A tale of two fractals

A. A. Kirillov

Department of Mathematics, The University of Pennsylvania, Philadelphia, PA 19104-6395 *E-mail address: kirillov@math.upenn.edu*

To Ben and Lisa

The author deeply thanks the Erwin Schroödinger International Institute for Mathematical Physics (ESI) where this work was started, the Max Planck Institute in Bonn (MPI), and the Institute des Hautes Etudes Scientifique (IHES) where it was accomplished.

I also thank the referees for helpful remarks.

Contents

Introduction

The proposed book is devoted to a phenomenon of fractal sets, or simply **fractals**. It is known more than a century and was observed in different branches of science. But only recently (approximately, last 30 years) it became a subject of mathematical study.

The pioneer of the theory of fractals was B. Mandelbrot. His book [Man82] appeared first at 1977 and the second enlarged edition was published at 1982. After that the serious articles, surveys, popular papers and books about fractals are counted by dozens (if not hundreds); since 1993 a special journal "Fractals" is published by World Scientific. So, what is a reason to write one more book?

First, it turns out that in spite of the vast literature, many people, including graduate students and even professional mathematicians, have only a vague idea about fractals.

Second, in many popular books the reader finds a lot of colorful pictures and amazing examples but no accurate definitions and rigorous results. On the contrary, the articles written by professionals are, as a rule, too difficult for beginners and often discuss very special questions without motivation.

Last and may be the most important reason is my belief that the endeavor of independent study of the Geometry, Analysis and Arithmetic on fractals is one of the best ways for a young mathematician to acquire an active and stable knowledge of basic mathematical tools.

This subject also seems to me an excellent opportunity to test his/her ability to creative work in mathematics. I mean here not only the solution of well-posed problems, but recognition a hidden pattern and formulating new fruitful problems.

My personal interest in fractals originates from the lecture course I gave in the University of Pennsylvania in 1995 according to the request of our undergraduate students. I repeated this course in 1999, 2003 and 2005. In 2004 I had an opportunity to expose the material in four lectures during the Summer School near Moscow organized for high school seniors and first year university students who were winners of the Russian Mathematical Olympiad. I was surprised by the activity of the audience and by the quickness of assimilating all necessary information.

In this book we deliberately restrict ourselves by only two examples of fractals: Sierpiński and Apollonian gaskets. We describe and rigorously formulate several problems coming from the study of these fractals. Most of them can be formulated and solved independently but only the whole collection gives an understanding of the world of fractals.

Some of these problems are more or less simple exercises, some are relatively new results and a few are unsolved problems of unknown difficulty.

The solution (and even formulating and understanding) of all problems requires some preliminary background. It contains, in particular, the following:

- Elements of Analysis: functions of one variable, differential and integral calculus, series.
- Elements of Linear Algebra: real and complex vector spaces, dimension, linear operators, quadratic forms, eigenvalues and eigenvectors. Coordinates and inner products.
- Elements of Geometry: lines, planes, circles, discs and spheres in $\mathbb{R}^3$. Basic trigonometric formulae. Elements of spherical and hyperbolic geometry.
- Elements of Arithmetic: primes, relatively prime numbers, gcd (greatest common divisor), rational numbers, algebraic numbers.
- Elements of Group Theory: subgroups, homogeneous spaces, cosets, matrix groups.

All this is normally contained in the first two or three years of mathematical curriculum. I consider the diversity of necessary tools and their interconnection as a great advantage of the whole problem and as a characteristic feature of modern mathematics.

Several words about the style of exposition. I tried to avoid two main dangers: to be dull explaining too much details in most elementary form and to be incomprehensible using very effective but sometimes too abstract modern technique. It is to the reader to judge how successful is this endeavor.

I also tried to communicate a non-formal knowledge of mathematical tools which distinguishes (almost all) professionals from most of beginners. Sometimes one phrase explains more than a long article[1]

So, from time to time, I use intentionally some "high-altitude" notions, explaining each time what they mean in simplest situations.

Additional information is included in the text in the form of short "Info's". The end of an Info is marked by the sign $\diamondsuit$.

I use also "Remarks" as another form of additional information. The end of a Remark is marked by the sign $\heartsuit$.

The end of a proof (or the absence of proof) is marked by the sign $\square$.

[1]In my personal experience it happened when I tried to understand induced representations, spectral sequences, intersection homology, etc...

Part 1

Sierpiński gasket

CHAPTER 1

Definition and general properties

1.1. First appearance and naive definition

I will not describe early manifestation of fractals in Natural Science (such as investigations of seashore length, cauliflower and snowflake forms etc); there are enough examples in popular expositions (see e.g. the pioneering book [?M] or the nice recent book [LGRE00]).

For mathematicians, the most simple and most known example of fractals is the famous **Cantor set**. The acquaintance with the Cantor set is a good test to see a difference between those who really understand Real Analysis and those who have formally passed a Calculus exam. We do not go now into details of this example here, but in Section 1.2 we come back to it and show that it is a part of the general theory of self-similar fractals.

Much more interesting examples of fractals exist on the plane $\mathbb{R}^2$. Here we shall consider in detail one special example.

Many people know so-called **Pascal triangle** whose entries are binomial coefficients $\binom{n}{k}$. It looks as follows

								1												
							1		1											
						1		2		1										
					1		3		3		1									
				1		4		6		4		1								
			1		5		10		10		5		1							
		1		6		15		20		15		6		1						
	1		7		21		35		35		21		7		1					
...		...		...		...		...		...		...		...		...				

It is very easy to continue this triangle since every entry is a sum of two entries above it.

Now, let us replace these numbers by their residues mod 2. In other words, we put 0 instead of every even number and 1 instead of every odd

number. We get the following picture

$$
\begin{array}{ccccccccccccccccc}
 & & & & & & & 1 & & & & & & & & & \\
 & & & & & & 1 & & 1 & & & & & & & & \\
 & & & & & 1 & & 0 & & 1 & & & & & & & \\
 & & & & 1 & & 1 & & 1 & & 1 & & & & & & \\
 & & & 1 & & 0 & & 0 & & 0 & & 1 & & & & & \\
 & & 1 & & 1 & & 0 & & 0 & & 1 & & 1 & & & & \\
 & 1 & & 0 & & 1 & & 0 & & 1 & & 0 & & 1 & & & \\
1 & & 1 & & 1 & & 1 & & 1 & & 1 & & 1 & & 1 & & \\
\ldots & & \ldots & & \ldots & & \ldots & & \ldots & & \ldots & & \ldots & & \ldots & & \ldots
\end{array}
$$

How one can describe this picture? Observe that this triangle of size 8 contains 3 identical triangles of size 4 (left, upper and right); each of these triangles contains 3 identical triangles of size 2 which consist of 3 one's. The remaining place is occupied by zeros.

Let us try to imagine what happens if we continue our triangle up to 2^N lines for some big number N. If we contract the triangle to the size of a book page and replace 1's by black dots and 0's by white dots, we get a picture like this:

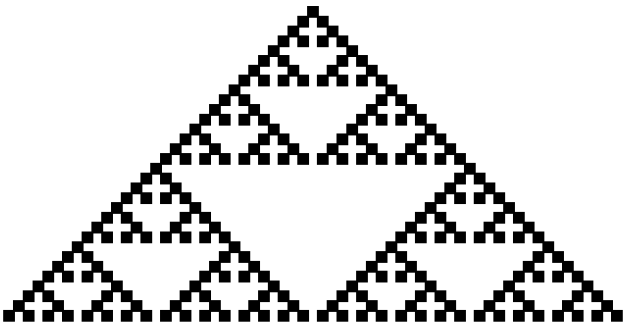

FIGURE 1.1. Pascal triangle mod 2

Here the whole triangle contains 3 triangles of half-size, which look similar to the whole thing. The space bounded by these triangles is filled by white dots.

It is rather clear that when N is going to infinity, our picture approaches a certain limit.[1] This limit is so-called **Sierpiński gasket** discovered in 1916 by Polish mathematician Waclaw Sierpiński.

Another appearance of the same set is related to the following problem of linear algebra. Let E_N be a $N \times N$ matrix with entries from the simplest finite field $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$ given by

$$(E_N)_{i,j} = \begin{cases} 1 & \text{if} \quad i < j \\ 0 & \text{otherwise.} \end{cases}$$

[1]See Info A below for the rigorous definition of a limit in this situation.

According to the general theory, this matrix is similar to a Jordan normal block J_N with $(J_N)_{i,j} = \begin{cases} 1 & \text{if} \quad j = i+1 \\ 0 & \text{otherwise.} \end{cases}$ Let us try to find the matrix A_N which establish the similarity: $E_N A_N = A_N J_N$. It turns out that A_N can be chosen so that it looks as follows:

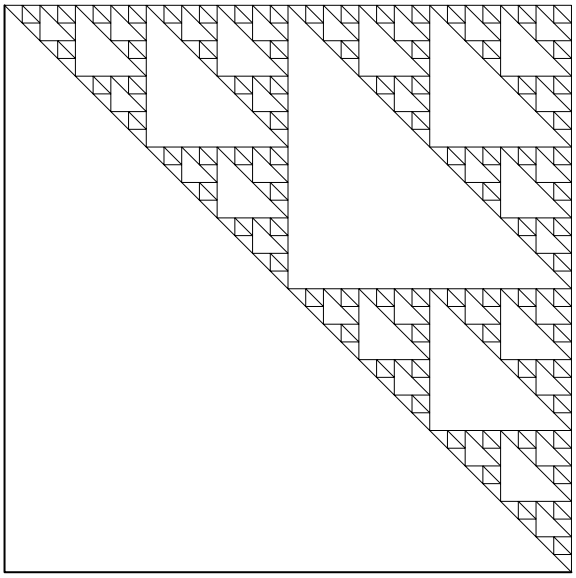

FIGURE 1.2. Pascal triangular matrix

We leave to a reader to explain this phenomenon and find the connection of A_N to Pascal triangle.

To go further we need to generalize the notion of a limit, the main notion in Analysis, so that it can be applied not only to numbers but to the objects of arbitrary nature. In particular, we want to give a meaning to the expression: "the sequence of sets $\{X_n\}$ converges to some limit set X".

The corresponding domain of mathematics is called the theory of metric spaces. Using this theory, we can define fractals (which are rather complicated sets) as limits of some sequences of more simple sets.

Info A. Metric spaces

We start with some general definitions which later will be specialized and explained on many examples. For some readers the text below will look too abstract and difficult for remembering and understanding. But you will see that the notions introduced here are very useful in many situations. They allow to treat uniformly the problems which seem completely different.

A.1.

DEFINITION A.1. A metric space is a pair (M, d) where M is a set and $d : M \times M \longrightarrow \mathbb{R}$ is a function which for any two points x and y defines the **distance** $d(x, y)$ between x and y so that the following axioms are satisfied:

1. Positivity: For all $x, y \in M$ the quantity $d(x, y)$ is a non-negative real number which vanishes iff[2] $x = y$.

2. Symmetry: $d(x, y) = d(y, x)$ for all $x, y \in M$.

3. Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in M$.

The original examples of metric spaces are: the real line $(\mathbb{R}, d)$ where the distance is defined by

$$d(x, y) = |x - y| \tag{A.1}$$

the plane $(\mathbb{R}^2, d)$ with the usual distance between $x = (x_1, x_2)$ and $y = (y_1, y_2)$:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}, \tag{A.2}$$

the 3-dimensional space $(\mathbb{R}^3, d)$ with the usual distance

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}. \tag{A.3}$$

DEFINITION A.2. We say that a sequence $\{x_n\}$ in M is **convergent**, or has a **limit**, if there exist $a \in M$ such that $d(x_n, a) \to 0$ when $n \to \infty$.

DEFINITION A.3. A sequence $\{x_n\}$ is called **fundamental**, or **Cauchy sequence**, if it has the property:

$$\lim_{m,n\to\infty} d(x_m, x_n) = 0. \tag{A.4}$$

For example, any convergent sequence is a Cauchy sequence. The converse is not always true. For instance, in the ray $\mathbb{R}_{>0}$ of all positive numbers with usual distance (A.1.1) the sequence $x_n = \frac{1}{n}$ is fundamental but has no limit.

DEFINITION A.4. A metric space (M, d) is called **complete** if every fundamental sequence in M has a limit.

In our book we shall consider mostly complete metric spaces. In particular, the examples (A.1.1-3) above are complete metric spaces according to well-known theorem of Real Analysis.

DEFINITION A.5. A subspace X of a metric space (M, d) is called **closed** in M if it contains all its limit points, i.e. the limits of sequences $\{x_n\} \subset X$.

[2]A standard mathematical abbreviation for the expression "if and only if".

EXERCISE 1. Let $(M,\, d)$ be a complete metric space and X be a subset of M. Than $(X,\, d)$ is itself a metric space.

Show that $(X,\, d)$ is complete if and only if the set X is closed in M.

HINT. This is simply a test on knowing and understanding the definitions. Formulate accurately what is done and what we have to prove and you will obtain a proof.

Warning. If this exercise does not seem easy for you, try again or discuss it with your instructor.

A.2.

DEFINITION A.6. A map f from a metric space $(M,\, d)$ to itself is called **contracting** if there is a real number $\lambda \in (0,\, 1)$ such that

$$d\big(f(x),\, f(y)\big) \le \lambda \cdot d(x,\, y) \qquad \text{for all} \quad x,\, y \in M. \tag{A.5}$$

We shall use the following

THEOREM (Theorem on contracting maps). *Assume that M is a complete metric space and f is a contracting map from M to itself. Then there exists a unique fixed point for f in M, i.e. the point x satisfying $f(x) = x$.*

The proof of this theorem is rather short and very instructive. Moreover, it gives a simple method to construct the fixed point. So, we give this proof here

PROOF. Let x_0 be an arbitrary point of M. Consider the sequence $\{x_n\}_{n\ge 0}$ defined inductively by $x_n = f(x_{n-1})$ for $n \ge 1$.

We claim that this sequence is convergent. For this end we show that $\{x_n\}$ is a Cauchy sequence. Indeed, let $d(x_0,\, x_1) = d$. Then, from A.5 we get

$$d(x_1,\, x_2) \le \lambda \cdot d, \quad d(x_2,\, x_3) \le \lambda^2 \cdot d, \quad \dots \quad d(x_n,\, x_{n+1}) \le \lambda^n \cdot d.$$

Therefore, for any $m < n$ we have $d(x_m,\, x_n) \le \sum_{k=m}^{n-1} \lambda^k \cdot d \le \frac{\lambda^m}{1-\lambda} \cdot d$. Hence

$$\lim_{m,n\to\infty} d(x_m,\, x_n) \to 0$$

and we are done.

Since M is complete, our Cauchy sequence has a limit which we denote x_∞.

Now, the function f, being contracting, is continuous. Therefore, $f(x_\infty) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x_\infty$, i.e. x_∞ is a fixed point.

Finally, if we had two fixed points x and y, then $d(x,\, y) = d\big(f(x),\, f(y)\big) \le \lambda \cdot d(x,\, y)$. It is possible only if $d(x,\, y) = 0$, hence $x = y$. □

This theorem, in particular, solves the following toy problem, given on some mathematical Olympiad for middle school students.

PROBLEM 1. A boy came out of his house and went to school. At a half-way he changed his mind and turned to a playground. But, passing half a way, he turned to a cinema. On the half-way to a cinema he decided again to go to school etc.

Where will he come continuing moving this way?

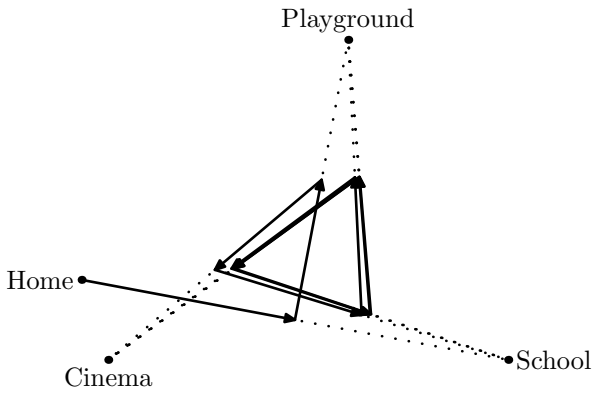

FIGURE A.3. Lazy boy

A.3.

DEFINITION A.7. A metric space (M, d) is called **compact** if every sequence $\{x_n\}$ of points in M has a convergent subsequence.

DEFINITION A.8. A subset $S \in M$ is called a **ϵ-net** in M if for any $m \in M$ there is a point $s \in S$ such that $d(m, s) < \epsilon$.

THEOREM (Theorem on ϵ-net). *A metric space (M, d) is compact iff it is complete and for any $\epsilon > 0$ there is a finite ϵ-net in M.*

EXERCISE 2. Show that a subset X in $\mathbb{R}$, $\mathbb{R}^2$ or $\mathbb{R}^3$ is compact iff it is closed and bounded.

HINT. If a subset X is not closed or unbounded, then you can construct a sequence of points in X without converging subsequences.

If X is bounded, then it is contained in a segment, or in a square, or in a cube of size R for R big enough. Using the theorem on ϵ-net, show that a segment, a square and a cube are compact. Then show that a closed subset of a compact set is itself a compact set.

$\diamondsuit$

1.2. Definition of self-similar fractals

Now we introduce the main technical tool to deal with a wide class of fractals.

Let M be a metric space. We denote by $\mathbb{K}(M)$ the collection of all non-empty compact subsets of M. We want to define a distance between two compact sets so that $\mathbb{K}(M)$ were itself a metric space. For this we define first the distance $d(x, Y)$ between a point x and a compact set Y:[3]

$$d(x, Y) := \min_{y\in Y} d(x, y). \tag{1.2.1}$$

Now, the distance between two sets X and Y is defined by

$$d(X, Y) := \max_{x\in X} d(x, Y) + \max_{y\in Y} d(y, X). \tag{1.2.2}$$

More detailed expression for the same distance is

$$d(X, Y) := \max_{x\in X} \min_{y\in Y} d(x, y) + \max_{y\in Y} \min_{x\in X} d(x, y) \tag{1.2.3}$$

This definition looks rather cumbersome but if you think a bit, how to define the distance between two sets, so that axioms 1 – 3 were satisfied, you find that (1.2.2) or (1.2.3) is a simplest possible definition.

On figure 1.4 the first and second terms in 1.2.3 are the lengths of segments AB and CD respectively.

