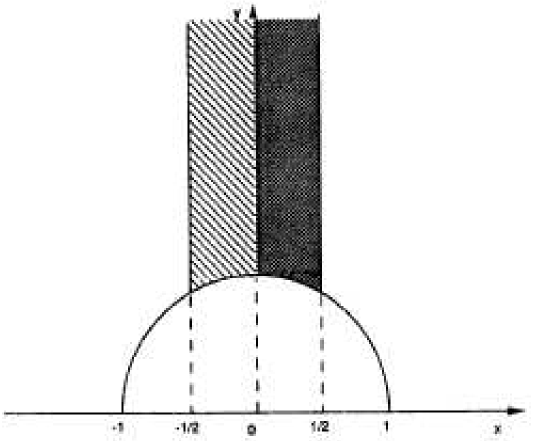


Figure 1.9: The figure represents the fundamental domain of $\mathrm{PSL}(2,\mathbb{Z})$ and $\mathrm{PGL}(2,\mathbb{Z})$ (one of the two hatched parts). Artin billiard corresponds to the geodesic flow inside one of the two fundamental domains for $\mathrm{PGL}(2,\mathbb{Z})$.

$$\sigma^k(\ldots, \widetilde{m}_2, \widetilde{m}_1, \widetilde{n}_1, \widetilde{n}_2, \ldots) = (\ldots, \widehat{m}_2, \widehat{m}_1, \widehat{n}_1, \widehat{n}_2, \ldots) \tag{1.109}$$

for some integer k, where σ is the shift in the space Σ of two-sided infinite sequences of positive integers. Thus we obtain a *coding* map $\psi : SQ \to \Sigma$.

Now it is possible to show that the ergodicity of the geodesic flow $\{T^t\}$ with respect to the Riemannian volume μ on SQ is equivalent to the ergodicity of σ with respect to induced measure $\psi_*\mu$. This, in turn, is equivalent to the ergodicity of σ in the space Σ_+ of one-sided infinite sequences of positive integers (Σ is the natural extension of Σ_+) with respect to the measure ν which is the projection of $\psi_*\mu$. The ergodicity of the latter can be studied with the help of the Gauss map

$$T(x) = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor \tag{1.110}$$

It is easy to see that if $x = [n_1, n_2, \ldots]$, then $T(x) = [n_2, n_3, \ldots]$ so that T is conjugate to the shift σ in Σ_+; let χ be the corresponding conjugacy map. The measure $\chi_*^{-1}\nu$ coincides with the Gauss measure on $[0,1]$, whose density $\frac{dx}{\ln 2\,(1+x)}$ is $T-$invariant. As we said, this measure is ergodic with respect to T, and consequently the geodesic flow on Q is ergodic.

We have just seen that there is a deep mathematical relation between the geodesic flow on a fundamental domain of $\mathrm{PGL}(2,\mathbb{Z})$ and the Gauss map. We will see in Part II that this relation can be *physically* realized in the dynamics of general relativity close to the cosmological singularity, in particular the Gauss map will describe the asymptotic evolution of a Bianchi IX universe whereas the geodesic flow for $\mathrm{PGL}(2,\mathbb{Z})$ will describe the asymptotic evolution of a generic inhomogeneous universe. From the Artin theorem, perhaps it is not so surprising that the asymptotic evolution occurs inside the fundamental Weyl chamber of the hyperbolic Kac-Moody algebra $\mathrm{HA}_1^{(1)}$, whose Weyl group is precisely $\mathrm{PGL}(2,\mathbb{Z})$. Finally, all the numerical simulations (see later on) suggest the behavior of the generic singularity is captured by a Bianchi IX cosmological model: as we have just seen, this is supported also by the Artin theorem.

As we mentioned, for a free motion generated by discrete groups the periodic orbits correspond to conjugacy classes of hyperbolic transformations; this means that if A and B are two hyperbolic transformations of the discrete group, then A and BAB^{-1} define the same periodic orbit. The length l of this orbit can be expressed in terms of the trace of these matrices

$$2\cosh\frac{l}{2} = \mathrm{Tr}A \tag{1.111}$$

for positive hyperbolic transformations. If the group contains also hyperbolic reflections, as in the case of $\mathrm{PGL}(2,\mathbb{Z})$, then a different formula holds for negative hyperbolic transformations A

$$2\sinh\frac{l}{2} = \mathrm{Tr}A \tag{1.112}$$

Now, the trace of a hyperbolic transformation fixes the hyperbolic length of the orbit, but there may be different orbits with the same length corresponding to hyperbolic transformations not conjugate. Thus it makes sense to speak of the *degeneracy* of lengths of periodic orbits. Let us denote by $g(l)$ the multiplicity of periodic orbits with fixed length l. We consider now the case of an *arithmetic* group, like $\mathrm{PSL}(2,\mathbb{Z})$ or $\mathrm{PGL}(2,\mathbb{Z})$; then [25] the mean multiplicity for an arbitrary arithmetic group is

$$\langle g\rangle = \frac{2}{C_0}\frac{e^{l/2}}{l} \tag{1.113}$$

where the constant C_0 depends on the group. Thus, for an arithmetic system, we have an exponential degeneration for the multiplicities of the lengths of periodic orbits.

For generic systems, one usually does not expect such a degeneracy, except for eventual symmetries of the model. For example, systems with time-reversal invariance have in general mean multiplicity equal to 2, which corresponds to the same orbit run in two different directions. Arithmetic systems are exceptional, since they display exponentially large multiplicity of periodic orbits; in this case, one speaks of *arithmetical chaos.* Note, however, that for any Riemann surface this degeneracy is *unbounded* [126], but the degeneracies connected with this theorem are much smaller then exponential. It is this exponential degeneracy which made me think of a possible link with the (multiplicities of the) roots of a hyperbolic Kac-Moody algebra, which we explore a little bit in the following chapter.

Let us also observe that for the large degeneracy of length of periodic orbits in arithmetical systems seems to have no importance in the classical dynamics: they are chaotic as any other model on compact negatively curved surfaces. Does the arithmetic property characterizes the quantum behavior of the system? The answer to this question is affirmative and the deep reason for that is the existence of the Hecke operators for arithmetic groups: this leads to anomalous statistics (see Appendix B) and to the arithmetic quantum unique ergodicity theorem, which we briefly describe in the following section.

1.7 Quantum Unique Ergodicity

Figure 1.10: The tree of quantum chaos as drawn by A. Terras [146]: in this thesis we explain only a part of this figure.

Let $Q = \Gamma\backslash\mathbb{H}$ be a hyperbolic manifold and $\mathcal{F}$ a fundamental domain for Γ. Suppose first Q is compact, and let $\{\phi_n\}_{n=0}^{\infty}$ be the eigenfunctions of the Laplacian, which form an orthonormal basis for $L^2(Q,\mu)$. Then a classical result by H. Weyl for the Dirichlet problem for the Laplacian Δ says that

$$\sharp\{n : \lambda_n \leq N\} \sim \frac{\mu(\mathcal{F})}{4\pi} N\,, \quad n \to \infty \tag{1.114}$$

where the eigenvalues λ_n are counted with their multiplicity. If the geodesic flow on Q is ergodic, then one can go further and prove the *quantum ergodicity theorem*, originally due to A. I. Shnirelman. The theorem states that if the flow is ergodic, then there exist a density one sequence of integers $(j_k)_{k\in\mathbb{N}}$ such that for each Borel subset $B \subset \mathcal{F}$, one has

$$\lim_{k\to\infty} \int_B |\phi_{j_k}(z)|^2 \, d\mu(z) = \frac{\mu(B)}{\mu(\mathcal{F})} \tag{1.115}$$

Let us explain in more details. First, a strictly monotonic sequence of integers $(j_k)_{k\in\mathbb{N}}$ is said to be of *density* a $(0 \leq a \leq 1)$ if the following is true

$$\lim_{J\to\infty} \frac{1}{J} \sharp\{k \in \mathbb{N} |\, j_k \leq J\} = a \tag{1.116}$$

So saying that a sequence is density a means that the fraction of integers that belong to the sequence equals a. Now, having normalized to 1 the eigenfunctions, in particular $\int_{\mathcal{F}} |\phi_{j_k}|^2 d\mu(z) = 1$, it is cleat that each eigenfunction defines a probability measure μ_j with density $|\phi_j|^2$ on $\mathcal{F}$

$$\mu_j(B) = \int_B |\phi_j|^2 d\mu(z) \tag{1.117}$$

The Shnirelman theorem says that there exist a density one sequence of integers j_k so that the measure μ_{j_k} converge to the normalized Lebesgue measure on $\mathcal{F}$, the standard hyperbolic measure. Since we can write

$$\frac{\mu(B)}{\mu(\mathcal{F})} = \int_B \frac{1}{\mu(\mathcal{F})} d\mu(z) \tag{1.118}$$

the Shnirelman theorem, deleting the integrals, can also be re-written as

$$\lim_{k\to\infty} |\phi_{j_k}|^2 = \frac{1}{\mu(\mathcal{F})} \tag{1.119}$$

This is statement of *equidistribution*, that is one often says that the eigenfunctions equidistribute because the probability densities $|\phi_{j_k}|^2$ tendo to a constant, independent of any point $z \in \mathcal{F}$. Thus, for non exceptional λ_n, the mass of ϕ_n can never localize to, say, just a finite number of closed geodesics on $\Gamma\backslash\mathbb{H}$.

This theorem has been improved by Y. Colin de Verdiere [32] and S.Zelditch [160], in particular Zelditch [161] showed that eigenfunctions still equidistribute for non-compact manifolds with $\Gamma = \mathrm{PSL}(2,\mathbb{Z})$ or its congruence subgroups.

The presence of an exceptional set is clearly a bit troubling. In fact, the so-called *scarring effect* has been observed in stadium-like domains in $\mathbb{R}^2$. It happens that for numerous n, the topography of ϕ_n is found to contain clear ridges of mass of scars, situated roughly along what would appear to be closed geodesics. The location of these scars change with n (see the book by Gutzwiller [67] and the paper by Heller [77]). Scars is what is left of periodic orbits.

The situation is different for hyperbolic arithmetic manifolds. Let us state the important results. Let $X = \Gamma\backslash\mathbb{H}$ a compact hyperbolic surface, with Γ a discrete compact subgroup of $\mathrm{PSL}(2,\mathbb{R})$, and SX the unit tangent bundle. As we said, the geodesic flow on SX is an Anosov flow and displays chaotic features. We want to address the question of the behavior of the eigenfunctions ϕ_n and eigenvalues λ_n of Δ, which is a quantization of the Hamiltonian generating the geodesic flow, in the semi-classical limit, i.e. $n \to \infty$. The central question is whether the ϕ_n's behave like random waves or if they display some localization related to classical trajectories. In the case of billiards in the Bunimovich stadium (which is a chaotic domain in $\mathbb{R}^2$), Heller found that certain states are enhanced on a finite union of periodic unstable orbits and he called this phenomenon scarring. In the case of hyperbolic manifolds, the numerical evidence points to eigenstates behaving like random waves, *although* the semi-classical limit is chaotic as well. Besides, regarding the statistics of the spectrum, one must distinguish between arithmetic and non-arithmetic manifolds.

As before, define the probability measures μ_j on X

$$d\mu_j = |\phi_j(z)|^2 \, d\text{vol}(z) \tag{1.120}$$

where $d\text{vol}(z)$ is the Riemannian volume element on X. It is well known that these probability measures give the probability density for finding a particle in the state ϕ_j at the point z. We say that ν is a *quantum limit* if it is the limit in the weak* topology of the sequence μ_j. The Shnirelman-Colin de Verdiere-Zelditch theorem says that if the geodesic flow is ergodic, then $\mu_j \to d\text{vol}(z)$ for *almost all* j. If fact, one can define an appropriate extension $\widetilde{\mu}_j$ of μ_j to the phase space T^*X and show that any limit $\widetilde{\nu}$ of $\widetilde{\mu}_j$ is invariant under the geodesic flow. From one side, this restricts the set of the possible quantum limits $\widetilde{\nu}$, but on the other side it is known that for chaotic systems the set of these invariant measures is large and complicated, as are its typical members. The simplest and most localized such measure is the arc-length measure supported on a union of periodic geodesics. The question is if these can occur as quantum limits and it is related of course to Heller's scarring in its strongest form. Following [131], we can say that a subsequence μ_{j_k} is said to *scar strongly* to a closed subset $S \subset X$ if $\mu_{j_k} \to \nu$ and $\emptyset \neq$ singsupp $\nu \subset S$. Now, if Γ is arithmetic, we said that there exist a commutative self-adjoint algebra of Hecke operators which commute with Δ. Hence we may assume that ϕ_j are also Hecke eigenfunctions of the Hecke operators. This is probably automatic since the spectrum of Δ is very probably simple (even in the case of $X = X(1)$ which is non-compact). The existence of Hecke operators allows to prove that scarring on closed geodesics is impossible for arithmetic surface X, i.e.: if X is an arithmetic hyperbolic surface, ν a quantum limit and σ the support of its singular part ν^s, and σ is contained in the union of a finite numbers of points and closed geodesics, then $\sigma = \emptyset$. This means that ν is absolutely continuous with respect to $d$vol. This a first step towards proving that the μ_j are individually equidistributed. In fact, we have the following important *Quantum Unique Ergodicity conjecture* due to Rudnick and Sarnak

QUE Conjecture [131] *Let X be a compact manifold of negative curvature. Then the measures μ_j converge to dvol.*

If this conjecture is true, it is remarkable, because it asserts that at the quantum level and in the semiclassical limit, there is no manifestation of chaos from this point of view. In particular, one would have quantum unique ergodicity, that is only one possible quantum limit, while classical unique ergodicity, i.e. uniqueness of the invariant measure for the Hamiltonian flow, is never satisfied for chaotic systems.

The proof of the conjecture for arbitrary manifolds (i.e. not necessarily arithmetic) is still out of reach, but progress has been made for arithmetic hyperbolic surfaces thanks to E. Lindenstrauss [102]. He has shown, using Ratner's theorems on unipotent flows, that for a compact arithmetic quotient X the quantum unique ergodicity conjecture is true. Moreover, the conjecture is essentially proven also in the case of the modular surface $X(1)$.

To conclude, we can say that for the modular surface (which is the subject of this thesis) there is no scarring.

1.8 Notes and Comments on Chapter 1

The literature on (integrable or chaotic) dynamical systems is now huge. A very good reference for the theory of integrable systems is the classical book by V. I. Arnold [5]. The discovers of chaos were H. Poincaré (the influence of Poincaré is very deep also today, although often forgotten: there are still a lot of published or unpublished works due to the French mathematician that await continuation by the next generation of mathematicians, see [6]) and J. Hadamard[9].

[9] "...each stable trajectory can be transformed, by an infinitely small variation in the initial conditions, into a completely unstable trajectory extending to infinity, or, more generally, into a trajectory of any of the types given in the general discussion: for example, into a trajectory asymptotic to a closed geodesic."

Remember that given a dynamical systems

$$\dot{x} = F(x) \quad x = (x_1, \dots, x_n) \tag{1.121}$$

any solution of this differential equation corresponds to a trajectory (a *flow*) in some $n-$dimensional phase space. An old program dating back to Poincaré seeks to determine the general behavior of the solution $x(t)$ as t goes to ∞, but the problem is solved only when $n = 2$. In this latter case, trajectories can not intersect in phase space and the two-dimensional topology limits the asymptotic behavior to two generic states: either trajectories approach a stable attractor (stationary solution) or a limit cycle (periodic solution) after an infinite time. In the first case the attractor has dimension 0 (a point), in the second it has dimension 1 (a closed curve). When $n > 2$, trajectories can cross and develop complicated knotted configurations without actually intersecting. The detailed behavior of the solution is not known for $n \geq 3$. A theorem of D. Ruelle and F. Takens [134] says that in the higher dimensional case the fate of generic trajectories is to approach a non-empty, finite (measure) region of the phase space, containing neither attracting points nor limit cycles and in which neighbouring trajectories rapidly diverge from each other when evolved backwards or forwards in time. Trajectories will enter this attracting set and then wander around it in chaotic fashion. Ruelle and Takens termed this set a *strange attractor.* A strange attractor is defined to be a set which attracts all nearby solution trajectories and which has the structure of $M \times C$ where M is a smooth manifold and C is a Cantor set with non-integral Hausdorff dimension. Such a fractal structure emerges in the limit of Einstein's theory to the Big-Bang singularity. It is well known that this work by Ruelle and Takens opened the door to the modern theory of turbulence: in fact they showed that Landaus'theory of turbulence was wrong, because it was mainly based on the assumption that turbulence or chaos derived from the excitation of a large number of degrees of freedom. The common belief was that the presence of random behavior in a deterministic system derived from prescribing random initial data or exciting a large number of degrees of freedom. These are sufficient conditions for the onset of chaos, but they are *not* necessary. There are very simple dynamical systems which are intrinsically chaotic, for example iterated maps of the interval, with regular initial data, no stochastic forcing and a minimal number of degree of freedom.

Hadamard's great achievement was that he could prove that all trajectories in his system (the free motion on a compact negatively curved manifold) are unstable and that neighboring trajectories diverge in time at a rate $e^{\omega t}$ where $\omega = \sqrt{2E/mR^2}$ is the Lyapunov exponent (R is the length scale fixed by the constant negative curvature, $\mathcal{K} = -1/R^2$). Thus he was the first who could show that the long-time behavior of a dynamical system can be very sensitive to the initial conditions and therefore unpredictable, even though the system is governed by deterministic laws (Newton's equations). Today this sensitivity to initial data is recognized as the most striking property of systems with *deterministic chaos* (an expression coined by Chirikov, Zaslavskii, and Ford [143]). It appears that Hadamard should be considered the true discover of chaos, without ever using the word chaos in his works.

The word *chaos* was used much later with the discovery of Lorenz of the "butterfly effect" (recently S. Luzzatto et al have shown that the Lorenz attractor is mixing). Classical texts on ergodic theory are the books by V.I. Arnold and A. Avez [4], I. P. Cornfeld, S. V. Fomin and Ya. G. Sinai [33] and by P. Walters [156]; the book edited by L. A. Bunimovich, Ya. P. Pesin and Ya. G. Sinai [142] is very complete and full of information (but no details of proofs) and is highly recommended. The book edited by C. Series et al [13] covers many topics related to this thesis (hyperbolic geometry, ergodic theory, symbolic dynamics). Modern texts are the book by B. Hasselblatt and A. Katok [69], which is a kind of encyclopedic treatise, the book by M. Pollicott and M. Yuri [124]. Finally the book in progress by T. Ward and M. Einsiedler [45] is a very promising book and it should contain also topics about unipotent flows, equidistribution theory, quantum unique ergodicity etc.

Good and quick introductions to the subject are the reviews by J. P. Eckmann and D. Ruelle [44], and many lecture notes and papers by L. S. Young available on her web site

`http://www.cims.nyu.edu/~lsy/`

About ten years ago, she developed the notion of Markov towers, which are more powerful of Markov partitions in deriving statistical properties of dynamical systems.

The notion of entropy was introduced by A. N. Kolmogorov (and refined by Ya. G. Sinai) after the work of Shannon on information theory. The metric entropy was the first example of a numerical invariant for dynamical systems and it allowed different dynamical systems to be distinguished. In particular the American mathematician D. Ornestein showed that Bernoulli shifts are isomorphic iff they have the same metric entropy, a news which arrived as a shock for the Russian school [90].

Many people claim the chaos theory is a kind of a new science, a new paradigm and that its implications will be as important as the revolutions of quantum mechanics and general relativity have been. This is very likely true. I suggest other easy readings on chaos I found very useful: the book by D. Ruelle [133], the classical text by J. Gleick [61] (a *must*) and for Italian readers [10].

For billiards the book by S. Tabachnikov [144] is a good introduction (it mostly deals with for integrable billiards and classical geometry): this book and a previous one on billiards can be found on his web site

`http://www.math.psu.edu/tabachni/`

For the reader who wants to learn all the details of the proofs (especially concerning chaotic billiards) we suggest the new book by N. Chernov and R. Markarian [28]. The first chapters of this book and a previous one on ergodic billiards can be found on Chernov's web site

`http://www.math.uab.edu/chernov/`

The talk by Chernov "Chaotic Billiards" at the MSRI is very good:

`http://www.msri.org/communications/vmath/VMathVideos/VideoInfo/2958/show_video`

We nave not even mentioned the first example of chaotic billiard, which is also the most famous one, the *Sinai billiard.* It has a convex scatterer, thus it is a *dispersive* billiard (see [55] for an introduction). After this seminal work of Sinai in 1963, only a few years ago N. Simanyi has shown that Boltzmann's ergodic hypothesis is true for typical hard ball systems and typical hard disks systems (for almost every initial configuration!). This is a great step in proving Boltzmann's ergodic hypothesis [26] in full generality. Recently, G. Gallavotti has suggested to extend the ergodic hypothesis to the *chaotic hypothesis* [56], that is to regard chaotic dynamical systems not only ergodic, but fundamentally hyperbolic (say Anosov).

For quantum chaos standard books are the classical book by M. Gutzwiller [67] (a very informative book), a volume of École de Physique Les Houches [59], the book by K. Nakamura [119] and the book by F. Haake [68] (and all the references inside all of them). Sir M. Berry has *all* his papers on his web site, also the very first ones

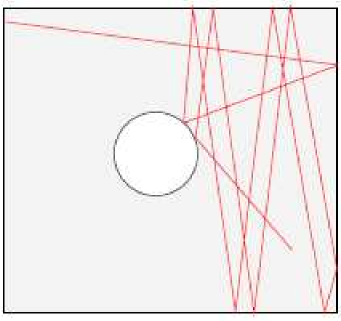


Figure 1.11: A typical Sinai billiard

`http://www.phy.bris.ac.uk/people/berry_mv/index.html`

The old review by N. L. Balas and A. Voros [9] is excellent, but it does not contain the phenomenon of arithmetical chaos which is instead described for mathematicians in [136] by P. Sarnak (who also coined the term arithmetical chaos) and for physicists by E. B. Bogomolny, B. Georget, M. J. Giannoni and C. Schmit [25]. The problem of scarred states in quantum mechanics was faced for the first time in a paper by E. J. Heller [77], see also the book by Gutzwiller; results on the absence of scarred states for the situation we are interested in were derived from P. Sarnak et al in a series of papers [103]-[104], [131]. The paper by D. Hejhal [76] on the topography of Maass waveforms is also useful.

The bibliography on the Selberg trace formula, its applications and and its generalizations à la Arthur [7] is huge too. The original paper by A. Selberg is [141]; the Selberg trace formula was then interpreted as a sum over periodic geodesics by Huber [79] (in German). We suggest the classical treatise by D. Hejhal in two volumes [75], the books by J. Fischer [52], the books by H. Iwaniec [85]-[86], the book by A. Terras [145], the paper by H. McKean [111] and especially the one by D. Hejhal [74].

The Schnirelman theorem appears in [140], but see also the papers by Y. Colin de Verdiére [32] and S. Zelditch [160].

Chapter 2

Kac-Moody Algebras

> It is a well kept secret that the theory of Kac-Moody algebras has been a disaster.
>
> *The idea of locality*
> V. G. Kac

In this chapter we review the theory of Kac-Moody algebras (briefly touching also Borcherds algebras), focusing our attention on the hyperbolic Kac-Moody algebras $\mathrm{HA}_1^{(1)}$ and E_{10}, the canonical hyperbolic extensions of A_1 and E_8. We state a theorem which codes a part of the imaginary root lattice of $\mathrm{HA}_1^{(1)}$ in terms of the periodic orbits on the modular surface $X(1)$ =PSL$(2,\mathbb{Z})\backslash\mathbb{H}$; this is possible because the positive Weyl group of $\mathrm{HA}_1^{(1)}$ is precisely PSL$(2,\mathbb{Z})$. We speculate also on a new possible interpretation of the Selberg trace formula and the Selberg zeta-function for PSL$(2,\mathbb{Z})$ as a sum over the root lattice of the hyperbolic Kac-Moody algebra $\mathrm{HA}_1^{(1)}$.

The easiest way to think about Kac-Moody algebras is to consider them as generalizations of classical simple Lie algebras. It is known that any complex finite-dimensional simple Lie algebra can be put in the following form (the so-called *Weyl-Cartan* basis)

$$[H_i, H_j] = 0 \quad [H_i, E_\alpha] = \alpha^{(i)}\, E_\alpha \tag{2.1}$$

$$[E_\alpha, E_{-\alpha}] = \sum_i \alpha^{(i)}\, H_i$$

$$[E_\alpha, E_\beta] = N_{\alpha\beta}\, E_{\alpha+\beta} \text{ if } \alpha+\beta \text{ is a root} \neq 0$$

The numbers appearing in the right hand side in the previous formulae can be chosen to be all integer by a change of basis, the result being the so-called *Serre-Chevalley* form:

$$[h_i, h_j] = 0 \quad [h_i, e_j] = a_{ij}\, e_j \quad [h_i, f_j] = -a_{ij}\, f_j \tag{2.2}$$

$$[e_i, f_j] = \delta_{ij}\, h_i$$

where a_{ij} is called the *Cartan matrix* of the algebra and the other commutators are given by the Serre-Chevalley relations: (ad $e_i)^{1-a_{ij}}\,(e_j) = 0$, (ad $f_i)^{1-a_{ij}}\,(f_j) = 0$ for $i \neq j$.

It is also true also the viceversa: every definite positive Cartan matrix A (in the sense that every principal minor is > 0) defines a complex finite-dimensional simple Lie algebra (Serre theorem). This is the starting point to search for generalizations of the classical Lie algebras.

2.1 Overview of Kac-Moody Algebras

At the end of the '60s, V. G. Kac and R. V. Moody independently and for different reasons extended the Serre-Chevalley result to the case of a more general matrix, considering in particular the case of semi-positive definite and indefinite matrices. One obtains infinite-dimensional algebras to which some of the finite-dimensional structure theory can be applied.

In this section we review basic facts about Kac-Moody algebras, for more details see the book by Kac [87] and the last section of this chapter.

Let I be a finite set let $A = a_{ij}$ be a *generalized Cartan matrix* [1], that is a $|I| \times |I|$ matrix subject to the conditions

$$a_{ii} = 2\,, \quad a_{ij} \text{ is a non-positive integer for } i \neq j\,, \quad a_{ij} = 0 \Leftrightarrow a_{ji} = 0\,. \tag{2.3}$$

We say that A is *indecomposable* when it is not possible to put A in a diagonal block form by reordering the set I (otherwise the Kac-Moody algebra $\mathfrak{g}(A)$, to be defined below, decomposes into a direct sum of Kac-Moody algebras associated to the indecomposable components of A) and *symmetrizable* when there exists an invertible diagonal matrix D such that DA is symmetric.

[1] In the following we will avoid to use expressions like generalized Cartan matrix or Borcherds-Cartan matrix and just say Cartan matrix, the type of algebra will be clear from the context.

Let $\mathfrak{h}$ be a complex vector space whose dimension is $|I|$ + corank A, and $\mathfrak{h}^*$ its dual. Then there exist linearly independent indexed sets

$$\Pi := \{\alpha_i\} \subset \mathfrak{h}^* \,, \quad \text{and } \Pi^\vee := \{h_i\} \subset \mathfrak{h} \,, \tag{2.4}$$

such that $\alpha_j\,(h_i) = a_{ij}$ $(i, j \in I)$. The α_i (h_i) are called *simple roots* (*dual simple roots*). The sets Π and $\Pi^\vee$ are uniquely determined by A up to isomorphism.

Then the *Kac-Moody algebra* $\mathfrak{g}(A)$ associated to A is the complex Lie algebra generated by $\mathfrak{h} \cup \{e_i, f_i\}$ with defining relations:

$$[e_i, f_j] = \delta_{ij}\, h_i \,, \quad [h, h'] = 0 \text{ for } h, h' \in \mathfrak{h} \tag{2.5}$$

$$[h, e_i] = \alpha_i(h)\, e_i \,, \quad [h, f_i] = -\alpha_i(h)\, f_i$$

$$(\text{ad } e_i)^{1-\,a_{ij}}\, e_j = 0 \,, \quad (\text{ad } f_i)^{1-\,a_{ij}}\, f_j = 0 \;\text{ for } i \neq j \,.$$

$\mathfrak{h}$ is the maximal abelian subalgebra of $\mathfrak{g}(A)$ (*Cartan subalgebra*). The derived subalgebra $\mathfrak{g}'(A) := [\mathfrak{g}(A), \mathfrak{g}(A)]$ is generated by the elements e_i, f_i and $\mathfrak{g}(A) = \mathfrak{g}'\,(A) + \mathfrak{h}$. The center of $\mathfrak{g}(A)$ is $\mathfrak{c} := \{h \in \mathfrak{h} |\, \alpha_i(h) = 0 \text{ for all } i \in I\}$ and has dimension corank A. The order $|I|$ of the Cartan matrix is also called the *rank* of the algebra, not to be confused with the rank of the matrix A.
Remark: Usually the Kac-Moody algebra is defined to be the quotient of our $\mathfrak{g}(A)$ by the sum of all ideals intersecting $\mathfrak{h}$ trivially. The two definitions coincide when A is symmetrizable (because any ideal of $\mathfrak{g}(A)$ either contains $\mathfrak{g}'\,(A)$ or is contained in $\mathfrak{c}$ provided that A is symmetrizable and indecomposable). We will only be concerned with symmetrizable (and indecomposable) matrices, so we prefer this definition.