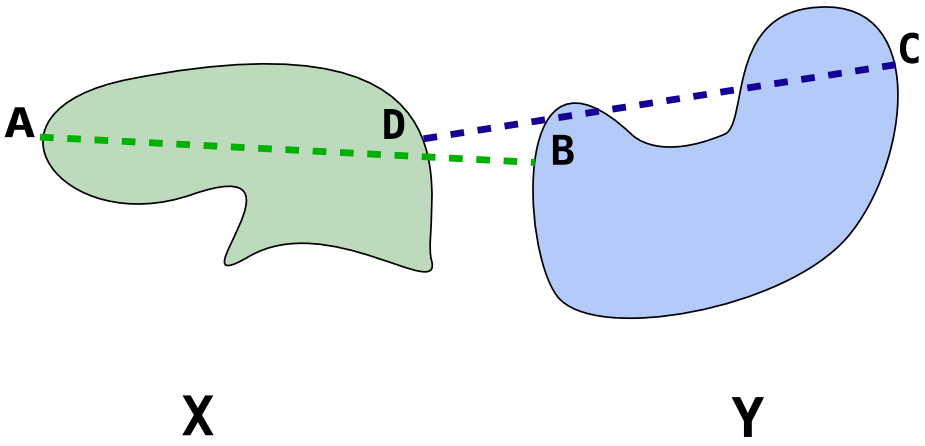

FIGURE 1.4. Hausdorff distance

EXERCISE 3. Prove that the minimum in (1.2.1) and maximum in (1.2.2) always exist.

[3]The sign ":=" used below denotes that the right hand side of the equation is a definition of the left hand side.

HINT. Use the compactness of sets X and Y.

EXERCISE 4. Compute the distance a) between the boundary of a square with side 1 and its diagonal; b) between a unit circle and the disc bounded by this circle.

ANSWER. a) $\frac{1+\sqrt{2}}{2}$ b) 1.

THEOREM 1.1. *If the metric space M is complete (resp. compact), then the space $\mathbb{K}(M)$ is complete (resp. compact) as well.*

Hint. Let $\{X_n\}$ be a sequence of compact subsets in M which forms a Cauchy sequence of points in $\mathbb{K}(M)$. Consider the set X of those points $x \in M$ for which there exists a sequence $\{x_n\}$ such that $x_n \in X_n$ and $\lim\limits_{n\to\infty} x_n = x$. Show that X is the limit of $\{X_n\}$ in $\mathbb{K}(M)$. (And, in particular, show that X is compact and non-empty.)

For the second statement use the theorem on ϵ-net.

Assume now that a family of contracting maps $\{f_1, f_2, \ldots, f_k\}$ in M is given. Define the transformation $F : \mathbb{K}(M) \longrightarrow \mathbb{K}(M)$ by

$$F(X) = f_1(X) \cup f_2(X) \cup \cdots \cup f_k(X) \tag{1.2.4}$$

THEOREM 1.2. *The map F is contracting. Therefore, there is a unique non-empty compact subset $X \subset M$ satisfying $F(X) = X$.*

DEFINITION 1.9. The set X from theorem 2 is called a **homogeneous self-similar fractal set**. The system of functions $f_1, \ldots, f_k$ is usually called an **iterated function system** (i.f.s. for short), defining the fractal set X.

Sometimes, a more general definition is used. Namely, instead of (1.2.4) let us define the map F by the formula

$$F(X) = f_1(X) \bigcup f_2(X) \bigcup \cdots \bigcup f_k(X) \bigcup Y \tag{1.2.5}$$

where Y is a fixed compact subset of M. This generalized map F is also contracting because of the following fact.

EXERCISE 5. Show that the "constant" map f_Y which sends any $X \in \mathbb{K}(M)$ to $Y \in \mathbb{K}(M)$ is contracting.

Hence, the sequence $\{X_n := F^n(X) := F(F(\cdots F(X_0)\cdots)\}$ is convergent and its limit X is a fixed point for F in $\mathbb{K}(M)$.

DEFINITION 1.10. The set X which is a fixed point for a map 1.2.5 is called a **non-homogeneous self-similar fractal**.

Examples.

(1) **Cantor set** $C \subset [0,\, 1]$. Here $M = [0,\, 1]$, $\quad f_1(x) = \frac{1}{3}x$, $\quad f_2(x) = \frac{x+2}{3}$. It is instructive to look how C, the fixed point for F, is approximated by a sequence of sets $\{C_n\}$ defined by the recurrence $C_{n+1} = F(C_n)$.

Choose first $C_1 = [0,\, 1]$; then

$$C_2 = [0,\, 1/3] \cup [2/3,\, 1], \qquad C_3 = [0,\, 1/9] \cup [2/9,\, 1/3] \cup [2/3,\, 7/9] \cup [8/9,\, 1] \ldots$$

The sequence $\{C_n\}$ is decreasing: $C_{n+1} \subset C_n$ and the limit set is $C = \bigcap_{n \geq 1} C_n$.

Now put $C'_1 = \{0,\, 1\}$. Then

$$C'_2 = \{0,\, 1/3,\, 2/3,\, 1\}, \qquad C'_3 = \{0,\, 1/9,\, 2/9,\, 1/3,\, 2/3,\, 7/9,\, 8/9,\, 1\}, \ldots$$

The sequence $\{C'_n\}$ is increasing: $C'_{n+1} \supset C'_n$ and the limit set C is the closure of $C'_\infty := \bigcup_{n \geq 1} C'_n$. Note, that C'_∞ is not compact, therefore it is not a point of $\mathbb{K}(M)$.

The main feature of self-similar fractals is easily seen on this example: if we consider a piece of Cantor set under a microscope which increase all the sizes in 3^n times, we shall see exactly the same picture as by a naked eye.

(2) **I_α-fractal.** Let Y be the subset of $\mathbb{R}^2$ given by $x = 0,\ -1 \leq y \leq 1$. Fix a real number $\alpha \in (0,\, \frac{1}{\sqrt{2}})$ and define the maps

$$f_1(x,\, y) = (-\alpha y,\, \alpha x + 1); \qquad f_2(x,\, y) = (-\alpha y,\, \alpha x - 1). \tag{1.2.6}$$

The corresponding non-homogeneous self-similar fractal is shown on the Figure 1.5 where for typographic convenience the y-axis is horizontal.

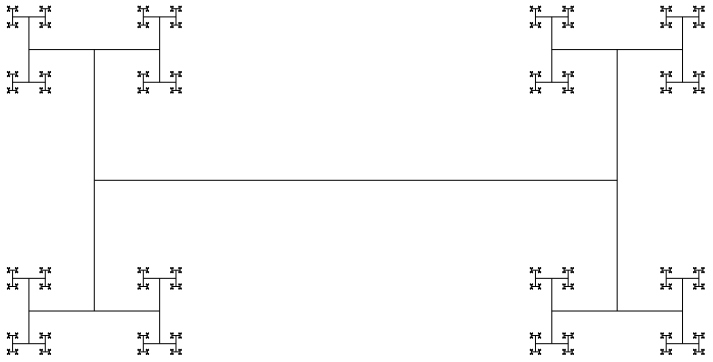

FIGURE 1.5. I_α-fractal for $\alpha = 0.5$

The first approximation $Y \bigcup f_1(Y) \bigcup f_2(Y)$ for small α looks like the capital letter I. It explains the name.

EXERCISE 6. Compute
a) The diameter D of I_α (as a subset of $\mathbb{R}^2$).

b) The length L of a maximal non-self-intersecting path on I_α.

ANSWER. a) $D = 2\frac{\sqrt{1+\alpha^2}}{1-\alpha^2}$; b) $L = \frac{2}{1-\alpha}$.

(3) **Sierpiński gasket** $\mathcal{S}$. Here $M = \mathbb{C}$, the complex plane.
Let $\omega = e^{\frac{\pi i}{3}}$ be a sixth root of 1. Define

$$f_1(z) = \frac{z}{2}, \quad f_2(z) = \frac{z+\omega}{2}, \quad f_3(z) = \frac{z+1}{2}.$$

DEFINITION 1.11. The fractal defined by the i.f.s. $\{f_1, f_2, f_3\}$ is called a **Sierpiński gasket**.

In this case there are three most natural choices for the initial set S_0.

First, take as S_0'' the solid triangle with vertices 0, ω, 1. Then the sequence $S_n'' = F^n(S_0)$ is decreasing and $\mathcal{S} = \lim_{n\to\infty} S_n'' = \bigcup S_n''$, see Figure 1.6.

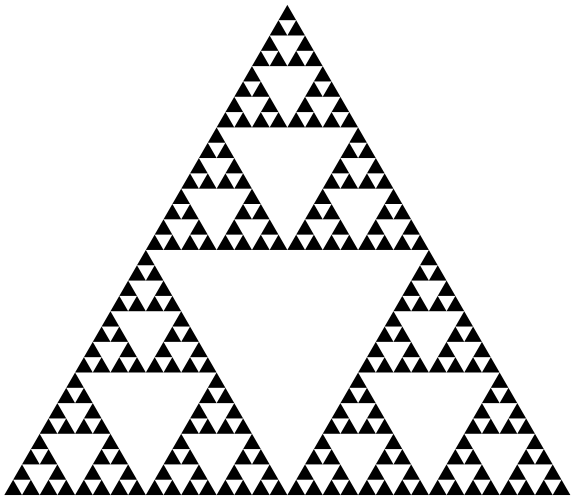

FIGURE 1.6. Approximation S_n''

Second, we put S_0' to be the hollow triangle with vertices in cubic roots of 1. Then the sequence $S_n' = F^n(S_0')$ is increasing and $\mathcal{S}$ is the closure of $S_\infty' = \bigcap_{n\geq 0} S_n'$.

EXERCISE 7. How many vertices, edges and hollow triangles are in S_n'?

Finally, let S_0 be the set of cubic roots of 1. Then $S_n = F^n(S_0)$ is a finite set. Here again $S_n \subset S_{n+1}$ and $\mathcal{S}$ is the closure of $S_\infty = \bigcap_{n\geq 0} S_n$.

We shall call the approximations $\{S''_n\}$, $\{S'_n\}$ and $\{S_n\}$ the 2-dimensional, the 1-dimensional and the 0-dimensional respectively. The first is an approximation from above and the other two are approximations from below.

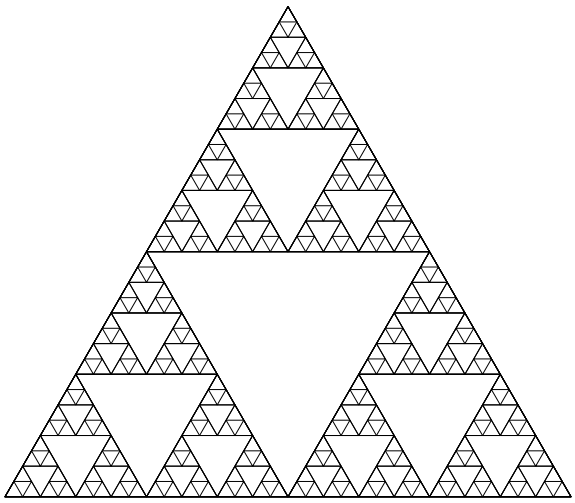

FIGURE 1.7. Approximation S'_n

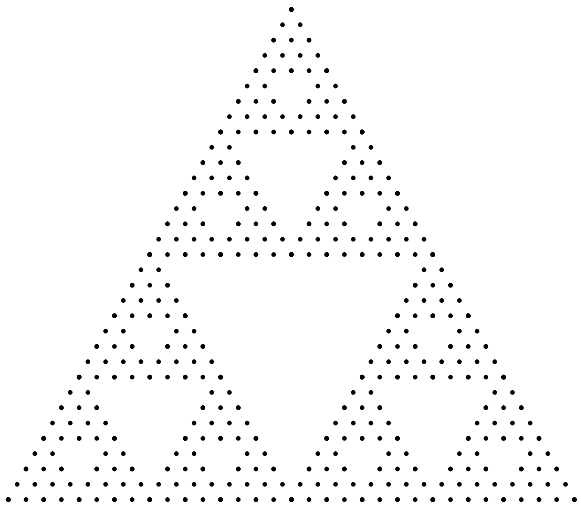

FIGURE 1.8. Approximation S_n

Sometimes it is more convenient to arrange Sierpiński gasket so that one side of it is horizontal. E.g., we can choose as initial 3 vertices the numbers 0, 1, $\frac{i\sqrt{3}}{2}$. Then the standard segment [0, 1] will be a subset of $\mathcal{S}$. Later on we mainly use this variant of Sierpiński gasket.

Info B. Hausdorff measure and Hausdorff dimension

We estimate the size of a curve by its length, the size of a surface by its area, the size of a solid body by its volume, etc. But how to measure the size of a fractal set?

A solution to this problem was proposed by F. Hausdorff in 1915. He defined for any real number $p > 0$ a measure μ_p of dimension p as follows.

Let X be a compact subset of $\mathbb{R}^n$. Then for any $\epsilon > 0$ it admits a finite covering by balls of radius ϵ. (The centers of these balls form a ϵ-net for X). Let $N(\epsilon)$ denote the minimal number of balls which cover X.

It is evident that $N(\epsilon)$ grows when ϵ decreases. Assume that it grows as some power of ϵ, namely, that the limit

$$\mu_p(X) := \lim_{\epsilon \to 0} \; N(\epsilon) \cdot \epsilon^p \tag{B.1}$$

exists. Then this limit is called the **Hausdorff p-measure** of X. We do not discuss here the general notion of a measure. For our goals it is enough the following

Proposition B.1. *The Hausdorff p-measure has the following properties:*

1. *Monotonicity: if $X \subset Y$, then $\mu_p(X) \le \mu_p(Y)$.*
2. *Subadditivity: if $X \subset \bigcup_{k=1}^{\infty} Y_k$, then*

$$\mu_p(X) \le \sum_{k=1}^{\infty} \mu_p(Y_i). \tag{B.2}$$

3. Additivity: if X_i, $1 \le i \le n$, are compact and $\mu_p\left(X_i \bigcap X_j\right) = 0$ for $i \ne j$, then

$$\mu_p\Big(\bigcup_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mu_p(X_i). \tag{B.3}$$

Actually, the first property formally follows from the second one, but we formulated it separately, because of its transparency and usefulness.

If the p-measure of X is different from 0 and ∞, then the number p is called the **Hausdorff dimension** of X.

Exercise 8. Show that if X has Hausdorff dimension d, then the limit (B.1) is equal to ∞ for $p < d$ and equal to 0 for $p > d$.

REMARK 1. There are several variants of this definition. Namely, instead of balls of radius ϵ one can use arbitrary sets of diameter ϵ, or, when $M = \mathbb{R}^n$, the cubes with a side ϵ.

Another variant: consider the covering of X by subsets X_k of different diameters $\epsilon_k \le \epsilon$ and instead of $N(\epsilon)$ investigate the quantity $\sum_k \epsilon_k^p$.

All these variants can lead to a different value of p-measure, but for "nice" examples, including self-similar fractals, define the same notion of dimension.

♡

In many cases it is not easy to prove that the limit (B.1) exists for a given set X, and still more difficult to compute it.

But often a weaker condition is satisfied and can be more easily checked:

$$N(\epsilon) \cdot \epsilon^p = O^*(1), \qquad \text{i.e.} \quad 0 < c \le N(\epsilon) \cdot \epsilon^p \le C < \infty \quad \text{for } \epsilon \text{ small enough} \tag{B.4}$$

In this case we also say that X has the **Hausdorff dimension** p. The constants c and C give the lower and upper estimates for the Hausdorff p-measure of X when this measure is defined.

EXERCISE 9. Show that the Hausdorff dimension of X, when it exists, can be given by the formula

$$d_H(X) = -\lim_{\epsilon \searrow 0} \frac{\log\, N(\epsilon)}{\log\, \epsilon}. \tag{B.5}$$

Examples. Let us find the Hausdorff dimensions of self-similar fractals defined above. In all cases we assume that not only the Hausdorff dimension but also the Hausdorff measure exists. It is not evident, but a persisting reader can try to prove it by him/herself.

Then we use a following simple arguments to compute it.

1. Cantor set C. Suppose, for some real number d the set C has finite non-zero Hausdorff measure $\mu_d(C)$. Now, C consists of two pieces which are similar to C with the coefficient $\frac{1}{3}$.

It is evident from the definition of d-measure that each of these two pieces of C have the measure $\left(\frac{1}{3}\right)^d \cdot \mu_d(C)$. Therefore, we get the equation $2 \cdot \left(\frac{1}{3}\right)^d = 1$ which implies $3^d = 2$, or

$$d = \log_3 2 = \frac{\log 2}{\log 3} \approx 0.63093....$$

2. I-fractal I_α. To compute the Hausdorff dimension of I_α we use the same scheme. Assume that $0 < \mu_d(I_\alpha) < \infty$ and recall the decomposition

$$I_\alpha = f_1(I_\alpha) \bigcup f_2(I_\alpha) \bigcup Y.$$

Since both $f_1(I_\alpha)$ and $f_2(I_\alpha)$ are similar to I_α with the coefficient α, we come to the equation $\mu_d(I_\alpha) = 2\alpha^d \mu_d(I_\alpha) + \mu_d(Y)$.

Note, that $1 \leq d \leq 2$, because I_α contains the segment Y of Hausdorff dimension 1 and is contained in a square of Hausdorff dimension 2.

Suppose $d > 1$. Then we have $\mu_d(Y) = 0$ according to Exercise 8; therefore $2 \cdot \alpha^d = 1$ and

$$d = \log_\alpha \frac{1}{2} = -\frac{\log 2}{\log \alpha} \tag{B.6}$$

The right hand side of (B.3) satisfies the inequality $1 \leq d \leq 2$ for $\alpha \in [\frac{1}{2}, \frac{1}{\sqrt{2}}]$.

EXERCISE 10. Prove that (B.6) gives the correct value for the Hausdorff dimension of I_α when $\alpha \in (\frac{1}{2}, \frac{1}{\sqrt{2}})$.

We leave to the reader to investigate the cases $\alpha = \frac{1}{2}$, $\alpha = \frac{1}{\sqrt{2}}$ and $\alpha \notin [\frac{1}{2}, \frac{1}{\sqrt{2}}]$.

$\diamondsuit$

CHAPTER 2

Laplace operator on Sierpiński gasket

A powerful mathematical method to study a certain set X is to consider different spaces of functions on X. For example, if X is a topological space, one can consider the space $C(X)$ of continuous functions; if X is a smooth manifold, the space C^∞ of smooth functions is of interest; for an homogeneous manifolds with a given group action, the invariant (and, more generally, covariant[1]) functions are considered and so on...

If M is a smooth manifold with additional structure(s), there are some naturally defined differential operators on M. The eigenfunctions of these operators are intensively studied and used in applications.

In the last century the vast domain of modern mathematics had arisen: the so-called **spectral geometry**. The main subject of it is to study spectra of naturally (i.e. geometrically) defined linear operators.

During the last two decades the spectral geometry included the analysis on fractal sets. We refer to the nice surveys [Str99, TAV00] and the original papers [Str00, MT95, Ram84, ?NS] for more details.

In this book we only briefly describe this theory and mainly restrict ourselves to the consideration of harmonic functions, i.e. eigenfunctions corresponding to the zero eigenvalue of the Laplacian.

Info C. Laplace operator and harmonic functions

C.1. Here we assume the acquaintance with elements of differential geometry on Riemannian manifolds. This section is not necessary for understanding the main text but gives the motivation for our study of Laplace operator and harmonic functions on fractal sets.

One of most famous differential operators on $\mathbb{R}^n$ is the **Laplace operator** Δ defined by

$$\Delta f = \sum_{k=1}^{n} \left(\frac{\partial}{\partial x^k} \right)^2 f.$$

The characteristic property of this operator is its invariance under the group E_n of rigid motions of $\mathbb{R}^n$. It is known that any differential operator on $\mathbb{R}^n$ which is invariant under E_n is a polynomial in Δ.

[1] I.e., functions which are transforming in a prescribed way under the action of the group. Details are explained in textbooks on Representation theory.

Actually, an analogue of this operator is defined for any Riemannian manifold M. Let $g = g_{i,j}(x)$ be the metric tensor on M defining the length of a tangent vector $v = \{v^k\}$ at a point x_0 by the formula

$$|v|^2 = \sum_{i,j} g_{i,j}(x_0) v^i v^j.$$

Traditionally, the inverse matrix to $||g_{i,j}||$ is denoted by $||g^{i,j}||$. Its geometric meaning is a quadratic form on the cotangent space, or a symmetric operator from cotangent to tangent space.

In particular, the differential of a function f at a point x_0 is a covector $df(x_0) = \sum_k \partial_k f dx^k$ where $\partial_k = \frac{\partial}{\partial x^k}$. Using the tensor $g^{i,j}$ we can "lift the index" and make from a covector df a vector v with coordinates $v^k = \sum_{j=1}^n g^{k,j} \partial_j f(x_0)$. This vector is called the **gradient** of f and is denoted by $\operatorname{grad} f$. So,

$$\operatorname{grad} f = \sum_{k=1}^n (\operatorname{grad} f)^k \partial_k = \sum_{j=1}^n g^{k,j} \partial_j f \partial_k.$$

On the other hand, on the space of vector fields on M there is a natural operation **divergence** which associate with a vector field v a function $\operatorname{div} v$. If we choose any local coordinate system $x^1, \dots, x^n$ such that $\det ||g_{i,j}|| = 1$ (such a system is called **unimodular**), then the divergence is given by a simple formula:

$$\operatorname{div} v = \sum_k \partial_k v^k.$$

Definition C.1. The **Laplace-Beltrami** operator Δ on M is defined by the formula

$$\Delta f = \operatorname{div} \operatorname{grad} f.$$

In appropriate local coordinates at given point x_0 the Laplace-Beltrami operator can be always written as a sum of second partial derivatives: $\Delta = \sum_k \partial_k^2$. But in general this expression can not hold in a whole neighborhood of x_0. The obstacle is the curvature of the metric on M.

There is another, more geometric, definition of the Laplace-Beltrami operator. Take an ϵ-neighborhood $U_\epsilon(x_0)$ of a point x_0. Then the integral of f over $U_\epsilon(x_0)$ has the following asymptotic behavior when $\epsilon \to 0$:

$$\int_{U_\epsilon(x_0)} f(x) d^n x = a_n \epsilon^n \cdot f(x_0) + b_n \epsilon^{n+2} \cdot (\Delta f)(x_0) + o(\epsilon^{n+2})$$

where $a_n = \frac{\pi^{n/2}}{\Gamma(1+\frac{n}{2})}$ is the volume of a unit ball in $\mathbb{R}^n$ and $b_n = \frac{n}{n+2} a_n$.

Thus, we can define the value $(\Delta f)(x_0)$ as the limit

$$(\Delta f)(x_0) = \lim_{\epsilon \to 0} \frac{1}{b_n \epsilon^{n+2}} \int_{U_\epsilon(x_0)} \big(f(x) - f(x_0)\big) d^n x \tag{C.1}$$

which certainly exists for all functions with continuous second partial derivatives.

DEFINITION C.2. A function satisfying the equation $\Delta f = 0$ is called **harmonic**.

It is known that on every manifold of constant curvature (e.g. on the Euclidean space $\mathbb{R}^n$, on the sphere S^n or on hyperbolic space H^n) harmonic functions are characterized by the property

$$\frac{1}{vol(U_\epsilon(x_0))}\int_{U_\epsilon(x_0)} f(x)d^n x = f(x_0),$$

i.e. the average over any spherical neighborhood is equal to the value in the center. This property has an important corollary.

THEOREM C.1 (Maximum principle). *Assume that M is a connected manifold with boundary. Then any non-constant real harmonic function on M attains its maximal value only on the boundary ∂M.*

It is known also that for any continuous function φ on the boundary ∂M there exists a unique harmonic function f on M such that $f|_{\partial M} = \varphi$. Moreover, for any point $m \in M$ there exists a probabilistic measure μ_m on ∂M such that $f(m) = \int_{\partial M} \varphi(x)d\mu(x)$. It is called **Poisson measure** and in case of smooth boundary is given by a density $\rho_m(x)$ which is a smooth function of $m \in M$ and $x \in \partial M$.

There is a simple physical interpretation of a harmonic function (as a stable heat distribution) and probabilistic interpretation of Poisson measure $\mu_m(A)$ (as a probability to reach boundary in a set A starting from m and moving randomly along M).

C.2. There exists a pure algebraic approach to the definition of the Laplace operator.

Suppose, in a real vector space V two quadratic forms Q_0 and Q_1 are given. Assume also that Q_0 is positive: $Q_0(v) > 0$ for all $v \neq 0$. Then we can introduce in V a scalar product

$$(v_1,\, v_2) := \frac{Q_0(v_1 + v_2) - Q_0(v_1) - Q_0(v_2)}{2} \tag{C.2}$$

If V is infinite-dimensional, we assume in addition that it is complete with respect to the norm $||v||^2 := (v,\, v) = Q_0(v)$. Thus, V is a real Hilbert space.

The completeness condition is easy to satisfy: we simply replace V, if necessary, by its completion $\overline{V}$ with respect to the given norm.

The other quadratic form Q_1 will be defined on the dense subspace $V \subset \overline{V}$. From the theory of operators in Hilbert spaces we know the

PROPOSITION C.1. *There exists a symmetric densely defined operator A in $\overline{V}$ such that*

$$Q_1(v) = (Av,\, v) \quad \textit{for all} \quad v \in Dom(A) \supset V_1.$$

REMARK 2. Sometimes, A is called a quotient of two forms Q_1 and Q_0. Indeed, any quadratic form Q defines the symmetric bilinear form $\widetilde{Q} : V \times V \to V$ by the formula

$$\widetilde{Q}(v_1,\, v_2) := \frac{Q(v_1 + v_2) - Q(v_1) - Q(v_2)}{2}$$

The bilinear form $\widetilde{Q}$ in its turn can be interpreted as a linear map $\widetilde{Q} : V \to V^*$. Namely, we define the functional $f = \widetilde{Q}(v_1)$ on V as $f(v_2) = \widetilde{Q}(v_1,\, v_2)$.

The operator A can be written as $A = \widetilde{Q}_0^{-1} \circ \widetilde{Q}_1$.

♡

The standard theorem about conditional extremum leads to the

COROLLARY. *The eigenvalues and unit eigenvectors of A are exactly the critical values and critical points of the function $Q_1(v)$ on the sphere*[2] *$Q_0(v) = 1$.*

C.3. We apply the general algebraic scheme described in C.2 to the following situation. Let M be a smooth Riemannian manifold, possibly with boundary. Denote by V the space of smooth functions on M with compact support restricted by some boundary conditions – see below.

There are two natural quadratic forms on V:

$$Q_0(v) = \int_M v^2(m)\, dm \qquad \text{and} \qquad Q_1(v) = \int_M |\operatorname{grad} v|^2 dm \tag{C.3}$$

where the measure m on M and the scalar square $|\operatorname{grad} v|^2$ are determined by the metric.

According to the general scheme there is an operator A on $\overline{V} = L^2(M,\, dm)$ such that

$$\int_M (\operatorname{grad} v_1,\, \operatorname{grad} v_2) dm = \int_M Av_1(m) \cdot v_2(m)\, dm. \tag{C.4}$$

On the other hand, an explicit computation using the Stokes formula gives for the left hand side the expression

$$\int_{\partial M} v_1 \partial_\nu v_2\, dn - \int_M \Delta\, v_1(m) \cdot v_2(m)\, dm \tag{C.5}$$

where ∂_ν is the normal derivative and dn is a measure on $\partial\, M$ as on a Riemannian manifold with a metric inherited from M.

[2]Another formulation: The eigenvalues and eigenvectors of A are the critical values and critical points of the function $Q(v) := \frac{Q_1(v)}{Q_0(v)}$ on $V \backslash \{0\}$.

Suppose, we restrict v by an appropriate boundary condition which forces the boundary integral in C.3.3 vanish. Then the operator $-\Delta$ will be exactly the ratio of Q_1 and Q_0.

Two special examples are widely known: the **Dirichlet problem** when the condition

$$v\,|_{\partial M} = 0 \tag{C.6}$$

is imposed, and **Neumann problem** when the boundary condition is

$$\partial_\nu v\,|_{\partial M} = 0 \tag{C.7}$$

In both cases $-\Delta$ is a non-negative self-adjoint operator in $L^2(M,\, dm)$ whose domain of definition consists of C^1-functions v on M satisfying boundary conditions and such that $\Delta v \in L^2(M,\, dm)$ in the sense of generalized functions.

The connection of the operator Δ with variational problems gives the remarkable physical interpretation of eigenvalues and eigenfunctions of the Laplace-Beltrami operator. Namely, the eigenvalues describe the frequencies and eigenfunctions determine the forms of small oscillations of the manifold M considered as an elastic membrane.

The question: "what can be the spectrum of a Laplace operator on a smooth compact manifold?" has given raise to the whole new domain in mathematics: the **spectral geometry**.

Since the fractal sets are playing essential role in some modern mathematical models of physical problems, the study of analogues of Laplace-Beltrami operators on fractals became very popular. We refer the interested reader to the surveys [TAV00, Str99] and papers cited there.

♢

2.1. Laplace operator on $\mathcal{S}_N$

In the first version of the book I wanted to describe in full detail the definition and computation of the spectrum of Laplace operator on $\mathcal{S}_N$ and on $\mathcal{S}$. After that I learned that this program was already realized by several physicists and mathematicians, see e.g. [?R, FS92, ?La]. Therefore, I decided not to repeat the result one more time but instead concentrate on some different and less known problems. So here I restrict myself to a short description of the rather interesting technique used in the study of the spectrum.

To define the analogue of a Laplace operator on Sierpiński gasket $\mathcal{S}$, we consider first the finite approximation $\mathcal{S}_N$ of $\mathcal{S}$.

First, let us try to follow the scheme used above. Let S_n be the n-th finite approximation to the Sierpiński gasket $\mathcal{S}$. Denote by V_n the set of real functions on S_n. Since S_n consists of $\frac{3^n+3}{2}$ points, V_n is a real vector space of dimension $d_n = \frac{3^n+3}{2}$.

Let us define two quadratic forms on V_n:

$$Q_0(v) = \sum_{s\in\mathcal{S}_n} v(s)^2; \qquad Q_1(v) = \sum_{s'\leftrightarrow s''} \big((v(s') - v(s'')\big)^2 \tag{2.1.1}$$

where the first sum is over all points of $\mathcal{S}_n$, and the second is over all pairs of neighbor points.

Clearly, these quadratic forms are discrete analogues of the quadratic forms defined by (C2) in Info C.

As in the case of ordinary Laplace operator we use Q_0 to define a scalar product in V_n:

$$(f_1,\, f_2) = \sum_{s\in S_n} f_1(s)f_2(s).$$

Then the second form can be written as

$$Q_1(f) = (\Delta_n f,\, f) \quad \text{where} \quad \big(\Delta_n f\big)(s) = k(s)f(s) - \sum_{s'\leftrightarrow s} f(s') \tag{2.1.2}$$

Here $k(s)$ denotes the number of points which are neighbors to s, i.e. $k(s) = 4$ for inner points and $k(s) = 2$ for boundary points.

We introduce two sorts of boundary conditions.

The **Dirichlet** boundary condition is the equation $f(s) = 0$ for $s \in \partial S_n$. The space $V_n^{(D)}$ of functions satisfying this condition has dimension $d_n - 3 = \frac{3^n-3}{2}$. The operator $\Delta_n^{(D)}$ in this space is given by 2.2.2 in all inner points s.

The Neumann boundary condition is the equation $2f(s) = f(s') + f(s'')$ where $s \in \partial S_n$ and s', s'' are two neighbor points to s. The space $V_n^{(N)}$ of functions satisfying this condition again has dimension $d_n - 3 = \frac{3^n-3}{2}$. The operator $\Delta_n^{(N)}$ in this space is given by 2.1.2 in inner points.

Both $\Delta_n^{(D)}$ and $\Delta_n^{(N)}$ are self-adjoint operators and their spectra are known explicitly (see, e.g. [FS92]).

To make things clear, we consider in detail the case $n = 2$.

Let first $V = V_2^{(D)}$. It is a 3-dimensional space of functions on S_2 whose values are shown in Figure 2.1

The operator $\Delta_2^{(D)}$ sends the triple of values $(x,\, y,\, z)$ into the new triple $(4x - y - z,\, 4y - x - z,\, 4z - x - y)$. In the natural basis this operator is given by the matrix $\begin{pmatrix} 4 & -1 & -1 \\ -1 & 4 & -1 \\ -1 & -1 & 4 \end{pmatrix}$. The eigenvalues can be easily computed using

LEMMA 2.1. *Let $n \times n$ matrix A have elements*

$$a_{ij} = \begin{cases} a & \text{if} \quad i = j \\ b & \text{if} \quad i \neq j. \end{cases}$$

Then A has the eigenvalue $a - b$ with multiplicity $n - 1$ and one more eigenvalue $a = (n - 1)b$.

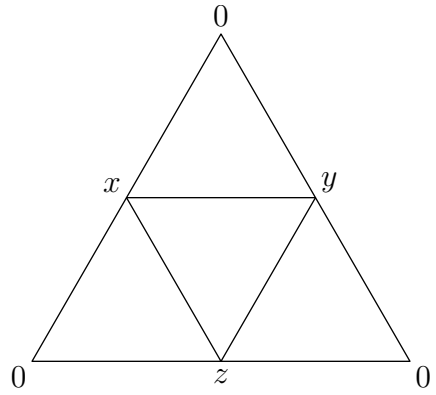

FIGURE 2.1. Functions on S_2 with Dirichlet condition

In our case we have a double eigenvalue 5 and simple eigenvalue 2. The corresponding eigenspaces consist of triples $(x,\, y,\, z)$ with $x+y+z=0$ and of triples $(x,\, y,\, z)$ with $x=y=z$.

It means that corresponding membrane (with fixed boundary) has two frequencies of oscillations such that their ratio is $\sqrt{\frac{5}{2}} \approx 1.581$.

Let now $V = V_2^{(N)}$. The values of functions from this space are shown in figure 2.2

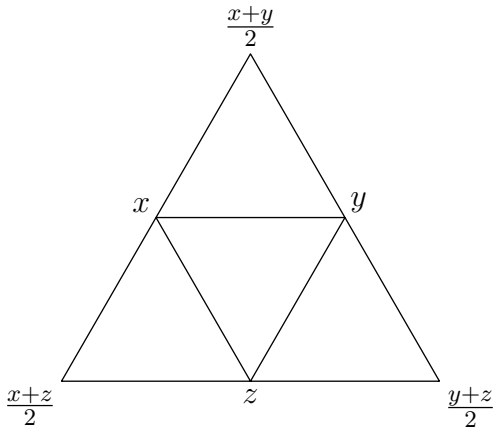

FIGURE 2.2. Functions on S_2 with Neumann condition

I leave you to check that the operator $\Delta_2^{(N)}$ sends the triple $(x,\, y,\, z)$ into the triple $\big(3x - \frac{3}{2}(y+z),\, 3y - \frac{3}{2}(y+z),\, 3z - \frac{3}{2}(y+z)\big)$. Therefore its matrix is $\begin{pmatrix} 3 & -\frac{3}{2} & -\frac{3}{2} \\ -\frac{3}{2} & 3 & -\frac{3}{2} \\ -\frac{3}{2} & -\frac{3}{2} & 3 \end{pmatrix}$. The spectrum of this matrix contains the double eigenvalue $4\frac{1}{2}$ and the single eigenvalue 0.

It means that corresponding membrane (with a free boundary) has one frequency of oscillations (slightly lower than the highest frequency in the first case) and one equilibrium state $x = y = z$.

2.2. Comparing spectra of Δ_n and of Δ_{n-1}

The computations we make in this section are rather dull and cumbersome, but they are necessary to get deep and beautiful results about the spectrum of the Laplace operator.

Let us denote by V_n^λ the space of functions satisfying

$$(4-\lambda)f(s) = \sum_{s\leftrightarrow s'} f(s') \tag{2.2.1}$$

for all inner points $s \in S_n$.