If we denote by $\mathfrak{n}_+$ $(\mathfrak{n}_-)$ the subalgebra of $\mathfrak{g}(A)$ generated by $\{e_i\}$ $(\{f_i\})$, we obtain the vector space decomposition (*triangular decomposition*)

$$\mathfrak{g}(A) = \mathfrak{n}_- \oplus \mathfrak{h} \oplus \mathfrak{n}_+ \,. \tag{2.6}$$

Furthermore, we have the root space decomposition of $\mathfrak{g}(A)$ with respect to its Cartan subalgebra:

$$\mathfrak{g}(A) = \bigoplus_{\alpha \in\, \mathfrak{h}^*} \mathfrak{g}_\alpha \tag{2.7}$$

where $\mathfrak{g}_\alpha := \{x \in \mathfrak{g}(A) | [h, x] = \alpha(h)\, x \text{ for all } h \in \mathfrak{h}\}$ is the *root space.* If $\alpha \neq 0$ and $\mathfrak{g}_\alpha \neq 0$, then α is called a *root* of *multiplicity* mult $\alpha :=$ dim

$\mathfrak{g}_\alpha$ (this is always finite). Note that $\pm\alpha_i$ are roots of multiplicity 1 since $\mathfrak{g}_{\alpha_i} = \mathbb{C}\, e_i$ and $\mathfrak{g}_{-\alpha_i} = \mathbb{C}\, f_i$. Let us denote by Δ the set of all roots.

The $\mathbb{Z}-$span Q of the set Π is called the *root lattice*, $Q = \bigoplus_{i\in I} \mathbb{Z}\, \alpha_i \supset \Delta$. Any lattice vector $\alpha = \sum_i k_i\, \alpha_i \in Q$ (root or not) has a *height* given by $\mathrm{ht}\, \alpha := \sum_i k_i$. With $Q_+ := \sum_i \mathbb{Z}_{\geq 0}\; \alpha_i$, we can introduce a partial ordering on $\mathfrak{h}^*$ by

$$\lambda \geq \mu \Leftrightarrow \lambda - \mu \in Q_+\, . \tag{2.8}$$

In this way we can define the *positive roots* as the ones in the set $\Delta_+ = \Delta \cap Q_+$. Then the *negative roots* belong to the set $\Delta_- = -\Delta_+$ and we have $\Delta = \Delta_+ \cup \Delta_-$ (disjoint union).

Given a $\mathfrak{g}(A)$, it is possible to define the *dual Kac-Moody algebra* $\mathfrak{g}(A^{tr})$. We will identify the Cartan subalgebra $\mathfrak{h}^\vee$ of $\mathfrak{g}(A^{tr})$ with $\mathfrak{h}^*$, so that the set of simple roots of $\mathfrak{g}(A^{tr})$ (resp. dual simple roots) is identified with $\Pi^\vee$ (resp. Π). Notions like $Q^\vee$, $\Delta^\vee$ etc are defined in an obvious way.

For each $i \in I$, let us define the *fundamental reflection* $r_i \in \mathrm{GL}(\mathfrak{h})$ by

$$r_i\,(h) := h - \alpha_i(h)\, h_i \tag{2.9}$$

with $h \in \mathfrak{h}$. Note that r_i operates *contragrediently* in $\mathfrak{h}^*$, i.e. $r_i\,(\alpha) := \alpha - \alpha(h_i)\, \alpha_i$. The *Weyl group* W is the subgroup of $\mathrm{GL}(\mathfrak{h})$ generated by the fundamental reflections r_i; we can identify r_i with $r_i^\vee$ and W with $W^\vee$ via the contragredient action. The operators (ad e_i) and (ad f_i) are locally nilpotent and $\widetilde{r}_i :=$ (exp ad e_i)(exp ad $(-f_i)$)(exp ad e_i) $\in$ Aut $\mathfrak{g}(A)$ and satisfies

$$\widetilde{r}_i(\mathfrak{g}_\alpha) = \mathfrak{g}_{\,r_i(\alpha)} \quad \text{and } \widetilde{r}_i|_{\mathfrak{h}} = r_i. \tag{2.10}$$

In particular, the root system Δ is W-invariant, mult $\alpha =$ mult $w(\alpha)$ for every $w \in W$ and r_i permutes the set $\Delta_+/\{\alpha_i\}$.

The Weyl group allows to distinguish roots into *real roots* and *imaginary roots.* A root is real if it is W-equivalent to a simple root (so its multiplicity is 1), otherwise it is imaginary. Thus we have $\Delta = \Delta^{re} \cup \Delta^{im}$. If $\alpha \in \Delta^{re}$, then $w(\alpha) = \alpha_i$ for some i and we can define the *dual root* $\alpha^\vee \in \Delta^\vee$

$$\alpha^\vee = w^{-1}(h_i) \in \mathfrak{h} \tag{2.11}$$

It is also possible to define reflections with respect to real roots: if $\alpha \in \Delta_+^{re}$ we define

$$r_\alpha(h) = h - \alpha(h)\alpha \tag{2.12}$$

so that $r_\alpha^2 = 1$, $r_\alpha(\beta) = \beta - \beta(\alpha^\vee)\alpha$ with $\beta \in \mathfrak{h}^*$, $wr_\alpha w^{-1} = r_{w(\alpha)}$ and $r_{\alpha_i} = r_i$.

The symmetrizability of the matrix A is a very important condition, being equivalent to the existence of a non-degenerate $\mathfrak{g}(A)$-invariant symmetric bilinear form $(,)$ on $\mathfrak{g}(A)$. The restriction of this form to $\mathfrak{h}$ is non-degenerate and W-invariant. Conversely, any non-degenerate W-invariant symmetric bilinear form $(,)$ on $\mathfrak{h}$ can be uniquely extended to a non-degenerate $\mathfrak{g}(A)$-invariant symmetric bilinear form $(,)$ on $\mathfrak{g}(A)$. We will always assume that A is symmetrizable. In this case we can choose a non-degenerate invariant bilinear form $(,)$ on $\mathfrak{g}(A)$ such that (h_i, h_j) is positive rational for all $i \in I$ (*standard bilinear form*) and identify $\mathfrak{h}$ with $\mathfrak{h}^*$ via $(,)$. A real root is then described by the condition $(\alpha, \alpha) > 0$, an imaginary one by $(\alpha, \alpha) \leq 0$. Furthermore, the generalized Cartan matrix can be written in the form

$$a_{ij} = 2\,\frac{(\alpha_i, \alpha_j)}{(\alpha_i, \alpha_i)} \tag{2.13}$$

as in the classical case, and the Weyl reflections are

$$r_\alpha(\lambda) = \lambda - (\lambda, \alpha^\vee)\alpha = \lambda - 2\frac{(\lambda, \alpha)}{(\alpha, \alpha)}\alpha \tag{2.14}$$

for $\alpha \in \Delta^{re}$, and for any root $\alpha \in \Delta$ we have $[\mathfrak{g}_\alpha, \mathfrak{g}_{-\alpha}] = \mathbb{C}\alpha$ which gives a non-degenerate pairing of $\mathfrak{g}_\alpha$ and $\mathfrak{g}_{-\alpha}$.

Let us set $\mathfrak{h}_\mathbb{R} = \{h \in \mathfrak{h} |\, \alpha_i(h) \in \mathbb{R} \text{ for all } i \in I\}$. This is a W-stable real subspace of $\mathfrak{h}$. We can define $\mathfrak{h}^*_\mathbb{R}$ similarly. The *fundamental Weyl chamber* is the set $C \subset \mathfrak{h}_\mathbb{R}$ defined by

$$C = \{h \in \mathfrak{h}_\mathbb{R} |\, \alpha_i(h) \geq 0\} \tag{2.15}$$

each $w(C)$ is a *chamber*, whose union gives the *Tits cone* $X = \cup_{w \in W} w(C)$. Also define the *imaginary cone* Z to be the closure of the convex hull of $\{0\} \cup \Delta^{im}_+$.

In particular, the Tits cone is a convex cone and the set C is a fundamental domain for the action of W on X, i.e. any orbit $W \cdot h$ with $h \in X$ intersect C in exactly one point; finally W operates simply transitively on chambers. In the finite-dimensional case, one has the following equivalent properties

$$|W| < \infty \Leftrightarrow X = \mathfrak{h}_\mathbb{R} \Leftrightarrow |\Delta| < \infty \tag{2.16}$$

The Weyl group of a Kac-Moody algebra is always a *Coxeter group* [2]. Yet, the Coxeter exponents of the Weyl group are *restricted* to the following values

$a_{ij}a_{ji}$	0	1	2	3	≥ 4
m_{ij}	2	3	4	6	∞

More on this will be said at the end of this chapter. Also note that different Kac-Moody algebras, say $\mathfrak{g}(A_1)$ and $\mathfrak{g}(A_1)$, may have the same Weyl group $W(A_1) \cong W(A_2)$. Since this is Coxeter group, it is identified by its Coxeter exponents. It may happen that $a_{ij}^{(1)}a_{ji}^{(1)} = a_{ij}^{(2)}a_{ij}^{(2)} = m_{ij}$, so the two Weyl groups are isomorphic.

Let us finish this section with some remarks about the subalgebras of Kac-Moody algebras. The structure of regular and singular subalgebras of (genuine) Kac-Moody algebras is much more involved compared to the finite-dimensional case, where a folding technique developed by Dynkin allows to find all the subalgebras.

Generally, indefinite Kac-Moody algebras contain infinitely many non-isomorphic subalgebras which are of indefinite type, of equal or less rank, and each of these subalgebras is infinite-dimensional too! In particular Feingold and Nicolai [51] found a very simple general construction which allows to locate the simple root system of an indefinite algebra inside the root system of a given indefinite Kac-Moody algebra; the resulting subalgebra may not be hyperbolic even if the starting algebra is hyperbolic. Their work also shows that indefinite algebras contain even Borcherds subalgebras, perhaps a surprising fact. More on this will be said in the sections dedicated to $\mathrm{HA}_1^{(1)}$ and E_{10}.

Embeddings of Borcherds algebras into Kac-Moody algebras were studied by S. Naito [120] who was the first to find locate Borcherds algebras in Kac-Moody algebras.

[2] A *Coxeter group* is a discrete group generated by n reflections r_i with the following defining relations

$$r_i^2 = 1\,, \quad (r_i r_j)^{m_{ij}} = 1 \tag{2.17}$$

where the *Coxeter exponents* m_{ij} are positive integers or ∞ (in this case we put $x^\infty = 1$). Every Coxeter group defines a Coxeter polytope, which is the polyhedron whose faces are pointwise-fixed by the fundamental reflections. When the Coxeter group is the Weyl group of a Kac-Moody algebra, this polyhedron corresponds to the *dual Weyl chamber* $C^\vee$, which is out main object of interest, i.e. the fundamental Weyl chamber of the dual Kac-Moody algebra. In the following, we simply use the term Weyl chamber to mean the polytope whose faces are orthogonal (with respect to the metric given by the Cartan matrix) to the simple roots of the algebra; thus it is contained in $\mathfrak{h}_{\mathbb{R}}^*$.

Let us define the *compact form* $\mathfrak{k}(A)$ of a Kac-Moody algebra $\mathfrak{g}(A)$. Denote by ω_0 the antilinear automorphism (*compact involution*) of $\mathfrak{g}(A)$

$$\omega_0(e_i) = -f_i\,, \omega_0(f_i) = -e_i\,, \omega_0(h) = -h \text{ for all } h \in \mathfrak{h}_{\mathbb{R}} \tag{2.18}$$

Then $\mathfrak{k}(A)$ is defined as the fixed point set of ω_0 and it is a real Lie algebra whose complexification gives $\mathfrak{g}(A)$. This definition of the compact form coincides with the usual one in the finite-dimensional case. Note also that this algebra is *not* necessarily a Kac-Moody algebra. A partial classification of compact real forms is given in [112] for affine Kac-Moody algebras and in [113] for hyperbolic Kac-Moody algebras. These papers are important in view of the fact in many cases of physical interests one has to consider split and not-split real forms of hyperbolic algebras to describe certain supergravity billiards [78].

2.1.1 On Kac-Moody groups: A remark on terminology

It is known that working with infinite dimensional Lie groups is a hard task (see [114]). For example, there is a good theory parallel to the theory of finite-dimensional Lie groups for infinite-dimensional Lie groups modelled on Banach algebras. But for more general topological vector spaces, these is no such a theory: most of theorems about Lie groups do not hold. It is possible to give numerous examples of Lie algebras which do not correspond to any Lie group and of Lie groups whose exponential maps are not locally bijective (which is really a bad behavior for physical applications). For example, if X is a real finite-dimensional compact smooth manifold, the group Diff(X) of all smooth diffeomorphisms $X \to X$ is a Lie group, whose Lie algebra is Vect(X), the vector space of all smooth vector fields on X endowed with the usual bracket operation. The exponential map exp: Vect$(X) \to$ Diff(X) assigns to every vector field the unique (X is compact) flow that it generates. The complexification of the Lie algebra Vect(X) does *not* correspond to any Lie group, that is there is no Lie group whose Lie algebra is $\text{Vect}_{\mathbb{C}}(X)$ [125].

In the case of Kac-Moody algebras the situation is the following. Affine algebras are realized as central extension of loop algebras, so every affine Lie algebra comes from a Lie group (a loop group). At the moment of writing this work, we do not know if this is true for indefinite Kac-Moody algebras: the question is open since their fist discovery (almost 40 years). No concrete realizations of these algebras are known. The case of a (genuine)

Borcherds algebra is even worse for the presence of imaginary simple roots. In the physical applications we have in mind [37]-[38],[47], the term Kac-Moody group is used to mean the infinite-dimensional Lie group (if any) corresponding to the infinite-dimensional Kac-Moody Lie algebra. In particular the Kac-Moody algebra is studied with a level decomposition with respect to a finite-dimensional Lie algebra, that is one sums infinite finite-dimensional representations of a classical Lie algebra (often su(n)) to recover the Kac-Moody algebra. This decomposition is of course infinite and "converges" at the level of the algebra; but, we do not know if it "converges" also to an infinite-dimensional Lie group whose Lie algebra is our beloved Kac-Moody algebra. In the applications, one usually truncates the decomposition at some low level and exponentiates the corresponding finite-dimensional representations to get a manageable Lie group and write a Lagrangian invariant under it. This Lagrangian often coincides with (part of) the Lagrangian of some important physical theory, like pure gravity in 4 dimensions or 11-dimensional supergravity.

Yet, it is possible to define in a very precise way what a *Kac-Moody group* is: it is not a manifold, its construction (due to J. Tits [147]-[148]) is based on the concepts of chambers and buildings in terms of a so-called BN-pair. It is this *discrete* structure which appears in the limit of general relativity to the cosmological singularity (not a smooth group!), and I believe that this direction should be explored better and in more details, because it should contain information about the birth of the universe.

2.2 Classification of Kac-Moody Algebras

Generalized Cartan matrices can be classified thanks to a theorem of E. Vinberg. The result if that only one of the following three possibilities holds:

- (Fin) $\mathrm{det}A \neq 0$; there exists $u > 0$ such that $Au > 0$; $Av \geq 0$ implies $v > 0$ or $v = 0$;

- (Aff) $\mathrm{corank}A = 1$; there exists $u > 0$ such that $Au = 0$; $Av \geq 0$ implies $Av = 0$;

- (Ind) there exists $u > 0$ such that $Au < 0$;$Av \geq 0$, $v \geq 0$ imply $v = 0$

We say that A is of *finite*, *affine* or *indefinite* type respectively. It turns out that generalized Cartan matrices of finite of affine type are always symmetrizable. Moreover, the algebras of finite type correspond precisely to finite-dimensional simple Lie algebras (this being equivalent also to $|W| < \infty$ or $|\Delta| < \infty$ or $(,)_{\mathfrak{h}_\mathbb{R}}$ positive-definite). Affine and indefinite Lie algebras are always infinite-dimensional[3].

2.2.1 Affine Kac-Moody Algebras

A matrix A is affine iff all its proper principal minors are positive and det $A = 0$ (thus corank $A = 1$). Affine Dynkin diagrams are listed in the book by Kac. Without entering the details, affine Kac-Moody algebras can be easily constructed from the finite-dimensional Lie algebras. In particular, there exist a standard mechanism of *affinization* of every simple Lie algebra X_N, which gives all untwisted affine algebras $\mathrm{X}_N^{(1)}$. The procedure consists in using the highest root θ of each X_N and to consider the lattice $\mathrm{II}^{1,1} \cong \mathbb{Z}^2$ with basis $k_+ = (1,0), k_- = (0,-1)$ and scalar product $\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$. From this one can define the vector $\alpha_0 = k_+ - \theta$, the affine simple root, which together with the simple roots of X_N gives the affine Dynkin diagram of $\mathrm{X}_N^{(1)}$. One can show that the only imaginary roots of these affine algebras are the integer multiples of k_+ (considered as vectors of $Q(X_N)\oplus\mathrm{II}^{1,1}$)

$$\Delta_+^{im} = \{nk_+ \,, n = 1, 2, \ldots\} \tag{2.19}$$

In particular, the imaginary roots of affine algebras are always isotropic ($k_+^2 = 0$). Finally, the multiplicities of imaginary roots for untwisted affine Lie algebras is $|I| = \mathrm{rk}\ A$. This result and many others can be deduced using a concrete realization of the affine algebras: they are isomorphic to central extensions of loop algebras. This also clarifies the geometric structure of these algebras: the corresponding Lie group in an infinite-dimensional loop group.

Affine Kac-Moody algebras were initially also called Euclidean Lie algebras, the term affine refers to the structure of their Weyl group, which is the so-called affine Weyl group of the underlying finite Lie algebra X_N.

[3] Note that the whole class of infinite-dimensional Lie algebras is perhaps too big to be classified; in the infinite-dimensional setting Levi's theorem (according to which any finite-dimensional Lie algebra is the semi-direct sum of a solvable Lie algebra, called radical, and a semi-simple Lie algebra) is not valid, so in particular there exist infinite-dimensional Lie algebras which do not admit a structure based on a (generalized or not) Cartan matrix. For example, the Virasoro algebra is another famous infinite-dimensional Lie algebra which does not belong to the class of Kac-Moody or Borcherds algebras.

2.2.2 Lorentzian and Hyperbolic Kac-Moody Algebras

Note that an indefinite matrix is not necessarily symmetrizable. There is not a general theory of indefinite Kac-Moody algebras, mostly because a concrete realization is missing. Indefinite Kac-Moody algebras are not of finite growth, that is the dimension of each imaginary root space (a homogeneous piece) is not of polynomial growth, but rather of exponential one. The indefinite Lie algebras we are interested in have a Cartan matrix A with det< 0 and are generically called *Lorentzian* Kac-Moody algebras; they can be defined by requiring that the Cartan matrix has Lorentzian signature $- + + \cdots +$. This is probably not enough to have an interesting theory of these algebras, as suggested by V. Nikulin and V. Gritsenko (see the comments at the end of this chapter). Anyhow, we will deal only with *hyperbolic* Kac-Moody algebras, which belong to the class of the indefinite algebras defined by the condition that any connected proper sub-diagram is of finite or affine type. A hyperbolic matrix is not necessarily symmetrizable, but if it is then the symmetrized matrix has Lorentzian signature $- + + \cdots +$.

It is possible to classify hyperbolic Kac-Moody algebras. The only rank-1 Kac-Moody algebra is A_1, which is the building block of all Kac-Moody algebras (Borcherds algebras need something more). In rank 2, the matrix $\begin{pmatrix} 2 & -a \\ -b & 2 \end{pmatrix}$ is finite, affine, or hyperbolic iff $ab \leq 3$, $ab = 4$ or $ab > 4$. Symmetrizable Kac-Moody algebras of rank 3 were classified by Yoshida in [159]. Finally the highest rank for hyperbolic Kac-Moody algebras is 10 and one can list all hyperbolic Dynkin diagrams.

There are not many results about hyperbolic Kac-Moody algebras: there is not a single case where the root system and the multiplicities are known explicitly! A theorem due to Moody states that in the (symmetrizable) hyperbolic case the imaginary roots are precisely all the vectors ($\neq 0$) in the root lattice with 0 or negative squared length

$$\Delta^{im} = \{\alpha \in Q |\, \alpha^2 \leq 0\} - \{0\} \quad \text{(hyperbolic symmetrizable)} \tag{2.20}$$

We will discuss in some details the hyperbolic algebras $\mathrm{HA}^1_{(1)}$ and E_{10} (both of them have a symmetric Cartan matrix). Regarding the multiplicities of imaginary roots for hyperbolic algebras, I. Frenkel (see the original reference in the book by Kac) made an interesting conjecture according to which

$$\dim\, \mathfrak{g}_\alpha \leq p_{rk-2}\left(1 - \frac{(\alpha, \alpha)}{2}\right) \tag{2.21}$$

where $p_{rk-2}(n)$ is the number of partitions of the integer n into parts of $(rk-2)$ colours (with rk the rank the the Cartan matrix of the corresponding hyperbolic algebra). This conjecture turned out to be wrong (see below). We can say in full generality that in all indefinite cases (hyperbolic or not), the multiplicity of imaginary roots is given by an *arithmetic function* (i.e. with values in $\mathbb{N}$) mult: $\Delta^{im} \to \mathbb{N}$; note that this is *not* necessarily a function of the squared norm of an imaginary root, but it may be a much more complicated function on the imaginary part of the root lattice. Understanding this function is the holy grail of theory of Kac-Moody algebras.

Let us finish this section with further comments on imaginary roots. For real roots $\alpha \in \Delta^{re}$, $\mathrm{mult}(\pm\alpha) = 1$ and $n\alpha$ is never a root for $n \neq \pm 1$ (i.e. $\mathrm{mult}(n\alpha) = 0$). For imaginary roots, $n\alpha$ is always a root for any non-zero integer $n \neq 0$ (imaginary roots are the real difference with respect to classical Lie algebras and also the major difficulty in understanding indefinite Kac-Moody algebras). Any *isotropic* (that is $(\alpha, \alpha) = 0$) root α is W-equivalent to an imaginary root of an affine Lie subalgebra, hence mult $\alpha < |I|$ (from the table in Kac and in [121] one can check for example that all the isotropic roots have multiplicities 1 and 8 for the algebras $\mathrm{HA}_1^{(1)}$ and E_{10} respectively, see below). For non-isotropic ($(\alpha, \alpha) < 0$) imaginary roots the situation changes drastically. In this case, $\oplus_{n>0}\,\mathfrak{g}_{n\alpha}$ is a free Lie algebra, $\mathrm{mult}(n\alpha)$ is a non-decreasing sequence and moreover $\lim_{n\to\infty} \frac{\ln mult(n\alpha)}{n}$ exists and is positive. Here we follow [89]. For any positive imaginary root $\alpha = \sum_i k_i\alpha_i$, we define the *Kac-Peterson function*

$$\psi : \alpha \in \Delta_+^{im} \longrightarrow \psi(\alpha) := \limsup_{n\to\infty} \frac{\ln\ \mathrm{mult}\ (n\alpha)}{n} \tag{2.22}$$

More specifically, Kac and Peterson have shown that:

- the limit exists without the sup: $\psi(\alpha) = \lim_{n\to\infty} \frac{\ln\ \mathrm{mult}\ (n\alpha)}{n}$; if $(\alpha, \alpha) < 0$, then $\psi(\alpha) = \sup_{n\geq 1} \frac{\ln\ \mathrm{mult}\ (n\alpha)}{n}$
- if $(\alpha, \alpha) = 0$, $\psi(\alpha) = 0$; if $(\alpha, \alpha) < 0$, then $0.48 < \psi(\alpha) \leq \mathrm{ht}(\alpha) \ln \mathrm{ht}(\alpha) - \sum_i k_i \ln k_i$
- if n is a positive integer, $\psi(n\alpha) = n\psi(\alpha)$; ψ is $W-$invariant, $\psi(w(\alpha)) = \psi(\alpha)$ for any $w \in W$
- if $\alpha, \beta, \alpha + \beta \in \Delta_+^{im}$, then $\psi(\alpha + \beta) \geq \psi(\alpha) + \psi(\beta)$

If A is indecomposable, the ψ function extends uniquely to a *concave function* on the interior of the imaginary cone Z such that $\psi(t\alpha) = t\psi(\alpha)$ for $t > 0$.

To prove this, one first considers free Lie algebras. In fact, let L be a free abelian group on generators $\beta_1, \dots, \beta_r$, let $L_+ = \sum_i \mathbb{Z}_+\beta_i$ and $J = J_1 \cup \cdots \cup J_r$ a disjoint union of non-empty finite sets. Let $a = \bigoplus_{\alpha \in L} \mathfrak{a}_\alpha$ be a free Lie algebra on generators e_j $(j \in J)$ graded by deg $e_j = \beta_i$ for $j \in J_i$. For $\alpha = \sum k_i\beta_i \in L_+$ and $k = \sum k_i$, define the function

$$\psi_0(\alpha) = k \ln k - \sum_i k_i \ln(k_i/|J|) \tag{2.23}$$

Then for all $\alpha \in L_+ \backslash \{0\}$, one has

$$\lim_{n\to\infty} \frac{\ln(1 + dim\, \mathfrak{a}_{n\alpha})}{n} = \psi_0(\alpha) \tag{2.24}$$

For a free Lie algebra $\mathfrak{a}$ on N generators $e_1, \dots, e_N$ of linearly independent degrees $\alpha_1, \dots, \alpha_N$ and $\alpha = \sum_i k_i\alpha_i$ with all $k_i > 0$, one has

$$dim\, \mathfrak{a}_{n\alpha} \sim \, C(\alpha)\, n^{(-N+1)/2}\, e^{n\psi_0(\alpha)} \qquad \text{as } n \to \infty \tag{2.25}$$

where $C(\alpha) = (2\pi)^{(1-N)/2}(\sum_i k_i)^{-1/2} \prod_i k_i^{-1/2}$. This led Kac and Peterson to the following conjecture: with A indecomposable and α in the interior of the imaginary cone Z, there exists $C(\alpha) > 0$ such that

$$mult(n\alpha) \sim C(\alpha)\, n^{-(|I|+1)/2}\, e^{n\,\psi(\alpha)} \qquad \text{as } n \to \infty \tag{2.26}$$

The multiplicities of imaginary roots for indefinite algebras is still mysterious: there is not a single case where one can compute them analytically.

The fact that the Kac-Peterson function is zero on isotropic roots and positive on non-isotropic imaginary roots and the presence of some ln functions reminds an entropy function (consider also that its asymptotics is similar to Margulis asymptotics for periodic orbits especially in the arithmetic case and that $-\psi$ is convex). In fact, it is known that affine algebras have to do with integrable dynamical systems, and we have said that isotropic roots are always conjugate to imaginary roots of affine subalgebras. And for integrable systems, the metric entropy is zero, exactly as the Kac-Peterson function. The fact that $\psi(\alpha) > 0$ on non-isotropic roots and a relation we describe in the next sections between periodic orbits and imaginary roots make very reasonable a relation between hyperbolic algebras and chaotic (say hyperbolic) dynamical systems.

2.3 The Character Formula and the Denominator Identity

For any $\Lambda \in \mathfrak{h}^*$, there exist an irreducible $\mathfrak{g}(A)-$module $L(\Lambda)$, unique up to isomorphism, satisfying

There exist a non-zero vector $v_\Lambda \in L(\Lambda)$ such that $\mathfrak{n}_+(v_\Lambda) = 0$ and $h(v_\Lambda) = \Lambda(h)v_\Lambda$ for all $h \in \mathfrak{h}$.

$L(\Lambda)$ is called the *irreducible highest weight module* with *highest weight* Λ.

We have the *weight space decomposition* of $L(\Lambda)$ with respect to $\mathfrak{h}$

$$L(\Lambda) = \bigoplus_{\lambda \in \mathfrak{h}^*} L(\Lambda)_\lambda \tag{2.27}$$

where $L(\Lambda)_\lambda = \{v \in L(\Lambda) | h(v) = \lambda(h)v, \text{ for all } h \in \mathfrak{h}\}$. Now let us define the *Konstant function* $K(\beta)$ through the formal expansion

$$\prod_{\alpha \in \Delta_+} \left(1 - e^{-\alpha}\right)^{\text{mult } \alpha} = \sum_{\beta \in \mathfrak{h}^*} K(\beta) e^{-\beta} \tag{2.28}$$

$K(\beta)$ is the number of partitions of β into a sum of positive roots, where each root is counted with its multiplicity, since $(1-e^{-\alpha})^{-1} = 1 + e^{-\alpha} + e^{-2\alpha} + \cdots$. We say that $\lambda \in \mathfrak{h}^*$ is a *weight* of $L(\Lambda)$ if $L(\Lambda)_\lambda \neq 0$ and we put $\text{mult}_\Lambda(\lambda) :=$ dim $L(\Lambda)_\lambda$; the latter is always finite. We denote by $P(\Lambda)$ the set of weights of $L(\Lambda)$.

For any $\lambda \in \mathfrak{h}^*$, define the function e^λ on $\mathfrak{h}$ by $e^\lambda(h) := e^{\lambda(h)}$. This allows to introduce the *character* $\text{ch}_{L(\Lambda)}$ of $L(\Lambda)$

$$h \to \text{ch}_{L(\Lambda)}(h) = \sum_{\lambda \in \mathfrak{h}^*} \text{mult}_\Lambda(\lambda)\, e^{\lambda(h)} \tag{2.29}$$

This is a function defined on a certain domain Y_Λ of all $h \in \mathfrak{h}$ such that the series converges absolutely. One can show that Y_Λ is convex from the convexity of $|e^\lambda|$, and that the convergence of $\text{ch}_{L(\Lambda)}$ is uniform on compact subsets of the interior of Y_Λ, thus $\text{ch}_{L(\lambda)}$ is holomorphic on the interior of Y_Λ.

We say that $\lambda \in \mathfrak{h}^*$ is an *integral weight* if $\lambda(h_i)$ in integral for all $i \in I$. An integral weight is *dominant* if $\lambda(h_i) \geq 0$ and *regular dominant* if $\lambda(h_i) > 0$. We denote by P, P^+, P^{++} the sets of integral, dominant, regular dominant weights respectively. Note that $Q \subset P$. Fix also $\rho \in \mathfrak{h}^*$ such that $\rho(h_i) = 1$ (the Weyl vector).