Let us choose a function $f \in V_n^{(\lambda)}$. Assume that the restriction of f on $\mathcal{S}_{n-1}$ is not identically zero. Consider in details a piece of $\mathcal{S}_n$ around the point where $f \neq 0$. We write the values of f on the corresponding points (values which do not matter marked by question marks):

$$\begin{array}{ccccccccc} & & & & ? & & & & \\ & & & ? & & ? & & & \\ & & y & & ? & & z & & \\ & u & & q & & r & & v & \\ b & & p & & x & & s & & c \end{array}$$

According to our hypothesis, $x \neq 0$. Moreover, since $f \in V_n^\lambda$, we have a family of equations:

$$\begin{gathered} (4-\lambda)x = p+q+r+s; \\ (4-\lambda)u = b+y+p+q; \qquad (4-\lambda)v = c+z+r+s; \\ (4-\lambda)p = b+u+q+x; \qquad (4-\lambda)q = y+u+p+x; \\ (4-\lambda)r = z+v+s+x; \qquad (4-\lambda)s = c+v+r+x \end{gathered} \tag{2.2.2}$$

Adding last four equations, we get

$$(4-\lambda)(p+q+r+s) = (p+q+r+s)+(b+y+z+c)+2(u+v)+4x \tag{2.2.3}$$

and adding two previous ones, we obtain

$$(4-\lambda)(u+v) = (p+q+r+s)+(b+y+z+c). \tag{2.2.4}$$

From (2.2.3), (2.2.4) we can express $(p+q+r+s)$ and $(u+v)$ in terms of $(b+y+z+c)$ and x. Then the first equation of (2.2.2) gives

$$(\lambda-6)(b+y+z+c) = (\lambda-6)(4-\lambda)(1-\lambda)x. \tag{2.2.5}$$

We come to the alternative: either $\lambda = 6$, or the function f (more precisely, its restriction to $\mathcal{S}_{n-1}$) belongs to V_{n-1}^μ where

$$4-\mu = (4-\lambda)(1-\lambda), \quad \text{or} \quad \mu = \lambda(5-\lambda). \tag{2.2.6}$$

The first important consequence of this alternative is

THEOREM 2.2. *The restriction of any harmonic function on $\mathcal{S}_n$ to $\mathcal{S}_{n-1}$ is also harmonic.*

Indeed, for harmonic functions $\lambda = 0$ and $\mu = \lambda(5 - \lambda)$ is also zero.

This fact leads to a natural definition of harmonic functions on $\mathcal{S}_\infty$.

DEFINITION 2.3. A function on $\mathcal{S}_\infty$ is called **harmonic** if its restriction on every $\mathcal{S}_n$ is harmonic.

2.3. Eigenfunctions of Laplace operator on S_N

Here we consider briefly the spectrum of the operators $\Delta_n^{(D)}$ with a goal to construct a Laplace operator $\Delta^{(D)}$ on $\mathcal{S}$.

First we have to study the so-called dynamics of the polynomial $P(\lambda) = \lambda(5-\lambda)$. Namely, for any number μ we call a μ-string any sequence $\mu_k,\ k = 0, 1, 2...$ such that $\mu_0 = \mu$ and $P(\mu_k) = \mu_{k-1}$ for $k \geq 1$.

We want to extend a function $f \in V_n^{\mu_n}$ so that extended function belong to $f \in V_{n+1}^{\mu_{n+1}}$. From (2.2.6) we know that it is possible only if μ_n and μ_{n+1} are in the same μ-string.

Conversely, for any μ-string $\{\mu_k\}$ we can construct a function f on $\mathcal{S}_\infty$ such that its restriction to $\mathcal{S}_n$ (which can be zero!) belongs to $V_n^{\mu_n}$ for all n.

So, the problem is: is such function f on $\mathcal{S}_\infty$ uniformly continuous, hence can be extended by continuity to $\mathcal{S}$? When this is the case, we can consider the extended function $\widetilde{f}$ as an eigenfunction for the Laplacian on the whole gasket and define the corresponding eigenvalue as a limit of suitably renormalized sequence $\{\mu_n\}$.

In this book we consider in detail only the case $\mu_n = 0$ where the function f is harmonic on $\mathcal{S}_\infty$.

CHAPTER 3

Harmonic functions on Sierpiński gasket

In this chapter we consider in more details the harmonic functions on Sierpiński gasket $\mathcal{S}$. Note, that a harmonic function satisfying Dirichlet boundary condition must be zero, and a harmonic function satisfying Neumann boundary condition must be a constant. So, we consider here harmonic functions whose restrictions on the boundary are subjected to no conditions.

Recall that the boundary points of $\mathcal{S}$ are 0, 1, $\omega = \frac{1+i\sqrt{3}}{2}$. So the segment [0, 1] of real line is a part of $\mathcal{S}$ and we can consider the restrictions of harmonic functions on this segment as ordinary real-valued functions on [0, 1]. It turns out that these functions have a very non-trivial analytic and number-theoretic behavior.

3.1. First properties of harmonic functions

We start with the following fact.

LEMMA 3.1. *The vector space $\mathcal{H}(\mathcal{S}_\infty)$ of all harmonic functions on $\mathcal{S}_\infty$ has dimension 3. The natural coordinates of a function $f \in \mathcal{H}(\mathcal{S}_\infty)$ are the values of this function at three boundary points.*

PROOF. From linear algebra we know that if an homogeneous system of linear equations has only the trivial solution, then the corresponding inhomogeneous system has the unique solution for any right hand part. It follows that $\dim \mathcal{H}(\mathcal{S}_n) = 3$ for all $n \geq 1$. Hence, any harmonic function on $\mathcal{S}_n$ has a unique harmonic extension to $\mathcal{S}_{n+1}$, hence, to $\mathcal{S}_\infty$. □

We need also the following simple observation

FIGURE 3.1. The ratio 1:2:2

LEMMA 3.2. *Let x, y, z be three neighbor points of $\mathcal{S}_m$ which form a regular triangle. Put $\alpha = \frac{y+z}{2}$, $\beta = \frac{x+z}{2}$, $\gamma = \frac{x+y}{2}$. Then α, β, γ also form a regular triangle and are neighbor points in $\mathcal{S}_{m+1}$ (see Fig. 3.1). For any harmonic function f on $\mathcal{S}_{m+1}$ we have*

$$f(\alpha) = \frac{f(x) + 2f(y) + 2f(z)}{5}, \quad f(\beta) = \frac{2f(x) + f(y) + 2f(z)}{5}, \quad f(\gamma) = \frac{2f(z) + 2f(y) + f(x)}{5}. \tag{3.1.1}$$

The informal meaning of this result is: the neighbor points have twice bigger impact than the opposite one.

Now we can prove the important result:

THEOREM 3.1. *Any harmonic function on $\mathcal{S}_\infty$ is uniformly continuous, hence has a unique continuous extension to $\mathcal{S}$.*

PROOF. Let f^c_{ab} be the harmonic function on $\mathcal{S}_\infty$ with the boundary values

$$f(0) = a, \quad f(1) = b, \quad f(\omega) = c.$$

Let us call the **variation** of a function f on a set X the quantity

$$\operatorname{var}_X f = \sup_{x,y \in X} |f(x) - f(y)|.$$

From the Maximum principle we conclude that

$$\operatorname{var}_{\mathcal{S}} f^c_{ab} = \max\{|a-b|, |b-c|, |c-a|\}.$$

From Lemma 3.2 and by induction on n we derive easily that for any two neighbor points x, y in $\mathcal{S}_n$ we have

$$|f^c_{ab}(x) - f^c_{ab}(y)| \le \operatorname{var} f \cdot \left(\frac{3}{5}\right)^n \le const \cdot d(x, y)^\beta, \quad \beta = \log_2 \frac{5}{3}.$$

Hence, the function f^c_{ab} belongs to some Hölder class. Therefore, it is uniformly continuous and can be extended by continuity to $\mathcal{S}$. We keep the same notation f^c_{ab} for the extended function. □

3.2. The functions χ, ϕ, ψ, ξ

Denote by u^c_{ab} the restriction of the harmonic function f^c_{ab} on the segment $[0, 1]$ which is the horizontal side of $\mathcal{S}$.

The following relations are rather obvious and follow from the natural action of the permutation group S_3 on $\mathcal{S}$ and on $\mathcal{H}(\mathcal{S})$:

$$u^c_{ab}(t) = u^c_{ba}(1-t); \qquad u^c_{ab}(t) + u^a_{bc}(t) + u^b_{ca}(t) \equiv a + b + c. \tag{3.2.1}$$

It follows that the values of any harmonic function at any point of $\mathcal{S}_n$ can be expressed in terms of a single function $\phi := u^0_{01}$.

EXERCISE 11. Derive from 3.2.1 that

$$u^c_{ab}(t) = c + (b-c)\phi(t) + (a-c)\phi(1-t). \tag{3.2.2}$$

Therefore, it is interesting to obtain as many information as possible about the nature of the function ϕ. Technically, it is convenient to introduce three other functions:

$$\begin{aligned} \chi(t) &:= u_{01}^{-1}(t) = -1 + 2\phi(t) + \phi(1-t), \\ \psi(t) &:= u_{01}^{1}(t) = 1 - \phi(1-t), \\ \xi(t) &:= u_{01}^{2}(t) = 2 - \phi(t) - 2\phi(1-t). \end{aligned} \tag{3.2.3}$$

The reason to introduce these four functions is the following. Let $\mathcal{H}$ denote the space of real-valued functions on $[0,\,1]$ spanned by restrictions of harmonic functions on $\mathcal{S}$. (It is worth to mention, that $\mathcal{H}$ is spanned by any two of the above functions χ, ϕ, ψ, ξ and a constant function.)

Consider two transformations of the segment $[0,\,1]$: $\alpha_0(t) = \frac{t}{2}$ and $\alpha_2(t) = \frac{1+t}{2}$. They induce the linear operators of functions:

$$\big(A_0 f\big)(t) = f\left(\frac{t}{2}\right) \qquad \text{and} \qquad \big(A_1 f\big)(t) = f\left(\frac{1+t}{2}\right).$$

It turns out that both linear operators A_0 and A_1 preserve the 3-dimensional subspace $\mathcal{H}$. Moreover, both operators have in $\mathcal{H}$ three different eigenvalues 1, $\frac{3}{5}$, $\frac{1}{5}$.

The corresponding eigenfunctions are 1, ψ, χ for A_0 and 1, $1-\xi$, $1-\phi$ for A_1.

In other words, if we introduce vector-functions

$$\vec{f}(x) = \begin{pmatrix} \psi(x) \\ \chi(x) \\ 1 \end{pmatrix} \quad \text{and} \quad \vec{g}(x) = \begin{pmatrix} \phi(x) \\ \xi(x) \\ 1 \end{pmatrix}, \tag{3.2.4}$$

then the following relations hold

$$\vec{f}\left(\frac{t}{2}\right) = A_0\vec{f}(t), \quad \vec{g}\left(\frac{1+t}{2}\right) = A_1\vec{g}(t), \quad \vec{f}(1-t) = T\vec{g}(t) \tag{3.2.5}$$

where

(3.2.6)

$$A_0 = \begin{pmatrix} 3/5 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 3/5 & 0 & 2/5 \\ 0 & 1/5 & 4/5 \\ 0 & 0 & 1 \end{pmatrix}, \quad T = \begin{pmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Exercise 12. Using relations 3.2.5, 3.2.6, compile the table of values of functions χ, ϕ, ψ, ξ at the points $k/8$, $k = 0, 1, \ldots, 7, 8$.

From 3.2.5 we derive several remarkable properties of the functions introduced above. For example, we can describe the behavior of these functions near all dyadic points r of the form $r = \frac{k}{2^n}$.

Lemma 3.3. *All four functions χ, ϕ, ψ and ξ increase strictly monotonically from 0 to 1 on $[0,1]$.*

PROOF. Since $\phi(t) = \frac{\xi(t)+2\chi(t)}{3}$ and $\psi(t) = \frac{2\xi(t)+\chi(t)}{3}$, it is enough to prove that $\xi(t)$ and $\chi(t)$ are strictly increasing. Let $0 \le t < s \le 1$. We have to show that $\xi(t) < \xi(s)$ and $\chi(t) < \chi(s)$. Let us introduce the vector-function $\vec{h}(t) := \begin{pmatrix} \xi(t) \\ \chi(t) \\ 1 \end{pmatrix}$.

From 3.2.5 we derive the following transformation rules for $\vec{h}$:

$$\vec{h}\left(\frac{t}{2}\right) = B_0\vec{h}(t); \qquad \vec{h}\left(\frac{1+t}{2}\right) = B_1\vec{h}(t) \tag{3.2.7}$$

where

$$B_0 = \begin{pmatrix} 3/5 & 1/5 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \qquad B_1 = \begin{pmatrix} 1/5 & 0 & 4/5 \\ 1/5 & 3/5 & 1/5 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.2.8}$$

Consider now the binary presentations of t and s:

$$t = 0.t_1t_2\ldots t_k\ldots, \qquad s = 0.s_1s_2\ldots s_k\ldots.$$

We can assume that $t_i = s_i$ for $i < m,\ t_m = 0,\ s_m = 1$.

Applying 3.2.7 several times, we get

$$\vec{h}(t) = B_{t_1}\cdots B_{t_{k-1}}A_0\vec{f}(z), \qquad \vec{h}(s) = B_{t_1}\cdots B_{t_{k-1}}B_1\vec{f}(w)$$

for some $z \in [0,\ 1),\ w \in (0,\ 1]$. Since B_i have nonnegative coefficients, it is enough to verify that $B_1\vec{h}(w) > B_0\vec{f}(z)$. (Here we write $\vec{a} > \vec{b}$ if the first two coordinates of $\vec{a}$ are bigger than the corresponding coordinates of $\vec{b}$.)

But

$$B_1\vec{h}(w) = \begin{pmatrix} 1/5 & 0 & 4/5 \\ 1/5 & 3/5 & 1/5 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \xi(t) \\ \chi(t) \\ 1 \end{pmatrix} > \begin{pmatrix} 0.8 \\ 0.2 \\ 1 \end{pmatrix}$$

while

$$B_0\vec{f}(z) = \begin{pmatrix} 3/5 & 1/5 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \xi(z) \\ \chi(z) \\ 1 \end{pmatrix} < \begin{pmatrix} 0.8 \\ 0.2 \\ 1 \end{pmatrix}.$$

□

THEOREM 3.2. *For all $x \in [0, 1]$ we have the relations*

$$A^{-1}x^\alpha \le \psi(x) \le Ax^\alpha, \qquad B^{-1}x^\beta \le \chi(x) \le Bx^\beta$$
$$\text{with } A = \frac{5}{3}, \quad \alpha = \log_2\frac{5}{3}, \qquad B = 5, \quad \beta = \log_2 5. \tag{3.2.9}$$

PROOF. Since $\frac{3}{5} \le \psi(x) \le 1$ for $\frac{1}{2} \le x \le 1$, we conclude from the first relation that

$$\left(\frac{3}{5}\right)^{n+1} \le \psi(x) \le \left(\frac{3}{5}\right)^n \qquad \text{for} \quad \frac{1}{2^{n+1}} \le x \le \frac{1}{2^n}.$$

But for the given value of α we have also

$$\left(\frac{3}{5}\right)^{n+1} \le x^{\alpha} \le \left(\frac{3}{5}\right)^{n} \qquad \text{for} \quad \frac{1}{2^{n+1}} \le x \le \frac{1}{2^{n}}.$$

This implies the first statement of the theorem. The second can be proved in the same way. □

As a corollary of Theorem 3, we obtain

$$u'(r) = +\infty. \tag{3.2.10}$$

where u is any one from the functions χ, ϕ, ψ, ξ and $r = \frac{k}{2^n}$ is any dyadic number with only two exceptions: $\chi'(0) = \xi'(1) = 0$ (see Fig.3.2).

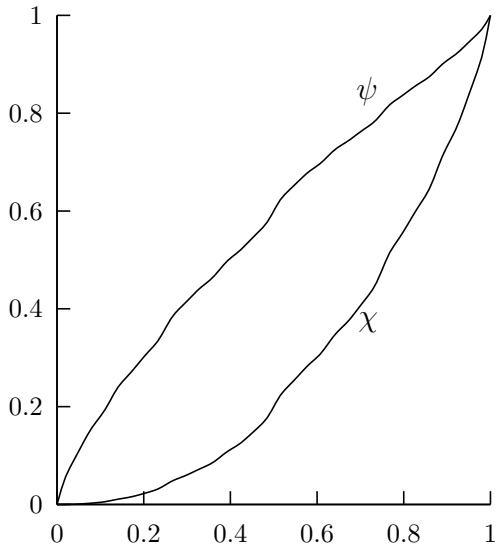

FIGURE 3.2. Functions χ, ϕ, ψ, ξ.

On the other hand, the functions χ, ϕ, ψ, ξ, being strictly monotone, have a finite derivative at almost all points of the interval [0, 1].

PROBLEM 2. Compute explicitly the derivative $u'(t)$ whenever it is possible (e.g. at all rational points).

The next interesting feature of $u(t)$ is that one can compute explicitly the integral of this function over any interval with dyadic ends. For instance, we have

LEMMA 3.4.

$$\int_0^1 u^c_{a,b}(t)dt = \frac{3a + 3b + c}{7}. \tag{3.2.11}$$

On the other side, the Corollary above suggests that t is, maybe, not a good parameter for functions u^c_{ab}. A more natural choice for the independent parameter x and a function $y(x)$ is

$$x = \phi + \psi - 1 = \chi + \xi - 1; \qquad y = \xi - \psi = \psi - \phi = \phi - \chi. \tag{3.2.12}$$

When t runs from 0 to 1, x increases from -1 to 1, while y grows from 0 to $\frac{1}{5}$ and then decays again to 0. The alternative definition is: $x = u^0_{-1,1}$, $y = u^1_{0,0}$.

THEOREM 3.3. *The quantity y is a differentiable function of x.*

A more precise statement is

THEOREM 3.4. *The derivative $y' = \frac{dy}{dx}$ is a continuous strictly decreasing function of x.*

EXERCISE 13. Show that the derivative $y'(x)$ satisfies the equations

$$y'\left(x\left(\frac{t}{2}\right)\right) = \frac{3y'(x(t)) + 1}{3y'(x(t)) + 5}, \qquad y'\left(x\left(\frac{1+t}{2}\right)\right) = \frac{3y'(x(t)) - 1}{5 - 3y'(x(t))}.$$

Hint. Prove and use the relations

(3.2.13)

$$x\left(\frac{t}{2}\right) = \frac{1}{2}x(t) + \frac{3}{10}y(t) - \frac{1}{2}; \qquad y\left(\frac{t}{2}\right) = \frac{1}{10}x(t) + \frac{3}{10}y(t) + \frac{1}{10}$$

$$x\left(\frac{1+t}{2}\right) = \frac{1}{2}x(t) - \frac{3}{10}y(t) + \frac{1}{2}; \qquad y\left(\frac{1+t}{2}\right) = -\frac{1}{10}x(t) + \frac{3}{10}y(t) - \frac{1}{10}.$$

We come back to this in Part II.

The next two problems are open.

PROBLEM 3. Compute the moments

$$m_n := \int_{-1}^{1} x^n y dx. \tag{3.2.14}$$

PROBLEM 4. Compute the Fourier coefficients

$$c_n := \int_{-1}^{1} e^{-\pi i n x} y dx. \tag{3.2.15}$$

3.3. Extension and computation of $\chi(t)$ and $\psi(t)$

There is a method of quick computing the values of $\chi(t)$ at binary fractions. Namely, we know that $\chi(t)$ satisfies relations[1]

$$\chi(2t) = 5\chi(t), \qquad \chi\left(\frac{1+t}{2}\right) + \chi\left(\frac{1-t}{2}\right) = \frac{2 + 3\chi(t)}{5}. \tag{3.3.1}$$

[1]The simplest way to derive these equation is to compare the boundary values of both sides, taking into account that they are harmonic functions.

We can use the first relation 3.3.1 to extend χ to the whole real line, putting
(3.3.2)

$$\chi(t) := 5^N \chi(2^{-N}|t|) \qquad \text{where } N \text{ is big enough for } \quad 0 \le 2^{-N}|t| \le 1.$$

Then the second equation for $t = \frac{k}{2^n}$ can be rewritten in the form

$$\chi(2^n + k) + \chi(2^n - k) - 2\chi(2^n) = 3\chi(k) \qquad \text{for} \quad 0 \le k \le 2^n. \tag{3.3.3}$$

Let us introduce the operator of second difference

$$\big(\Delta_k^2 f\big)(t) = \frac{f(t+k) - 2f(t) + f(t-k)}{2}.$$

Then we can write

$$\big(\Delta_k^2 \chi\big)(2^n) = 3\chi(k) \quad \text{for} \quad 0 \le k \le 2^n. \tag{3.3.4}$$

It is easy to derive from (3.3.4) the following statement.

THEOREM 3.5. *For any integer k the value $\chi(k)$ is also an integer and $\chi(k) \equiv k \mod 3$.*

The relation (3.3.4) allows not only compute the values $\chi(k)$ for integer k but also formulate the following

Conjecture 1. Let $\beta = \log_2 5 = 2.3219281...$ The ratio $\frac{\chi(t)}{t^\beta}$ attains a maximal value 1.044... at the point $t_{\max} \approx \frac{8}{15}$ and a minimal value 0.912... at the point $t_{\min} \approx \frac{93}{127}$.

A similar approach allows to compute the values of extended function ψ at integral points. The key formula is the following analog of (3.3.4):

$$\big(\Delta_k^2 \psi\big)(2^n) = -\frac{1}{3}\chi(k) \quad \text{for} \quad 0 \le k \le 2^n. \tag{3.3.5}$$

In the table below we give the values of $\chi(k)$ and values of $\psi(k)$ (multiplied by $3^6 = 729$ to make them integral). We also show the first differences $\Delta\psi(k) := \psi(k) - \psi(k-1)$ for the function $\psi(k)$ and the second differences $\Delta_1^2\chi(k)$ for the function $\chi(k)$.

Note that the first differences $\Delta\psi(k)$ manifest a symmetry in the intervals $[2^l,\, 2^{l+1}]$. This symmetry is due to the relation

$$\psi(3+t) + \psi(3-t) = 2\psi(3) = \frac{40}{3} \qquad \text{for} \quad |t| \le 1 \tag{3.3.6}$$

In particular, putting $t = \frac{k}{16}$, $0 \le k \le 16$, we get

$$\psi(48+k) + \psi(48-k) = \frac{25000}{729}.$$

The same symmetry is observed for φ:

$$\varphi\left(\tfrac{1}{4} + t\right) + \varphi\left(\tfrac{1}{4} + t\right) = 2\varphi\left(\tfrac{1}{4}\right) \qquad \text{for} \qquad |t| \le \tfrac{1}{4}. \tag{3.3.7}$$

All this suggest the search of minimal "wavelets" such that graphs of all basic functions can be built from affine images of these wavelets.

The candidates are the graphs of χ on $[\frac{1}{2},\, 1]$ and of ψ on $[\frac{3}{4},\, 1]$.

TABLE 3.1. Table of values of $\chi(k)$, $2^6\psi(k)$ and their second differences

k	$\chi(k)$	$\frac{1}{3}\Delta^2\chi$	$3^6\psi(k)$	$3^6\Delta\psi$		k	$\chi(k)$	$\frac{1}{3}\Delta^2\chi$	$3^6\psi(k)$	$3^6\cdot\Delta\psi$
1	1	1	729	729		34	3745	-11	9985	245
2	5	1	1215	486		35	3965	5	10191	206
3	12	2	1620	405		36	4200	-2	10400	209
4	25	1	2025	405		37	4429	-11	10597	197
5	41	1	2403	378		38	4625	5	10755	158
6	60	2	2700	297		39	4836	26	10916	161
7	85	5	2997	297		40	5125	1	11125	209
8	125	1	3375	378		41	5417	-23	11331	206
9	168	-2	3744	369		42	5640	-2	11480	149
10	205	1	74005	261		43	5857	17	11617	137
11	245	5	4239	234		44	6125	5	11775	158
12	300	2	4500	261		45	6408	-2	11936	161
13	361	1	4761	261		46	6685	17	12085	149
14	425	5	4995	234		47	7013	53	12255	170
15	504	14	5256	261		48	7500	2	12500	245
16	625	1	5625	369		49	7993	-47	12745	245
17	749	-11	5991	366		50	8345	-11	12915	170
18	840	-2	6240	249		51	8664	14	13064	149
19	925	5	6453	213		52	9025	1	13225	161
20	1025	1	6675	222		53	9389	-11	13383	158
21	1128	-2	6888	213		54	9720	14	13520	137
22	1225	5	7065	177		55	10093	53	13669	149
23	1337	17	7251	186		56	10625	5	13875	206
24	1500	2	7500	249		57	11172	-33	14084	209
25	1669	-11	7749	249		58	11605	1	14245	161
26	1805	1	7935	186		59	12041	41	14403	158
27	1944	14	8112	177		60	12600	14	14600	197
28	2125	5	8325	213		61	13201	1	14809	209
29	2321	1	8547	222		62	13805	41	15015	206
30	2520	14	8760	213		63	14532	122	15260	245
31	2761	41	9009	249		64	15625	1	15625	365
32	3125	1	9375	366		65	16721	-119	$15989\frac{2}{3}$	$364\frac{2}{3}$
33	3492	-38	9740	365		66	17460	-38	$16233\frac{1}{3}$	$243\frac{2}{3}$

We leave to the reader to observe other patterns in this table and prove corresponding statements. For example, look at the values of $\Delta\psi$ at the points 2^n, $\quad 2^n \pm 1$, $\quad 2^n + 2^{n-1}$ and $2^n + 2^{n-1} + 1$.

It is also interesting to study p-adic behavior of $\chi(t)$ and the possible extension of $\chi(t)$ to a function from $\mathbb{Q}_2$ to $\mathbb{Q}_5$.

Finally, we recommend to draw a graph of the function $k \to \Delta\psi(k)$ on the interval $[2^n + 1,\ 2^{n+1}]$ and think about its limit when n goes to ∞.

Info D. Fractional derivatives and fractional integrals

The derivative of order n is defined as the n-th iteration of ordinary derivative. Sometime the indefinite integral $\int_0^x f(t)dt$ is called the anti-derivative of f, or the derivative of order -1. One can also define the derivative of order $-n$ as the n-th iteration of the anti-derivative. The explicit form of this operation is

$$f^{(-n)}(x) = \int_0^x dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} f(t_n)dt_n.$$

This iterated integral can be written as n-dimensional integral

$$\int_{\Delta_x} f(t_n)dt_1dt_2\cdots dt_n$$

where Δ_x is the simplex in $\mathbb{R}^n$ with coordinates t_1, t_2, $\dots t_n$ given by the inequalities

$$0 \le t_1 \le t_2 \le \cdots \le t_n \le x.$$

If we change the order of integration, we can rewrite this integral in the form

$$\int_{\Delta_x} f(t_n)dt_1dt_2\cdots dt_n = \int_0^x vol\Delta_x(t)f(t)dt = \int_0^x \frac{(x-t)^{n-1}}{(n-1)!}f(t)dt. \tag{D.1}$$

Here $\Delta_x(t)$ is the $(n-1)$-dimensional simplex which is obtained as the intersection of Δ_x and the hyperplane $t_n = t$.

Now we observe that the factor $\frac{(x-t)^{n-1}}{(n-1)!}$ make sense not only for $n \in \mathbb{N}$ but for any real n. So, we replace n by α and define an anti-derivative of order α, or a derivative of order $-\alpha$ by the formula

$$f^{(-\alpha)}(x) = \int_0^x \frac{(x-t)^{\alpha-1}}{\Gamma(\alpha)}f(t)dt. \tag{D.2}$$

Of course, we have to precise, what kind of functions we allow to consider and how to understand this integral when the integrand has singularity at 0. For the beginning it is enough to assume that our functions are defined and smooth on $(0,\ \infty)$ and also vanish at zero together with several derivatives.

Exercise 14. 19. Denote by $\Phi_\beta(x)$ the function $\frac{x^{\beta-1}}{\Gamma(\beta)}$. Show that

$$\Phi_\beta^{(-\alpha)}(x) \ = \ \Phi_{\beta-\alpha}(x). \tag{D.3}$$

HINT. Use the B-function of Euler given by

$$B(\alpha,\,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

and the identity

$$B(\alpha,\,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Note also the connection of fractional derivatives with the convolution operation on $\mathbb{R}_+$:

$$(f_1*f_2)(x)=\int_0^x f_1(t)f_2(x-t)dt.$$

Namely, the derivative of order α is just a convolution with $\Phi_{-\alpha}$ while integral of order α is a convolution with Φ_α.

$\diamondsuit$

3.4. Some arithmetic properties of basic functions

As was shown in 3.3, the function $\chi(t)$ takes integer values in integer points. Such functions often have interesting arithmetic properties. For convenience we extend this function to the whole line $\mathbb{R}$ by the rules:

$$\chi(2t)=5\chi(t),\qquad \chi(-t)=\chi(t) \tag{3.4.1}$$

The extended function still takes integer values in integer points.

We also extend the functions ψ, ϕ, ξ to the positive half-line $\mathbb{R}_+$ by the rules

$$\psi(2t)=\frac{5}{3}\psi(t),\qquad \phi(t)=\frac{\chi(t)+\psi(t)}{2},\qquad \xi(t)=\frac{3\psi(t)-\chi(t)}{2} \tag{3.4.2}$$

We can consider these functions as boundary values of harmonic functions defined on the **infinite Sierpiński gasket** bounded by the rays $x\ge 0,\ y=0$ and $x\ge 0,\ y=\frac{x\sqrt{3}}{2}$.

We want to study the local behavior of χ in a vicinity of some dyadic number $r=\frac{k}{2^n}$. In view of 3.4.1, it is sufficient to consider only odd positive integers $k=2m+1$.

THEOREM 3.6. *For any odd k and any $\tau\in[0,\,1]$ we have*

$$\chi(k\pm\tau)=\chi(k)+\Delta_2\cdot\chi(\tau)\pm\Delta_1\cdot\big(2\chi(\tau)+3\psi(\tau)\big) \tag{3.4.3}$$

where $\Delta_2=\frac{\chi(k-1)+\chi(k+1)-2\chi(k)}{2}$, $\quad\Delta_1=\frac{\chi(\frac{k+1}{2})-\chi(\frac{k-1}{2})}{2}$.

COROLLARY. *For any n and any odd k and odd $l<2^n$ we have*[2]

[2]Note that the number $3^{n+1}\psi(l)$ is an integer when $l<2^n$.

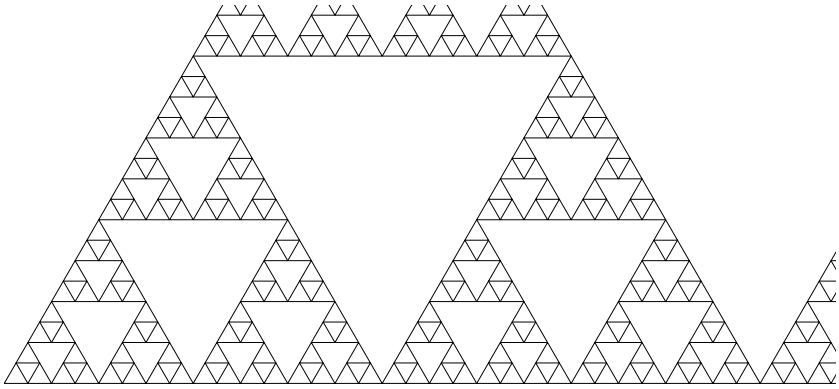

FIGURE 3.3. Infinite Sierpiński gasket.

$$\chi(2^n k + l) \equiv \chi(2^n k - l) \mod \big(2\chi(l) + 3^{n+1}\psi(l)\big) \tag{3.4.4}$$

and

$$\chi(2^n k + l) + \chi(2^n k - l) - 2\chi(2^n k) \equiv 0 \mod \chi(l) \tag{3.4.5}$$

Some particular cases:

a) $n = 1$, $k = 2m + 1$, $l = 1$: $\chi(4m + 3) \equiv \chi(4m + 1) \mod 11$

b) $n = 2$, $k = 2m + 1$, $l = 3$: $\chi(8m + 7) \equiv \chi(8m + 1) \mod 84$

c) $k = 1$: $\chi(2^n + l) \equiv \chi(2^n - l) \mod \big(2\chi(l) + 3^{n+1}\psi(l)\big)$ (actually, it is not only congruence but even equality since in this case $2\Delta_1 = 1$.)

PROOF OF THE THEOREM. Consider the triangular piece of the infinite gasket which is based on the segment $[k - 1, k + 1]$. It is shown on Figure 3.4.

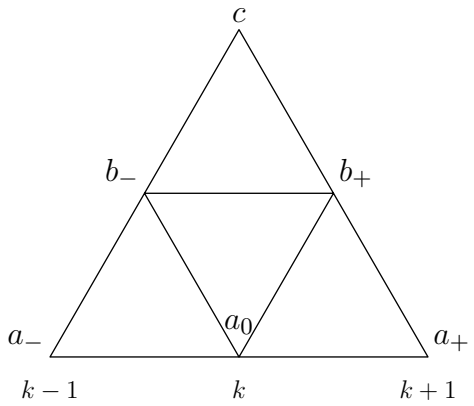

FIGURE 3.4. A fragment of infinite Sierpiński gasket.

We denote the values of χ at the points $k-1$, k, $k+1$ by $a_-,\ a,\ a_+$ respectively. Then the values $b_+,\ b_-,\ c$ in remaining vertices, shown on Figure 3.3, can be uniquely determined from the equations:

$$5a = 2a_- + 2a_+ + c, \qquad 5b_\pm = 2a_\pm + 2c + a_\mp.$$

The result is

$$c = 5a - 2a_- - 2a_+, \qquad b_+ = 2a - \frac{3a_- + 2a_+}{5}, \qquad b_- = 2a - \frac{2a_+ + 3a_-}{5}.$$

Consider now the functions $g_\pm : \tau \to \chi(k \pm \tau)$. Knowing the boundary values of corresponding harmonic functions on pieces of $\mathcal{S}$, we can write:

$$g_\pm(\tau) = a + \frac{a_\pm + b_\pm - 2a}{2} \cdot \psi(\tau) + \frac{a_\pm - b_\pm}{2} \cdot \chi(\tau).$$

To prove the theorem it remains to note that

$$\frac{a_\pm + b_\pm - 2a}{2} = \pm\frac{3}{10}(a_+ - a_-) = \pm 3 \cdot \Delta_1$$

and

$$\frac{a_\pm - b_\pm}{2} = \frac{a_- + a_+ - 2a}{2} \pm \frac{1}{5}(a_+ - a_-) = \Delta_2 \pm 2\Delta_1.$$

□

Proof of the corollary. Put $\tau = \frac{l}{2^n}$ in (3.4.1). Then we get

$$\chi(2^n k + l) - \chi(2^n k - l) = 5^n\big(\chi(k + \tfrac{l}{2^n}) - \chi(k - \tfrac{l}{2^n})\big) =$$
$$2 \cdot 5^n \Delta_1\big(2\chi(\tfrac{l}{2^n}) + 3\psi(\tfrac{l}{2^n})\big) = 2 \cdot \Delta_1 \cdot \big(2\chi(l) + 3^{n+1}\psi(l)\big).$$

Since $2\Delta_1 \in \mathbb{Z}$, we have proved (3.4.4) The congruence (3.4.5) can be proved in a similar way. □

3.5. Functions $x(t)$, $y(t)$ and $y(x)$

Theorem 3.6 suggests that apparently t is not a good parameter for basic functions. A more natural choice for the independent parameter x and a function $y(x)$ is:

(3.5.1) $$x = \phi + \psi - 1 = \chi + \xi - 1; \qquad y = \xi - \psi = \psi - \phi = \phi - \chi$$

The alternative definition: $x = u^{\,0}_{-1,1}, \quad y = u^{\,1}_{0,0}$.

When t runs from 0 to 1, the value of x increases from -1 to 1, while the value of y grows from 0 at 0 to $\frac{1}{5}$ at $\frac{1}{2}$ and then decays again to 0 at 1.

All basic functions are easily expressed in terms of x and y:

(3.5.2)
$$\chi = \frac{x + 1 - 3y}{2}, \quad \phi = \frac{x + 1 - y}{2}, \quad \psi = \frac{x + 1 + y}{2}, \quad \xi = \frac{x + 1 + 3y}{2}$$

The another advantage of this choice is the nice behavior of x and y with respect to operator T : $Tx = -x,\ Ty = y$.

The disadvantage is the more complicated behavior with respect to A_1 and A_2. Namely, if we introduce the vector function $\vec{h}(t) = (x(t),\, y(t),\, 1)^t$, then we get the following transformation rules:

$$\vec{h}\left(\frac{t}{2}\right) = C_0\vec{h}(t), \qquad \vec{h}\left(\frac{1+t}{2}\right) = C_1\vec{h}(t) \tag{3.5.3}$$

where

$$C_0 = \frac{1}{10}\begin{pmatrix} 5 & 3 & -5 \\ 1 & 3 & 1 \\ 0 & 0 & 10 \end{pmatrix}, \qquad C_1 = \frac{1}{10}\begin{pmatrix} 5 & -3 & 5 \\ -1 & 3 & 1 \\ 0 & 0 & 10 \end{pmatrix} \tag{3.5.4}$$

Both quantities x and y are originally functions of $t \in [0,\, 1]$. Since x defines a bijection $[0,\, 1] \to [-1,\, 1]$, we can consider the map

$$\widetilde{y} := y \circ x^{-1} : [-1,\, 1] \to [0,\, 1].$$

Often we will not distinguish between y and $\widetilde{y}$ and write simply $y(x)$.