Then for a symmetrizable Cartan matrix A, one has the following *character and denominator formulas*

$$\begin{aligned}\left(\sum_{w\in W}(\det\, w)\, e^{w(\rho)}\right)\, \mathrm{ch}_{L(\Lambda)} &= \sum_{w\in W}(\det\, w)\, e^{w(\Lambda+\rho)} \qquad (2.30)\\ \sum_{w\,\in W}(\det w)\; e^{w(\rho)-\rho} &= \prod_{\alpha\,\in\Delta^+}\left(1-e^{-\alpha}\right)^{mult\,\alpha}\end{aligned}$$

2.4 Generalized Kac-Moody algebras à la Borcherds

A kind of last generalization of Kac-Moody algebras are the so-called *Borcherds algebras*, which allow for the presence of *imaginary simple roots* in the root system. The starting point as usual is a Cartan matrix which satisfy certain conditions, in particular the main differences with respect Kac-Moody algebras are that in a Borcherds algebra

I may be countably infinite rather that finite
a_{ii} may not be positive and need not lie in $\mathbb{Z}$
$2a_{ij}/a_{ii}$ is only assumed to lie in $\mathbb{Z}$ when $a_{ii} > 0$

thus the Cartan subalgebra may be infinite-dimensional and the a_{ii}'s zero or negative on the diagonal of the Cartan matrix correspond to the squared norms of imaginary simple roots.

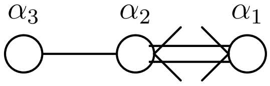
We do not insist too much on their definition here. Many statements about Kac-Moody algebras continue to hold, the Weyl group is defined only in terms of fundamental reflections with respect *only* to real simple roots and there is a denominator identity which accounts for imaginary simple roots (whose multiplicity can be bigger than 1, although they are simple)

$$e^{\rho}\prod_{\alpha\in\Delta_+}(1-e^{-\alpha})^{mult\,\alpha} = \sum_{w\in W}(\det\, w)\, w\left(e(\rho)\sum_{\Psi}(-1)^{\Psi}e(-\sum\Psi)\right) \qquad (2.31)$$

where Ψ runs over all finite subsets of mutually orthogonal imaginary simple roots and $(\rho, \alpha_i) = 1$ only for real simple roots.

2.5 The hyperbolic Kac-Moody algebra $\mathrm{HA}_1^{(1)}$

The hyperbolic Kac-Moody algebra $\mathrm{HA}_1^{(1)}$ is defined by the following Dynkin diagram [4]



or equivalently by its Cartan matrix

$$\begin{pmatrix} 2 & -2 & 0 \\ -2 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \tag{2.32}$$

This is in a certain sense the simplest hyperbolic Kac-Moody algebra of rank 3, being the canonical hyperbolic extension of A_1 [5]. Let us write down the scalar products between the simple roots $(\alpha_i, \alpha_j) = a_{ij}$ for future convenience

$$(\alpha_1, \alpha_2) = -2\,, \quad (\alpha_1, \alpha_3) = 0\,, \quad (\alpha_2, \alpha_3) = -1 \tag{2.33}$$

Its Weyl group W is generated by the reflections r_1, r_2, r_3 subject to the Coxeter relations

$$r_i^2 = (r_1 r_3)^2 = (r_2 r_3)^3 = 1 \tag{2.34}$$

and it is thus isomorphic to the extended modular group $\mathrm{PGL}(2, \mathbb{Z})$. Its even subgroup W^+ is generated by $r_2 r_1$ and $r_1 r_3$ and is isomorphic to $\mathrm{PSL}(2, \mathbb{Z})$ (in fact one can put $T = r_2 r_1$ and $S = r_1 r_3$ with T, S the standard generators of $\mathrm{PSL}(2, \mathbb{Z})$, as we have already said).

General results about indefinite Kac-Moody algebras are lacking; still something is known for $\mathrm{HA}_1^{(1)}$. As we said, Moody's theorem says that the

[4] This algebra is often denoted by $\widehat{A_1^{(1)}}$ or A_1^{++} (the latter is used especially in the physical literature).

[5] The canonical hyperbolic extension consists in adding a single link to the affine root of an untwisted affine algebra. The algebras so obtained are also known as over-extended or $\mathcal{G}^{++}$ algebras in the physical literature. One can go further and add another single link the the last one of a $\mathcal{G}^{++}$ algebra and get a (still) Lorentzian very-extended $\mathcal{G}^{+++}$. Note that this is not the only way to produce indefinite algebras starting from finite-dimensional ones. A different, non-standard mechanism of extension (studied in [53]) allows to produce higher rank indefinite algebras using the fundamental weights instead of the highest root.

imaginary roots are precisely all the vectors in the root lattice with squared length 0 or < 0

$$\alpha = k_1\alpha_1 + k_2\alpha_2 + k_3\alpha_3 \in \Delta^{im} \Leftrightarrow \alpha^2 = 2(k_1^2 + k_2^2 + k_3^2) - 4k_1k_2 - 2k_2k_3 \leq 0 \tag{2.35}$$

The structure of its subalgebras was investigated recently by Feingold and Nicolai [51], who found that it contains all rank-2 symmetric indefinite Kac-Moody algebras, all rank-3 algebras with Cartan matrix $\begin{pmatrix} 2 & -m & 0 \\ -m & 2 & -2 \\ 0 & -2 & 2 \end{pmatrix}$ with $m > 2$ and also Borcherds algebras.

Perhaps the most important and inspiring paper about $\mathrm{HA}_1^{(1)}$ was written by Feingold and Frenkel [50], who found that its Weyl group W is isomorphic to the discrete group $\mathrm{PGL}(2,\mathbb{Z})$, which contains the modular group $\mathrm{PSL}(2,\mathbb{Z})$ (isomorphic to the positive Weyl group W^+) as a subgroup of index 2. They also found that the denominator identity can be interpreted as a Siegel modular form on the Siegel upper-half plane (which is a matrix generalization of the hyperbolic plane, that is one replaces points with matrices). This is another example of some modularity property contained in the denominator identity of Kac-Moody algebras. The isomorphism constructed by Feingold and Frenkel realizes the root system of the algebra $\mathrm{HA}_1^{(1)}$ in the following way. Let us introduce another basis in $\mathfrak{h}^*$ by

$$\gamma_1 = \alpha_1/2 \quad \gamma_2 = -\alpha_1 - \alpha_2 - \alpha_3 \quad \gamma_3 = -\alpha_1 - \alpha_2 \tag{2.36}$$

and let $S(2,\mathbb{C})$ the complex symmetric 2×2 matrices. Let us define the map

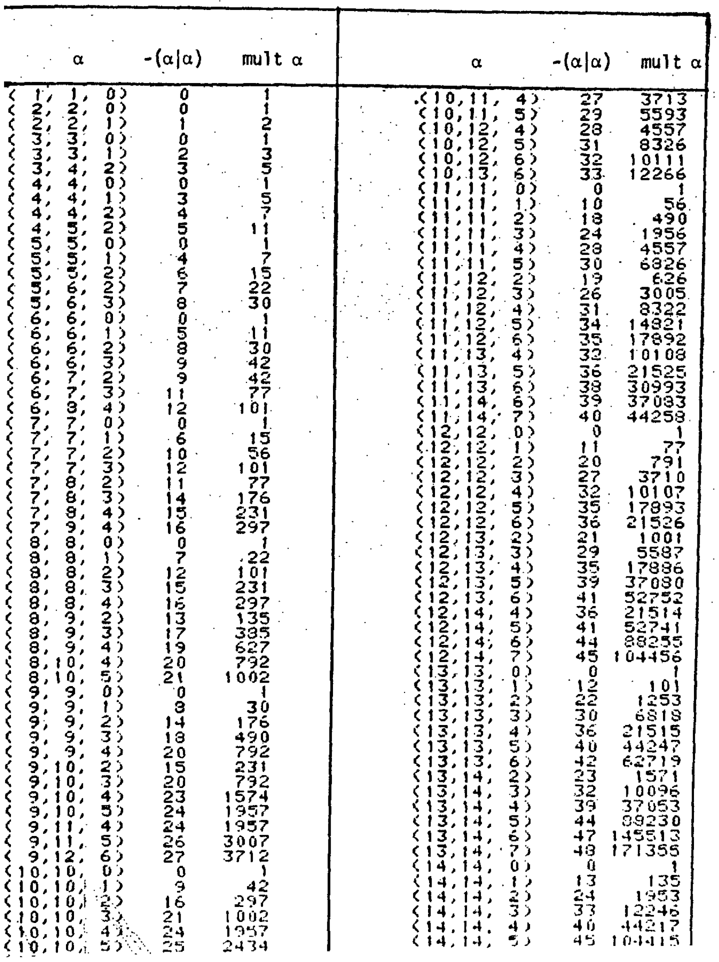
$$\nu : (z_1\gamma_1 + z_2\gamma_2 + z_3\gamma_3) \in \mathfrak{h}^* \to \begin{pmatrix} z_3 & z_1/2 \\ z_1/2 & z_2 \end{pmatrix} \in S(2,\mathbb{C}) \tag{2.37}$$

and the group homeorphism $\overline{\nu} : W \to \mathrm{PGL}(2,\mathbb{Z})$ determined by $\overline{\nu}(r_i) = W_i(z)$ where W_i are the standard generators of $\mathrm{PGL}(2,\mathbb{Z})$ (the hyperbolic reflections in the three sides of the fundamental domain). Then Feingold and Frenkel show that the map ν is a vector space isomorphism $\mathfrak{h} \simeq S(2,\mathbb{C})$ and lattice isomorphism $Q \simeq S(2,\mathbb{Z})$. $\overline{\nu}$ extends to a group isomorphism $W \simeq \mathrm{PGL}(2,\mathbb{Z})$ and $W^+ \simeq \mathrm{PSL}(2,\mathbb{Z})$. In particular, the real and imaginary roots can be realized, simply by their definition, as

$$\begin{aligned} \nu(\Delta^{re}) &= \{g \in S(2,\mathbb{Z}) |\ \det g \geq 0\} \\ \nu(\Delta^{im}) &= \{g \in S(2,\mathbb{Z}) |\ \det g = -1\} \end{aligned} \tag{2.38}$$

Some of the multiplicities of the imaginary roots were computed on a computer with the help of Peterson's recurrent formula and are listed in the book by Kac, page 215 [6]. We report the table here for convenience

[6] From the general theory it is known that isotropic roots are $W-$equivalent to imaginary roots of an affine subalgebra of $\mathrm{HA}_1^{(1)}$. This affine subalgebra must have rank 2, so $\dim \mathfrak{g}_\alpha = 1$ for all α isotropic, and we do not consider them in the following discussion.



α	-(α|α)	mult α	α	-(α|α)	mult α
⟨1, 1, 0⟩	0	1	⟨10, 11, 4⟩	27	3713
⟨2, 2, 0⟩	0	1	⟨10, 11, 5⟩	29	5593
⟨2, 2, 1⟩	1	2	⟨10, 12, 4⟩	28	4557
⟨3, 3, 0⟩	0	1	⟨10, 12, 5⟩	31	8326
⟨3, 3, 1⟩	2	3	⟨10, 12, 6⟩	32	10111
⟨3, 4, 2⟩	3	5	⟨10, 13, 6⟩	33	12266
⟨4, 4, 0⟩	0	1	⟨11, 11, 0⟩	0	1
⟨4, 4, 1⟩	3	5	⟨11, 11, 1⟩	10	56
⟨4, 4, 2⟩	4	7	⟨11, 11, 2⟩	18	490
⟨4, 5, 2⟩	5	11	⟨11, 11, 3⟩	24	1956
⟨5, 5, 0⟩	0	1	⟨11, 11, 4⟩	28	4557
⟨5, 5, 1⟩	4	7	⟨11, 11, 5⟩	30	6826
⟨5, 5, 2⟩	6	15	⟨11, 12, 2⟩	19	626
⟨5, 6, 2⟩	7	22	⟨11, 12, 3⟩	26	3005
⟨5, 6, 3⟩	8	30	⟨11, 12, 4⟩	31	8322
⟨6, 6, 0⟩	0	1	⟨11, 12, 5⟩	34	14821
⟨6, 6, 1⟩	5	11	⟨11, 12, 6⟩	35	17892
⟨6, 6, 2⟩	8	30	⟨11, 13, 4⟩	32	10108
⟨6, 6, 3⟩	9	42	⟨11, 13, 5⟩	36	21525
⟨6, 7, 2⟩	9	42	⟨11, 13, 6⟩	38	30993
⟨6, 7, 3⟩	11	77	⟨11, 14, 6⟩	39	37083
⟨6, 8, 4⟩	12	101	⟨11, 14, 7⟩	40	44258
⟨7, 7, 0⟩	0	1	⟨12, 12, 0⟩	0	1
⟨7, 7, 1⟩	6	15	⟨12, 12, 1⟩	11	77
⟨7, 7, 2⟩	10	56	⟨12, 12, 2⟩	20	791
⟨7, 7, 3⟩	12	101	⟨12, 12, 3⟩	27	3710
⟨7, 8, 2⟩	11	77	⟨12, 12, 4⟩	32	10107
⟨7, 8, 3⟩	14	176	⟨12, 12, 5⟩	35	17893
⟨7, 8, 4⟩	15	231	⟨12, 12, 6⟩	36	21526
⟨7, 9, 4⟩	16	297	⟨12, 13, 2⟩	21	1001
⟨8, 8, 0⟩	0	1	⟨12, 13, 3⟩	29	5587
⟨8, 8, 1⟩	7	22	⟨12, 13, 4⟩	35	17886
⟨8, 8, 2⟩	12	101	⟨12, 13, 5⟩	39	37080
⟨8, 8, 3⟩	15	231	⟨12, 13, 6⟩	41	52752
⟨8, 8, 4⟩	16	297	⟨12, 14, 4⟩	36	21514
⟨8, 9, 2⟩	13	135	⟨12, 14, 5⟩	41	52741
⟨8, 9, 3⟩	17	385	⟨12, 14, 6⟩	44	88255
⟨8, 9, 4⟩	19	627	⟨12, 14, 7⟩	45	104456
⟨8, 10, 4⟩	20	792	⟨13, 13, 0⟩	0	1
⟨8, 10, 5⟩	21	1002	⟨13, 13, 1⟩	12	101
⟨9, 9, 0⟩	0	1	⟨13, 13, 2⟩	22	1253
⟨9, 9, 1⟩	8	30	⟨13, 13, 3⟩	30	6818
⟨9, 9, 2⟩	14	176	⟨13, 13, 4⟩	36	21515
⟨9, 9, 3⟩	18	490	⟨13, 13, 5⟩	40	44247
⟨9, 9, 4⟩	20	792	⟨13, 13, 6⟩	42	62719
⟨9, 10, 2⟩	15	231	⟨13, 14, 2⟩	23	1571
⟨9, 10, 3⟩	20	792	⟨13, 14, 3⟩	32	10096
⟨9, 10, 4⟩	23	1574	⟨13, 14, 4⟩	39	37053
⟨9, 10, 5⟩	24	1957	⟨13, 14, 5⟩	44	89230
⟨9, 11, 4⟩	24	1957	⟨13, 14, 6⟩	47	145513
⟨9, 11, 5⟩	26	3007	⟨13, 14, 7⟩	48	171355
⟨9, 12, 6⟩	27	3712	⟨14, 14, 0⟩	0	1
⟨10, 10, 0⟩	0	1	⟨14, 14, 1⟩	13	135
⟨10, 10, 1⟩	9	42	⟨14, 14, 2⟩	24	1953
⟨10, 10, 2⟩	16	297	⟨14, 14, 3⟩	33	12246
⟨10, 10, 3⟩	21	1002	⟨14, 14, 4⟩	40	44217
⟨10, 10, 4⟩	24	1957	⟨14, 14, 5⟩	45	104415
⟨10, 10, 5⟩	25	2434			

Figure 2.1: The table shows a list of some imaginary roots for $\mathrm{HA}_1^{(1)}$. (k_1, k_2, k_3) denotes the root $\alpha = k_1\alpha_1 + k_2\alpha_2 + k_3\alpha_3$, $\alpha \in C^\vee$ ($C^\vee$ is the dual Weyl chamber, i.e. the Weyl chamber in the $\mathfrak{h}^*$ space) and $(\alpha|\alpha) := \frac{1}{2}\sum a_{ij}k_ik_j$.

Looking at the table, one can note that these multiplicities are close the values of the classical partition function $p(n)$ [7]. If we identify the integer n

[7] $p(n)$ is the number of partitions of the integer n, where a *partition* of n is a finite non-decreasing sequence of positive integers $p_1, \cdots, p_k$ whose sum is n (we put $p(0) = 1$ and $p(n) = 0$ if n is a negative integer). The first values of $p(n)$ are

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$p(n)$	1	1	2	3	5	7	11	15	22	30	42	56	77	101	135	176

n	16	17	18	19	20	21	22	23	24	25	26	27
$p(n)$	231	297	385	490	627	792	1002	1255	1575	1958	2436	3010

n	28	29	30	31	32	33	34	35	36	37
$p(n)$	3718	4565	5604	6842	8349	10143	12310	14883	17977	21637

The following result is the culmination of an intense research effort that took place in the first half of the twentieth century

$$p(n) = \frac{1}{\pi\sqrt{2}} \sum_{k=1}^{\infty} A_k(n)\sqrt{k} \left[\frac{d}{dx} \frac{\sinh \frac{\pi}{k}\sqrt{\frac{2}{3}\left(x - \frac{1}{24}\right)}}{\sqrt{\left(x - \frac{1}{24}\right)}} \right]_{x=n} \tag{2.39}$$

where

$$A_k(n) = \sum_{h \mod k,\, (h,k)=1} \omega_{h,k}\, e^{-2\pi i n h/k}$$

and $\omega_{h,k}$ is a certain 24-th root of unity. This formula is not one of those mathematical formulas that elicits the response "Just as I expected!" and it is due to the genius of Hardy, Ramanujan and Rademacher. The formula is not only an asymptotic series, it is a finite, exact formula for $p(n)$. It can be shown that if we sum the first $c\sqrt{n}$ terms in this expansion for some constant c, then the nearest integer to that sum will be the exact value of p(n) [3]! The method that they used to find and to prove the validity of their formula is called the *circle method*, because the successive terms in the expansion arise from singularities of the generating function in a certain ordering of the rational points on the unit circle. The circle method is nowadays considered one of the most difficult problem in mathematics. By taking only the first term of this expansion, we obtain the *asymptotic* behavior of $p(n)$

$$p(n) \sim \frac{1}{4\sqrt{3}\,n}\, e^{\pi\sqrt{2n/3}} \tag{2.40}$$

which shows that the growth of $p(n)$ is sub-exponential. Probably, the fact that the partition function $p(n)$ is asymptotically sub-exponential rather that exponential is a rapid way to discard it as a (complete) multiplicity function.

with $\left(1-\frac{(\alpha,\alpha)}{2}\right)$, then some of these multiplicities obey

$$\text{mult } \alpha = p\left(1-\frac{(\alpha,\alpha)}{2}\right) \qquad \text{for some imaginary roots} \tag{2.41}$$

a fact already checked in [50] for all imaginary roots of the form $\begin{pmatrix} n & 0 \\ 0 & 1 \end{pmatrix}$ for which mult $\begin{pmatrix} n & 0 \\ 0 & 1 \end{pmatrix} = p(n)$ (using Feingold-Frenkel realization of the root lattice). But there are also cases which violate this relation; for example, reading from the previous table, the roots $(8,9,4)$ and $(11,12,2)$ have both squared length -38 but

$$\text{mult } (8,9,4) = 627 = p(20)\,, \quad \text{mult } (11,12,2) = 626 = p(20) - 1 \tag{2.42}$$

This example also shows that there are imaginary roots with the same squared norm but with different multiplicity, because, as we said, in general the multiplicity of an imaginary root is not a function of α^2.

In the listed cases, one has always mult $\alpha = p\left(1-\frac{(\alpha,\alpha)}{2}\right)$ or mult $\alpha < p\left(1-\frac{(\alpha,\alpha)}{2}\right)$, which experimentally confirms Frenkel's conjecture. We will see that for E$_{10}$ computer calculations support the reverse inequality. The way mult α fails to be equal to $p\left(1-\frac{(\alpha,\alpha)}{2}\right)$ is perhaps interesting. Let us also call the *defect* of a root the difference $p\left(1-\frac{(\alpha,\alpha)}{2}\right) - \text{mult } \alpha$. In fact, one can check that in all the cases listed one always has

$$p\left(1-\frac{(\alpha,\alpha)}{2}\right) - \text{mult } \alpha = \text{ integer combinations of } \left\{p\left(1-\frac{(\beta_i,\beta_i)}{2}\right)\right\}_i \tag{2.43}$$

where β_i is a finite set of imaginary roots with $\beta_i^2 < \alpha^2$. Let us illustrate this point with some examples. Comparing the table in [87] and the values of $p(n)$, one can read the following

root	mult	defect	
(8,9,4)	627	0	
(11,12,2)	626	1	
(8,10,4)	792	1	
(12,12,2)	791	0	
(13,13,2)	1253	2	
(9,10,4)	1574	1	
(13,14,2)	1571	4	$p(1)+p(3), p(2)+p(2)$
(9,10,5)	1957	1	
(11,11,3)	1956	2	
(14,14,2)	1953	3	
(10,10,5)	2434	2	
(14,15,2)	2429	7	
(9,11,5)	3007	3	
(11,12,3)	3005	5	
(9,12,6)	3712	6	$p(1)+p(4), p(3)+p(3)$
(10,11,4)	3713	5	
(12,12,3)	3710	8	$p(1)+p(5), p(3)+p(4)$
(10,12,4)	4557	8	$p(1)+p(5), p(3)+p(4)$
(10,11,5)	5593	11	
(12,13,3)	5587	17	$p(2)+p(7)$
(11,11,5)	6826	16	$p(1)+p(7), p(4)+p(6)$
(13,13,3)	6818	24	$p(2)+p(8)$
(10,12,5)	8326	23	$p(1)+p(8)$
(11,12,4)	8322	27	$p(4)+p(8)$
(10,12,6)	10111	32	$p(2)+p(9)$
(11,13,4)	10108	35	$p(4)+p(9)$
(12,12,4)	10107	36	$p(1)+p(4)+p(9), p(3)+p(6)+p(8)$
(13,14,3)	10096	47	$p(4)+p(10), p(2)+p(7)+p(9), p(3)+p(5)+p(7)+p(8)$

As shown, the defect always corresponds to an exact value of $p(n)$ (for example $1, 2, 3, 5, 7, 11$) or to (different) combinations of different values of $p(n)$ (we exclude all the trivial sums $4 = 1+1+1+1 = 4p(1)$, $6 = 1+1+1+1+1+1 = 6p(1)$ etc).

All these considerations suggests a likely number theoretic interpretation for the multiplicities of imaginary roots of $\mathrm{HA}_1^{(1)}$ and motivate the following *conjecture about the explicit form of the multiplicity function*

Conjecture 1 *The multiplicities of the imaginary roots for the hyperbolic algebra $HA_1^{(1)}$ are given by the classical partition function $p(n)$ or mult $\alpha = p(n) - \sum_{i=1}^{l} m_i$ where each term m_i is the value of the partition function $p(m)$ for some m.*

To prove or disprove this, one must understand how the root α is related to the integers n and m_i, in particular if n is always $\left(1 - \frac{(\alpha,\alpha)}{2}\right)$ and which are the roots β_i corresponding to the integers $m_i = p\left(1 - \frac{(\beta_i,\beta_i)}{2}\right)$. It is also likely that the number of terms entering the defect increases with $-\alpha^2$. Indeed, in full generality, studying this defect phenomenon $\left| dim\, \mathfrak{g}_\alpha - p_{rk-2}\left(1 - \frac{(\alpha,\alpha)}{2}\right)\right|$ could help understanding the problem of imaginary roots.

Feingold and Frenkel also showed that the multiplicity of the imaginary roots of the form $\begin{pmatrix} n & 0 \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} n & 1 \\ 1 & 2 \end{pmatrix}$ is given respectively by $p'(2n+1)$ and $p'(2n)$, where $p'(n)$ is a modified partition function

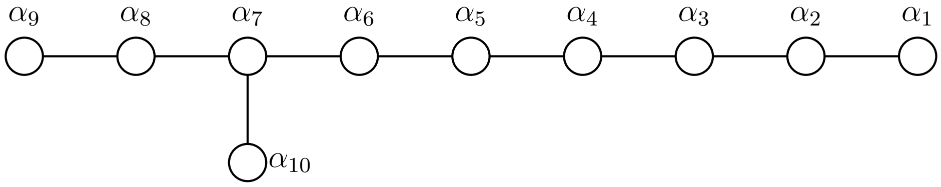
$$\sum_{n\geq 0} p'(n)\, t^n = \left[\prod_{n\geq 1}(1-t^n)^{-1}\right](1 - t^{20} + t^{22} - t^{24} + \ldots) \tag{2.44}$$

and $p(n) = p'(n)$ for $0 \leq n \leq 19$, for higher values of n we have the defect phenomenon we have described. Evidently, partition functions know about the multiplicities of imaginary roots for this algebra. Regarding these multiplicities, Feingold and Frenkel suggested a number-theoretical meaning connected with ideal classes of imaginary quadratic fields. I do not know of any development in this direction (except for computer calculations). As we have mentioned, Feingold and Frenkel found a relation between the characters of the algebra and the works on automorphic forms by Siegel and Maass, and we have also described the way Maass automorphic forms solve the Laplacian problem for $\mathrm{PGL}(2,\mathbb{Z})$ which is isomorphic to W.

Finally, using a kind of generalized vertex operator, a vertex operator construction for $\mathrm{HA}_1^{(1)}$ has been built in [107] by A. Sciarrino and V. Marotta, who also developed some attempts of realizations of Borcherds algebras in [108].

2.6 The hyperbolic Kac-Moody algebra E_{10}

There exist 4 rank-10 hyperbolic algebras, E_{10}, BE_{10}, CE_{10} and DE_{10} (see the book by Kac, page 57). Of these E_{10} is a distinguished one because its root lattice $Q(E_{10})$ is the unique even Lorentzian self-dual lattice $II^{1,9} = Q(E_8)\oplus II^{1,1}$ in dimension 10 (this is probably the most striking evidence for any eventual role played by E_{10} in string theory). Furthermore, E_{10} is the only hyperbolic symmetric matrix with determinant -1. It is given by the following Dynkin diagram



Simple roots linked by a segment have scalar products -1, otherwise they are orthogonal. The simple roots $\{\alpha_3, \dots, \alpha_{10}\}$ give an E_8 algebra, and together with affine root α_2 (often denoted as α_0 or α_{+1}) one has the affine algebra E_9 according to the standard mechanism of affinization of a finite Lie algebra. Finally with the root α_1 (often denoted by α_{-1} or α_{+2}), one obtains the hyperbolic algebra E_{10} according to the canonical hyperbolic extension of a finite Lie algebra.

Following Feingold and Frenkel's approach for $HA_1^{(1)}$, attempts to understand E_{10} in terms of a level decomposition with respect to its affine $E_9 = E_8^{(1)}$ algebra were made by Kac, Moody and Wakimoto [88]. They found the following

$$\dim \mathfrak{g}_\alpha = \begin{cases} p_8\left(1 - \frac{(\alpha,\alpha)}{2}\right) & \text{if } \alpha \text{ is of level 0 or 1} \\ \xi\left(3 - \frac{(\alpha,\alpha)}{2}\right) & \text{if } \alpha \text{ is of level 2} \end{cases} \tag{2.45}$$

where the level of a root α is the number of times the affine root α_1 appears in the decomposition of the root α in basis of simple roots and $p_8(n)$ [8] is the

[8] The generating function for the partition function is the *Euler ϕ function*

$$\sum_{n\geq 0} p(n)q^n = \prod_{n\geq 1}(1-q)^{-1} := \phi(q) \tag{2.46}$$

while in general

$$\sum_{n\geq 0} p_k(n)q^n = \frac{1}{\phi(q)^k} \tag{2.47}$$

number of partitions of the integer n into parts of 8 colours

$$\frac{1}{\phi(q)^8} = \sum_{n\geq 0} p_8(n) q^n \tag{2.49}$$

and

$$\frac{1}{\phi(q)^8}\left[1 - \frac{\phi(q^2)}{\phi(q^4)}\right] = \sum_{n\geq 0} \xi(n) q^n \tag{2.50}$$

In particular expanding the second series as

$$q^2 \left[\sum p_8(n)q^n + q^4 \sum p_8(n)q^n - q^6 \sum p_8(n)q^n + \ldots\right] \tag{2.51}$$

we see that $\xi(6) = p_8(4) + 1 > p_8(4)$ which disproves Frenkel's conjecture about multiplicities of imaginary roots for hyperbolic algebras. Indeed, for E_{10}, in all the cases known, the reverse inequality occurs (compare the tables in [88] and [121]), which somehow goes in the opposite direction with respect to $\mathrm{HA}_1^{(1)}$. Anyhow, the same kind of defect phenomenon (with a + sign) occurs for E_{10}

Conjecture 2 *The multiplicities of the imaginary roots for the hyperbolic algebra E_{10} are given by the partition function $p_8(n)$ or mult $\alpha = p_8(n) + \sum_{i=1}^{l} m_i$ where each term m_i is the value of the partition function $p_8(m)$ for some m.*

We can check this conjecture looking at the table in [121]. The multiplicities of some imaginary roots are

- up to level $l = 8$: 8, 44, 192
- $l = 9$: 8, 44, 192, 727
- $l = 10$: 8, 44, 192, 727, 2472, 7749
- $l = 11$: 8, 44, 192, 727, 2472, 7749

I do not know of any asymptotic formula for $p_8(n)$, although there exist exact formulae for $p_k(n)$ for small k

$$p_2(n) = \lfloor (n+1)/2 \rfloor\,, \quad p_3(n) = \{(n+3)^2/12\} \tag{2.48}$$

where $\lfloor x \rfloor$ is the largest integer not exceeding x and $\{x\}$ (only here) is the nearest integer to x.

- $l = 12$: 8, 44, 192, 726, 727, 2472, 7747, 7749, 22725
- $l = 13$: 8, 44, 192, 726, 727, 2472, 7747, 7749, 22712, 22725, 63085, 167116
- $l = 14$: 8, 44, 192, 726, 727, 2472, 7747, 7749, 22712, 22725, 63085, 167116, 167133
- $l = 15$: 8, 44, 192, 726, 727, 2464, 2472, 7747, 22712, 22725, 63020, 63085, 167099, 167116, 425156, 425227, 1044218

If we now write the first values of $p_8(n)$

$$1, 8, 44, 192, 726, 2464, 7704, 22528, 62337, 164560, 417140, 1020416 \tag{2.52}$$

we can make the following combinations:

$$727 = 1 + 726\,, \quad 2472 = 8 + 2464\,, \quad 7749 = 1 + 44 + 7704 \tag{2.53}$$

A first analysis of subalgebras of E_{10} was carried in a joint paper with my supervisor A. Sciarrino [53], where we found the following theorem

Theorem 1 *The indefinite Kac-Moody algebras of rank 10 described by the Dynkin diagrams, obtained by adding to the diagram of the affine algebra E_9, a dot, connected with a simple link to the $j-$th dot of E_9 $(j \neq 2)$, is a subalgebra of E_{10}.*

In the same paper, we also prove a similar statement for E_{11}, which has a physical relevance too [157], and show that the simply-laced over-extended and very-extended Kac-Moody algebras contain all the non-simply laced ones by a straightforward generalization of Dynkin's folding.

E_{10} has often been indicated as the ultimate symmetry for string theory. I believe that if E_{10} has anything to do with string theory in 10 dimensions or M-theory, a decisive role must be played by fermions too. From this point of view, one should instead consider the *Fake Monster Lie Superalgebra* constructed by Nils R. Scheithauer [139]. This is a generalized Kac-Moody superalgebra realized as the physical states of a 10 dimensional superstring moving on a torus, thus it contains all the superstring spectrum. If E_{10} has any role for superstrings in 10 dimensions, one should be able to locate E_{10} inside the fake Monster Superalgebra or viceversa; this would give a further hint of an E_{10} symmetry in string theory. Indeed, the situation could be more subtle. One could find that E_{10} is (or is not) a subalgebra of the fake Monster superalgebra, but the role of an E_{10} symmetry could be different, for example E_{10} could be just a symmetry of the space of solutions of a theory (full or bosonic); this means that different solutions of the theory are transformed into each other through Weyl reflections of E_{10} or something like that. Anyhow, understanding the way E_{10} is or is not related in the fake Monster superalgebra is not only important from a mathematical point of view, but it could also shed new light on the possible role played by E_{10} and clarify the meaning of such a symmetry. A further consideration is the following. Recently, there has been a lot of interest in $\mathrm{K}(\mathrm{E}_{10})$, the (formal) compact real form of E_{10} defined through the Chevalley involution as described before. If E_{10} is a symmetry of string theory in 10 dimensions, fermions should live in a spinorial representation of $\mathrm{K}(\mathrm{E}_{10})$ [9]. Given the role of the fake Monster superalgebra, there must be a relation between $\mathrm{K}(\mathrm{E}_{10})$ and fake Monster superalgebra if $\mathrm{K}(\mathrm{E}_{10})$ has to contain fermions inside. This is indeed an interesting way to explore, using for example Naito's theorems [120] about embeddings of Kac-Moody algebras into Borcherds algebras. A first simple observation is the following. From [139], one knows that the simple roots of the Fake Monster superalgebra are the zero norm vectors in the closure of the positive cone of $\mathrm{II}^{9,1}$. Since it is possible to write down explicitly the denominator identity (which turns out to be an automorphic form of weight 4 for a subgroup of $\mathrm{O}_{10,2}(\mathbb{R})$), one knows the multiplicity of the roots of the Fake Monster superalgebra, in particular the simple roots have multiplicities 8, which is exactly the multiplicity of all isotropic (i.e. zero norm) roots of E_{10} as one can check from the table in [121].

2.7 Imaginary roots and periodic geodesics

In this section we focus our attention on the hyperbolic algebra $\mathrm{HA}_1^{(1)}$, in particular we find a correspondence between its imaginary roots and the periodic geodesics inside the fundamental domain of its positive Weyl group $W^+ \cong \mathrm{PSL}(2,\mathbb{Z})$. This relation seems to be new and due to the author. Let us also stress that this seems to be the first *geometric interpretation*

[9] In order to include the fermions, one can think to consider hyperbolic Kac-Moody superalgebras, but note that the maximum rank for them is 6, whereas it is 10 for hyperbolic algebras. A classification of hyperbolic Kac-Moody superalgebras has been obtained very recently by L. Frappat and A. Sciarrino in [54].

of the imaginary roots of an indefinite Kac-Moody algebra, but we are not claiming this is the general case. Moreover, this relation also gives a physical interpretation of the imaginary roots: these correspond to periodic orbits, thus to classical periodic solutions of a certain theory. The Weyl reflections with respect to the real roots transform solutions into other solutions, so we could conclude that the hyperbolic algebra $\mathrm{HA}_1^{(1)}$ is a symmetry of the space of the solutions of a particular physical theory. This theory is classical Einstein gravity in 4 dimensions close to a the cosmological singularity, as we will describe in Part II of this thesis.

As before, let T and S be the standard generators for $\mathrm{PSL}(2,\mathbb{Z})$, that is:

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \tag{2.54}$$

$$T(z) = z + 1\,, \quad S(z) = \frac{-1}{z} \tag{2.55}$$

Let W_1, W_2 and W_3 be the three hyperbolic reflections in the sides of a fundamental domain of $\mathrm{PGL}(2,\mathbb{Z})$, that is the three fundamental Weyl reflections [10]

$$W_1(z) = -\overline{z}\,, \quad W_2(z) = -\overline{z} + 1\,, \quad W_3(z) = \frac{1}{\overline{z}} \tag{2.56}$$

then we have [11]

$$T = W_2 W_1 \neq W_1 W_2 = T^{-1} \tag{2.57}$$

$$S = W_1 W_3 = W_3 W_1 \tag{2.58}$$

First let us associate to any W_i the corresponding simple root

$$W_i \in \mathrm{PGL}(2,\mathbb{Z}) \to \alpha_i \in \Delta(\mathrm{HA}_1^{(1)}) \tag{2.59}$$

Let $n_1, \ldots, n_m \geq 2$ be integers. Then the matrix

$$A = T^{n_1} S\, T^{n_2} S \cdots T^{n_m} S \tag{2.60}$$

is always hyperbolic (with a positive trace), reduced and its arithmetic code is $(A) = (n_1, \ldots, n_m)$ (this is proved in [92]); with this we mean that $(n_1, \ldots, n_m)$

[10] We use the letter W_i instead of the previous one R_i to stress that these are Weyl reflections, i.e. the domain derives from as Weyl chamber.

[11] Also $W_2 W_3 = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix} \neq W_3 W_2 = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}$

is the *minus* '-' continued fraction expansion of the fixed points of the invariant geodesic of A, the fixed points are irrational quadratics so the continued fraction expansion is eventually periodic.

Now, let us write the matrix A in terms of W_1, W_2, W_3

$$A = [(W_2 W_1)^{n_1}\ W_1 W_3]\ [(W_2 W_1)^{n_2}\ W_1 W_3] \cdots [(W_2 W_1)^{n_m}\ W_1 W_3] \quad (2.61)$$

and count the number of hyperbolic reflections in its expression

$$\begin{aligned} \sharp W_1 &= (n_1 + n_2 + \cdots + n_m) + m && (2.62)\\ \sharp W_2 &= (n_1 + n_2 + \cdots + n_m) && (2.63)\\ \sharp W_3 &= m && (2.64) \end{aligned}$$

We can associate to each A the following lattice vector $\alpha_{(A)}$ just counting the number of Weyl reflections in A and putting this number equal to the coefficient of the corresponding simple root, i.e.

$$\alpha_{(A)} := (n_1 + \cdots + n_m + m)\, \alpha_1 + (n_1 + \cdots + n_m)\, \alpha_2 + m\, \alpha_3 \quad (2.65)$$

we say that $\alpha_{(A)}$ is *coded* by the primitive hyperbolic matrix A.

Then all the (reduced) hyperbolic matrices in PSL$(2, Z)$, that is all primitive periodic geodesics in the modular domain, give rise to imaginary roots for HA$_1^{(1)}$, thus we can establish the following theorem

Theorem 2 $\alpha_{(A)}$ *is an imaginary root for* $HA_1^{(1)}$

The proof is very simple. Since HA$_1^{(1)}$ is hyperbolic, it is enough to check that all such vectors have squared length 0 or negative (because this is a necessary and sufficient condition for symmetrizable hyperbolic root systems). Let us call $n := \sum_i^m n_i$ and note that $n \geq 2m$ (because each n_i is ≥ 2). Thus we have to calculate the norm of the vector

$$\alpha_{(A)} = (n + m)\, \alpha_1 + n\, \alpha_2 + m\, \alpha_3 \quad (2.66)$$

Remembering the scalar products between the simple roots, one easily arrives at the following expression

$$\alpha_{(A)}^2 = 4\, m^2 - 2\, mn \quad (2.67)$$

and since $n \geq 2m$ one always has

$$\alpha^2_{(A)} \leq 0 \tag{2.68}$$

Note that the isotropic case $\alpha^2_{(A)} = 0$ occurs for the minimum value $n = 2m$. Let us now analyze some specific cases, considering also $n_i = 0, 1$.

For $m = 1$, we have $A(n_1) = T^{n_1}S = \begin{pmatrix} -n_1 & 1 \\ 1 & 0 \end{pmatrix}$, which is hyperbolic only for $n_1 \geq 3$. For $n_1 = 0$, we have $A(n_1 = 0) = S = W_1W_3$, to which we would associate the lattice vector $\alpha_1 + \alpha_3$, which is not a root. For $n_1 = 1$, we have $A(n_1 = 1) = TS = W_2W_1W_1W_3$, to which we would associate $2\alpha_1 + \alpha_2 + \alpha_3$ which has squared norm $+2$ and it is a real root [12]. For $n_1 = 2$, we have $A(n_1 = 2) = T^2S = (W_2W_1)^2W_1W_3$, to which we would associate $3\alpha_1 + 2\alpha_2 + \alpha_3$ which is a isotropic imaginary roots. For $n_1 \geq 3$, $A(n_1)$ is always hyperbolic with Tr $A(n_1) = -n_1$ and the lattice vector $(n_1 + 1)\alpha_1 + n_1\alpha_2 + \alpha_3$ is always an imaginary roots with *strictly* negative squared norm $2(2 - n_1) < 0$. In this case the periodic geodesic γ invariant under $A(n_1 \geq 3)$ has hyperbolic length $l_\gamma = 2\cosh^{-1}\left(\frac{n_1}{2}\right)$ and we can express this geometrical quantity in terms of the squared norm of the root, because $\frac{n_1}{2} = \frac{1}{2}\frac{4-\alpha^2}{2}$, thus $l_\gamma = 2\cosh^{-1}\left(\frac{4-\alpha^2}{4}\right)$.

Let us now consider the case $m = 2$, then we have $A(n_1, n_2) = T^{n_1}ST^{n_2}S = \begin{pmatrix} n_1n_2 + 1 & -n_1 \\ -n_2 & 1 \end{pmatrix}$, which is hyperbolic for $n_1, n_2 \geq 1$. For $n_2 = 0$, $A(n_1, n_2 = 0) = T^{n_1}SS = (W_2W_1)^{n_1}W_1W_3W_1W_3$, to which we would associate $(n_1 + 2)\alpha_1 + n_1\alpha_2 + 2\alpha_3$ which has squared norm $4(2 - n_1)$. It is clear that for $n_1 = 0$ too, we obtain $2(\alpha_1 + \alpha_3)$ which is not a root, as previously stated. For $n_1 = 1$, we have $3\alpha_1 + \alpha_2 + 2\alpha_3$ which is not a root. For $n_1 \geq 2$, $(n_1 + 2)\alpha_1 + n_1\alpha_2 + 2\alpha_3$ is always an imaginary root (isotropic for $n_1 = 2$), even if the matrix $A(n_1 \geq 2, n_2 = 0)$ is never hyperbolic. Now let us consider the cases $n_1 = n_2 = 1$, with $A(n_1 = 1, n_2) = TSTS$, to which we would associate $4\alpha_1 + 2\alpha_2 + 2\alpha_3$, which is not a root (as stated for $A(n_1 = 1) = TS$). Now, let us fix $n_2 = 1$, and let us increase n_1. For $n_1 = 2$, we have $A(n_1 = 2, n_1 = 1) = T^2STS$, to which we would associate

[12] The real roots of hyperbolic Kac-Moody algebras are

$$\Delta^{re} = \{\alpha = \sum_j k_j\alpha_j \in Q |\, \alpha^2 > 0 \text{ and } k_j\, \alpha_j^2/\alpha^2 \in \mathbb{Z} \text{ for all} j\}$$

$5\alpha_1 + 3\alpha_2 + 2\alpha_3$ which is not a root, even if $A(n_1 = 2, n_1 = 1)$ is hyperbolic. For $n_1 \geq 3$, $A(n_1 \geq 3, n_1 = 1) = T^{n_1}STS$ is always hyperbolic and the vector $(n_1 + 3)\alpha_1 + (n_1 + 1)\alpha_2 + 2\alpha_3$ is always an imaginary root (isotropic for $n_1 = 3$). For bigger values of n_1, n_2, $A(n_1, n_2)$ is always hyperbolic and the corresponding vector is an imaginary root as in the theorem. Note in the case $m = 2$, we have Tr $A(n_1, n_2) = n_1 n_2 + 2$, thus for $n_1, n_2 \geq 2$ the length of a periodic orbit is $l_\gamma = 2\cosh^{-1}(\frac{n_1 n_2 + 2}{2})$, while the squared norm of the root α_A is $4[2 - (n_1 + n_2)]$ and I do not see a way to express l_γ in terms of a quantity which identifies the hyperbolic algebra.

Note that in general the correspondence is not 1-1. For example, the root $(8, 6, 2)$ can arise from the combinations $n_1 = 3, n_2 = 3$ or $n_1 = 2, n_2 = 4$ or $n_1 = 4, n_2 = 2$ which give different hyperbolic matrices but the same imaginary root. Moreover, any non-primitive periodic geodesic (i.e. a periodic geodesic run k times) gives an imaginary root which is a $k-$th multiple of the imaginary root corresponding to the primitive periodic geodesic. In fact, given a non-primitive hyperbolic matrix A^k, one can form the vector $\alpha_{(A^k)} := k\,\alpha_{(A)}$ which has squared norm $k^2\,\alpha_{(A)}^2 \leq 0$ and so it is again an imaginary root, in agreement with the general fact that integer multiples of imaginary roots are still imaginary roots.

This seems to be the first result which codes the root system of an indefinite Kac-Moody algebra and in particular it gives a *geometric* flavor to the imaginary roots.

This construction gives all the imaginary roots with any value for coeff$(\alpha_3) = m$ and coeff$(\alpha_1) = (n + m) >$ coeff$(\alpha_2) = n$. It is clear that there exist other imaginary roots which do not have this form, for example the ones with coeff(α_2) =coeff(α_3). The question is, of course, if it is possible to obtain all the imaginary roots (or even all the roots) from a construction similar to this, for example using Weyl reflections in real roots, or using a different presentation for the matrices of PGL$(2, \mathbb{Z})$. In fact, up to now, we have only considered hyperbolic matrices in PSL$(2, \mathbb{Z})$, while the Weyl group is PGL$(2, \mathbb{Z})$. Thus if we consider also matrices with det $= -1$, following [25], we can write any hyperbolic matrix in PGL$(2, \mathbb{Z})$ as

$$A = W_3\,(W_1 W_2)^{n_1}\,W_3\,(W_1 W_2)^{n_2}\,\cdots\,W_3\,(W_1 W_2)^{n_m} \qquad (2.69)$$

Exactly as above, the numbers n_i give the (usual) continued fraction expansion of the hyperbolic fixed point in $(0, 1)$ determined by the primitive closed geodesic invariant under A. Defining $n = n_1 + n_2 + \cdots + n_m$ as before, again all the roots of the form

$$\alpha_{(A)} = n\alpha_1 + n\alpha_2 + m\alpha_3 \tag{2.70}$$

are imaginary provided that each n_i is greater than 1. For example, in the case where we have only the first block, $A(n_1) = \begin{pmatrix} 0 & 1 \\ 1 & -n_1 \end{pmatrix}$, which is hyperbolic for $n_1 \geq 3$, the corresponding root

$$\alpha_{(A)} = n_1\alpha_1 + n_1\alpha_2 + \alpha_3 \tag{2.71}$$

is imaginary as soon as $n_1 \geq 1$. This gives all the roots with coeff(α_1) =coeff(α_2) and coeff$(\alpha_3) = 1$ (and their multiple integers considering $A(n_1)^k$). This further observation strongly supports the idea that a better presentation exists in order to code the root system.

Finally, one can try to compare the multiplicities of these imaginary roots with the multiplicities of the periodic orbits from which they derive, since, as we said, for arithmetic systems, there is an exponential degeneration for the multiplicities of the lengths of periodic orbits (with fixed trace). This is compatible also with the asymptotic behavior of the Kac-Peterson function.

2.8 A new interpretation for the Selberg Trace Formula and the Selberg Zeta-function?

In analogy with the Riemann zeta function which is defined as a product over the primes, the Selberg zeta function $Z(s)$ is defined as the product over all primitive periodic orbits (ppo) for the motion on the hyperbolic surface considered

$$Z(s) = \prod_{ppo} \prod_{m=0}^{\infty} (1 - e^{-l_p\,(s+m)}) \tag{2.72}$$

where l_p is the hyperbolic length of the primitive orbit, s is a complex parameter and m counts how many times an orbit is run. Note that for the Selberg zeta-function, it does not exist an equivalent expression in terms of a sum over something, while classical $L-$functions are given equivalently by an Eulerian product over the primes or a sum (Appendix A). This makes the Selberg zeta-function somehow different from classical $L-$functions.

According to our result, primitive periodic orbits inside the standard modular domain for PSL$(2,\mathbb{Z})$ give rise to imaginary roots for HA$_1^{(1)}$, so it is reasonable to suppose that the Selberg zeta function may also be expressed a product over a suitable subset of imaginary roots

$$Z(s) \sim \prod_{imaginary\, roots\, coded\, by\, ppo} (1 - e^{-l_p(A_{ij})(s+m)}) \tag{2.73}$$

where $l_p(A_{ij})$ is the hyperbolic length as a function of the Cartan matrix A_{ij} of $\mathrm{HA}_1^{(1)}$. Of course, one has to find the relation between the length l_p of a ppo and some quantity given in terms of the Cartan matrix of $\mathrm{HA}_1^{(1)}$ (this has been done in the previous section for all imaginary roots coming from the matrices $A(n_1) = T^{n_1}S$ with $n_1 \geq 3$).

This expression reminds the product part of the denominator identity

$$\prod_{\alpha \in \Delta_+} \left(1 - e^{-\alpha}\right)^{mult\, \alpha} \tag{2.74}$$

although in the latter the exponent is mult α, whereas it is 1 in the Selberg zeta function. We repeat that Feingold and Frenkel showed that a certain subspace $\mathcal{M}'_k$ of weight k $\mathrm{PSL}(2,\mathbb{Z})-$invariant $A_1^{(1)}-$characters is isomorphic to the space $\mathcal{M}_k^2$ of genus 2 Siegel modular forms of weight k. It would be remarkable if part of the denominator identity contained the Selberg zeta function too, or even information on the Maass waveforms with respect to the Weyl group $W \simeq \mathrm{PGL}(2,\mathbb{Z})$.

These speculations apply also to the Selberg trace formula, which was first derived by Selberg as a sum over primitive hyperbolic conjugacy classes, then interpreted by Huber a sum over periodic orbits. We suggest a new possible interpretation in terms of the sum over the root system of a Lie algebra. In saying this I am probably influenced by the following words in M. Berger's book [18] (page 391, chapter 9): "...There are at least two ways to compute the spectra of the remaining KP^n. One is to use a very general formula due to Hermann Weyl, and valid for all symmetric spaces. But the formula is explicit only in the sense that it is a summation over the roots of a certain Lie algebra. To get explicit expressions is hard. The other way is to use the general link between periodic geodesics and the spectrum, a quite deep result (unavoidably using the wave equation)which we will meet in 9.9 ...". The second formula is the Selberg trace formula, but I never heard about the first one. I had the opportunity to meet M. Berger at IHES on March 2007 and I asked him about this Weyl formula, but he did not remember exactly it. Then I asked P. Cartier and we spent a couple of hours in the library of IHES looking for something like that in Weyl's collected works. Most of them are in German, I do not speak German, Cartier does. We did not find anything. Then I wrote to G. Besson (following Berger's suggestion) and to V. S. Varadarajan, but this formula did not show up. It does not likely exist.

Anyhow, trying to relate the Selberg trace formula to a certain sum over the root system of a Lie algebra is very interesting in my opinion and probably it is already in the mind of some more talented mathematician. We note that this would be possible if the *whole* root system (simple, real and imaginary roots) could be coded by using parabolic, elliptic and hyperbolic transformations of $\mathrm{PSL}(2,\mathbb{Z})$, because in the Selberg trace formula all three kinds of contributions appear.

2.9 Notes and Comments on Chapter 2

Kac and Moody introduced the algebras that carry their names in a different way. It is instructive to remind how they were led to the discovery of this beautiful part of mathematics. For the theory of finite-dimensional Lie algebras we suggest the books by Humphreys [80] or Varadarajan [150].

Let us remember that the Weyl groups corresponding to the finite-dimensional simple Lie algebras are precisely the finite crystallographic Coxeter groups. Moody asked what is the class of Lie algebras which correspond more generally to any Coxeter group (most Coxeter groups are infinite). The partial answer to Moody's question is that the Lie algebras corresponding to the (possibly infinite) *crystallographic* Coxeter groups are the Kac-Moody algebras. Note that we still do not know which are (or if exist) the Lie algebras corresponding to the *non-crystallographic* Coxeter groups, that is the ones which have Coxeter exponents m_{ij} different from $2, 3, 4, 6, \infty$ [81].
Kac's road to these algebras was quite different (mostly using the machinery of filtered and graded Lie algebras developed by Guillemin, Singer and Sternberg). Let $\mathfrak{g}$ be a complex Lie algebra. By a $\mathbb{Z}-$grading we mean that we can write the underlying vector space $\mathfrak{g}$ as $\mathfrak{g} = \oplus_{n=-\infty}^{\infty}\mathfrak{g}_n$ such that $[\mathfrak{g}_n, \mathfrak{g}_m] \subseteq \mathfrak{g}_{n+m}$ for all $n, m \in \mathbb{Z}$. We call $\mathfrak{g}$ a simple $\mathbb{Z}-$graded Lie algebra if in addition $\mathfrak{g}$ does not contain any non-trivial $\mathbb{Z}-$graded ideal.
It is probably hopeless to classify all simple $\mathbb{Z}-$graded Lie algebras, there are too many of them. However, decades earlier, Cartan had studied vector fields on polynomial algebras and found four infinite families that were simple $\mathbb{Z}-$graded, with the dimension dim $\mathfrak{g}_n$ bounded above by some polynomial

in n. We say that these $\mathbb{Z}-$graded algebras have polynomial growth. Kac conjectured that if $\mathfrak{g}$ is a simple graded Lie algebra of finite growth, then $\mathfrak{g}$ is isomorphic to one of the following algebras:

- a finite-dimensional Lie algebra, or
- a loop algebra, or
- a Cartan algebra, or
- the Virasoro algebra.

This conjecture has been shown recently by O. Mathieu. Note also that for affine algebras the dimension of each root space is bounded by a unique constant.

The standard reference is the book by Kac [87]; other references (which contain also the theory of Borcherds algebras and more advanced topics) are the more recent books by M. Wakimoto [154], R. Carter [27], T. Gannon [57], U. Ray [128], and A. Pressley and G. Segal [125] for loop algebras (also known as current algebras in the physics literature). The theory of generalized Kac-Moody algebras developed by R. E. Borcherds at the end of the '80s allows for a the presence of *imaginary simple roots* in the simple root system. These algebras play a key role in the Borcherds proof of the Monstrous Moonshine Conjecture, together with the notion of *vertex algebras* introduced by Borcherds too. For this topic see the beautiful book by Gannon [57] (the original papers by Borcherds are all available on his web site `http://math.berkeley.edu/~reb/`). The work of Borcherds also shows that in a certain sense these generalized Kac-Moody algebras represent the last possible generalization through the h_i, e_i, f_i formalism and their definition through generators and relations is essentially the same as the definition given by Kac for arbitrary (i.e. not Cartan-like) matrices. Kac and Peterson [89] showed that, in the affine case, the denominator identity gives a modular form for some $\Gamma(N)$, indeed a vector-valued Jacobi form.

The work of Borcherds also shows that there are interesting relations between generalized Kac-Moody algebras, automorphic forms and hyperbolic reflection groups, but there is no general theory unfortunately. In particular, reversing the point of view, V. Gritsenko and V. Nikulin [66] (and references therein) asked (and found) which are the good Lorentzian Kac-Moody algebras which admit an *automorphic correction*, that is whose denominator identity can be put in relation with an automorphic form (the improved algebras have different Cartan matrices and are often indefinite or generalized Kac-Moody algebras, or super-algebras). They used a variant of Borcherds lift to find automorphic forms.

What we have shown in this chapter is that there is a relation between dynamical quantities (periodic geodesics) and algebraic quantities (root systems) for billiards flows in particular domains corresponding to the projections of hyperbolic Weyl chambers on the hyperbolic plane. If this correspondence is deep and generalizable to more indefinite Kac-moody algebras, then one could hope for a (still lacking) geometric interpretation for the imaginary roots, especially for their multiplicities. I believe that the multiplicities of imaginary roots is related to the length spectrum of periodic geodesics on the corresponding hyperbolic surface (thus to the eigenvalues of the Laplacian problem too through the Selberg trace formula). General links between graded infinite dimensional Lie algebras and dynamical systems were studied, for example, by A. Vershik in [151]. Is it possible that our analysis is related to Vershik's work. We hope to investigate this point in the future. The message is that hyperbolic Kac-moody algebras (or even all the indefinite ones?) have to do with *chaotic* dynamical systems (in particular the ones exhibiting hyperbolicity), whereas it is known that affine algebras are deeply related to the theory of *integrable* dynamical systems (KdV equation, Calogero-Sutherland models etc). In fact, the Kac-Peterson function is *zero* on isotropic imaginary roots and *positive* on non-isotropic imaginary roots. *This function resembles an entropy function*, which is zero for integrable systems and positive for chaotic systems; besides, its concavity and the ln functions which it contains and which remember a metric entropy could motivate the notion of an *algebraic entropy.* Also the asymptotics is similar to Margulis asymptotics. As isotropic imaginary roots are $W-$equivalent to imaginary roots of affine subalgebras (which describe integrable systems), one can say that non-isotropic imaginary roots are related to hyperbolic dynamical systems. This new idea deserves certainly further investigation.

Part II

Physical Applications

Chapter 3

The Mixmaster Universe and Beyond

> Thoughtland, Fletch, the cosmos. Pure mentation. Abstract possibility. Infinite dimensions. The class of all sets. God's mind. The pre-geometric substratum. Hilbert space. Penultimate reality. White . . .
>
> *Master of Space and Time*
> R. Rucker

In this chapter, we describe the BKL's approach to the study of a cosmological singularity in terms of a never-ending sequence of Kasner eras and we state DHN's recent result about the asymptotic dynamics of general relativity in terms of a billiard motion in the Weyl chamber of the algebra $\mathrm{HA}_1^{(1)}$ (we comment on this specific billiard law). We carry on a quantum analysis of the problem, derive the properties of the wave functions of the system in this limit and prove the absence of scarred states in quantum cosmology in the billiard representation. We suggest also an interpretation of the imaginary roots of $\mathrm{HA}_1^{(1)}$ as periodic solutions to Einstein's equations in the asymptotic regime.

3.1 General Considerations

The following considerations follow verbatim [117] (page 813), and, although written in 1972, are still very modern and explain some features of the problem of the cosmological singularity. I could not have found better words.

The cosmological singularity involves infinite curvature and infinite density. What abhors is the fact that these infinities occurred at a finite proper time in the past and would occur again at some finite proper time in the future. The prediction of a singularity would be more tolerable if the infinite densities could be pushed to an infinitely distant past. In this case, the universe could find its natural state to be one of expansion, so every finite density will have been experienced at some suitably remote past time, but infinite density becomes a formal abstraction never realized in the course of evolution.

To push infinite curvature out of the finite past might be achieved in two ways. One way is to change the physical law which require the singularity, perhaps stating the laws of gravity in a proper quantum language. Nowadays, it is not clear at all if quantum geometry can actually remove the singularity problem [1].

Another way to discard the singularity is to accept the mathematics of the classical Einstein equations but reinterpret it in terms of an infinite past time. A coordinate transformation such that $t = \ln \tau$ moves the singularity from $\tau = 0$ to $t = -\infty$; but an arbitrary coordinate is without significance. The problem is that the singularity occurs at a *finite proper time* in the past, and proper time is the most physically significant, most physically real

[1] On the question of whether theories containing gravity and matter, like string theory, may actually resolve singularities there is at the moment no consensus. Moreover, the meaning of resolving a singularity is not clear, too (what should replace a singularity?). "Many people believe that the resolution of the problem of singularities will come from the modifications of the Einstein equations due to Quantum Gravity at the Planck scale , but this is by no means obvious. The necessary modifications could, in principle, have nothing to do with quantum mechanics. It might for example entail the introduction of higher curvature terms ..."[60]. The same kind of criticism can be applied to the common belief that amplitudes in supergravity theories are, generally, infinite. Recently, works of M. Green et al. [64] (see also the works of Zvi Bern et al. [19]) have shown that some amplitudes which were considered infinite are actually *finite*. If true, this means that the popular saying that it is not possible to quantize gravity using only QFT techniques is not reliable any more. The discovery of novel cancellations in the calculation of amplitudes (not predicted by traditional superspace power-counting arguments) suggests that supergravity theories may be perturbatively finite theories of quantum gravity.

time we know. It corresponds to the ticking of physical clocks and measures the natural rhythms of actual events. To reinterpret finite past time as infinite, one must attack proper time on precisely these grounds and claim it inadequately physical. On a local basis, where special relativity is valid, no challenge to the physical significance of proper time can succeed. It is on a *global* scale that the physical primacy of proper time needs to be reviewed.