The claim that x is a better parameter is supported by the following fact

THEOREM 3.7. *The derivative $y' = \frac{dy}{dx}$ exists and is a continuous strictly decreasing function of x.*

We leave the proof to the reader as a rather non-trivial exercise. In my opinion, the best way to prove the theorem is to show that y is a concave function in x, i.e.

$$y\left(\frac{x_1 + x_2}{2}\right) \;>\; \frac{y(x_1) + y(x_2)}{2} \tag{3.5.5}$$

EXERCISE 15. Show that the derivative $y'(x)$ satisfies the equations

$$y'(x(\tfrac{t}{2})) = \frac{3y'(x(t)) + 1}{3y'(x(t)) + 5}, \qquad y'(x(\tfrac{1+t}{2})) = \frac{3y'(x(t)) - 1}{5 - 3y'(x(t))} \tag{3.5.6}$$

HINT. Use the relations (3.5.4).

The relations (3.5.4) allow to compute the derivative $y'(x)$ explicitly in some points (knowing that the derivative exists).

E.g., if we put $t = 0$ in the first relation, we get the equation $y'(0) = \frac{3y'(0)+1}{3y'(0)+5}$, or $3y'(0)^2 + 2y'(0) - 1 = 0$.

This quadratic equation has two roots: $\frac{1}{3}$ and -1. But since $y(-1) = 0$ and $y(-1+\epsilon) > 0$, only the first root is suitable. So, we get $y'(-1) = \frac{1}{3}$.

In the same way, putting $t = 1$ in the second relation, we get $y'(1) = -\frac{1}{3}$.

The graphs of the functions $y(x)$ and $y'(x)$ are shown on Figure 3.5

The method used above can be applied to compute $y'(x)$ for any x of the form $x(t)$ with a rational t. Indeed, any rational number r can be written as an eventually periodic dyadic fraction. It follows that r has the form $r = \frac{k}{2^m(2^n-1)}$ where n is the length of the period and m is the number of digits before the period starts.

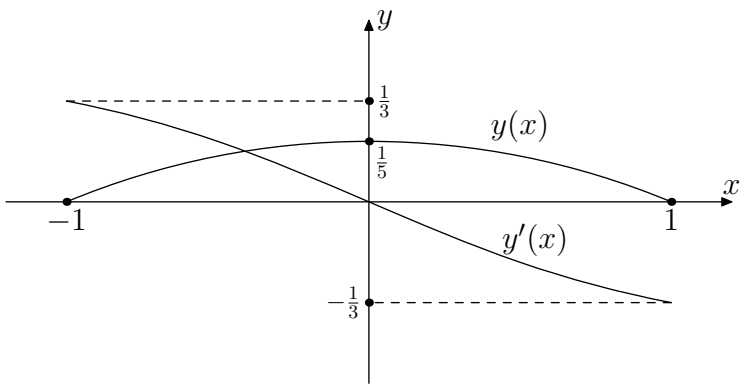

FIGURE 3.5. The graphs of the functions $y(x)$ and $y'(x)$

E.g., $\frac{5}{6} = 0.11010101... = 0.1(10) = \frac{5}{2(2^2-1)}$.

The number $r' = \frac{k}{2^n-1}$ is a fixed point of some transformation of the form $\alpha := \alpha_{i_1}\alpha_{i_2}\cdots\alpha_{i_n}$ (see section 3.2). And the number r is the image of r' under some transformation of the form $\alpha' := \alpha_{j_1}\alpha_{j_2}\cdots\alpha_{j_m}$.

Geometrically the transformation α is the contraction with center at r' and ratio 2^{-n}. It follows that under this contraction the functions $x - x(r')$ and $y - y(r')$ are transformed linearly by some 2×2 matrix with rational coefficients. It gives a quadratic equation for the derivative $y'(x)$ at the point $x(r')$. The value of $y'(x(r))$ can be computed using (3.5.6).

EXERCISE 16. Find $x\left(\frac{5}{6}\right)$, $y\left(\frac{5}{6}\right)$ and the value of $y'(x)$ at $x\left(\frac{5}{6}\right)$.

The next problem is open.

PROBLEM 5. Let $\Gamma \subset \mathbb{R}^2$ be the graph of the function $y(x)$. It contains a big subset X of points with rational coefficients. E.g., all the points which correspond to the rational values of the parameter t belong to X.

It is very interesting to study the closure $\overline{X}_p$ in the p-adic topology (see Info G below).

3.6. Harmonic image of $\mathcal{S}$

In conclusion of the first part of the book we show how Sierpińsky gasket is related to the Apollonian gasket – the main subject of the second part.

Let us introduce a complex harmonic function $z = f_{-1,1}^{i\sqrt{3}}$ on $\mathcal{S}$. The boundary values of this function form an equilateral triangle. The whole image of $\mathcal{S}$ is shown on figure 3.6.

We see that the image of $\mathcal{S}$ under the harmonic map to $\mathbb{C}$ looks as a part of the another famous fractal, the so-called Apollonian gasket. The second part of the book is devoted to the detailed study of Apollonian gaskets from different points of view.

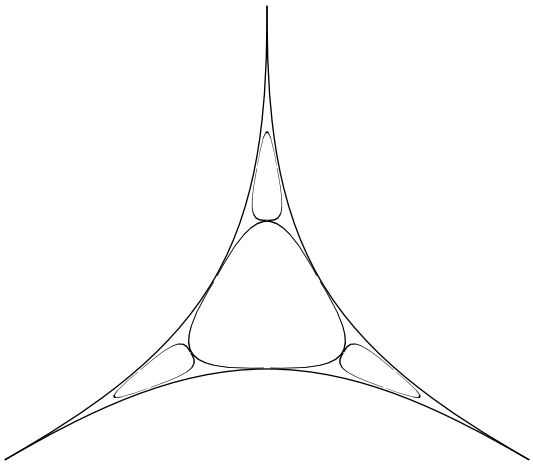

FIGURE 3.6. Harmonic image of $\mathcal{S}$.

The ultimate problem, however, is to explore the similarity of these two sorts of fractals to better understand each of them.

3.7. Multidimensional analogs of $\mathcal{S}$

Sierpiński gasket has natural analogs in higher dimensions. They are self-similar fractal sets in $\mathbb{R}^n$ defined by the system of contractions

(3.7.1)

$$f_i(x) = \frac{x + p_i}{2} \quad \text{where} \quad p_i \in \mathbb{R}^n,\ 1 \le i \le n+1 \quad \text{are not in one hyperplane.}$$

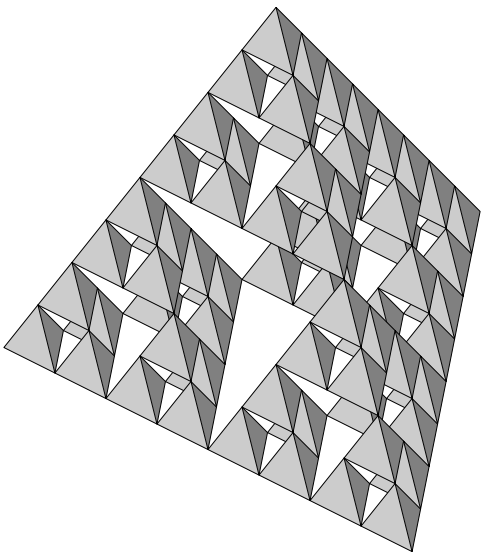

FIGURE 3.7. 3-dimensional Sierpiński gasket

It is not difficult to show that n-dimensional Sierpiński gasket has the Hausdorff dimension $\log_2(n+1)$.

EXERCISE 17. Define a projection of (2^n-1)-dimensional Sierpiński gasket to a n-dimensional plane in such a way that almost all points of the image have a unique preimage.

The theory of harmonic functions on many-dimensional gasket is completely parallel to the theory described above. We mention some facts from this theory. We choose one edge of the initial n-simplex $\{p_1, p_2, \dots, p_{n+1}\}$, say, p_1p_2, identify it with the standard segment $[0, 1]$ and restrict all harmonic functions to this edge.

LEMMA 3.5. *The restriction of a harmonic function f to the edge p_1p_2 depends only on the values $f(p_1)$, $f(p_2)$ and on the sum $\sum_{k=3}^{n+1} f(p_k)$.*

HINT. Use the symmetry of the restriction with respect to permutations of points $p_3, \dots, p_{n+1}$.

Corollary. The restrictions of harmonic functions on $\mathcal{S}$ to any edge p_ip_j form a 3-dimensional space.

Let $f^c_{a,b}$ denote any harmonic function on $\mathcal{S}$ satisfying $f(p_1) = a$, $f(p_2) = b$ and $\sum_{k=3}^{n+1} f(p_k) = c$. The restriction of this function to the segment $[p_1, p_2]$ is a uniquely defined function of the parameter $t \in [0, 1]$. We denote it by $u^c_{a,b}(t)$.

We define basic functions by

$$\chi(t) = u^{-1}_{0,1}(t), \quad \phi(t) = u^0_{0,1}(t), \quad \psi(t) = u^{n-1}_{0,1}(t), \quad \xi(t) = u^n_{0,1}(t) \tag{3.7.2}$$

and the functions x, y by

$$x(t) = u^0_{-1,1}(t), \qquad y(t) = u^1_{0,0}(t). \tag{3.7.3}$$

Then

$$x = \chi + \xi - 1 = \phi + \psi - 1, \qquad y = \phi - \chi = \xi - \psi = \frac{\psi - \phi}{n-1}.$$

Note also, that $u^{n-1}_{1,1}(t) \equiv 1$.

Main relations:

$$\chi(2t) = (n+3)\cdot\chi(t), \qquad \psi(2t) = \frac{n+3}{n+1}\cdot\psi(t); \tag{3.7.4}$$

$$\begin{gathered} \chi(1+\tau) + \chi(1-\tau) = 2 + (n+1)\chi(\tau) \\ \chi(1+\tau) - \chi(1-\tau) = 2\,\frac{n+1}{n}\psi(\tau) + \frac{(n-1)(n+2)}{n}\chi(\tau); \end{gathered} \tag{3.7.5}$$

$$\begin{aligned}\psi(1+\tau)+\psi(1-\tau) &= 2-\frac{n-1}{n+1}\chi(\tau)\\ \psi(1+\tau)-\psi(1-\tau) &= \frac{2}{n}\,\psi(\tau)+\frac{(n-1)(n+2)}{n(n+1)}\chi(\tau).\end{aligned} \tag{3.7.6}$$

These relations allow to develop the arithmetic theory of basic functions for any[3] integer n parallel to the case $n=2$.

In particular, the function $\chi(t)$ always takes integer values at integer points.

Some values of n are of special interest.

When $n=1$, we get $\chi(t)=t^2$, $\quad\phi(t)=\psi(t)=t$, $\quad\xi(t)=2t-t^2$.

When $n=0$, we obtain $y=0$, hence, $\chi(t)=\phi(t)=\psi(t)=\xi(t)$ and this function satisfies the relations

$$\chi(2t)=3\chi(t),\quad \chi(2^m+k)+\chi(2^m-k)=2\cdot 3^m+\chi(k). \tag{3.7.7}$$

To analyze the structure of χ it is useful to introduce the function

$$f(k):=\chi(k+1)-2\chi(k)+\chi(k-1)\quad\text{for any integer}\quad k>0. \tag{3.7.8}$$

THEOREM 3.8. *The function $f(k)$ possesses the properties:*

$$f(2k)=f(k),\qquad f(2^n+k)+f(2^n-k)=f(k)\quad \textit{for}\quad 0<k<2^n. \tag{3.7.9}$$

The detailed investigation of this function is very interesting and I would highly recommend it for an independent study.

For $n=-1$ we have $\chi(t)=t$ and it is not clear how to define other basic functions.

Finally, for $n=-2$, we obtain $\chi(k)=\begin{cases}1 & \text{if}\quad k\not\equiv 0 \mod 3\\ 0 & \text{if}\quad k\equiv 0 \mod 3.\end{cases}$

Similar formulas hold for other basic functions in this case.

We leave to readers to consider other negative values for n and find interesting facts.

Info E. Numerical systems

E.1. Most of real numbers are irrational, so they can not be written as a ratio of two integers. Moreover, real numbers form an uncountable set, therefore, we can not label them by any "words" or "strings" which contain only finite number of digits.

On the other hand there are many numerical systems which allow to write all real numbers using infinite words containing only finite or countable set of digits. The well-known examples are usual decimal and binary systems.

[3] I do not know geometric interpretation of these functions for $n\le 0$ as harmonic functions of some kind.

Recall that a digital numerical system S contains the following data:

• A real or complex base b, $|b| > 1$,

• A set of real or complex digits $D = \{d_1, d_2, \dots\}$ which usually contains the number 0.

To any semi-infinite sequence of the form

$$a = a_n a_{n-1} \cdots a_1 a_0 . a_{-1} a_{-2} \cdots a_{-n} \cdots , \quad a_k \in \mathbb{Z}_+,$$

the system S associates the number

(E.1.1) $\quad val(a) = \sum_{-\infty}^{n} d_{a_k} \cdot b^k$.

In a **standard** numerical system b is a positive integer m and digits are $d_j = j \in X_m = \{0, 1, \dots, m-1\}$. It is well-known that any non-negative real number x can be written in the form

(E.1.2) $\quad x = val(a) = \sum_{-\infty}^{n} a_j \cdot b^j$.

More precisely, every non-negative integer N can be uniquely written as $val(a)$ with the additional condition $a_k = 0$ for $k < 0$.

And any real number from the interval $[0, 1]$ can be almost uniquely written as $val(a)$ with the condition $a_k = 0$ for $k \geq 0$. The non-uniqueness arises from the identity

(E.1.3) $\quad \sum_{k \geq 1} (m-1) \cdot m^{-k} = 1$.

The usual way to avoid this ambiguity is to never use the infinite sequence of the digit $m - 1$.

Motivated by this example, we call for any numerical system S the **whole** numbers those which can be written in the form (E.1.2) with $a_k = 0$ for $k < 0$ and **fractional** number those which can be written in the same form with $a_k = 0$ for $k \geq 0$. The set of whole numbers is denoted by $W(S)$, while the set of fractional numbers – by $F(S)$.

For a standard system S we have $W(S) = \mathbb{Z}_+$, $F(S) = [0, 1]$.

E.2. The non-standard systems are more interesting.

Exercise 18. Consider the system S with the base $b = -2$ and digits $\{0, 1\}$. Check that for this system $W(S) = \mathbb{Z}$ and $F(S) = [-\frac{2}{3}, \frac{1}{3}]$. Show that any real number can be almost uniquely written in the form (E.1.2).

Exercise 19. Introduce a system S with the base $b = 1 + i$ and digits $\{0, 1\}$. Check that here $W(S) = \mathbb{Z}[i]$, the set of so-called **Gaussian integers** of the form $a + ib$, $a, b \in \mathbb{Z}$. As for $F(S)$, it is a fractal compact set of dimension 2, determined by the property

$$F = \frac{1-i}{2} \left(F \bigcup (1 + F) \right). \tag{E.2.1}$$

Here, as always, when an arithmetic operation is applied to a set, it means that it is applied to each element of the set. The picture of this set is shown on the figure E.8 (taken from the cover-sheet of the book [Edg90]).

FIGURE E.8. The set F

EXERCISE 20. Let $\omega = e^{\frac{2\pi i}{3}}$ be the cubic root of 1. Does there exist a system S with a base and digits from $\mathbb{Z}[\omega]$ for which $W(S) = \mathbb{Z}[\omega]$? What is $F(S)$ for such a system?

E.3. There is one more interesting numerical system related to the notion of **continuous fraction**. Let $k = \{k_1, k_2, \dots\}$ be a finite or infinite system of positive integers. We associate to k the number

$$val(k) = \cfrac{1}{k_1 + \cfrac{1}{k_2 + \cfrac{1}{k_3 + \dots + \cfrac{1}{k_n}}}} \quad \text{if the sequence } k \text{ is finite,} \tag{E.1}$$

or the limit of the expression (E.1) where $n \to \infty$ if the sequence k is infinite.

It is well-known that the limit in question always exists. Moreover, every irrational number from (0, 1) is the value of the unique infinite continuous fraction. As for rational numbers from (0, 1), they can be values of two different finite continuous fractions: $k = \{k_1, \dots, k_{n-1}, 1\}$ and $k' = \{k_1, \dots, k_{n-1} + 1\}$.

There is a simple algorithm to reconstruct a sequence k with a given $val(k)$. Namely, denote by $[x]$ the so-called **whole part** of a real number x. By definition, it is a maximal integer $n \le x$. By $\{x\}$ we denote the **fractional** part of x which is $x - [x]$.

Now, for any $x \in (0, 1)$ we define consecutively:

$$x_1 = \frac{1}{x},\ k_1 = [x_1];\ x_2 = \frac{1}{\{x_1\}},\ k_2 = [x_2],\ \dots,\ x_n = \frac{1}{\{x_{n-1}\}},\ k_n = [x_n], \dots$$

For a rational x this process stops when for some n we have $\{x_{n+1}\} = 0$. Then the continuous fraction $k = \{k_1, \dots, k_n\}$ has value x.

For an irrational x the process never stops and we get an infinite continuous fraction k with value x.

Example. Let $k_n = 2$ for all n. Then $x = val(k)$ evidently satisfies the equation $\frac{1}{x} = 2 + x$, hence $x^2 + 2x - 1 = 0$ and $x = -1 \pm \sqrt{2}$. Since $x \in (0, 1)$ we conclude $x = \sqrt{2} - 1$. So, the square root of 2 is given by an infinite continuous fraction:

$$\sqrt{2} \ = \ 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \dots}}}},$$

hence is not a rational number.

This result[4] was known to Pythagoras and kept in secret because it undermined the faith in the power of (rational) numbers.

There are a few cases when the value of an infinite continuous fraction can be expressed in terms of known functions. I know of two such cases.

First, if the fraction in question is **pure periodic**, i.e. when the number k_n depends only of a residue $n \mod m$ for some m, or **mixed periodic**, when this property holds starting with some number n_0.

In this case the number $val(k)$ satisfies a quadratic equation with rational coefficients and can be written explicitly. The converse is also true: any real root of a quadratic equation with rational coefficients (which has the form $\frac{a+\sqrt{b}}{c}$, a, b, $c \in \mathbb{Z}$), can be written in the form of a periodic continuous fraction.