Let us consider the following statement "The cosmological singularity occurred ten thousand million years ago" and take time to mean the proper time along the worldline of the solar system. Then the statement would have a most direct physical significance if it meant that the earth had completed 10^{10} orbits around the sun since the beginning of the universe. But proper time is not that closely tied to actual physical phenomena. The statement merely implies that those 5×10^9 which the earth may have actually accomplished give a standard of time which is to be extrapolated in prescribed ways, thus giving theoretical meaning to the order 5×10^9 years which are asserted to have preceded the formation of the solar system. A hardier standard clock changes the details of the argument, but not its qualitative conclusion. To interpret 10^{10} years in terms of the SI seconds assigns a past history containing some 3×10^{27} oscillations of a hyperfine transition in neutral Cesium. But again the critical early ticks of the clock are missing. The time needed for stellar nucleosynthesis to produce the first Cesium disqualifies this clock on historical grounds, and the still earlier high temperatures nearer the singularity would have ionized all Cesium even if this element has predated stars.

The conclusion is that *proper time near the singularity is not a direct counting of simple and actual physical phenomena, but an elaborate mathematical extrapolation.* Each actual clock has its ticks discounted by a suitable factor, 3×10^7 seconds per orbit from the earth-sun system, 1.1×10^{-10} seconds per oscillation for the Cesium transition. No single clock (because of its finite size and strength) is conceivable all the way back to the singularity, so a statement about the proper time since the singularity involves the concept of an infinite sequence of successively smaller and sturdier clocks with their ticks then discounted and added. Finite proper time then need not imply that any finite sequence of events was possible. It may describe a necessarily infinite number of events (ticks) in any physically conceivable history, converted by mathematics into a finite sum by the action of a non-local convergence factor, the discount applied to convert ticks into proper time.

Here one has the conceptual inverse of Zeno's paradox. One rejects Zeno's suggestion that a single swing of a pendulum is infinitely complicated (being composed of a half period plus a quarter period, plus 2^{-n} ad infinitum) because the terms in his infinite series are mathematical abstractions, not physically achieved discrete acts in a drama that must be played out. By a comparable standard, one should ignore as a mathematical abstraction the finite sum of the proper time series for the age of the universe, if it can be proved that there must be an infinite number of discrete acts played out during its past history. In both cases, finiteness would be judged by counting the number of discrete ticks on realizable clocks, not by assessing the weight of unrealizable mathematical abstractions.

Whether the universe is infinitely old by this standard remains to be determined. The quantum influences remain to be determined. The decisive question is whether each present epoch event is subject to the influence of infinitely many previous discrete events. In that case statistical assumptions (large numbers, random phase) could enter in stronger ways into theories of cosmology. The Mixmaster cosmological model *does* have an infinite past history in this sense, since each bounce from one Kasner-like motion to another is a recognizable cosmological event, of which infinitely many must be realized between any finite epoch and the singularity.

3.2 The Kasner solution

All the cosmological observations confirm that the universe is homogeneous and isotropic to high accuracy on large scales. The question is: why is the unverse so symmetric? After all, homogeneity and isotropy is only a very idealized situation. We would like to understand what would have happened if the universe had started out highly irregular, so we allow large deviations from the symmetry of the FLRW universes and put asymmetries into only a few degrees of freedom.

The prototype for cosmological models with great asymmetry in a few degrees of freedom is the the *Kasner metric* or *Kasner solution*

$$ds^2 = -dt^2 + t^{2p_1}\, dx^2 + t^{2p_2}\, dy^2 + t^{2p_3}\, dz^2 \tag{3.1}$$

where the Kasner exponents p_i are constants satisfying

$$\sum_{i=1}^{3} p_i = \sum_{i=1}^{3} p_i^2 = 1 \tag{3.2}$$

Each t =constant hypersurface is a flat 3-dimensional space. The worldlines of constant x, y, z are timelike geodesics along which galaxies or other matter, treated as test particles, can be imagined to move. This model represents an expanding universe since the volume element

$$\sqrt{-\mathbf{g}} = \sqrt{{}^{(3)}\mathbf{g}} = t \tag{3.3}$$

is increasing. It is a homogeneous universe, anisotropically expanding universe (because in each space direction the three scale factors t^{2p_i} expand at a different rate). The relations (3.2) require that one of the p_i, say p_1 to be non-positive:

$$-\frac{1}{3} \leq p_1 \leq 0 \tag{3.4}$$

A consequence of this is that if black-body radiation were emitted at one time t and never subsequently scattered, later observers would see blue shifts near one pair of antipodes on the sky and red shifts in most other directions. The fundamental cosmological question is why the FLRW metrics should be a more accurate approximation to the real universes than this Kasner metric is. We can ask what would become of a universe that starts our near $t = 0$ with a form described by the Kasner metric. This metric is an exact solution of Einstein's equations in vacuum. It approximates a situation where the matter terms are negligible by comparison with typical non-zero components of the Riemann tensor. In the case of a pressureless fluid, the curvature of empty spacetime dominates both the geometry and the expansion rate at early times $t \rightarrow 0$, but after some characteristic time t_m the matter terms become more important and the metric reduces asymptotically to the homogeneous, isotropic model with $k = 0$, i.e. the Kasner model with matter becomes isotropic in old age.

This example illustrates the possibility that the universe might achieve a measure of isotropy and homogeneity in old age, even if it were born in a highly irregular state. Whether the symmetry of our universe can be explained along these lines is not yet clear. The model universe just mentioned is only a hint, especially since the critical parameter t_m can be given any value whatsoever.

This mechanism can also be described by ascribing the the anisotropic motion of empty spacetime an effective energy density ρ_{aniso}, which enters the G_{00} component of the Einstein equations on an equal footing with the matter-energy density, and thereby helps to account for the expansion of the universe

$$H^2 = \left(\frac{1}{3}\frac{d}{dt}\ln\sqrt{^{(3)}\mathbf{g}}\right)^2 = \frac{8\pi}{3}\left(\rho_{aniso} + \rho_{matter}\right) \tag{3.5}$$

The anisotropy energy density is found to have an equation of state

$$\rho_{aniso} \propto^{(3)} \mathbf{g}^{-1} = (\text{volume})^{-2} \tag{3.6}$$

while

$$\rho_{matter} \propto^{(3)} \mathbf{g}^{-\gamma/2} = (\text{volume})^{-\gamma} \tag{3.7}$$

where $\gamma = 1, 4/3, 5/3$ for pressureless matter, a radiation fluid, a non-relativistic ideal gas respectively. This arrangement of the Einstein equations allows one to think of the anisotropy motions as being adiabatically cooled by the expansion of the universe, just as the thermal motion of an ideal gas would be.

The conclusion is that, in principle, the mechanism of adiabatic cooling of anisotropy (together with other dissipative mechanisms which convert anisotropy energy in thermal energy and considering especially the quantum pair production effect through virtual quanta near the initial singularity) could explain the high homogeneity and isotropy of the present universe even if were born in a very irregular state.

The model universe considered above is homogeneous although anisotropic. It is also crucial to study inhomogeneous cosmological models, in which the metric has a non-trivial dependence on the space coordinate. The first attempt to understand the behavior of inhomogeneous and anisotropic solutions of Einstein equations had been developed by Belinskii, Khalatnikov and Lifshitz. Rather than truncating the Einstein theory by limiting attention to specialized situations where exact solutions can be obtained, they have sought to study the widest possible class of solutions, but to describe their behavior only in the immediate neighborhood of the singularity. These studies give a greatly enhanced significance to some of the exact solutions, by showing that phenomena found in them are in fact typical of much broader classes of solutions. In the first large class of solutions studied, it was found that near the singularity solutions containing matter showed no features not already found in the vacuum solutions. Furthermore, space derivatives in the Einstein equations become negligible near the singularity in these solutions with the consequence that a metric of the Kasner form described the local behavior of spacetime near the singularity, but with a different set of p_i values possible at each point of the singular hypersurface. Subsequently, broadened studies of solutions near a singularity showed that the Mixmaster universe is a still better homogeneous prototype for singularity behavior than the Kasner metric.

The *Mixmaster universe* is a generalization (still homogeneous) of the Kasner metric (3.1) to the case where the Kasner exponents p_i are functions of time. A convenient parametrization, due to BKL (see section 3.3), is the following

$$\begin{aligned} p_1(u) &= \frac{-u}{1+u+u^2} \\ p_2(u) &= \frac{1+u}{1+u+u^2} \\ p_3(u) &= \frac{u+u^2}{1+u+u^2} \end{aligned} \tag{3.8}$$

where u is a parameter greater than 1. As one extrapolates backward in time toward the singularity, one finds that the expansion rates in the three principal directions correspond to those of the Kasner metric, with p_i values corresponding to some fixed u parameter. In these Mixmaster models, the metric is not independent on the space coordinates.

The Kasner-like behavior at fixed u can persist through many decades of volume expansion before effects of the spatial derivatives of the metric come into play. The role played by the space curvature is brief and decisive. The expansion is converted from a type corresponding to a parameter value $u = u_0$ to a type corresponding to the value $u = -u_0$ (which is equivalent to the value $u = u_0 - 1$ after relabelling the axes). Extrapolating still farther back toward the singularity, one finds a previous period with $u = u_0 - 2$. Throughout an entire sequence $u = u_0, u_0 - 1, u_0 - 2, u_0 - 3 \ldots$, with $u_0 \gg 1$, nearly the entire volume expansion is due to expansion in the 3-direction, whereas the 1- and 2-directions change very little, alternating at each step between expansion and contraction. Sufficiently far in the past, however, such a sequence leads to a value of u between 0 and 1. This value can be interpreted as the starting point for another, similar sequence, through the transformation $u \to 1/u$, which interchanges the names of axes 2 and 3.

The extrapolation of the universe's evolution back toward the singularity at $t = 0$ therefore shows an extraordinarily complex (quasi-periodic) behavior, in which similar but not precisely identical sequences of behavior are repeated infinitely many times. In the generic example to which the BKL methods lead, one has a metric whose asymptotic behavior near the singularity is at each spatial point on the singular hypersurface described by a Mixmaster-type behavior, but with the principal axes of expansion changing their directions as well as their roles (as characterized by the u parameter) at each step, and with the Mixmaster parameters spatially variable.

We can ask if there are important solutions or classes of solutions, relevant to the cosmological problem, with asymptotic behavior not described by the BKL generic case, i.e.: are there any other generic types of behavior near singularities? The answer to this question is negative in the case of quiescent singularities (as shown rigorously in [1]) and seems to be negative also in the case of oscillating (chaotic) singularities, as indicated by the numerical simulations [17], by the theoretical work of Uggla et al. [73], and of H. Ringstrom [129].

3.3 BKL's metric approach

As we mentioned, some of the simplest nonlinear dynamical systems can have very complicated (even stochastic) behavior in spite of the fact that the equations are deterministic.

We will show that the evolution of the relativistic cosmological models towards the singularity undergoes *spontaneous stochastization.* The statistical parameters of this evolution can be calculated exactly. The knowledge of the source of stochasticity makes it possible to develop a quantitative statistical theory with appreciable completeness.

The evolution of a generic singularity can be described as an infinite succession of interchanging Kasner epochs with a certain law of replacement of the Kasner exponents when passing over from one epoch to the next one. This kind of behavior was first discovered for a vacuum homogeneous model of the Bianchi type VIII and IX and then generalized to the presence of matter. This latter introduces a new property in the evolution of the model: rotation of the Kasner axes (i.e. directions to which the scale factors a, b, c refer) during the interchange of Kasner epochs, but the law of interchange of the exponents remains the same. The solutions for Bianchi IX and VIII homogeneous models serve as a prototype for the construction of the general solution of the Einstein equations in the case of a generic cosmological singularity.

The law of replacement of the Kasner exponents remains the same also in the general inhomogeneous case. This law leads to an important property:

spontaneous stochastization of the behavior of the model on approach to singularity and the loss of memory of the initial conditions prescribed at some instant of time $t = t_0 > 0$. Thus stochasticity turns out to be a general property of relativistic cosmological models in the neighborhood of the singularity.

For the sake of concreteness, let us consider the Bianchi IX homogeneous model (in vacuum), which is governed by the the following equations

$$\begin{aligned} 2\,\alpha'' &= (b^2 - c^2)^2 - a^4 \\ 2\,\beta'' &= (a^2 - c^2)^2 - b^4 \\ 2\,\gamma'' &= (a^2 - b^2)^2 - c^4 \end{aligned} \tag{3.9}$$

$$\alpha'\beta' + \alpha'\gamma' + \beta'\gamma' - \frac{1}{4}\,(a^4 + b^4 + c^4 - 2a^2b^2 - 2a^2c^2 - 2b^2c^2) = 0 \tag{3.10}$$

where $a(t), b(t), c(t)$ are the 3 scale factors and α, β, γ are their natural logarithms respectively. The prime $'$ denotes the derivative with respect to a time variable τ related to the synchronous time t by the equation

$$dt = abc\,d\tau \tag{3.11}$$

The equation (3.10) contains only the first derivatives and thus plays the role of an additional restriction, imposed on the initial conditions for equations (3.9). It is easy to verify that the derivative of the expression (3.10) with respect to τ is indeed identically zero due to equations (3.9); thus if the solution of (3.9) satisfies the condition (3.10) in an initial instant of time, the latter will always be satisfied.

From a formal point of view, we deal with a deterministic dynamical model, governed by a system of three ordinary differential equations with one additional condition, so the phase space is actually not 6- but 5-dimensional. Apart from the profound cosmological significance of this system, we encounter here a specific mode of spontaneous stochastization of a deterministic system!

Let us denote by p_1, p_2, p_3 the Kasner exponents arranged in a fixed order with respect to their magnitude: $p_1 < p_2 < p_3$. These three numbers, subject to the Kasner relations (3.2), can be parameterized in the form

$$p_1(u) = -u/f(u), \quad p_2(u) = (1+u)/f(u), \quad p_3(u) = (u+u^2)/f(u) \tag{3.12}$$

$$f(u) = 1 + u + u^2$$

where the real parameter $u \in [1, \infty)$. The values $0 < u < 1$ can be reduced again to the interval $[1, \infty)$ using the formulas

$$p_1(1/u) = p_1(u), \quad p_2(1/u) = p_3(u), \quad p_3(1/u) = p_2(u) \tag{3.13}$$

As u decreases monotonically from ∞ to 1, the exponent p_1 decreases monotonically, while p_2, p_3 increase monotonically in the ranges

$$-\frac{1}{3} \leq p_1 \leq 0, \quad 0 \leq p_2 \leq \frac{2}{3}, \quad \frac{2}{3} \leq p_3 \leq 1 \tag{3.14}$$

so the exponent p_1 is always negative, while p_2 and p_3 are always positive, and $p_3 > p_2$.

The *Kasner regime* is a solution of (3.9)-(3.10) when all terms in the right-hand side can be neglected; we call a *Kasner epoch* the time interval during which it is admissible. Such an interval is certainly short with decreasing t since the right-hand side of eqs. (3.9) always contain an increasing term. For instance, if the negative exponent refers to the function $a(t)$ ($p_a = p_1$), the perturbation of the Kasner regime will be due to the terms α^4; the remaining terms decrease with decreasing t. This perturbation leads after a brief transitional period to an establishment of a new Kasner epoch with the following rule of replacement of the exponents: if

$$p_a = p_1(u), \quad p_b = p_2(u), \quad p_c = p_3(u) \tag{3.15}$$

then

$$p_a^{new} = p_2(u-1), \quad p_b^{new} = p_1(u-1), \quad p_c^{new} = p_3(u-1) \tag{3.16}$$

The function $a(t)$ acquires a positive exponent and starts to decrease (with decreasing t, the singularity is at $t = 0$); the function $b(t)$ acquires a negative exponent and starts to increase, the function $c(t)$ continues to decrease.

The subsequent evolution with the increasing function $b(t)$ leads in an analogous way to the next interchange of the Kasner epochs and so on. The successive interchanges according to the rule (3.16), accompanied by a bouncing of the negative exponent between the functions $a(t)$ and $b(t)$, continues as long as the integral part of the initial value of u is exhausted, that is until u becomes less than unity. The value $u < 1$ transforms into $u > 1$ according to (3.13); at this moment either the exponent p_a or p_b is negative and p_c becomes the smaller one of the two positive exponents ($p_c = p_2$). The next sequence of changes will bounce the negative exponent between the functions c and a or between c and b. For an arbitrary irrational initial value of u the process continues indefinitely.

Thus the evolution of the model on approaching the singularity consists of successive periods (which we call *eras*) during which two of the scale functions oscillate and the third one decreases monotonically. On passing from one era to another the monotonic decrease is transferred to another of the three scale functions (see the picture).

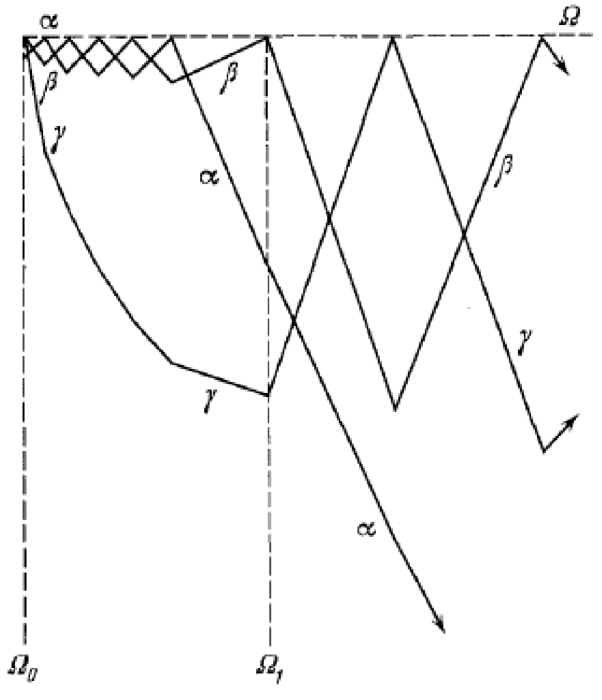


Figure 3.1: Evolution of the logarithms of the scale factors in terms of the logarithmic time $\Omega = -\ln t$.

To each $s-$th era there corresponds a series of values of the parameter

u starting with a certain largest one, $u_s^{(max)}$ and reaching the smallest one, $u_s^{(min)} < 1$, via the values $u_s^{(max)} - 1, u_s^{(max)} - 2, \ldots$. We can put

$$u_s^{(max)} = k_s + x_s, \quad u_s^{(min)} = x_s \tag{3.17}$$

where

$$k_s = \left\lfloor u_s^{(max)} \right\rfloor, \quad x_s = \left\{ u_s^{(max)} \right\} \tag{3.18}$$

are the integer and fractional part of $u_s^{(max)}$ respectively. The number k_s determines the length of the era measured in terms of the number of Kasner epochs it contains. For the next era

$$u_{s+1}^{(max)} = 1/x_s, \quad k_{s+1} = \lfloor 1/x_s \rfloor \tag{3.19}$$

The sequence of the lengths of the successive eras has a character of a *random* process. The *source of stochasticity* is just the rule (3.19). This rule states that if the entire infinite sequence begins with a certain initial value $u_0^{(max)} = k_0 + x_0$, then the lengths of the eras $k_0, k_1, k_2, \ldots$ are the numbers in the continued fraction expansion

$$u_0^{(max)} = k_0 + \cfrac{1}{k_1 + \frac{1}{k_2 + \cdots}} \tag{3.20}$$

As we said in Part I, this expansion is related to the Gauss map (1.33), which is highly chaotic. In particular, T also satisfies the criteria for Poincaré recurrence, this means that each Kasner solution is visited an arbitrary large number of times during the infinite sequence of oscillations.

In particular, this dynamical system has a metric entropy given by

$$h(T_{Gauss}) = \frac{\pi^2}{6\ \ln 2} \tag{3.21}$$

and it is isomorphic to a Bernoulli shift with the same entropy.

Note that the Gauss map accounts only for the transitions between successive Kasner epochs. One can do better, in fact following [34], it is possible also to consider the oscillations in two of the scale factors inside a single Kasner epoch and include them in a more complete map which describes the discrete evolution (still approximated). This can be realized through the *Farey map*, whose entropy is given by

$$h(T_{Farey}) = 2 \ln 2 \tag{3.22}$$

This map accounts for oscillations (one pair of axes oscillates while the third one decreases monotonically) and bounces (when the roles of the three axes are interchanged and a different axis decreases monotonically), according to a chaotic Farey tale [35]. These papers use fractal techniques (which are observer independent) to show that the Mixmaster universe is indeed chaotic.

Note that the Farey tree appears also in recent works on the entropy of black holes based on the string theory approach [42]-[43]. These papers explore somehow the idea of *spacetime* modular invariance, which is exactly what we deal with in this thesis.

In Part I, we have described Artin's theorem which relates the ergodicity of the geodesic flow on $X(1)$ to the ergodicity of the Gauss map. We have here an example of a physical system described by the Gauss map: the asymptotic behavior of a *homogeneous* Bianchi IX universe. We can ask if there is a similar physical systems described by the geodesic flow on $X(1)$. As we see in the next sections, such a system exists and it is precisely the asymptotic behavior of a generic *inhomogeneous* singularity. Thus, Artin's result supports the conjectures that the behavior of a generic singularity is somehow well described by a Bianchi IX homogeneous cosmological model.

3.4 DHN's Approach

In this section we describe the result due to DHN in the study of a general cosmological (spacelike) singularity without any symmetry assumption for the metric [2].

As usual we consider only the case of pure gravity in 4 dimensions. For more details, proofs and the case of higher dimensional theories see the paper [37].

DHN's result is the following. The BKL oscillatory behavior of pure gravity in 4 spacetime dimensions (in particular the billiard representation) is valid also for other physical theories. In the case also studied by BKL (pure gravity in 4 dimensions), the dynamics of Einstein equations close to the cosmological singularity is equivalent to a *null* geodesic motion inside a billiard given by a Coxeter polytope in a 3-dimensional Minkowski space. What is remarkable is that *the billiard is the Weyl chamber of the hyperbolic algebra* $HA_1^{(1)}$ (whose Weyl group is $\mathrm{PGL}(2,\mathbb{Z})$). The reflections at the walls are elastic. Each Kasner epoch is represented by the null geodesic segment between two successive reflections. In particular, given a Kasner epoch and the wall where this epoch crashes/ends, the following one is obtained by Weyl reflection with respect to simple root orthogonal to that face of the Weyl chamber. In other words, Weyl reflections with respect to simple roots send null geodesic segments into null geodesic segments, i.e. transform Kasner solutions into Kasner solutions (with different values of the Kasner

[2] If accidental symmetries are present in the metric, the analysis is *not* valid. For example, it is known that the Schwarzschild solution

$$ds^2 = -\left(1-\frac{2m}{r}\right)dt^2 + \left(1-\frac{2m}{r}\right)^{-1} dr^2 + r^2 d\Omega^2 \tag{3.23}$$

has a spacelike singularity at $r=0$; moreover, inside the horizon ($r<2m$) the r coordinate is time-like (there is a minus sign in front of dr^2). If we take the limit $r\to 0$, we obtain

$$\lim_{r\to 0} ds^2 = \frac{2m}{r}\,dt^2 - \frac{r}{2m}\,dr^2 + r^2\,d\Omega^2 \tag{3.24}$$

which is a Kasner metric

$$-d\tau^2 + \tau^{-2/3}d\sigma^2 + \tau^{4/3}(\sin(\theta)d\overline{\phi})^2 \tag{3.25}$$

once we put $\tau = \frac{2r^{3/2}}{3\sqrt{2m}}$, $\sigma = (4m/3)^{1/3}t$, $\overline{\theta} = (9m/2)^{1/3}\theta$ and $\overline{\phi} = (9m/2)^{1/3}\phi$, i.e. the Schwarzschild solution corresponds (in the neighbourhood of the singularity) to a *single* Kasner epoch, not to a never-ending succession of Kasner eras.

Finally, the analysis does not apply to time-like (like the one in the Reissner-Nordström solution, the charged black hole) or null singularities where a causal decoupling of spatial points does not occur. The question about the general behavior of not-space-like singularities is still open.

exponents). Note that in this formalism, only *null* geodesics are physical and correspond to Kasner solutions. The role of the spacelike and timelike geodesic flow inside the billiard is not clear. Note that the walls of the Weyl chamber are timelike, that is their orthogonal vectors are spacelike, since for the simple roots one has $(\alpha_i, \alpha_i) = 2 > 0$. Because of this, every reflection conserves the null character of the velocity vector.

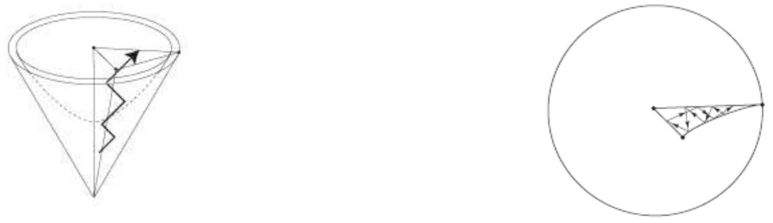


Figure 3.2: Null geodesic motion inside the Weyl chamber of $\mathrm{HA}_1^{(1)}$ and its projection on the Poincaré disc. The basis of the billiard is chaotic, being a non-compact region of finite hyperbolic area.

Let us stress that the motion occurs in a *Minkowskian* (or pseudo-Riemannian) 3-dimensional space, not in a Euclidean space. This space is indeed $\mathfrak{h}^*_{\mathbb{R}}$, in the notation of Chapter 2, endowed with the metric given by the Cartan matrix of $\mathrm{HA}_1^{(1)}$. The walls are the hyperplanes orthogonal to the simple roots, the incoming trajectories are null and the reflected ones are null too, as we have just said. Thus the billiard flow is a *null flow* in a pseudo-Riemannian space. This situation is different from the typical billiards which are embedded in *Riemannian* spaces. Indeed, in pseudo-Riemannian manifolds, one has three kinds of geodesics: timelike, null, spacelike. Consequently, one should first define (and this is not trivial) the corresponding geodesic flows, then study then billiard flows, which have different properties from the usual Euclidean billiard, because each reflection depends on the character of the wall and on the character on the incoming trajectory. We know that the Weyl group of $\mathrm{HA}_1^{(1)}$ is $\mathrm{PGL}(2, \mathbb{Z})$, and the flow on the standard (extended or not) modular domain on the hyperbolic plane is chaotic, being an Anosov flow. *But this does not imply the chaoticity of the null billiard flow in the full*

Weyl chamber. In fact, the latter could be less chaotic or even integrable [3]. Let us conclude saying that there are no results (perhaps not even studies) on geodesic/billiard flows in pseudo-Riemannian manifolds, and this is an interesting topic of future research.

Let us also observe that the metric entropy of the geodesic flow inside the fundamental domain of PGL$(2,\mathbb{Z})$ can be explicitly calculated[4]

$$h(\{S^t\} \text{ on } \mathrm{PGL}(2,\mathbb{Z})) = 1 \tag{3.26}$$

thus the billiard representation on the hyperbolic plane (or disc) and the BKL approach based on the Gauss map are *not* equivalent, that is they are not isomorphic as dynamical systems as their entropies are different. In fact, the first describes the behavior of a generic inhomogeneous singularity, the second the fate of a Bianchi IX homogeneous universe.

3.5 Quantum Birth of the Universe

Following [63], let us consider the Universe close to the initial singularity. We have mentioned many times that the physics of this process is generically captured by a Bianchi IX homogeneous metric

$$ds^2 = -dt^2 + g_{ij}(t)\,\omega^i\omega^j \tag{3.27}$$

[3] For example, the Bunimovich stadium is a 2-dimensional chaotic Euclidean billiard, but if we consider a 3-dimensional Euclidean billiard raising the stadium as a basis in the $z-$direction, then the billiard is integrable, because of a translation symmetry in the $z-$direction. That is, a 3-dimensional billiard with a chaotic basis is not necessarily chaotic. Our case is even more complicated, because the billiard (a Coxeter polytope) does not live in a Euclidean space, but in a Minkowskian space.

[4] if we put the Gaussian curvature $\mathcal{K} = -1$

with t the standard cosmic time. Limiting ourselves to the case of closed non-rotating universes [5], ω^i are a basis of 1-forms on the 3-sphere

$$\begin{aligned} \omega^1 &= \cos\psi\, d\theta + \sin\psi\, \sin\theta\, d\phi \\ \omega^1 &= \sin\psi\, d\theta - \cos\psi\, \sin\theta\, d\phi \\ \omega^3 &= d\psi + \cos\theta\, d\phi \end{aligned} \tag{3.28}$$

$$0 \le \theta < \pi\,, \quad 0 \le \phi < 2\pi\,, \quad 0 \le \psi < 4\pi$$

and we use Misner's parametrization

$$g_{ij}(t) = a^2(t)\ \left(e^{2\beta(t)}\right)_{ij} \tag{3.29}$$

$$\beta = \text{diag}\ (\beta_+ + \sqrt{3}\beta_-, \beta_+ - \sqrt{3}\beta_-, -2\beta_+)$$

with $\mathrm{Tr}\beta = 0$. Let us put g =det g_{ij} and R the scalar curvature of the full metric (both of them are functions of time t). The usual Einstein-Hilbert action

$$S = \frac{1}{16\pi G} \int L(t)\, dt \tag{3.30}$$

with Lagrangian $L(t) = (4\pi)^2 R(t)\sqrt{g(t)}$ can be expressed in the coordinates a, β_+, β_- in the following form

$$\frac{1}{12\pi^2}\, L = a^3 \left(\frac{-\dot{a}^2}{a^2} + \dot{\beta_+}^2 + \dot{\beta}_-^2 \right) - a\left[V(\beta_+, \beta_-) - 1\right] \tag{3.31}$$

where V is the potential

$$V(\beta_+, \beta_-) = \frac{1}{3}\ \text{Tr}\ \left(1 - 2e^{-2\beta} + e^{4\beta}\right) \tag{3.32}$$

Let us now change the temporal coordinate, $dt = N(t')dt'$; we will mostly use

$$dt = 12\pi^2 a^3\, dt_f \tag{3.33}$$

[5] If one considers Bianchi IX with rotation of axes, then the billiard is identified with a fundamental domain for $\Gamma_0(2)$: in fact M. Marcolli [105] shows that every geodesic on the hyperbolic surface $X(\Gamma_0(2))$ (which is also a 1-dimensional complex curve) not ending at cusps determines a Mixmaster universe. She also suggests to study the flow on the (singular) quotient space $\Gamma_0(2)\backslash\mathbb{H}$ considering the latter a (non-singular) non-commutative space in the sense of A. Connes rather than a classical (singular) topological space.

Consequently L is changed too, in particular with respect to the previous choice we have

$$L_f = -\left(\frac{\dot{a}}{a}\right)^2 + \dot{\beta_+}^2 + \dot{\beta}_-^2 - \left(\frac{a}{a_0}\right)^4 [V(\beta_+, \beta_-) - 1] \tag{3.34}$$

with $a_0 = 1/2\pi\sqrt{3}$ and the dot is the time derivative with respect to t_f. We thus have a model of minisuperspace with metric tensor $\widetilde{\mathbf{G}}$; the kinetic term of L, $\dot{\mathbf{g}} \cdot \widetilde{\mathbf{G}} \cdot \dot{\mathbf{g}}$, is subject to conformal transformations $\widetilde{\mathbf{G}}' = N^{-1}\widetilde{\mathbf{G}}$ by redefinition of time. The choice (3.33) we have made leads to a flat realization of minisuperspace (the f in the subscript stands for flat).

Let us now study the quantum system. Following Misner, we ask that the free choice of time coordinate must be preserved in the quantum theory, thus all conformally equivalent realizations of minisuperspace must be quantum-mechanically equivalent. This is simplified by using the present flat realization and writing the gravitational action $S = \frac{1}{16\pi G}\int L_f(t_f)dt_f$ for the Schrödinger equation as

$$L_f(t_f) = \int \frac{da}{a}\, d\beta_+\, d\beta_- \left[\left|\frac{\partial\psi}{\partial \ln a}\right|^2 - \left|\frac{\partial\psi}{\partial\beta_+}\right|^2 - \left|\frac{\partial\psi}{\partial\beta_-}\right|^2 - \left(\frac{a}{a_0}\right)^4 [V(\beta_+, \beta_-) - 1]\,|\psi|^2\right] \tag{3.35}$$

The corresponding wave equation is

$$\frac{\partial^2\psi}{\partial\Omega^2} - \frac{\partial^2\psi}{\partial\beta_+^2} - \frac{\partial^2\psi}{\partial\beta_-^2} + e^{-4\Omega}\,[V(\beta_+, \beta_-) - 1]\,\psi = 0 \tag{3.36}$$

where we have introduced a new (time) coordinate $\Omega = -\ln(a/a_0)$ (the authors of [63] consider the presence of a positive cosmological constant λ too). To ensure conformal invariance in the present 3-dimensional minisuperspace model the wave function ψ_c corresponding to the use of cosmic time t must be related to the wave function ψ for general time $dt' = dt/N$ via $\psi = N^{1/4}\psi_c$. In the flat case

$$\psi_c = (12\pi^2 a^3)^{-1/4}\psi \tag{3.37}$$

We want now to understand equation (3.36) in the limit $(-\Omega) \to -\infty$ (the initial singularity). It is very useful to perform the following coordinate transformation

$$\begin{aligned} \beta_+ &= \ln\frac{1}{\rho}\,\sinh\zeta\,\cos\phi \\ \beta_- &= \ln\frac{1}{\rho}\,\sinh\zeta\,\sin\phi \\ \Omega &= \ln\frac{1}{\rho}\,\cosh\zeta \end{aligned} \tag{3.38}$$

The wave equation becomes

$$-\frac{1}{\rho(\ln\rho)^2}\frac{\partial}{\partial\rho}\left[\rho(\ln\rho)^2\right]\frac{\partial\psi}{\partial\rho}+\frac{1}{\rho^2(\ln\rho)^2}\left[\Delta_{LB}-U(\rho,\zeta,\phi)\right]\psi=0 \tag{3.39}$$

where Δ_{LB} is the Laplace-Beltrami operator

$$\Delta_{LB}=\frac{1}{\sinh\zeta}\frac{\partial}{\partial\zeta}\left[\sinh\zeta\,\frac{\partial}{\partial\zeta}\right]+\frac{1}{\sinh^2\zeta}\frac{\partial^2}{\partial\phi^2} \tag{3.40}$$

and $U(\rho,\zeta,\phi)$ is a potential term containing the potential V

$$U(\rho,\zeta,\phi)=(\ln\rho^2)\left[\rho^{4\cosh\zeta}(V-1)\right] \tag{3.41}$$

For fixed ζ we have

$$-\infty<-\Omega<0\Leftrightarrow 0<\rho<1 \tag{3.42}$$

Note that up to here we have done no approximations, the wave equation (3.39) is directly derived from the original Bianchi IX metric only by coordinate transformations. Let us now consider the asymptotic limit $\rho\to 0^+$. Then the potential U vanishes inside and is $+\infty$ outside the triangular domain bounded by

$$\tanh\zeta=-\frac{1}{2}\sec\left(\phi+m\,\frac{2\pi}{3}\right)\qquad m=0,\pm1 \tag{3.43}$$

This makes possible to factorize solutions of equation (3.39) as

$$\Psi=\psi(\rho)\phi(\zeta,\varphi) \tag{3.44}$$

and one has first to consider the eigenvalue problem for the Laplacian

$$-\Delta_{LB}\,\phi(\zeta,\varphi)=\lambda\,\phi(\zeta,\varphi) \tag{3.45}$$

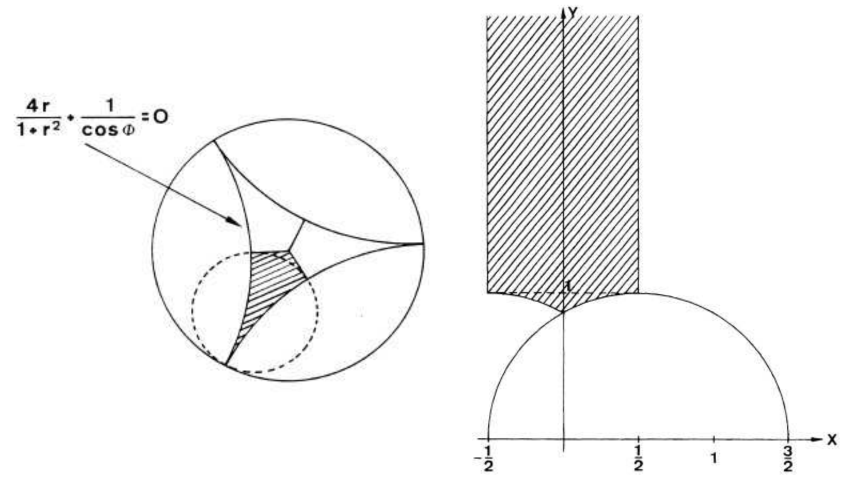


Figure 3.3: Billiard table for a Bianchi IX universe on the Poincaré disc and the hyperbolic plane (from [63]). This domain is non-compact, but still of finite hyperbolic area, thus it is a chaotic domain.

with suitable boundary conditions at the walls.

This equation is better solved mapping the problem on the Poincaré disk through the transformation

$$r = \tanh \frac{\zeta}{2} \tag{3.46}$$

which leads to a quantum billiard problem on the Poincaré disk in the coordinates (r, φ). Classically trajectories are broken geodesics reflected at the boundary and as we said many times the motion is known to be strongly chaotic. Finally, it is useful to go to the Poincaré upper-half plane (with coordinates (x, y)) using the fractional linear transformation

$$x + iy = \frac{3^{1/2}}{2} \frac{-iz\, e^{i\pi/6} + i}{z\, e^{i\pi/6} + 1}\,, \quad z = r\, e^{i\varphi} \tag{3.47}$$

The Laplacian problem on the hyperbolic plane becomes

$$-y^2 \left(\frac{\partial}{\partial x^2} + \frac{\partial}{\partial y^2} \right) Z(x, y) = \lambda\, Z(x, y) \tag{3.48}$$

which is the problem studied by Maass as we said in Part I. Dirichlet boundary conditions are not good in this case, since the domain is not compact, and one must use Neumann boundary conditions. We explore the consequence of this in the next section.

3.6 The wave function of the Universe

As we said, the general inhomogeneous case is modelled on a billiard problem inside the fundamental Weyl chamber of the hyperbolic algebra $\mathrm{HA}_1^{(1)}$. As in the previous section, we can decompose the motion in a chaotic motion inside a fundamental domain for $\mathrm{PGL}(2, \mathbb{Z})$ plus a radial part. This means that the angular part of the wave function of the early universe (considering only the case of pure gravity in 4 dimensions) is an automorphic L-function (indeed a Maass waveform) for the modular group. Actually, since the projected billiard is $\mathrm{PGL}(2, \mathbb{Z})$ and not $\mathrm{PSL}(2, \mathbb{Z})$, one should instead consider the spectral problem with Neumann boundary conditions

$$\begin{aligned} -\Delta\phi &= \lambda\phi \\ \phi &\in L^2(\mathcal{F}_3, \mu) \\ \partial_n \phi|_{\partial\mathcal{F}_3} &= 0 \end{aligned} \tag{3.49}$$

where $\mathcal{F}_3$ is the halved standard modular domain (remember that one of the angle is $\pi/3$), Δ is the hyperbolic Laplacian and μ is the usual measure on the hyperbolic plane. Thus solutions to the Neumann problem for $\mathrm{PGL}(2,\mathbb{Z})$ are given by *even* Maass cusp forms for $\mathrm{PSL}(2,\mathbb{Z})$. The precise statement, then, is that *the wave function of the universe is an even Maass cusp form for the modular group* $\mathrm{PSL}(2,\mathbb{Z})$. Actually, more is true. In fact, since $\mathrm{PSL}(2,\mathbb{Z})$ is arithmetic and the cuspidal spectrum is very likely simple, one can diagonalize the hyperbolic Laplacian and the Hecke operators simultaneously (see Appendix A). Thus it turns out that *the wave function is a Maass-Hecke eigenform.* This is an even more interesting statement. In fact, Hecke eigenvalues are multiplicative (see Appendix A), i.e. $\lambda(mn) = \lambda(m)\lambda(n)$, and this should put some conditions on the physical interpretation of them. These eigenvalues are the energy levels too (because they are eigenvalues of Δ), thus if we denote by E_n the Hecke eigenvalue λ_n and take the logarithms, we have

$$\ln E_{mn} = \ln(E_m E_n) = \ln(E_m) + \ln(E_n) \tag{3.50}$$

for m, n co-primes. This should give some information on the entropy of the system, since we expect this proportional to the logarithms of the density of states, according the Boltzmann formula. In recent years, in string theory, a lot of work has been done on counting the entropy of black holes (see for example [123] for a review). It turns out that in some cases this entropy is counted by the Fourier coefficients of certain automorphic functions. In the present case, the question would be the following: *is the gravitational entropy computable from the Fourier coefficients of Maass-Hecke eigenforms*? Another remark is the following. In the semi-classical limit, the expected statistics for the level spacing distributions for $X(1)$ (or E_{10}) is the Poisson distribution. This should, in principle, allow to compare with the observations.

The inclusion of matter changes the shape of the billiard and also increases the number of dimensions of the billiard. The most interesting case is perhaps supergravity theory in 11 dimensions, which is believed to be a kind of ultimate theory unifying all fundamental interactions. The supergravity billiard is the fundamental Weyl chamber for E_{10}, thus in this case the statement is that the wave function is a Maass waveform with respect to the (discrete) Weyl group of E_{10} (which is not known). In this case, we can not speak safely of Maass cusp forms, and we must use only the general term Maass waveform, because we do not know if the residual spectrum is empty as in the cases of each $\Gamma(N)$. Besides, the discrete spectrum could also be degenerate. The existence of Maass forms with respect to $W(E_{10})$ should be guaranteed by the fact $W(E_{10})$ is arithmetic [152].

Usually in quantum mechanics, one has a discrete and a continuous spectrum for a self-adjoint Hamiltonian and these are separated. Bound states are the proper eigenfunctions of the discrete spectrum, whereas the continuous part is interpreted as a free motion. Our model of quantum cosmology is different from this usual situation. In fact, from the Selberg theory for the automorphic Laplacian inside a fundamental domain of some $\Gamma(N)$, we know that the discrete spectrum is *not* separated from the continuous one $[\frac{1}{4}, \infty)$ (whose improper eigenfunctions are given by the Eisenstein series), but it is *embedded* in the continuous part [6]. Moreover, apart from the trivial eigenvalue λ_0 and the corresponding constant eigenfunction ϕ_0, the first eigenvalue is very likely $\lambda_1 = \frac{1}{4}$. Is the ground state of quantum cosmology given by a constant eigenfunction? Remember that the full wave function Ψ is a solution of some kind of wave equation [7] in the Coxeter 3-dimensional billiard. What we call the angular part is a Maass cusp form (i.e. it is zero at the cusps) on the hyperbolic plane. As we wrote above, the remaining part depends of the coordinate ρ. We used the term radial part, but this is not quite correct physically, because the ρ variable is a *time coordinate*, thus $\psi(\rho)$ should give the time evolution of the wave function close to the cosmological singularity. I do not expect that the $\rho-$dependence may change the features of the physical spectrum, which is thus contained in the angular part ϕ (for Bianchi IX this function is calculated in an approximated way in [63]).

We believe this says something about the nature of quantum gravity, i.e. this suggests that *quantum gravity/cosmology is a non-trivial mixing of discrete and continuous concepts*, whereas it is commonly believed that

[6] There are cases in scattering problems where points of the discrete spectrum lie in the continuous spectrum (I thank prof. G. Marmo for informing me on that), but our situation is distinguished, because the *whole* discrete spectrum is embedded in the continuous one.

[7] We have considered a mini-superspace model following Misner's insight. It is known the physical interpretation of the Wheeler-DeWitt equation is problematic, because it contains a second-order differential operator in the time variable. Even if one could in principle think of another wave equation to describe the quantum system, I believe this should still be of the second order in time, considering the fact the the classical evolution is described by *null* segments, i.e. rays of lights reflected at the walls. Moreover, I believe that the angular part should still satisfy the eigenvalue problem for the hyperbolic Laplacian, thus the Maass problem is relevant anyway.

quantum gravity means discretization of spacetime.

It would be interesting to understand if other approaches to the problem of quantum gravity like the formalism of loop quantum gravity or string theory say something like that.

Let us now come to the interpretation of the imaginary roots of the hyperbolic algebra $\mathrm{HA}_1^{(1)}$. As we showed in Part I, all the periodic orbits for $X(1)$ can be put in correspondence with a subset of imaginary roots of $\mathrm{HA}_1^{(1)}$. In the DHN formalism, we have seen that the simple roots define reflections in the walls, in particular each null geodesic is interrupted at a wall and then reflected through Weyl reflections

$$w_{\alpha_i}(\beta) = \beta - 2\,\frac{(\alpha_i, \beta)}{(\alpha_i, \alpha_i)}\,\alpha_i \tag{3.51}$$

(compare with the general billiard reflection formula in Part I, where the vectors orthogonal to each face of the billiard have norm 1, in our case $(\alpha_i, \alpha_i) = 2$). Remember that each null segment between two reflections corresponds to a Kasner solution. This means that the fundamental Weyl reflections transform Kasner solutions into Kasner solutions, i.e. *the Weyl group seems to act on the space of solutions of Einstein's equations*. Indeed, this Weyl group should be an algebraic technique to generate solutions to Einstein equations [8]. It is also very likely that the other Weyl reflections, defined in terms of real (not simple) roots also transform solutions into solutions (see comments at the end of this chapter). The imaginary roots do not define reflections, *but* they are related to periodic orbits of the billiard flow for $X(1)$. Thus, as the real roots transform solutions of Einstein equations, the imaginary roots can be interpreted as *periodic (non-singular) solutions to Einstein equations*. As the number of ppo is infinite, this gives an algebraic proof that in the BKL limit to the singularity, *there are infinitely many periodic solutions to Einstein equations*. Note that the cosmological singularity (the singular hyper-surface) has not disappeared: in fact, the billiard table has a cusp at infinity. If we had found some compact billiard table, the interpretation would have been a little bit problematic. The interpretation of the imaginary roots as periodic solutions is reasonable considering also that Margulis asymptotics for the number of periodic orbits is very similar to the

[8] This is not new. R. Geroch found an infinite-dimensional Lie group transforming solutions to Einstein theory among them in the case of pure gravity reduced to 3=2+1 dimensions. The *Geroch group* was then proved to be the loop group $\mathrm{A}_1^{(1)}$.

asymptotics of imaginary roots. Besides, the positivity of the Kac-Peterson function on imaginary roots is, in my opinion, another indication of that, since this function resembles an entropy function a lot.

This is an attempt of an answer to the question “what is the role of imaginary roots in the Kac-Moody formulation of gravity/string theories?”, at least for the imaginary roots coded by the periodic orbits and for the case of $\mathrm{HA}_1^{(1)}$ and supports the idea the algebra $\mathrm{HA}_1^{(1)}$ is a symmetry of the space of solutions of pure gravity in 4 dimensions, as it contains also infinitely many periodic solutions. This seems to be reasonable at least in the billiard regime. If the hyperbolic Kac-Moody algebra is a hidden symmetry of the *full* theory, the role of real and imaginary roots can be different; in fact writing a Lagrangian formally invariant under an infinite-dimensional group, at low levels the roots have the right symmetry properties in order to be associated to the *fields* of the theory, not to solutions.

Note that periodic solutions to Einstein equations were also found by S. W. Hawking in [72] in the case of gravity coupled to scalar fields. See also the papers by D. N. Page [122] and by A. Yu. Kamenshchik [97].

Finally, remember that each ppo γ_0 is given by the reduced matrix

$$A = T^{n_1} S \, T^{n_2} S \, \cdots \, T^{n_m} S \tag{3.52}$$

or in terms of Weyl reflections by

$$A = [(W_2 W_1)^{n_1} \; W_1 W_3] \; [(W_2 W_1)^{n_2} \; W_1 W_3] \cdots [(W_2 W_1)^{n_m} \; W_1 W_3] \tag{3.53}$$

This means that if we consider an alphabet of 3 letters $\{W_1, W_2 W_3\}$ with a grammar given by the commutation rules of the W_i's (see Part I), then we can construct infinitely many periodic solutions to Einstein equations with just these three letters subject to the previous constraint. That is, in this regime, the periodic solutions can be coded through the Weyl reflections. The shortest periodic orbits derive from hyperbolic matrices A with Tr $A = 3$; the corresponding hyperbolic length is

$$l_{min} = 2\cosh^{-1}(\mathrm{Tr}A/2) = 2\cosh^{-1}(3/2) = 2\ln\left(1 + \frac{1+\sqrt{5}}{2}\right) \tag{3.54}$$

Thus, the length of the shortest periodic orbit is related to the golden ratio, and, of course,one can make numerous aesthetic comments about that. Anyhow, note that the same formula appears in a paper by A. Yu. Kamenshchik [97] in a different context (they consider scalar fields too).

The difficult part is how to write down *explicitly* the metric of these periodic solutions, because our result is only an existence results and uses the tools of symbolic dynamics to code these periodic solutions. The interesting and physical part is how these periodic metrics look like.

3.7 Scarred States in Quantum Cosmology and Counting of Quantum States

In [12], Barrow and Levin have analyzed the case of a finite universe studying the case of a universe emerging from a compact octagon on the hyperbolic plane. They found a fractal structure which, in the classical to quantum transition, can persist in the form of scars, ridges of enhanced amplitude in the semiclassical wave function. They conclude that if the universe is finite and negatively curved, the cobweb of luminous matter might be a residue of primordial quantum scars.

Of course, we can never know if our universe is finite or not. Yet, in the case of a generic singularity, we have seen that the interesting fundamental domain is the one for $\mathrm{PGL}(2,\mathbb{Z})$, for which *there is no scarring.* Thus the conclusions by Barrow and Levin do not apply, and it remains to understand the physical (cosmological) meaning for the absence of scarred states. Again, results in this direction given by loop quantum gravity or string theory would be interesting, in order to confirm or discard the roles of scarred states in quantum cosmology.

Note one more thing. Our analysis is limited to the case of pure gravity in 4 dimensions, thus the conclusion about the absence of scarring states in quantum cosmology is valid only in this context. Thus one may think that the inclusion of matter or extra dimensions could change the situation. If we consider the case of supergravity theory in 11 dimensions (or a not well defined quantum version of it, like M-theory or whatever), then the cosmological billiard is the Weyl chamber of E_{10} (the role of the fermions is still matter of debate). Its Weyl group is not known, but it is known [152] that is it still *arithmetic.* Thus for the hyperbolic manifold $W(\mathrm{E}_{10})\backslash\mathbb{H}^9$ the quantum unique ergodicity theorem should be true and, again, there is no scarring effect. 11-dimensional supergravity is a candidate theory to describe the universe and all of its interactions. The message is that, with the knowledge we have today, it does seem that *in the early universe scarred* *states are absent.* And, in my opinion, quantum gravity has to cope with quantum chaos.

Finally, regarding the asymptotic number of quantum states, let us write again the Selberg result for $\mathrm{PSL}(2, \mathbb{Z})$

$$N_{\Gamma(1)}(\lambda) = \sum_{0<\lambda_j\leq\lambda} 1 \sim \frac{\mu(X(1))}{4\pi}\,\lambda \tag{3.55}$$

for $\lambda \to \infty$. This gives the asymptotic number of eigenvalues/eigenfunctions, thus *it counts the asymptotic number of discrete states in quantum cosmology.* Again, a confirmation of this from other approaches to the problem to quantum cosmology would be extremely interesting. For E_{10}, there is no such a result.

3.8 Notes and Comments on Chapter 3

The singularity theorems of S. W. Hawking and R. Penrose state that (under very general and reasonable assumptions) the solution to Einstein equations is not singularity free (see [71], [70], [155]). This is a result in differential topology (using Morse theory for Lorentzian manifolds) and does not account for effects of quantum physics at some small length scale: the general belief is that quantum effects must be incorporated in order to cope with gravity at the Planck energy scale ($E_P \sim 10^{19}$ GeV) [9]. However, the beautiful theorems due to Hawking and Penrose do not say anything about the nature of the singularity.

The first attempt to understand what happens in the case of a cosmological singularity dates back to the Russian physicists V. A. Belinski, I. M. Khalatnikov and E.M. Lifshitz (BKL). In a series of works (see [15]-[16] for reviews and other older references), after a wrong statement by Khalatnikov and Lifshitz about the absence of a singularity in the solutions to Einstein equations, BKL showed that the generic solution to Einstein equations in vacuum admits never ending oscillations in the spacetime metric and exhibits chaotic behavior. Their analysis was limited to the case of pure gravity in 4 dimensions and to homogeneous cosmological models. The coupling of the gravitational field to a scalar field was also studied [14] and the result

[9] There is always the problem of the *falsification* of a theory of quantum gravity, since it deals with Planck scale physics.

was that oscillations disappear: chaos is replaced by a monotonic power law evolution (Kasner-like solution). Other studies, mostly due to C. W. Misner [115]-[116], J. D. Barrow [11], [29] focused on the case of Bianchi IX, renamed by Misner the " Mixmaster Universe " for its chaotic behavior. A definitive answer to the question of chaoticity of the Mixmaster universe was given by N. Cornish and J. Levin [34]-[35]. In fact, in general relativity, because of the general covariance of the theory, care must be used in order to state the presence or the absence of chaos only using the positivity of Lyapunov exponents, because these quantities can depend on the particular choice of the coordinate systems. Lyapunov exponents are not reliable indicators of chaos in general relativity. Good discussions can be found in the book edited by A. Coley and D. Hobill [31]. Cornish and Levin used coordinate independent, fractal techniques to show that the Mixmaster universe is indeed chaotic. A fractal set of self-similar universes is uncovered by numerically solving Einstein's equations. These universes form fractal boundaries in the space of initial conditions. Such fractal partitions are the result of a chaotic dynamics.

The analysis of BKL has been improved in the last very years by T. Damour, M. Henneaux and H. Nicolai [37] (DHN) especially in the case of supergravity/string theories. These authors extend previous results to the case of pure gravity in $D \geq 4$ dimensions, the bosonic sector of the low energies effective actions of string theories in 10 dimensions and the bosonic sector of supergravity in 11 dimensions. Under *no* symmetry assumptions (unlike the homogeneous cases previously discussed), using only the existence of a foliation of spacetime by spacelike hypersurfaces Σ_t in the limit to the singularity, they show that the generic solution to (generalized, with matter) Einstein equations is still oscillatory in the asymptotic limit (except for pure gravity for $D \geq 11$ where it becomes Kasner-like). In particular, they show that the asymptotic dynamics is equivalent to a geodesic motion of a massless particle which moves at the speed of light as a free ball inside a billiard (in an auxiliary Minkowski space) with elastic reflections on the walls [10]. The shape of the billiard and the dimension of this auxiliary Minkowski space depend on

[10] The billiard representation on the hyperbolic plane is originally due to Misner [115], Chitre [30] and appears also in various works of A. A. Kirillov, V. D. Ivashchuk, V. N. Melnikov [99]-[100], [83]-[84], G. Imponente and G. Montani [82]. The analysis of DHN deals with more degrees of freedom at the same time (gravitational field, matter) and is more general. The analysis makes use of the Hamiltonian formalism for general relativity and of an Iwasawa decomposition for the metric.

the theory (number of spacetime dimensions, scalar fields, p-forms). For any "relevant" physical theory, it has been found that the billiard can always be identified with the fundamental Weyl chamber of a suitable hyperbolic Kac-Moody algebra. Excellent reviews of these approaches are the PhD thesis of my friend S. De Buyl [41] and the very recent review by M. Henneaux, D. Persson, P. Spindel [78].

Finally, the work of C. Uggla et al [2], [49] has shown that in a generic cosmological model *asymptotic silence holds*, that is particle horizons shrink to zero along all timelines in the limit to the singularity (in Uggla's [149] words at the last MG11 " Everybody dies alone "). These works also give some evidence to the fact that spatial derivatives become dynamically insignificant along generic timelines, supporting the initial BKL's conjecture. A past attractor, the *cosmological billiard attractor*, has been identified in [73], but this billiard is not the same as the Kac-Moody billiard. A good background reading to understand these works is the book by J. Wainright and G. F. R. Ellis [153] on dynamical systems in cosmology.

All of these analyses support the fact that the generic solution to Einstein equations in vacuum in the asymptotic limit towards a cosmological singularity is oscillatory and chaotic; moreover, in this limit, the spatial points decouple and the evolution of each spatial point is Mixmaster-like. This has also been checked trough many numerical simulations, see the recent report by B. K. Berger [17]. Unfortunately, a rigorous proof of the BKL limit is lacking in the chaotic case (infinite oscillations of the metric), whereas there are rigorous results in the Kasner-like case (gravity coupled to scalar fields, sub-critical systems) [1], [39]. A mathematical proof of the BKL limit is a very important point, especially regarding the possibility of ignoring the spatial gradients, whose occurrence transforms the nonlinear *partial* differential equations of general relativity in nonlinear *ordinary* differential equations in the time variable t (ordinary differential equations which mimic homogeneous models). The idea is to prove in a rigorous way that this limit exists and corresponds to the BKL limit. Some experts in nonlinear PDEs could like the problem and solve it for us. The reward is a journey into the cosmological singularity! Any volunteer?

In this work, we have studied the case of pure gravity in 4 dimensions, which is described by the null-geodetic motion inside the Weyl chamber of $\mathrm{HA}_1^{(1)}$. We deal with this case only because it can shed some light on gravity in the asymptotic limit (where quantum effects and/or other unknown phenomena should be considered) and also because $\mathrm{HA}_1^{(1)}$ is the simplest hyperbolic Kac-Moody algebra about which we know something more with respect to other hyperbolic or Lorentzian Kac-Moody algebras like E_{10} or E_{11}. The key hypothesis is that the symmetry (the Weyl chamber of $\mathrm{HA}_1^{(1)}$) survives any kind of transition from classical to quantum (or whatever) phenomena in approaching the singularity. We support this assumption (which can be completely wrong) because we believe that in a neighbourhood of the cosmological singularity it is more natural to describe the physics in terms of discrete and arithmetic mathematical objects rather that in terms of smooth objects. In our case, this means that rather considering the infinite-dimensional Lie group [11] corresponding to the infinite dimensional Lie algebra (an indefinite Kac-Moody algebra) we want instead to study what happens in the Weyl chamber of the algebra (which is a well defined discrete structure): in a certain sense this approach is complementary to the one developed by DHN, P. West, F. Englert and L. Houart, and others, based on the level decomposition of the corresponding Kac-Moody algebra with respect to a finite-dimensional Lie algebra.

We have shown in particular that the *imaginary* roots of the algebra $\mathrm{HA}_1^{(1)}$ can be interpreted as periodic solutions of the billiard flow in its chamber once projected on the hyperbolic plane. As concerns the *real* roots, at the moment I think that their roles is to define Weyl reflections which transform solutions of Einstein equations into other solutions. This is basically the point of view adopted in [47], where the question is faced in relation to supergravity theory and BSP solutions.

So what is the role of imaginary roots?

We believe the other part of the algebra, i.e. the imaginary roots spaces, which in some sense is the biggest one, is associated to periodic solutions.