In the second case the sequence $\{k_n\}$ is an arithmetic progression or some modification of it. We only cite three examples

$$\tanh 1 = \frac{e^2-1}{e^2+1} = \cfrac{1}{1+\cfrac{1}{3+\cfrac{1}{5+\cfrac{1}{7+\cfrac{1}{9+\dots}}}}}; \qquad \tanh\frac{1}{2} = \frac{e-1}{e+1} = \cfrac{1}{2+\cfrac{1}{6+\cfrac{1}{10+\cfrac{1}{14+\cfrac{1}{18+\dots}}}}};$$

$$e = 2 + \cfrac{1}{1+\cfrac{1}{2+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{4+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{6+\dots}}}}}}}}$$

E.4. It turns out that all numerical systems described above are particular cases of the following general scheme. Fix a set $D \subset \mathbb{Z}$ of "digits". To any digit $d \in D$ we associate a real or complex $n \times n$ matrix A_d. Choose also a row n-vector f and a column n-vector v.

Then to any semi-infinite sequence of digits $a = \{a_1,\, a_2 \dots\}$ we associate the number

$$val(a) = f \cdot (A_{a_1} A_{a_2} \cdots) \cdot v$$

in the case when the infinite product make sense.

Let us explain the relation to previously described numerical systems. Let $A_a = \begin{pmatrix} m & 0 \\ a & 1 \end{pmatrix}$, $0 \le a \le m-1$. Then

$$A_{a_n} \cdots A_{a_1} A_{a_0} = \begin{pmatrix} m^{n+1} & 0 \\ \sum_{j=0}^{n} a_j m^j & 1 \end{pmatrix}.$$

So, if we put $f = (0,\, 1)$, $v = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, we get

E.4.1 $val(a_0, a_1 \dots, a_n) = a_0 + a_1 m + \cdots + a_n m^n = f \cdot A_{a_0} A_{a_1} \cdots A_{a_n} \cdot v.$

[4]More precisely, its geometric interpretation, showing that the diagonal of a square is not commensurable with its side.

Let now $A_k = \begin{pmatrix} k & 1 \\ 1 & 0 \end{pmatrix}$. Consider the matrices:

$$A_k = \begin{pmatrix} k & 1 \\ 1 & 0 \end{pmatrix}, \quad A_k A_l = \begin{pmatrix} kl+1 & k \\ l & 1 \end{pmatrix}, \quad A_k A_l A_m = \begin{pmatrix} klm+m+k & kl+1 \\ lm+1 & l \end{pmatrix}$$

and compare them with continuous fractions:

$$\frac{1}{k}\ ; \qquad \frac{1}{k+\frac{1}{l}} = \frac{l}{kl+1}; \qquad \frac{1}{k+\frac{1}{1+\frac{1}{m}}} = \frac{lm+1}{klm+m+k}.$$

This comparison suggests the general identity:

LEMMA E.6. *The value of a continuous fraction can be computed by the formula:*

$$val(k) = \cfrac{1}{k_1 + \cfrac{1}{k_2 + \cfrac{1}{k_3 + \cdots + \cfrac{1}{k_n}}}} = \frac{(A_{k_1} \cdot A_{k_2} \cdots A_{k_n})_{21}}{(A_{k_1} \cdot A_{k_2} \cdots A_{k_n})_{11}}. \tag{E.2}$$

CHAPTER 4

Applications of generalized numerical systems

4.1. Application to the Sierpiński gasket

First, let us try to label the points of $\mathcal{S}$. Consider the alphabet with 3 digits: $-1,\ 0,\ 1$. To any finite word $a = a_1 a_2 \dots a_n$ in this alphabet we associate the complex number

$$val(a) = \frac{\epsilon^{a_1}}{2} + \frac{\epsilon^{a_2}}{4} + \dots + \frac{\epsilon^{a_n}}{2^n} \quad \text{where} \quad \epsilon = e^{2\pi i/3}.$$

We also associate the number 0 to the empty sequence.

It is easy to understand that the numbers $val(a)$ for all 3^n sequences of length n situated in the centers of the 3^n triangles of rank $n-1$, complementary to $\mathcal{S}$.

Exercise 21. For any infinite sequence a let us denote $a^{(n)}$ the sequence of first n digits of a. Show that

a) the sequence $val(a^{(n)})$ has a limit when $n \to \infty$. We denote this limit as $val(a)$;

b) the point $val(a)$ belongs to $\mathcal{S}$;

c) $val(a) = val(b)$ iff one sequence can be obtained from another by substituting the tail of the form $xyyyy\dots$ by the tail $yxxxx\dots$.

Exercise 22. Which infinite sequences correspond

a) to boundary points? b) to points of segments joining the boundary points?

c) to vertices of $\mathcal{S}_n$? d) to segments, joining the vertices of $\mathcal{S}_n$?

4.2. Application to the question mark function

The so-called **question mark function** is a function defined by Minkowski in 1904 for the purpose of mapping the quadratic irrational numbers in the open interval (0, 1) into rational numbers of (0, 1) in a continuous, order-preserving manner. Later, in 1938, this function was introduced by A. Denjoy for arbitrary real numbers.

By definition,[1] the function $?(\cdot)$ takes a number a represented by a continued fraction

$$a = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots + \cfrac{1}{a_k} + \ldots}}}$$

to the number

$$?(a) \;:=\; \sum_k \frac{(-1)^{k-1}}{2^{a_1+\cdots+a_k-1}} \;=\; 0.\overbrace{0\ldots0}^{a_1}\overbrace{1\ldots1}^{a_2}\overbrace{0\ldots0}^{a_3}\ldots$$

For example, $?\left(\frac{\sqrt{2}}{2}\right) = 0.11001100\ldots = \frac{4}{5}$, $?\left(\frac{e^2-1}{e^2+1}\right) = \sum_{k\geq 0} 2^{-k^2}$.

We shall say more about this function in the second part of the book. Here we only observe that this is one more example of a function which is naturally defined using generalized numerical systems.

[1]Better to say: By one of possible definitions (see below).

Part 2

Apollonian Gasket

Introduction

In this part of the book we consider another remarkable fractal: a so-called Apollonian gasket $\mathcal{A}$. It seems rather different from the Sierpiński gasket $\mathcal{S}$. For example, it is not a self-similar fractal, though for any $k \geq 0$ it can be represented as a union of $3k + 2$ subsets homeomorphic to $\mathcal{S}$.

Nevertheless, there are deep and beautiful relations between both fractals and our goal, only partly achieved here, is to reveal these relations.

Many of facts discussed below are of elementary geometric nature. However, in modern educational programs the Euclidean geometry occupies a very small place and we can not rely on the information acquired at school. Therefore, sometimes we use more sophisticated tools to get the desired results.

As in the first part, we study our gasket from different points of view: geometric, group-theoretic and number theoretic. The interplay of all three approaches makes the subject very interesting and promising.

CHAPTER 5

Apollonian gasket

5.1. Descartes' theorem

We start with a simply looking geometric problem:

Describe all configurations of four pairwise tangent circles on a plane

Examples of such configurations are shown below on Fig. 5.1. We include the cases when one of the circles degenerates to a straight line (a circle with an infinite radius) and the case when one of the circles is tangent to others from inside (we shall interpret it later as a circle with a negative radius).

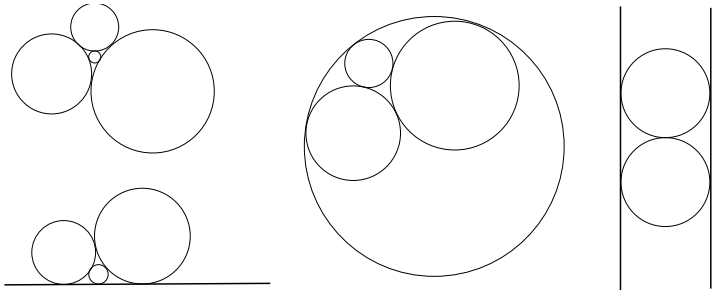

FIGURE 5.1. Quadruples of tangent circles

There exist some other configurations which we want to exclude. They are shown on Fig. 5.2. Here all four circles have a common tangency point, finite or infinite. The reason why these configurations are excluded will be clear when we make the formulation more precise and pass from circles to discs.

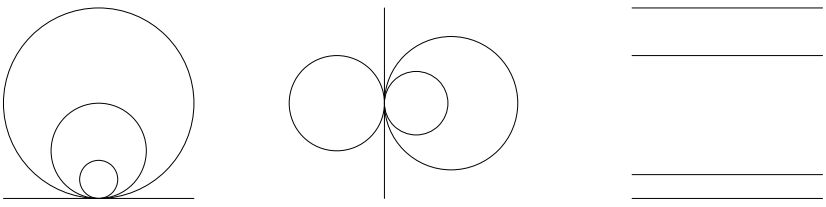

FIGURE 5.2. "Wrong quadruples"

It turns out that the complete and clear solution of this problem uses tools from several different domains in mathematics. Moreover, the problem has natural many-dimensional analogues and requires a more precise and slightly modified formulation. Here we outline an elementary approach which already show us the necessity of refinements and modifications.

To approach our problem, make one step back and consider a triple of pairwise tangent circles. There are three kinds of such triples – see Fig.**??**.

Note, that the triangle formed by the points of tangency is acute in the case a), right in the case b) and obtuse in the case c).

In the case a) it is rather obvious that our three circles can have arbitrary positive radii r_1, r_2, r_3. Indeed, let O_1, O_2, O_3 be the centers of circles in question. We can always construct the triangle $O_1O_2O_3$ since its sides are known: $|O_iO_j| \ = \ r_i + r_j$ and satisfy the triangle inequality:

$$|O_iO_j| + |O_jO_k| = (r_i + r_j) + (r_j + r_k) \geq r_i + r_k = |O_iO_k|. \tag{5.1.1}$$

In the case c) we have $|O_1O_2| = r_1+r_2$, $|O_2O_3| = r_3-r_2$, $|O_3O_1| = r_3-r_1$ and $r_1 + r_2 \leq r_3$. There is a way to unite a) and c) in a general formula

$$|O_iO_j| \ = \ |r_i + r_j|. \tag{5.1.2}$$

For this we have only to replace r_3 by $-r_3$. Then (5.1.2) will be satisfied if $r_1 + r_2 \leq |r_3|$, or $r_1 + r_2 + r_3 \leq 0$.

In the case b) the center O_3 is situated at infinity. We put $r_3 = \infty$ and (5.1.2), suitably interpreted, is still satisfied.

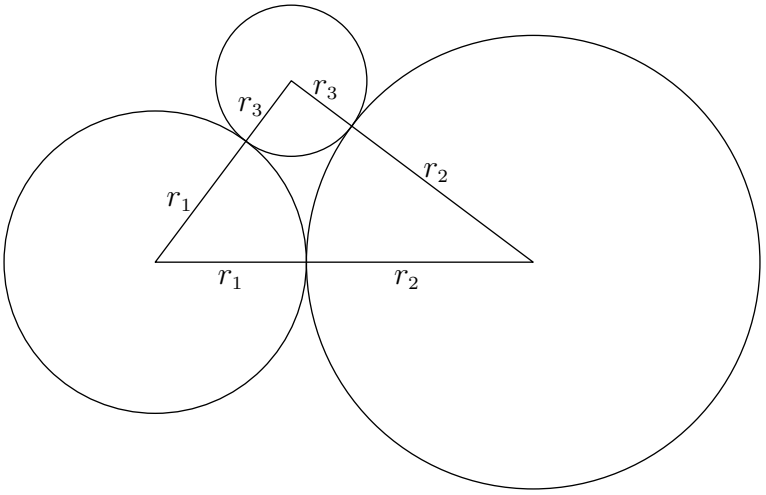

FIGURE 5.3. Triples of tangent circles a)

If four circles are pairwise tangent, their radii r_1, r_2, r_3, r_4 are not arbitrary but must satisfy some equation. This equation and/or some of its

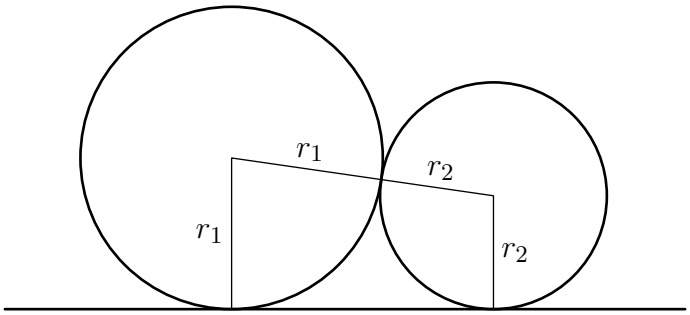

FIGURE 5.4. Triples of tangent circles b)

consequences were apparently known in Ancient Greece more than two thousand years ago.

More recently, the condition was explicitly written by René Descartes, the famous French mathematician and philosopher of the first half of 17-th century.

The Descartes equation looks simpler if we replace the radii r_i by the inverse quantities

$$c_i := r_i^{-1}, \qquad 1 \leq i \leq 4.$$

The geometric meaning of the quantity c_i is the curvature of the circle with the radius r_i.[1]

The equation in question looks as follows:

$$(c_1 + c_2 + c_3 + c_4)^2 \ - \ 2(c_1^2 + c_2^2 + c_3^2 + c_4^2) \ = \ 0. \tag{5.1.3}$$

[1]The reason, why curvatures are better than radii, will be explained later, when we develop a group-theoretic approach to the problem.

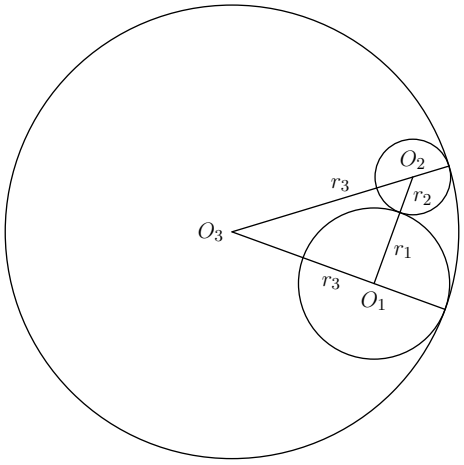

FIGURE 5.5. Triples of tangent circles c)

We leave to geometry fans the challenge to recover the proof of the Descartes theorem using the high school geometry. The following exercise and the Fig. 5.6 can help.

EXERCISE 23. Find the common formula for the area of the triangle $O_1O_2O_3$ above which is true for cases a) and c).

HINT. Use the Heron formula.

ANSWER. $S = \sqrt{r_1r_2r_3(r_1+r_2+r_3)}$. Not, that the expression under the root sign is always positive.

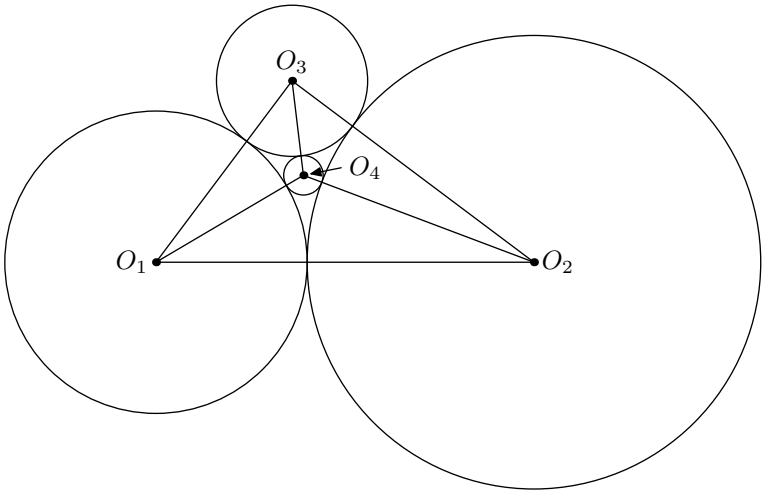

FIGURE 5.6. Towards the proof of Descartes theorem

There is a special case of the Descartes theorem which is much easier to prove. Namely, assume that one of the four circles degenerates to a

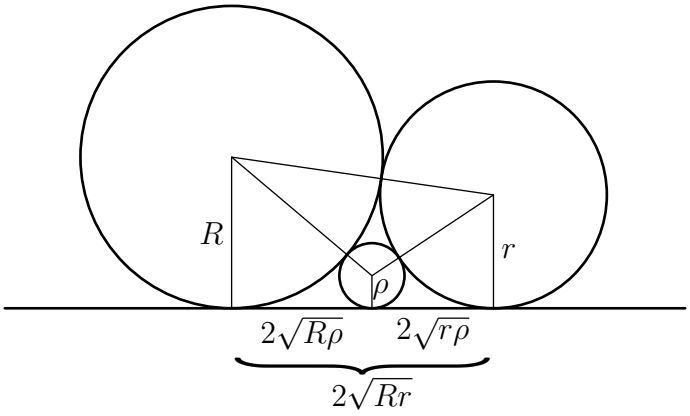

FIGURE 5.7. Degenerate Descartes equation

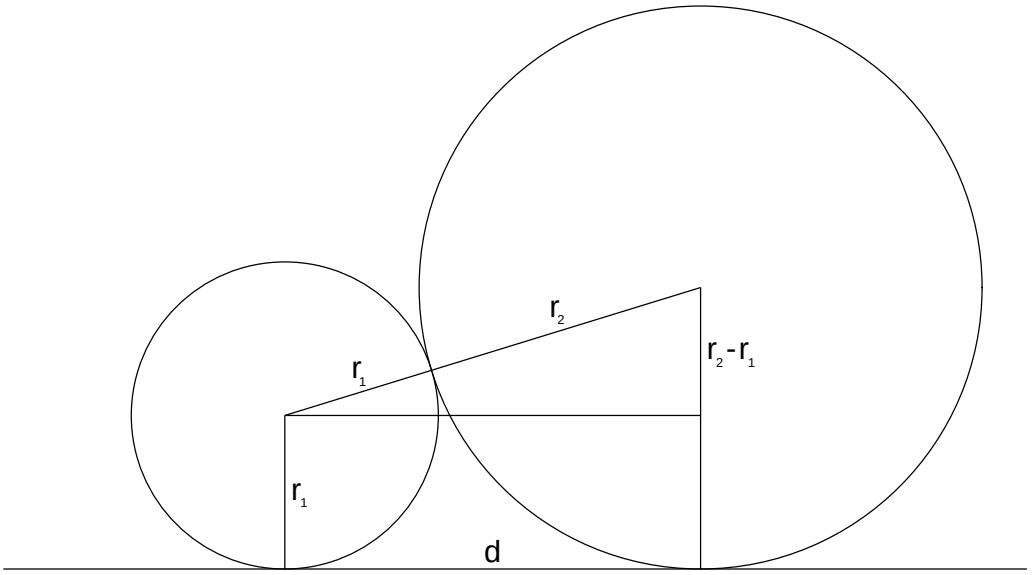

FIGURE 5.8. Degenerated triple with $d = 2\sqrt{r_1 r_2}$

straight line. Let, for example, $c_4 = 0$ so that the relation between remaining curvatures is:

$$(c_1 + c_2 + c_3)^2 - 2(c_1^2 + c_2^2 + c_3^2) = 0. \tag{5.1.4}$$

Fortunately, the left hand side of (5.1.4) can be decomposed into simple factors. For this end we rewrite it in the form of quadratic polynomial in c_1:

$$-c_1^2 + 2c_1(c_2 + c_3) - c_2^2 + 2c_2c_3 - c_3^2$$

This quadratic polynomial has the roots $c_2+c_3\pm 2\sqrt{c_2c_3} = (\sqrt{c_2}\pm\sqrt{c_3})^2$. Therefore, it can be written as

$$-\left(c_1 - (\sqrt{c_2} + \sqrt{c_3})^2\right)\left(c_1 - (\sqrt{c_2} - \sqrt{c_3})^2\right) =$$
$$(\sqrt{c_1} + \sqrt{c_2} + \sqrt{c_3})(-\sqrt{c_1} + \sqrt{c_2} + \sqrt{c_3})(\sqrt{c_1} - \sqrt{c_2} + \sqrt{c_3})(\sqrt{c_1} + \sqrt{c_2} - \sqrt{c_3}).$$

It follows that (5.1.4) is true iff at least one of the following equations are satisfied:

$$\sqrt{c_1} \pm \sqrt{c_2} \pm \sqrt{c_3} = 0, \quad \text{or} \quad \sqrt{r_2r_3} \pm \sqrt{r_1r_2} \pm \sqrt{r_1r_3} = 0. \tag{5.1.5}$$

Actually, the signs depend on the relative sizes of radii. E.g., when $r_1 \geq r_2 \geq r_3$, we have $\sqrt{r_1r_2} = \sqrt{r_2r_3} + \sqrt{r_3r_1}$. You can easily verify this relation using figures 5.7 and 5.8.