It is clear that if we believe the billiard representation for gravity close to the singularity, then the periodic orbits must have a role, since these are classical solutions to the Hamiltonian flow on the hyperbolic plane. Remember also that these trajectories are very important because they are dense,

[11] The fact that in the study of the singularity it is the Weyl chamber which appears (and so naturally, the Kac-Moody group) as a symmetry can be a hint that chambers, buildings, apartments and all that are the right language to describe physics in the proximity of a singularity. This of course suggests a discretization of spacetime in this regime and somehow reminds of the approaches to quantum gravity based on triangulations of spacetime, though in our case the Weyl chamber does not live on the spacetime as we have a kind of holographic description.

thus they can approximate any other trajectory. It is reasonable that these periodic geodesics are solutions to Einstein's equations in the BKL limit and that they are related to the imaginary roots. The Weyl reflections are used to code these periodic orbits. The symbolic coding of cosmological solutions is not new, see [34]-[35] and [97] for example. But the fact that we can use an alphabet of Weyl reflections to code periodic orbits supports the role of the algebra $\mathrm{HA}_1^{(1)}$ as a symmetry of the gravitational theory in the limit to the singularity.

As we have briefly mentioned, the entropy of certain black holes can be computed by the Fourier coefficients of certain automorphic forms. In particular [123], a holographic description allows to quantize BPS black holes. The case we have studied, the limit of gravity near a cosmological singularity, also reveals a kind of holography, since the behavior of Einstein equations in this regime is equivalent to a null motion in an auxiliary Minkowski space. And as we have seen, the wave function in this auxiliary description is an automorphic form with many arithmetic properties, exactly as the wave function of these black hole is an automorphic function. This is an example of what I have in mind for "The Automorphic Universe". These connections deserve to be investigated in depth, especially in relation to the many holography conjectures which are today stated.

Conclusions and Some Speculations

> The old problems have not been solved, but little by little they make less sense, they are forgotten, they disappear . . .
>
> *Chance and Chaos*
> D. Ruelle

In this thesis, we have assumed DHN's result according to which the dynamics of classical general relativity in 4 dimensions (and in absence of matter) is equivalent, in the BKL limit, to a null geodetic motion inside the fundamental Weyl chamber of the hyperbolic Kac-Moody algebra $\mathrm{HA}_1^{(1)}$ with elastic reflections at the walls (billiard representation). The evolution of the spacetime metric in this limit can be represented as an infinite succession of Kasner epochs: each Kasner epoch can be mapped to a geodesic segment inside the fundamental domain of $\mathrm{PGL}(2, \mathbb{Z})$, which is the Weyl group of $\mathrm{HA}_1^{(1)}$. This fundamental domain also contains infinitely many periodic orbits (which are solutions of the classical Hamiltonian flow) which we have put in relation with the imaginary roots of the hyperbolic Kac-Moody algebra $\mathrm{HA}_1^{(1)}$. These periodic orbits correspond to classical periodic solutions to Einstein equations and escape the cosmological singularity since they are periodic. Periodic orbits can be coded through symbolic dynamics, i.e. to each periodic orbit we can associate a finite set of letters from an alphabet. The letters are the Weyl reflections and the alphabet is given by the commutation rules of these reflections. At the moment, I can not think of a more primitive tool than symbolic dynamics (i.e. grammar) to describe gravity close to a cosmological singularity.

Regarding the question of integrability/chaoticity of the null billiard flow inside the Weyl chamber of $\mathrm{HA}_1^{(1)}$, one can not claim anything, as there are no results for such flows in pseudo-Riemannian manifolds. In fact, from the Anosov property of the flow on the basis of the billiard (the fundamental domain for $\mathrm{PGL}(2,\mathbb{Z})$), it does *not* follow the chaoticity for the dynamics in the full 3-dimensional Weyl chamber, since from the ergodicity of a dynamical system in a proper subset of the phase space one can not infer the ergodicity of the system in the full phase space. I do hope to study billiards and flows in pseudo-Riemannian manifolds in the near future, since these are the natural evolution of classical billiards towards considering the problem of ergodicity in special/general relativity. The next step, in fact, should be the study of relativistic billiards confined in some polyhedron in a Minkowski spacetime (think of a relativistic gas).

We have also carried on for the first time the quantum analysis for the billiard representation. The result is that, for pure gravity in 4 dimensions, the wave function of the universe (or better, the angular part projected on the hyperbolic plane) is an automorphic form for $\mathrm{PGL}(2,\mathbb{Z})$, precisely an even Maass cusp form for $\mathrm{PSL}(2,\mathbb{Z})$. Indeed, since $\mathrm{PSL}(2,\mathbb{Z})$ is arithmetic, *the wave function is a Maass-Hecke eigenform*, being also eigenfunction of the Hecke operators (which commute with the Hamiltonian). The arithmetic nature of $\mathrm{PGL}(2,\mathbb{Z})$ allows also to state that in the early universe *scarred states are absent*. This conclusion, true for pure gravity in 4 dimensions if we believe the billiard representation, is very likely valid also in the case of 11-dimensional supergravity close to the singularity, whose billiard is modelled on the E_{10} hyperbolic Kac-Moody algebra. Finally, we have pointed out that the Selberg trace formula for $\mathrm{PSL}(2,\mathbb{Z})$ gives a semiclassical quantization rule for pure gravity in 4 dimensions, as typically occurs for the Gutzwiller trace formula in quantum chaos. But the difference is that the Gutzwiller trace formula is divergent, whereas the Selberg trave formula is convergent. Thus, in the BKL limit to the singularity, a semiclassical quantization of gravity is well defined. Indeed, the Selberg trace formula is a kind of path-integral. This point of view is emphasized in the book by C. Grosche [65]. But remember that indeed we have an infinite number of semi-classical quantization rules, since the class of test functions entering the trace formula is very large. This is the best one can do in a rigorous/convergent way for the quantum systems whose semi-classical limit is a Hamiltonian flow for which the Selberg trace formula is valid.

Moreover, as it happens in all situations in which quantum chaos exists, that is whose underlying classical dynamical system is chaotic, one should face the question of decoherence. In our situation, this would be very important, since we are speaking of decoherence in quantum gravity/cosmology and this means to enter the debate whether it is more correct to consider quantum mechanics as a collapse of the wave function or a many-worlds interpretation. Some work on decoherence in quantum gravity and the role of time has been done recently by R. Gambini and J. Pullin.

This is what we have found for the case of pure gravity in 3+1 dimensions, where spatial points decouple and the dynamics involves only a single (time) variable. The question now is: what could be the analog for a quantum theory when all 4 (time *and* space) variables must be considered? A very simple (and possibly wrong) answer to this question can be the following. Classical general relativity in 4 dimensions is a covariant theory, where all physical quantities must be invariant under the diffeomorphism $\mathrm{GL}(4,\mathbb{R})$. Then one could suggest (and we do) that the corresponding quantum system is described by a wave function which is an automorphic form with respect to a discrete group which has something to do with $\mathrm{GL}(4,\mathbb{Z})$, which is the simplest choice just to start. Moreover, we have seen that the Selberg trace formula for $\mathrm{PGL}(2,\mathbb{Z})$ gives a natural and well-defined path-integral to quantize the system in the asymptotic limit. For $\mathrm{GL}(4,\mathbb{Z})$ (or some other group) one can invoke a generalized trace formula, the Arthur trace formula, which does the same job. One can also try to build a quantum gravity from quantum cosmology by invoking some lifting of automorphic forms, that is a technique which allows to extend our Maass waveform to be an automorphic function with respect to a higher-dimensional discrete group [12]. It is likely that in this lifting procedure, some discrete group is naturally selected. Everything is mathematically well defined and beautiful, and leads, unavoidably, to the

[12] Note that this proposal is different from the one contained for examples in [37] or in [157] or in [47] (and references therein). DHN's approach is based on the assumption (for which there is indeed evidence) that supergravity theory in 11 dimensions (or a quantum extension of it) exhibits a hidden E_{10} symmetry at the level of the Lagrangian, i.e. there should exist a formulation of this theory invariant under the infinite-dimensional Lie group $\mathrm{E}_{10}\backslash\mathrm{K}(\mathrm{E}_{10})$, where $\mathrm{K}(\mathrm{E}_{10})$ is the formal maximal compact subgroup of the E_{10} Lie group. In this setting, the discrete group one looks for to build the quantum theory is a discretized version of E_{10}, $\mathrm{E}_{10}(\mathbb{Z})$. Our proposal is different: we suggest to look at a discrete group which is, naively, related to $\mathrm{GL}(4,\mathbb{Z})$, *not* to some $\mathrm{HA}_1^{(1)}(\mathbb{Z})$. Building automorphic forms for $\mathrm{GL}(n,\mathbb{Z})$ groups (and subgroups) is a well-established technique, while automorphic forms for $\mathrm{E}_{10}(\mathbb{Z})$ or $\mathrm{HA}_1^{(1)}(\mathbb{Z})$, at the moment, are out of reach.

Langlands program. Very recently, E. Witten et at have given a physical interpretation of the geometric Langlands program in terms of gauge theory. This construction uses a lot Hecke eigensheaves, whose classical counterpart is Hecke operators and Hecke eigenfunctions. It is remarkable, in my opinion, that these objects appear in general relativity in a different context and with a different language. It could be another subtle indication of the so-called gauge-gravity correspondence.

Regarding this last point, nowadays, a lot of research is focused on holography conjectures (AdS-CFT etc). The general belief is that a gauge theory, i.e. a theory without gravity, is "dual" to a theory containing gravity. Many efforts are also concentrated to prove the integrability of a Yang-Mills theory, although all the arguments given so far are not conclusive. Of course, in some approximation scheme, for example large $N-$expansion, one obtains the integrability. Gauge theories are non-linear, thus it is reasonable to think that Yang-Mills theory is not integrable, although one knows stable solutions (solitons etc) to the non-linear equations. In fact, chaoticity in gauge theory must not be surprising, see the book by S. G. Matinyan et al [23]. In [138], G. K. Savvidy shows that Yang-Mills classical mechanics can be reformulated in terms of dynamical systems theory, in particular it has positive metric entropy and it is isomorphic to a Kolmogorov system, thus it has strong statistical properties (see also [23], from the analysis of the problem there is a billiard representation too). It would be very interesting to reformulate Yang-Mills theory in any regime as a dynamical system and try to compare it with general relativity in some specific situation. We thus suggest to study the problem of the gauge/gravity duality as a problem of *isomorphism* of gravity and gauge theories reformulated according to the dynamical systems theory, in particular looking at the correlation functions of these two systems for which one can say something for isomorphic dynamical systems.

Finally, the role of chaos in gauge theory should be investigated more, in view of possible implications for the problem of confinement (see [23]).

I do hope to study all these points in depth in the near future.

Appendix A

The zoo of L-functions

In this appendix, we give a glimpse at the beautiful theory of zeta-functions, L-functions, automorphic L-functions and all that. There are plenty of books and survey articles on the subject, we suggest the book by H. Davenport [40].

The *Riemann zeta-function*

$$\zeta(s) = \sum_{n=1}^{+\infty} \frac{1}{n^s} = \prod_p (1 - p^{-s})^{-1} \tag{A.1}$$

is the mother of all zeta-functions. It was first studied by Dirichlet for real s, while Riemann considered the case of complex s. The identity is equivalent to the unique factorization of integers into primes and was proven by Euler.

Applying Poisson summation formula and using Euler Γ-function, one arrives at the functional equation for $\zeta(s)$

$$\zeta^* := \pi^{-s/2}\, \Gamma\left(\frac{s}{2}\right) \zeta(s) = \zeta^*(1-s) \tag{A.2}$$

$\zeta^*(s)$ is called the *completed zeta-function*, it has a meromorphic continuation to the entire $s-$plane and it is analytic except for simple poles $s = 0, 1$. We observe that ζ^* has no zeros outside the *critical strip* $0 \leq \text{Re } s \leq 1$. In fact, since $\Gamma(s)$ is never zero and $\zeta(s)$ is analytic and non-zero in the region of convergence Re $s > 1$, the completes zeta-function $\zeta^*(s)$ is $\neq 0$ in Re $s > 1$; by the functional equation, the same is true for Re $s < 0$. Moreover, since $\Gamma(s)$ is analytic except for simple poles at $s = 0, -1, -2, \ldots$, $\zeta(s)$ is non-zero in Re $s < 0$ except for simple zeros at the negative even integers $s = -2, -4, -6, \ldots$, to make up for the simple poles of $\Gamma(s/2)$ at these points. These are called the *trivial zeros* of $\zeta(s)$. The non-critical ones are the zeros of $\zeta^*(s)$ and they all lie in the critical strip. The first few zeros were computed by Riemann himself, and all lie on the critical line Re $s = \frac{1}{2}$. They are $\rho_n 1/2 + iE_n$ with $E_1 = 14.13\ldots$, $E_2 = 21.02\ldots$, $E_3 = 25.01\ldots$ etc (by symmetry we only need to consider positive E). The famous *Riemann hypothesis* (RH) is that all non-trivial zeros of $\zeta(s)$ lie on the critical line Re $s = 1/2$.

The RH has been checked extensively and is widely believed to be true, though an explanation and proof are still missing to date.

If one defines the function

$$\Lambda(n) = \begin{cases} \ln p & \text{if } n = p^k \text{ for some } k \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{A.3}$$

and let h and g functions as in the Selberg trace formula, then

$$\sum_E h(E) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} h(r) \frac{\Gamma'}{\Gamma} \left(\frac{1}{4} + \frac{1}{2} ir \right) dr + h\left(\frac{i}{2}\right) + h\left(-\frac{i}{2}\right) + \tag{A.4}$$

$$-g(0) \ln \pi - 2 \sum_{n=1}^{\infty} \frac{\Lambda(n)}{\sqrt{n}} g(\ln n)$$

which is sometimes called *Weil explicit formula.* The sum is over all non-trivial zeros. Note the similarity between this formula and the Selberg trace formula for a co-compact group (see Part I). This is clearer is one puts

$$\Lambda(P) = \frac{\ln N(P_0)}{1 - N(P)^{-1}} \tag{A.5}$$

This similarity led Selberg to the definition of the *Selberg zeta-function* $Z(s)$ as a product over all primitive periodic orbits as $\zeta(s)$ is a product over all primes.

Generalizations of the Riemann zeta-function are the *Dirichlet* $L-$*functions*

$$L(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} \tag{A.6}$$

which converge for Re $s > 1$. $\chi(n)$ is a Dirichlet character modulo q, that is it is a function on the integers satisfying

- χ is $q-$periodic: $\chi(n+q) = \chi(n)$ where q is an integer > 1
- $\chi(n) = 0$ if n is not co-prime to q
- $\chi(n)$ is multiplicative: $\chi(nm) = \chi(n)\chi(m)$
- $\chi(1) = 1$

In particular $\chi(-1) = \pm 1$ and we say that χ is even (odd) if $\chi(-1) = 1$ ($\chi(-1) = -1$). In general, there are precisely $\phi(q)$ Dirichlet character modulo q. They satisfy some orthogonality relations. $L(s, \chi)$ is analytic for Re $s > 0$. By using factorization for primes and multiplicativity of χ, one shows that there is an Eulerian product

$$L(s, \chi) = \prod_p \left(1 - \frac{\chi(p)}{p^s}\right)^{-1} \tag{A.7}$$

Similarly, there is a functional equation connecting $L(s, \chi)$ with $L(s, \chi^{-1})$. The $L-$function associated to a non-trivial character χ has an analytic continuation, with no poles, and all its non-trivial zeros are in the critical strip. The generalization of the RH is that all non-trivial zeros are on the critical line; in this case too a proof is lacking.

Riemann zeta-function and Dirichlet $L-$functions all belong to a wide class of number theoretic objects called *automorphic $L-$functions.* We do not explain what they are, but we give an example: the $L-$functions attached to the eigenfunctions of the Laplacian on the modular domain. Let $\Gamma =$ PSL$(2, \mathbb{Z})$. The spectrum of Δ on the space of odd functions on $\Gamma\backslash\mathbb{H}$ is purely discrete, but on the even space there is continuous spectrum (the infinite interval $\left[\frac{1}{4}, \infty\right)$) and discrete spectrum. As we said in Part I, the corresponding eigenfunctions are the Maass waveforms, which are $\Gamma-$periodic eigenfunctions of the Laplacian, square-integrable on $\mathcal{F}(1)$.

The space of such forms splits up into odd/even forms under the symmetry $W_1 : z \to -\overline{z}$

$$\psi(-\overline{z}) = \pm\psi(z) \tag{A.8}$$

The discrete spectrum is embedded in the continuous one, so that the eigenvalues E satisfy $E = \frac{1}{4} + t^2 > \frac{1}{4}$. Since the translation $T : z \to z + 1$ is in Γ, an even/odd Maass waveform $\psi(z)$ has a Fourier expansion $\psi(z) =$

$\sum_n W_n(y)e^{2\pi inx}$. Considering that $\Delta\psi+E\psi=0$ and the square-integrability condition, one can derive a more explicit form

$$\psi(z)=\sum_{n\neq 0} a_\psi(n)\, y^{1/2}\, K_{it}(2\pi|n|y)\, e^{2\pi inx} \tag{A.9}$$

where $K_{it}(y)$ are modified Bessel functions; as $y\to\infty$, $K_{it}(y) << e^{-2\pi y}$. ψ is an eigenfunction with eigenvalue λ and t is such that $\lambda=\frac{1}{4}+t^2$. The coefficients $a_\psi(n)$ are the Fourier coefficients for $\psi(z)$. For even forms, $a_\psi(-n)=a_\psi(n)$, while for odd ones $a_\psi(-n)=-a_\psi(n)$. More explicitly, for Hecke triangle groups which have an obvious symmetry with respect to the imaginary axis, one can write

$$\psi=\sum_{n=1}^{\infty} c_n\, y^{1/2}\, K_{it}(2\pi ny)\left\{\begin{array}{l}\cos(2\pi inx)\\ \sin(2\pi inx)\end{array}\right. \tag{A.10}$$

depending on whether ψ is even or odd. Since Γ is arithmetic, there are additional symmetries, the *Hecke operators*. These are defined for $n>0$ as

$$T_n\psi(z):=\frac{1}{\sqrt{n}}\sum_{ad=n,b \mod n}\psi\left(\frac{az+b}{d}\right) \tag{A.11}$$

the sum going over all positive integers a,d with $ad=n$ and $0\le b<d$. The Hecke operators $\{T_n\}$ are a commutative algebra of self-adjoint operators on $L^2(\Gamma\backslash\mathbb{H})$ and commute with Δ and with the reflection W_1. Thus they preserve the even/odd eigenspaces of Δ and each eigenspace has a basis consisting of simultaneous eigenfunctions of all Hecke operators (remember also that for Γ all numerical evidences support that the cuspidal spectrum is simple). Such eigenfunctions are called *Maass-Hecke eigenforms.* Given such an eigenfunction ψ, $T_n\psi=\lambda_n\psi$, its Fourier coefficients are given by

$$a_\psi(n)=a_\psi(1)\lambda(n) \tag{A.12}$$

thus we can normalize the first Fourier coefficient $a_\psi(1)=1$ and then the $n-$th Fourier coefficient is the Hecke eigenvalue $\lambda(n)$. The Hecke eigenvalues are multiplicative

$$\begin{aligned}\lambda(mn) &= \lambda(m)\lambda(n) \quad m,n \text{ co-prime} \\ \lambda(p^k)\lambda(p) &= \lambda(p^{k+1})+\lambda(p^{k-1}) \quad p \text{ prime}\end{aligned} \tag{A.13}$$

Since the ψ are bounded in the fundamental domain, one can infer that for any Maass form $a_\psi(n) << n^{1/2}$. For Hecke eigenforms, it is conjectured that the Fourier coefficients are essentially bounded, more precisely that for prime p, $|\lambda(p)| \leq 2$ and consequently $|\lambda(n)| << n^\epsilon$ for any $\epsilon > 0$. This is the Ramanujan conjecture for Maass forms and is still open. Furthermore, it is believed that the signs of $\lambda(p)/\sqrt{p}$ are distributed according to the Sato-Tate law, i.e. equidistributed with respect to the semi-circle distribution

$$d\mu(x) = \frac{2}{\pi} \sqrt{1 - x^2} dx \tag{A.14}$$

The $L-$function attached to a normalized Maass-Hecke eigenform ψ with Fourier coefficients $a(n) = \lambda(n)$, $a(1) = 1$ is defined by

$$L(s, \chi) = \sum_{n=1}^{\infty} \frac{\lambda(n)}{n^s} \tag{A.15}$$

and it is absolutely convergent for Re $s > 1$. Since the Hecke eigenvalues $\lambda(n)$ are multiplicative, we have an Euler product expansion

$$L(s, \chi) = \prod_p \frac{1}{1 - \lambda(p) p^{-s} + p^{-2s}}\,, \qquad \text{Re } s > 1 \tag{A.16}$$

This $L-$function allows for an analytic continuation and satisfies a functional equation of the kind

$$L^*(s, \psi) = L^*(1 - s, \psi) \tag{A.17}$$

The trivial zeros are at $s = \pm it + k$, with $k = 0, -2, -4, \ldots$. For the modular group all the Laplace eigenvalues lie above 1/4, so t is real and there are no trivial zeros in the critical strip. All the non-trivial zeros $\rho_n = 1/2 + iE_n$ lie inside the critical strip and the analogue of the RH is that they all have Re $s = \frac{1}{2}$.

In the following appendix, we describe the spectral statistics of the zeros of $L-$functions with the tools of random matrix theory and make connections with quantum mechanics.

Appendix B

Random Matrix Theory

Random matrix theory was introduced to the theoretical physics community as a subject of intensive study by E. Wigner [158] in his work on nuclear physics in the 1950s (for random matrix theory see the excellent book by M. L. Mehta [118]). Wigner was concerned with scattering resonances for neutrons off heavy nucleii. For any given nucleus there could be hundreds, even thousands, of such resonances, and many researchers realized that the only hope to bring some order to the subject was through a statistical approach. It was Wigner, however, who first proposed that the local statistical behavior of the resonance levels (i.e. the energy levels of complex Hamiltonians describing many nucleons) be modelled by the local statistical behavior of the eigenvalues of a large random matrix.

Let us consider a sequence of numbers $x_1 \leq x_2 \leq \cdots \leq x_n \leq \cdots$, normalized so that $x_n \sim n$ as $n \to \infty$. We want to understand the fluctuations of the levels x_n from their mean. For instance, the *nearest-neighbor level spacings* are $s_n := x_{n+1} - x_n$, whose mean is unity. The *level spacing distribution* $P(s)$ measures the distribution of the spacings s_n

$$P(s) = \lim_{N\to\infty} \frac{1}{N} \sum_{n \leq N} \delta(s - s_n) \tag{B.1}$$

that is we want that for any test function $f \in C_0(0, \infty)$

$$\frac{1}{N} \sum_{n\leq N} f(s_n) \longrightarrow \int_0^\infty f(s)P(s)ds \quad \text{as } N \to \infty \tag{B.2}$$

The first example of such a sequence of numbers x_n is given by the zeros of the $L-$functions previously described, $\rho_n = \frac{1}{2} + iE_n$; here we assume also the relevant RH, thus the E_n are all real. We want to understand the fluctuations of these zeros in an interval $[E, 2E]$ with $E >> 1$. The number of these levels E_n in this interval is asymptotically equal to $d\,E \ln E/2\pi$ as $E \to \infty$, with $d = 1$ in the case of $\zeta(s)$ (a result already known to Riemann) and Dirichlet $L-$functions, while $d = 2$ for the $L-$functions attached to the Maass waveforms (thus the density of zeros for these automorphic $L-$functions is twice the density for $\zeta(s)$ or Dirichlet $L-$functions). A standard procedure allows to have a sequence of normalized levels $x_n \sim n$

$$x_n := \frac{d \ln E}{2\pi} E_n \tag{B.3}$$

this is known as *unfolding* the spectrum.

One model for such a sequence is to take x_n as random, uncorrelated numbers. In this case the level spacing distribution if $P(s) = e^{-s}$, a Poisson distribution. Other models come from *Random Matrix Theory.* For instance, we can take the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$ of an $N \times N$ Hermitian matrix H chosen from the *Gaussian Unitary Ensmble* (GUE), which is the set of Hermitian matrices endowed with a probability measure $d\mu(H) = c_N e^{-\mathrm{Tr}\ H^2} dH$. The Gaussian profile for the measure explains the term Gaussian ensemble; besides, the measure is invariant under unitary transformations, which explains the term unitary. Then one can form the unfolded eigenvalues $x_n := \frac{\sqrt{2N}}{2\pi}\lambda_n$; in the limit $N \to \infty$ (i.e. for large random matrices), the expected level spacing distribution of x_n is given in terms of a Fredholm determinant

$$P_{GUE}(s) = \frac{d^2}{ds^2} \det\,(I - Q_s) \tag{B.4}$$

where Q_s is the integral operator in $L^2(-1, 1)$ with kernel

$$Q_s(x, y) = \frac{\sin \pi(x-y)\, s/2}{\pi(x-y)} \tag{B.5}$$

For small s, $P_{GUE}(s) \sim \frac{\pi^2}{3} s^2$.

The same level spacing distribution arises if we consider the eigenphases of an $N \times N$ unitary matrix, chosen at random with respect to the Haar measure on the unitary group $\mathrm{U}(N)$ (Dyson's CUE). Similarly, if we take any of the families compact classical groups such as the unitary symplectic group $\mathrm{USp}(2N)$. In these *compact* examples, Katz and Sarnak [93]-[94] proved that the ensemble averages converge to $P_{GUE}(s)$ for a class of test functions. This is the only rigorous results about expected level spacings distributions, together with another similar result by Rudnick and Sarnak [132] which say that the $n-$level correlation functions for the zeros of $L-$functions for any cuspidal automorphic form (like Maass waveforms) agree with the GUE predictions, at least for a restricted class of test functions. In fact, it is easier to study correlations between all $n-$tuples of levels, the *$n-$level correlation functions*, rather than studying spacings between adjacent levels. For example, the *pair correlation function* ($n = 2$) of an unfolded sequence x_n is defined as

$$R_2(f, N) = \frac{1}{N} \sum_{j \neq k \leq N} f(x_j - x_k) \tag{B.6}$$

where f is an even test function. The goal is to understand the limit $N \to \infty$

$$R_2(f, N) \longrightarrow \int_{-\infty}^{+\infty} f(x) R_2(x) dx \tag{B.7}$$

For uncorrelated levels, we clearly have $R_2(x) = 1$, while for the GUE case, F. Dyson found that $R_2^{GUE}(x) = 1 - (\sin \pi x/\pi x)^2$. At this point it is good to remind an event which took place at the Institute for Advanced Study in Princeton in the early 1970s. H. Montgomery had been working for a number of years on the problem of the zeros of the Riemann zeta function. Assuming the RH, Montgomery rescaled (unfolded) the imaginary parts $\gamma_1 \leq \gamma_2 \leq \ldots$ of the zeros $\{1/2 + iE\}$ of $\zeta(s)$

$$E_j \to \widetilde{E_j} = \frac{E_j \, \ln E_j}{2\pi} \tag{B.8}$$

to have a mean spacing of 1. He obtained, modulo certain technical assumptions, an expression for the limiting form for the distribution of pairs of zeros

$$R(a, b) = \lim_{N \to \infty} \frac{1}{N} \sharp \left\{ \text{pairs } (j_1, j_2) \ : 1 \leq j_1 \, j_2 \leq N \, , \, \widetilde{E_{j_1}} - \widetilde{E_{j_2}} \in (a, b) \right\} \tag{B.9}$$

for any interval (a,b). Montgomery gave a talk on his work, but it turned out that Dyson was unable to attend his lecture. However, at tea that afternoon Montgomery met Dyson and told him about his work. Before he could describe his formulae for $R(a, b)$, Dyson astounded him by asking whether he found

$$R(a, b) = \int_a^b \left(1 - \left(\frac{\sin 2\pi u}{2\pi u} \right)^2 \right) du \tag{B.10}$$

which was exactly Montgomery's hard-earned result! When Montgomery asked Dyson how he knew this, Dyson made the extraordinary remark that this is what one would obtain if the zeros of the Riemann zeta function behaved like the eigenvalues of a random matrix chosen from a particular Hermitian ensemble, the Gaussian Unitary Ensemble (GUE).

It took the mathematicians another fifteen years to pick up on this extraordinary result. In 1987, A. Odlyzko computed the spacing distribution for the $\widetilde{E_j}$'s using highly accurate computations on millions of zeros of $\zeta(s)$ and confirmed Montgomery's result with extraordinary accuracy. Moreover, he also considered other statistics for the spacing distribution. In particular, he computed

$$\frac{1}{N}\sharp\{s_j := \widetilde{E}_{j+1} - \widetilde{E}_j\,,\ 1 \leq j \leq N : s_j \in (a,b)\} \tag{B.11}$$

and found to extreme accuracy that as $N \to \infty$

$$\frac{1}{N}\sharp\{s_j := \widetilde{E}_{j+1} - \widetilde{E}_j\,,\ 1 \leq j \leq N : s_j \in (a,b)\} \to \int_a^b p(x)\,dx \tag{B.12}$$

where $p(x)dx$ is the distribution $P_{GUE}(s)$ of normalized spacings of eigenvalues of large random matrices from GUE. As we mentioned, up to some technical restrictions, these results have now been verified rigorously by N. Katz, Z. Rudnick and P. Sarnak. It is now commonly believed that the spectral statistics of the zeros of these $L-$functions follow the GUE.

Let us now describe the second example of such a sequence x_n, in relation to the energy levels of a quantum system. As we know, the spectrum of a system $E_1 \leq E_2 \leq \cdots \leq E_n \leq \cdots$ is characterized by the level density

$$d(E) = \sum_n \delta(E - E_n) \tag{B.13}$$

or the counting function

$$N(E) = \sum_n \theta(E - E_n) \tag{B.14}$$

It is useful to divide these two functions into two parts, a smooth part and an oscillatory part, the latter having mean value zero

$$\begin{aligned} d(E) &= \overline{d}(E) + d^{osc}(E) \\ N(E) &= \overline{N}(E) + N^{osc}(E) \end{aligned} \tag{B.15}$$

$\overline{d}(E)$ does not depend on whether the underlying classical system is chaotic or not; in fact Weyl's law says that this quantity depends only on the volume of the systems, at first order. Therefore, if one wants to compare spectra of different systems, one must compare spectra with the same mean level density; as we already said, this is achieved unfolding the spectrum. That is, one simply replaces the energy eigenvalues E_i by the sequence $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq \cdots$ with

$$\lambda_i = \overline{N}(E_i) \tag{B.16}$$

This new sequence has mean density one, but it has the same fluctuations as the energy eigenvalues. At this point, one can compare the remaining variations between different spectra, the *spectral fluctuations*, i.e. fluctuations around the mean level density. These can be compared to theoretical distributions. As before, the simplest one is the Poisson distribution, which simply corresponds to uncorrelated levels. Another possibility is to consider the random matrix theory distributions. According to a conjecture due to Berry and Tabor [22], the spectral fluctuations of classically *integrable systems* should follow the Poisson law (the harmonic oscillator is an exception), while Bohigas, Giannoni and Schmit [24] have conjectured that classically *chaotic systems* which possess the time-reversal symmetry should correspond to the fluctuations of the Gaussian orthogonal ensemble GOE (that is an ensemble of symmetric matrices where the measure is invariant under orthogonal transformations). Chaotic systems without time-reversal symmetry should instead follow the GUE predictions [1]. The random matrix theory GUE and GOE present *level repulsion* at short distance and *rigidity* at long distance, while the Poisson distribution exhibits the *clustering* property.

However, it is known today that the predictions of random matrix theory agree only for short- and medium-range correlations of the quantum spectra, but fail completely for long-range correlations. This was analyzed by M. Berry using the semiclassical trace formula. Berry's semiclassical argu-

[1] From this point of view, one is tempted to say that the zeros of the Riemann zeta-function, which experimentally follow the GUE predictions, can be interpreted (when rotated of 90 degrees to make them real) as the eigenvalues of a classically chaotic Hamiltonian system which does not have time-reversal invariance. This system should be the prototype of chaotic system , like the harmonic oscillator is the father of all integrable systems. The spectral interpretation of the zeros of ζ was suggested long ago by Hilbert and Polya independently, and nowadays random matrix theory makes it very reliable. There is a big evidence the these zeros are vibrations of a dynamical system, but we do not know what is that is vibrating [20].

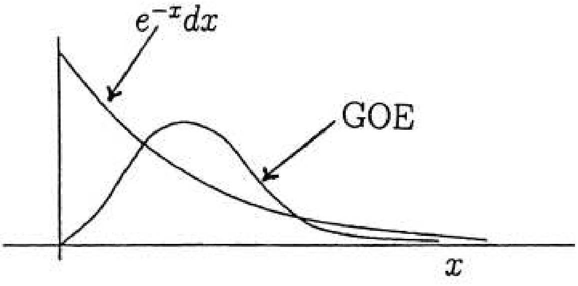


Figure B.1: The Poisson and the GOE distributions.

ments suggest that one of the commonly studied spectral statistics , the so-called Dyson-Metha spectral rigidity $\Delta_3(L)$, should saturate for large L in contrast to the logarithmic behavior predicted by random matrix theory. It thus appears that the properties of the spectral rigidity provide no universal signature of classical chaos in quantum mechanics.

Moreover, there are chaotic systems whose quantum spectra behave like a Poisson distribution (thus resembling more classical integrable systems) rather than following GUE/GOE predictions, that is they violate universality in energy level statistics even in the short-range regime. This is the case of *arithmetical chaos* [25]. As we said in Part I, arithmetical systems satisfy the Quantum Unique Ergodicity conjecture of Rudnick and Sarnak. Numerically, the spectrum of $X(1)$ follows a Poissonian law, which is *unexpected* because the geodesic flow is chaotic and one would expect a GOE-type behavior (because the system is time-reversal invariant). In this sense, arithmetical systems present anomalous statistics. And there is no doubt, again, that the reason for this anomaly is that the Laplacian Δ commutes with the Hecke operators; these make an arithmetical systems mimic an integrable systems at the quantum level, because there exists an infinite family of operators commuting with the Hamiltonian as it happens in classical integrable systems. This phenomenology about the spectrum of $X(1)$ is very fascinating, but at the moment very little can be proven.

On the subject of random matrix theory, zero of $L-$functions and quantum mechanics, there are a lot of very good video-lectures on the MSRI web site. On the top, I would suggest the lecture by F .J. Dyson,

F. J. Dyson, *Random matrices, neutron capture levels, quasicrystals and zeta-function zeros*,
`http://www.msri.org/publications/ln/msri/2002/rmt/dyson/1/index.html`

by E. Bogomolny

E. Bogomolny, *Spectral Statistics*,
`http://www.msri.org/publications/ln/msri/1999/random/bogomolny/1/index.html`

and some video-lectures by P. Sarnak

P. Sarnak, *Random matrix theory and zeroes of zeta functions - a survey*,
`http://www.msri.org/publications/ln/msri/1999/random/sarnak/3/index.html`
Random Matrix Models,
`http://www.msri.org/publications/ln/msri/1999/random/sarnak/1/index.html`,
0's of Zeta Functions and Random Matrices,
`http://www.msri.org/publications/ln/msri/1999/random/sarnak/2/index.html`

Appendix C

M. C. Escher and H. S. M. Coxeter

In this appendix, we explain why we have chosen the picture on the front page to represent the gist of our thesis.

It is a beautiful panting by M. C. Escher, drawn when he was in Holland (1952) [48]. The title of the work is *Gravity* and it represents an icosahedron. The why of this title is a mystery to me, but it clearly resembles the Coxeter billiard we have spoken in this thesis, although it is not the right one. I like thinking that Escher already had in his mind a picture of gravity in terms of polytopes. The influence of the mathematicians D. Hilbert and H.S.M. Coxeter on him was very deep as it is shown in his works. Coxeter dedicated a beautiful paper [36] to one of his paintings, *Circle Limit III*, explaining the underlying regular tessellation of that picture.

Bibliography

[1] L. Andersson, A. D. Rendall, *Quiescent cosmological singularity*, Comm. Math. Phys. 218 (2001) 479, `hep-th/0001047`

[2] L. Andersson, H. Van Elst, W. C. Lim, C. Uggla *Asymptotic silence of cosmological singularities*, Phys. Rev. Lett. **94** (2005) 051101 , `gr-qc/0402051`

[3] G. Andrews, *The Theory of Partitions*, The Encyclopedia of Mathematics and Its Applications Series, Addison-Wesley Pub. Co. (1976)

[4] V. I. Arnold, A. Avez, *Problèmes Ergodiques des Mécanique Classique*, Gaithier-Villars Editeur (1968)

[5] V. I. Arnold, *Mathematical Methods in Classical Mechanics*, Springer (1973)

[6] V. I. Arnold, *Forgotten and neglected theories of Poincaré*, Russ. Math. Survey **61** (2006) 1

[7] J. Arthur, *The trace formula and Hecke operators*, in Number Theory, Trace Formulas and Discrete Groups, Academic Press (1989) 11

[8] E. Artin, *Ein mechanisches System mit quasi-ergodischen Bahnen*, Abh. Math. Sem. d. Hamburgischen Universität **3** (1924) 170

[9] N. L. Balazs, A. Voros, *Chaos on the Psudoshpere*, Phys. Rep. **143** (1986) 109

[10] G. Bangone, G. Parisi, S. Ruffo et al, *Gli Ordini del Caos*, Manifestolibri (1991)

[11] J. D. Barrow, *Chaotic Behavior in General Relativity*, Phys. Rep. **85** (1982) 1

[12] J. D. Barrow, J. J. Levin, *Fractals and Scars on a compact octagon*, Class. Quant. Grav. **17** (2000) 61, `gr-qc/9909041`

[13] T. Bedford, M. Keane, C. Series (Eds) *Ergodic Theory, Symbolic Dynamics and Hyperbolic Space*, Oxford University Press (1991)

[14] V. A. Belinskii, I. M. Khalatnikov, *Effect of scalar and vector fields on the nature of the cosmological singularity*, Sov. Phys. JETP **36** (1973) 591

[15] V.A. Belinskii, I.M. Khalatnikov, E.M. Lifshitz, *Oscillatory approach to a singular point in the relativistic cosmology*, Adv. Phys. **19** (1970) 525

[16] V.A. Belinskii, I.M. Khalatnikov, E.M. Lifshitz, *A general solution of the Einstein equations with a time singularity*, Adv. Phys. **31** (1982) 639

[17] B. K. Berger, *Numerical Approaches to Spacetime Singularities* , Living Rev. Rel **4** (2002); see also Berger's talk at the KITP Miniprogram "The quantum nature of spacetime Singularities" (January 2007), http://online.kitp.ucsb.edu/online/singular_m07/

[18] M. Berger, *A Panoramic View of Riemannian Geometry*, Springer-Verlag (2002)

[19] Z. Bern, J. J. Carrasco, L. J. Dixon, H. Johansson, D. A. Kosower, R. Roiban, *Three-Loop Superfiniteness of $N = 8$ Supergravity*, `hep-th/0702112`

[20] M. Berry, *Riemann's zeta function: a model for quantum chaos?*, in Quantum Chaos and Statistical Nuclear Physics, Lecture Notes in Physics 263, Springer (1986) 1

[21] M. Berry, *Quantum Chaology, not Quantum Chaos*, Physica Scripta **40** (1989) 335

[22] M. Berry, M. Tabor, *Calculating the bound spectrum by path summation in action angle variables*, J. Phys. A **10** (1977) 371

[23] T. S. Biró, S. G. Matinyan, B. Müller, *Chaos and Gauge Field Theory*, World Scientific (1995)

[24] O. Bohigas, M. J. Giannoni, C. Schmit, *Characteristic of chaotic quantum spectra and Universality of level fluctuations laws*, Phys. Rev. Lett. **52** (1984) 1; *Spectral properties of the Laplacian and random matrix theory*, J. Physique Lett. **45** (1984) L-1015

[25] E. B. Bogomolny, B. Georgeot, M. J. Giannoni, C. Schmit, *Arithmetical Chaos*, Phys. Rep. **291** (1997) 219

[26] L. Boltzmann, *Über die mechanischen Analogien des zweiten Hauptsatzes der Thermodynamik*, Journal für die reine und angenwandte Mathematik (Crelles Journal) **100** (1887) 201

[27] R. Carter, *Lie Algebras of Finite and Affine Type*, Cambridge University Press (2005)

[28] N. Chernov, R. Markarian, *Chaotic Billiards*, American Mathematical Society (2006)

[29] D. F. Chernoff, J. D. Barrow, *Chaos in the Mixmaster universe*, Phys. Rev. Lett. **50** (1983) 134

[30] D. M. Chitre, *Investigations of the vanishing of a horizon for Bianchi IX (Mixmaster) Universe*, Doctoral Dissertation, University of Maryland (1972), unpublished

[31] A. Coley, D. Hobill (Eds), *Deterministic Chaos in General Relativity*, Plenum (1994)

[32] Y. Colin de Verdiére, *Ergodicité et fonctions propres du Laplacien*, Comm. Math. Phys. **102** (1985) 497

[33] I. P. Cornfeld, S. V. Fomin, Ya. G. Sinai, *Ergodic Theory*, Springer (1982)

[34] N. J. Cornish, J. J. Levin, *The Mixmaster universe is chaotic*, Phys. Rev. Lett. **78** (1997) 998, `gr-qc/9605029`

[35] N. J. Cornish, J. J. Levin, *The mixmaster universe: A chaotic Farey tale*, Phys. Rev. **D55** (1997) 7489, `gr-qc/9612066`

[36] H. S. M. Coxeter, *The Non-Euclidean Symmetry of Escher's Picture "Circle Limit III"*, Leonardo **12** (1979) 19

[37] T. Damour, M. Henneaux, H. Nicolai, *Cosmological Billiards*, Class. Quant. Grav. **20** (2003) R145, `hep-th/0212256`

[38] T. Damour, M. Henneaux, H. Nicolai, *E_{10} and a "small tension expansion" of M-Theory*, Phys. Rev. Lett. **89** (2002) 221601, `hep-th/0207267`

[39] T. Damour, M. Henneaux, A. D. Rendall, M. Weaver, *Kasner-like behavior for subcritical Einstein-matter systems*, `gr-qc/0202069`

[40] H. Davenport, *Multiplicative Number Theory*, Springer (1980)

[41] S. De Buyl, *Kac-Moody Algebras in M-theory*, `hep-th/0608161`

[42] R. Dijkgraaf, J. Maldacena, G. Moore, E. Verlinde, *A Black Hole Farey Tail*, `hep-th/0005003`

[43] R. Dijkgraaf, J. de Boer, M. C.N. Cheng, J. Manschot, E. Verlinde, *A Farey Tail for Attractor Black Holes*, JHEP **24** (2006) 611, `hep-th/0608059`

[44] J. P. Eckmann, D. Ruelle, *Ergodic Theory of Chaos and Strange Attractors*, Rev. Mod. Phys. **57** (1985) 617

[45] M. Einsiedler, T. Ward, *Ergodic Theory: with a view towards Number Theory*, http://www.mth.uea.ac.uk/ergodic/

[46] A. Einstein, *Zum Quantensatz von Sommerfeld und Epstein*, Ven. Deut. Phys. Ges. **19** (1917) 82

[47] F. Englert, L. Houart, A. Kleinschmidt, H. Nicolai, N. Tabti, *An E_9 multiplet of BPS states*, `hep-th/0703285`

[48] M. C. Escher, *Le magiche visioni di Escher*, Taschen (2003)

[49] G. F. R. Ellis, H. van Helst, C. Uggla, J. Wainwright, *Past attractor in inhomogeneous cosmology*, Phys. Rev. **D68**, (2003), 103502

[50] A. J. Feingold, I. Frenkel, *A Hyperbolic Kac-Moody Algebra and the theory of Sigel modular forms of genus 2*, Math. Ann. **263** (1983) 87

[51] A. J. Feingold, H. Nicolai, *Subalgebras of Hyperbolic Kac-Moody Algebras*, `math/0303179`

[52] J. Fischer, *An Approach to the Selberg Trace Formula via the Selberg Zeta-Function*, Springer (1987)

[53] L. A. Forte, A. Sciarrino, *Standard and nonstandard extensions of Lie algebras*, J. Math. Phys. **47** (2006) 013513

[54] L. Frappat, A. Sciarrino, *Hyperbolic Kac-Moody superalgebras*, `math-ph/0409041`

[55] G. Gallavotti, *Lectures on the Billiard*, in Dynamical Systems, Theory and Applications, Lect. Notes Phys. **38**, Springer (1975)

[56] G. Gallavotti, *Chaotic hypothesis: Onsager reciprocity and fluctuation-dissipation theorem*, Journal of Statistical Physics **84** (1996) 899

[57] T. Gannon, *Moonshine Beyond the Monster: The Bridge Connecting Algebra, Modular Forms and Physics*, Cambridge University Press (2006)

[58] P. Gaspard, D. Alonso, *$\hbar$ expansion for the periodic-orbit quantization for hyperbolic systems*, Phys. Rev. **A47** (1993) R3468

[59] M. J. Giannoni, A. Voros, P. Zinn-Justin (Eds) *Chaos and Quantum Physics*, Proc. Les Houches Summer School 1989, North Holland (1991)

[60] G. Gibbons, *Singularities*, Rapporteur Talk at 23rd International Solvay Conference in Physics "The Quantum Structure of Space and Time" (2005), http://www.solvayinstitutes.be/

[61] J. Gleick, *Chaos: Making a New Science*, Penguin Books (1988)

[62] S. Graffi, *Le Radici della Quantizzazione*, Quaderni di Fisica Teorica Università di Pavia (1993)

[63] R. Graham, P. Szépfalusy, *Quantum creation of a generic universe*, Phys. Rev. **D42** (1990) 2483

[64] M. B. Green, J. G.Russo, P. Vanhove, *Ultraviolet properties of Maximal Supergravity*, `hep-th/0611273`

[65] C. Grosche, *Path Integrals, Hyperbolic Spaces, and Selberg Trace Formulae*, World Scientific (1996)

[66] V. Gritsenko ,V. Nikulin, *On Classification of Lorentzian Kac-Moody Algebras*, math-QA/0201162

[67] M. Gutzwiller, *Chaos in Classical and Quantum Mechanics*, Springer (1990)

[68] F. Haake, *Quantum Signatures of Chaos*, Springer (2000)

[69] B. Hasselblatt, A. Katok, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press (1995)

[70] S. W. Hawking, G. F. R. Ellis, *The Large Scale Structure of Spacetime*, Cambridge University Press (1973)

[71] S. W. Hawking, R. Penrose, *The singularities of Gravitational Collapse and Cosmology*, Proc. R. Soc. London, Ser. A **314** (1970) 529

[72] S. W. Hawking, *Quantum Cosmology*, in Relativity, Groups and Topology II, B. S. DeWitt, R. Stora (Eds), North Holland (1984)

[73] J. M. Heinzle, C. Uggla, N. Rohr, *The Cosmological Billiard Attractor*, gr-qc/0702141

[74] D. Hejhal, *The Selberg Trace Formula and the Riemann Zeta Function*, Duke Math. J. **43** (1976) 441

[75] D. Hejhal, *The Selberg Trace Formula for PSL*$(2,\mathbb{R})$, Springer vol 1 (1979); vol 2 (1983)

[76] D. Hejhal, B. N. Rackner, *On the topography of Maass waveforms for PSL*$(2,\mathbb{Z})$, Experiment. Math. Volume **1** (1992) 275

[77] E. J. Heller, *Bound-State Eigenfunctions of Classically Chaotic Hamiltonian System: Scars of Periodic Orbits*, Phys. Rev. Lett. **53** (1984) 1516

[78] M. Henneaux, D. Persson, P. Spindel, *Spacelike Singularities and Hidden Symmetries of Gravity*, 0710.1818

[79] H. Huber, *Zur analytischen Theorie hyperboloschen Raumformen und Bewegungsgruppen*, Math. Ann. **138** (1959) 1

[80] J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag (1997)

[81] J. E. Humphreys, *Reflection Groups and Coxeter Groups*, Cambridge University Press (1992)

[82] G.Imponente, G. Montani, *On the Covariance of the Mixmaster Chaoticity*, Phys.Rev. **D63** (2001) 103501, `astro-ph/0102067`

[83] V. D. Ivashchuk, V. N. Melnikov, *Billiard representation for pseudo-Euclidean Toda-like systems of comological origin*, Regular and Chaotic Dynamics **1** (1996) 23

[84] V. D. Ivashchuk, V. N. Melnikov, *Billiard representation for multidimensional cosmology with multicomponent perfect fluid near the singularity*, Class. Quantum Grav. **12** (1995) 809

[85] H. Iwaniec, *Topics in Classical Automorphic Forms*, American Mathematical Society (1997)

[86] H. Iwaniec, *Spectral Methods of Automorphic Forms*, American Mathematical Society (2003)

[87] V. G. Kac, *Infinite dimensional Lie algebras* , third ed. [1], Cambridge University Press, Cambridge (1990)

[88] V. G. Kac, R. V. Moody, M. Wakimoto, *On E_{10}*, Differential Geometrical Methods in Theoretical Physics, Proceedings, NATO Advanced Research Workshop, 16th International Conference, Como, Amsterdam, Kluwer, 1988, Editors:K. Bleuler, M. Werner, pp. 109-128.

[89] V. G. Kac, D. H. Peterson, *Infinite-Dimensional Lie Algebras, Theta Functions and Modular Forms*, Advances in Mathematics 53 (1984) 125; *Affine Lie algebras and Hecke modular forms*, Bull. Amer. Math. Soc. **3** (1980) 1057

[1] The first two editions do not deal with Borcherds algebras.

[90] A. Katok, *Fifty years of Entropy in Dynamics: 1958-2007*, Journal of Modern Dynamics **1** (2007) 545

[91] S. Katok, *Fuchsian Groups*, University of Chicago Press (1992)

[92] S. Katok, *Coding of Closed Geodesics after Gauss and Morse*, Geometriae Dedicata 63 (1996) 123

[93] N. Katz, P. Sarnak, *Zeros of Zeta Functions and Symmetry*, Bull. Amer. Math. Soc. **36** (99) 1

[94] N. Katz, P. Sarnak, *Random Matrices, Frobenius Eigenvalue and Monodromy* American Mathematical Society (1999)

[95] I. M. Khalatnikov, E. M. Lifshitz, *Investigations in relativistic cosmology*, Adv. Phys. **12** (1963) 185

[96] I. M. Khalatnikov, K. M. Khanin, E. M. Lifshitz, L. N. Schur, Ya. G. Sinai, *On the stochasticity in relativistic cosmology* J. Stat. Phys. **38** (1985) 97

[97] I. M. Khalatnikov, A. Yu. Kamenshchik, *Chaos, Fractality and Topological Entropy in Cosmological Models with a Scalar Field*, Gravitation and Cosmology **6**, Supplement, (2000) 22

[98] A. Khinchin, *Continued Fractions*, The University of Chicago Press (1964)

[99] A. A. Kirillov, *Billiards in Cosmological Models*, Regular and Chaotic Dynamics **1** (1996) 13

[100] A. A. Kirillov, V. N. Melnikov, *Dynamics of inhomogeneities of the metric in the vicinity of a singularity in multidimensional cosmology*, Phys. Rev. **D52** (1995) 723

[101] P. D. Lax, R. S. Phillips, *Scattering Theory for Automorphic Functions*, Princeton University Press (1977); *Scattering theory for automorphic functions*, Bull. Amer. Math. Soc. **2** (1980) 261

[102] E. Lindenstrauss, *Invariant measures and arithmetic quantum unique ergodicity*, Ann. of Math. **163** (2006) 165

[103] W. Luo, P. Sarnak, *Number Variance for Arithmetic Hyperbolic Surfaces*, Comm. Math. Phys. **161** (1994) 419

[104] W. Luo, P. Sarnak, *Quantum ergodicity of eigenfunctions on* $PSL(2,\mathbb{Z})/\mathbb{H}^2$, Publ. Math. IHES **81** (1995) 207

[105] M. Marcolli, *Modular Curves, C^*-algebras and chaotic Cosmology*, `math-ph/0312035`

[106] G. A. Margulis, *Application of ergodic theory to the investigation of manifolds of negative curvature*, Funct. Anal. Appl. **4** (1969) 335

[107] V. Marotta, A. Sciarrino, *Vertex Opertator Realization and Representations of Hyperbolic Kac-Moody Algebra* $\widehat{A_1^1}$, J. Phys. **A**26 (1993) 1161

[108] V. Marotta, A. Sciarrino, *Realization of Borcherds Algebras*, Inter. J. Mod. Phys. **A10** (1995) 3921

[109] D. H. Mayer, *Relaxation properties of the mixmaster universe*, Phys. Lett. **A121** (1987) 390

[110] D.H. Mayer, *Continued fractions and related transformations* in [13]

[111] H. P. McKean, *Selberg's Trace Formula as applied to a compact Rieman surface*, Commun. Pure and Appl. Math. **15** (1972) 225

[112] H. B. Messaoud, G. Rousseau, *Classification des formes réelles des presque compact des algèbres de Kac-Moody affines*, Journal of Algebra **267** (2003) 443

[113] H. B. Messaoud, *Almost split real forms for hyperbolic Kac-Moody Lie algebras*, J. Phys. A: Math. Gen. **39** (2006) 13659

[114] J. Milnor, *Remarks on infinite-dimensional Lie groups*, Relativity, Groups and Topology II, B.S. De Witt and R. Stora (Eds), Elsevier (1984)

[115] C. W. Misner, *Mixmaster Universe*, Phys. Rev. Lett. **22** (1969) 1071

[116] C. W. Misner, *Minisuperspace*, in Magic without Magic: J. A. Wheeler, J. Klander (Ed), Freeman (1972)

[117] C. W. Misner, K. S. Thorne, J. A. Wheeler, *Gravitation*, W H Freeman and Co. (1973)

[118] M. L. Mehta, *Random Matrices*, Academic Press (1980)

[119] K. Nakamura, *Quantum versus Chaos - Questions emerging from Mesoscopic Cosmos*, Kluwer Academic Publishers (2002)

[120] S. Naito, *Embedding into Kac-Moody algebras and construction of folding subalgebras for generalized Kac-Moody algebras*, Japan. J. Math. (New Series) **18** (1992) 155 [2]

[121] H. Nicolai, T. Fischbacher, *Low Level Representations for E_{10} and E_{11}*, `hep-th/0301017`

[122] D. N. Page, *A fractal set of perpetually bouncing universes?*, Class. Quantum Grav. **1** (1984) 417

[123] B. Pioline, *Lectures on Black Holes, Topological Strings and Quantum Attractors*, Class. Quant. Grav. **23** (2006) S981, `hep-th/0607227`

[124] M. Pollicott, M. Yuri, *Dynamical Systems and Ergodic Theory*, Cambridge University Press (1998)

[125] A. Pressley, G. Segal, *Loop Groups*, Oxford Mathematical Monographs (1988)

[126] B. Randol, *The length spectrum of Riemann surface is always of unbounded multiplicity*, Proc. Amer. Math. Soc. **78** (1980) 455

[127] J. Ratcliffe, *Foundations of Hyperbolic Manifolds*, Springer (2006)

[128] U. Ray, *Automorphic Forms and Lie Superalgebras*, Kluwer Academic Publishers (2006)

[129] H. Ringstrom, *The Bianchi IX Attractor*, `gr-qc/0006035`

[130] R. Ruffini (Ed), *The Chaotic Universe*, World Scientific (2000)

[131] Z. Rudnick, P. Sarnak *The behaviour of eigenstates of arithmetic hyperbolic manifolds*, Comm. Math. Phys. **161** (1994) 195

[2] I thank prof. S. Naito for sending me some paper copies of his works on this subject.

[132] Z. Rudnick, P. Sarnak, *Zeros of Principal L-Functions and Random Matrix Theory*. Duke Math. J. **81** (1996) 269

[133] D. Ruelle, *Chance and Chaos*, Princeton University Press (1993)

[134] D. Ruelle, F. Takens, *On the Nature of Turbulence*, Commun. Math. Phys. **20** (1971) 167

[135] P. Sarnak, *Some Applications of Modular Forms*, Cambridge University Press (1990)

[136] P. Sarnak, *Arithmetic Quantum Chaos* Israel Math. Conf. Proc., Vol. 8, Ramat Gan (1995) 183 [3]

[137] P. Sarnak, *Spectra of Hyperbolic Surfaces*, Bull. Amer. Math. Soc. **40** (2003) 441

[138] G. K. Savvidy, *The Yang-Mills classical mechanics as a Kolmogorov K-system*, Phys. Lett. **130B** (1983) 303

[139] N. R. Scheithauer, *The Fake Monster Superalgebra*, Adv. Math. **151** (2000) 226

[140] A. Schnirelman, *Ergodic Properties of Eigenfunctions*, Usp. Math. Nauk. **29** (1974) 181

[141] A. Selberg, *Harmonic Analysis and Discontinuous Groups in Weakly Symmetric Riemannian Spaces with Applications to Dirichlet Series*, J. Indian Mat. Soc. **20** (1956) 47

[142] Ya. G. Sinai et al, *Dynamical Systems, Ergodic Theory and Applications*, Encycl. Math. Sciences **100**, Springer (2000)

[143] Ya. G. Sinai, *How Mathematicians and Physicists Found Each Other in the Theory of Dynamical Systems and in Statistical Mechanics*, in Ya. G. Sinai et al (Eds), *Mathematical Events of the Twentieth Century*, Springer (2005)

[144] S. Tabachnikov, *Geometry and Billiards*, American Mathematical Society (2005)

[3] I thank prof. P. Sarnak for sending me a copy of this paper.

[145] A. Terras, *Harmonic Analysis on Symmetric Spaces and Applications: Vol I*, Springer (1985)

[146] A. Terras, *Finite Quantum Chaos*, Amer. Math. Monthly **109** (2002) 121

[147] J. Tits, *Uniqueness and Presentation of Kac-Moody groups over fields*, J. Algebra **105** (1987) 542

[148] J. Tits, *Twin buildings and Groups of Kac-Moody type*, Groups, Combinatorics and Geometry, Cambridge University Press (1992) 249

[149] C. Uggla, *The nature of generic cosmological singularities*, Talk at Eleventh Marcel Grossmann Meeting on General Relativity, Freie Universität Berlin (2006), http://www.icra.it/MG/mg11/; `0706.0463`

[150] V. S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representation*, Springer-Verlag (1999)

[151] A. Vershik, *Graded Lie Algebras and Dynamical Systems*, `math/0203018`

[152] E. B. Vinberg (Ed), *Geometry II: Spaces of Constant Curvature*, Springer (1993)

[153] J. Wainright, G. F. R. Ellis (Eds), *Dynamical Systems in Cosmology*, Cambridge University Press (2005)

[154] M. Wakimoto, *Infinite-Dimensional Lie Algebras*, American Mathematical Society (2001)

[155] R. M. Wald, *General Relativity*, University of Chicago Press (1984)

[156] P. Walters, *An Introduction to Ergodic Theory*, Springer (1982)

[157] P. West, *E_{11} and M-Theory*, Class. Quant. Grav. **18** (2001) 4443, `hep-th/0104081`

[158] E. Wigner, *Random matrices in physics*, SIAM Review **9** (1967) 1

[159] M Yoshida, *Discrete reflection groups in a parabolic subgroup of $Sp(2,\mathbb{R})$ and symmetrizable hyperbolic generalized Cartan matrices of rank 3*, J. Math. Soc. Japan **36** (1984) 243

[160] S. Zelditch, *Uniform Distribution of eigenfunctions on Compact Hyperbolic Surfaces*, Duke Math. J. **55** (1987) 919

[161] S. Zelditch, *Mean Lindelof hypothesis and equidistribution of cusp forms and Eisenstein series*, J. Funct. Anal. **97** (1991) 1