In the next section we give the proof of more general result, using the matrix algebra and the geometry of Minkowski space. But before doing it we have to correct one inaccuracy in the previous exposition.

Namely, we did not take into account the sign of the curvature which may make the formula (5.1.3) incorrect. Indeed, let us check the equality (5.1.3) in the case shown on Fig. 5.9 below.

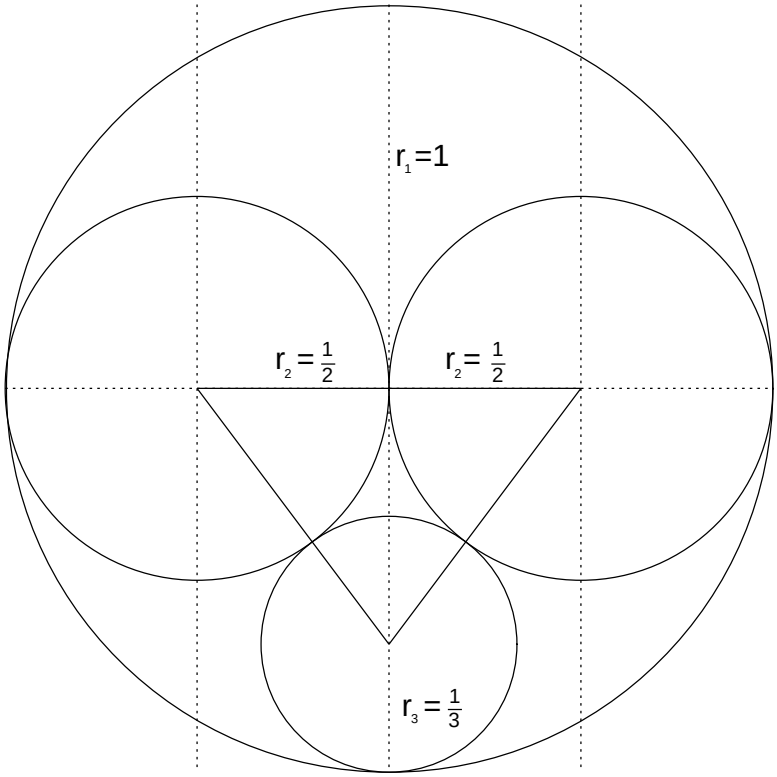

FIGURE 5.9. "Violation" of Descartes equation

If we take $c_1 = 1$, $c_2 = c_3 = 2$, $c_4 = 3$, we get the wrong equality

$$64 = (1 + 2 + 2 + 3)^2 = 2(1 + 4 + 4 + 9) = 36.$$

But if we put the value of c_1 equal to -1, then we get the correct equality

$$36 = (-1 + 2 + 2 + 3)^2 = 2(1 + 4 + 4 + 9) = 36.$$

Looking on the picture, we see that the circle of radius 1 is in a special position: the other circles touch it from inside. We have already seen, that in this case it is convenient to interpret it as a circle with negative radius -1.

To make the exposition rigorous, we need either introduce an orientation on our circles, or consider instead of circles the solid discs bounded by them.

The both possibilities are in fact equivalent. Indeed, any disc inherits an orientation from the ambient plane or sphere. And the boundary of an oriented disc has a canonical orientation. In our case it can be defined by a

simple "left hand rule": when we go along the circle in the positive direction, the surrounded domain must remain on the left.

In particular, the outer circle on Fig.6.3 bounds the domain which is complementary to the unit disc. So, we are forced to include the domains of this sort in the consideration.

Also, it seems natural to complete the plane $\mathbb{R}^2$ by an infinite point ∞. The new set $\overline{\mathbb{R}}^2$ can be identified with two-dimensional sphere S^2 using the stereographic projection (see Info F). Under this identification the "generalized discs" go the ordinary discs on S^2 which contain the North pole inside. Those discs which contain the North pole as a boundary point, correspond to half-planes in $\overline{\mathbb{R}}^2$.

So, we have determined our main object of study. It is the set $\mathcal{D}$ of discs on two-dimensional sphere S^2. To each disc $D \in \mathcal{D}$ there corresponds an oriented circle $C = \partial D$.

We can also identify $S^2 \simeq \overline{\mathbb{R}}^2$ with the extended complex plane $\overline{\mathbb{C}}$ and consider our discs and circles as subsets of $\overline{\mathbb{C}}$.

Let us say that two discs are **tangent** if they have exactly one common point. In terms of oriented circles it means a negative tangency, because the orientations of the two circles at the common point are opposite.

Now it is clear, why we excluded the configurations shown on Fig. 5.2: they do not correspond to a configuration of four pairwise tangent discs.

Let now C be an oriented circle of (ordinary) radius r on $\overline{\mathbb{C}}$. We say that C has the curvature $c = r^{-1}$ if C bounds an ordinary disc, the curvature $c = -r^{-1}$ if it is the boundary of a complement to a disc, and the curvature 0 if our circle is actually a straight line.

In particular, the outer circle on Fig. 6.3 corresponds to the complement to the open unit disc. Therefore, the curvature of the boundary is -1.

REMARK 3. Let us look in more details on the signs of numbers $\{c_i\}_{1\le i\le 4}$ which satisfy equation (5.1.3). Note first, that if the quadruple $(c_1,\, c_2,\, c_3,\, c_4)$ is a solution to (5.1.3), then so is $(-c_1,\, -c_2,\, -c_3,\, -c_4)$. But these two solutions are never realized simultaneously as quadruples of curvatures for tangent discs.

Further, the equation (5.1.3) can be written in the form

$$2(c_1 + c_2)(c_3 + c_4) = (c_1 - c_2)^2 + (c_3 - c_4)^2. \tag{5.1.6}$$

We see that either $c_1 + c_2 \ge 0$ and $c_3 + c_4 \ge 0$, or $c_1 + c_2 \le 0$ and $c_3 + c_4 \le 0$. Suppose that numeration is chosen so that $c_1 \ge c_2 \ge c_3 \ge c_4$. Then in the first case we have $|c_4| \le c_3 \le c_2 \le c_1$; in the second one $c_4 \le c_3 \le c_2 \le -|c_1|$.

Only in the first case our solution can be interpreted as a set of curvatures of four pairwise tangent discs. So, only this case will be considered below. Because of the note above, we do not loose any information about solutions to (5.1.3).

Thus, from now on we can assume that either

a) all numbers c_i are positive, or

b) three numbers are positive, the forth is negative and by absolute value smaller that others, or

c) three numbers are positive and the forth one is 0, or, finally,

d) two of c_i are positive and equal each other while other two are zeros.

It reflects the evident geometric fact: among four pairwise tangent discs at most two are unbounded.

♡

Info F. Conformal group and stereographic projection

F.1. In our exposition we consider the general n-dimensional case. But all arguments and computations are practically the same in all dimensions. So, the reader, not acquainted with the subject, can start with the case $n = 1$.

Let $\overline{\mathbb{R}}^n$ denote the set which arises from $\mathbb{R}^n$ by adding an infinite point ∞. There is a remarkable 1-1 correspondence between n-dimensional sphere S^n and $\overline{\mathbb{R}}^n$. This correspondence is called **stereographic projection**. Here we give its definition and list the main properties.

Let $\mathbb{R}^{n+1}$ be an Euclidean space with coordinates $(\alpha_0, \alpha_1, \dots, \alpha_n)$. The unit sphere $S^n \subset \mathbb{R}^{n+1}$ is given by the equation $\alpha_0^2 + \alpha_1^2 + \dots \alpha_n^2 = 1$. The point $P = (1, 0, 0, \dots, 0) \in S^n$ we call the North pole.

Let $\mathbb{R}^n$ be another Euclidean space with coordinates $(x_1, x_2, \dots, x_n)$. It is convenient to think of $\mathbb{R}^n$ as of subspace in $\mathbb{R}^{n+1}$ consisting of points with coordinates $(0, x_1, \dots, x_n)$.

Define a map s from $S^n \backslash P$ to $\mathbb{R}^n$ by the formula:

$$s(\alpha) = \left(0, \quad \frac{\alpha_1}{1-\alpha_0}, \quad \frac{\alpha_2}{1-\alpha_0}, \quad \dots, \quad \frac{\alpha_n}{1-\alpha_0}\right). \tag{F.1}$$

The inverse map has the form

$$s^{-1}(x) = \left(\frac{|x|^2-1}{|x|^2+1}, \quad \frac{2x_1}{1+|x|^2}, \quad \frac{2x_2}{1+|x|^2}, \quad \dots, \frac{2x_n}{1+|x|^2}\right) \tag{F.2}$$

where $|x|^2 = x_1^2 + x_2^2 + \dots + x_n^2$.

Exercise 24. Check that three points P, $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_n)$ and $s(\alpha) = \big(0, x_1(\alpha), x_2(\alpha), \dots, x_n(\alpha)\big)$ belong to one line in $\mathbb{R}^{n+1}$.

So, our map s geometrically is a projection of $S^n \backslash P$ from the point P to the coordinate plane $\mathbb{R}^n \in \mathbb{R}^{n+1}$ given by the equation $\alpha_0 = 0$.

Both algebraic and geometric definition of s do not make sense at the point P. We assume additionally that $s(P) = \infty \in \overline{\mathbb{R}}^n$. The so defined map s is a bijection between S^n and $\mathbb{R}^{n+1}$.

In conclusion of this Section we show that the stereographic projection indeed sends discs to discs. In general, a disc $D \in S^2$ is defined as an intersection of S^2 with a half-space given in coordinates α_0, α_1, ..., α_n by the linear inequality

$$p_0\alpha_0 + p_1\alpha_1 + \cdots + p_n\alpha_n + p_{n+1} \leq 0. \tag{F.3}$$

Note, that the hyperplane $p_0\alpha_0 + p_1\alpha_1 + \cdots + p_n\alpha_n + p_{n+1} = 0$ intersect the sphere S^2 along a non-trivial circle if and only if

$$p_{n+1}^2 - p_0^2 - p_1^2 - \cdots - p_n^2 \ < \ 0. \tag{F.4}$$

Therefore, it is natural to consider vector p as an element of $\mathbb{R}^{1,n+1}$ and denote the left hand side of F.4 by $|p|^2$.

Since the multiplication of p by a positive constant does not change the meaning of the inequality F.3, we can normalize p by the condition[2] $|p|^2 = -1$.

Expressing $\{\alpha_i\}$ in terms of the coordinates $\{x_j\}$ of the point $s(\alpha)$, we get the inequality defining $s(D)$ in the form

$$p_0(|x|^2 - 1) + 2p_1x_1 + \cdots + 2p_nx_n + p_{n+1}(|x|^2 + 1) \leq 0.$$

It can be rewritten in the form

$$a + (\vec{p}, \vec{x}) + c|\vec{x}|^2 \leq 0. \tag{F.5}$$

where $a = p_{n+1} - p_0$, $\quad c = p_{n+1} + p_0$, $\quad \vec{p} = (p_1, \ldots, p_n)$ and $\vec{x} = (x_1, \ldots, x_n)$.

Now we can use the equation $|p|^2 = ac - |\vec{p}|^2$ and the normalization $|p|^2 = -1$ to write our inequality as follows: It can be rewritten in the form

$$c \cdot \left|x + \frac{\vec{p}}{c}\right|^2 \ \leq \ c^{-1}. \tag{F.6}$$

The last inequality for $c > 0$ describes a disc on $\mathbb{R}^n$ with the center $-\frac{\vec{p}}{c}$ and the curvature c.

If $c < 0$, then F.6 describe the complement to a disc with the center $-\frac{\vec{p}}{c}$ and radius $-\frac{1}{c}$. We agree to associate to this generalized disc the negative curvature c.

Finally, if $c = 0$, then F.6 does not make sense and F.5 defines a half-space (the initial disc D in this case contains the north pole as a boundary point).

[2]Do not confuse $p \in \mathbb{R}^{1,n+1}$ with $\vec{p} \in \mathbb{R}^n$ introduced below.

F.2. There is a big group $Conf_n$ (or C_n for short) of **conformal maps** which acts on S^n and on $\overline{R}^n$ so that stereographic projection is a so-called **C_n-covariant map**, i.e. the following diagram is commutative:[3]

$$\begin{array}{ccc} S^n & \xrightarrow{s} & \overline{R}^n \\ {\scriptstyle g\cdot}\downarrow & & \downarrow{\scriptstyle g\cdot} \\ S^n & \xrightarrow[s]{} & \overline{R}^n \end{array} \tag{F.7}$$

where $g\cdot$ means the action of $g \in C_n$ and s is the stereographic projection.

The group C_n can be defined in several ways. We give here three equivalent definitions.

Geometric definition (for $n > 1$). We recall that a smooth map $g : \mathbb{R}^n \to \mathbb{R}^n$ has at any point $x \in \mathbb{R}^n$ a derivative $g'(x)$ which is a linear operator $g'(x) : \mathbb{R}^n \to \mathbb{R}^n$. We say that g is **conformal** if its derivative $g'(x)$ at any point x is a composition of rotation and dilation.

So, infinitesimally, conformal transformations preserve the form of figures. This explains the name "conformal".

For $n = 1$ the group of conformal maps in this sense is too big (infinite dimensional). Namely, it is the group of all smooth transformations of S^1.

Note, that for $n = 2$ the group of all holomorphic (or, complex-analytic) transformations of $\mathbb{C}$ is also infinite-dimensional. But for $\overline{\mathbb{C}}$ the situation is different: every conformal transformation of $\overline{\mathbb{C}}$ is fractional-linear (see below).

We now give another, group-theoretic, definition which for $n > 1$ is equivalent to the geometric one, but defines a finite dimensional Lie group in all dimensions.

Let $E_n(\mathbb{R})$ be the group of all rigid motions of $\mathbb{R}^n$, extended naturally on $\overline{R}^n$ so that the infinite point is fixed.

Let us call **inversion**, or else, **reflection in the unit sphere**, the map $Inv : \overline{R}^n \to \overline{R}^n$ given by the formula

$$Inv(x) = \begin{cases} \frac{x}{|x|^2}, & \text{if} \quad x \neq 0, \infty \\ \infty, & \text{if} \quad x = 0 \\ 0, & \text{if} \quad x = \infty. \end{cases} \tag{F.8}$$

Group-theoretic definition. We define first the **extended conformal group** $\overline{Conf}_n$ as the group generated by $E_n(\mathbb{R})$ and Inv.

[3]We say that a diagram consisting of sets and maps is **commutative**, if for any path composed from the arrows of the diagram the composition of corresponding maps depends only on the start and the end points of the path. In the case in question it means that $s \circ g = g \circ s$.

The group $\overline{Conf}_n$ consists of two connected components. The transformations from one component preserve the orientation of $\mathbb{R}^n$. This component contains $E_n(\mathbb{R})$ and all products of type $g_1 \circ Inv \circ g_2 \circ Inv \circ \cdots \circ g_{2n} \circ Inv$ with even number of involutions. It is itself a group and this is our **conformal group** $Conf_n$

The other component contains Inv and all products of the form $g_1 \circ Inv \circ g_2 \circ Inv \circ \cdots \circ g_{2n+1} \circ Inv$ with odd number of involutions. These transformations reverse the orientation of $\mathbb{R}^n$. They are called sometimes **conformal transformations of the second kind**.

Finally, the most working definition is the following, which we shall call **matrix definition**.

Let $\mathbb{R}^{1,n+1}$ denote the real vector space with coordinates $(x_0, x_1, \ldots, x_{n+1})$ endowed with the symmetric bilinear form $B(x, y) = x_0y_0 - x_1y_1 - \ldots - x_{n+1}y_{n+1}$. The group of linear transformations which preserve this form is called **pseudo-orthogonal** group and is denoted by $O(1, n+1; \mathbb{R})$. In the chosen basis the elements of the group are given by block-matrices of the form $\begin{pmatrix} a & \vec{b}^{\,t} \\ \vec{c} & D \end{pmatrix}$ where a is a real number, $\vec{b}^{\,t}$ is a row $(n+1)$-vector, $\vec{c}$ is a column $(n+1)$-vector and D is a $(n+1) \times (n+1)$ matrix, satisfying the relations

$$a^2 = 1 + |\vec{c}|^2, \qquad D^tD = 1_{n+1} + \vec{b}\,\vec{b}^{\,t}, \qquad D^t\vec{c} = \vec{b}a \tag{F.9}$$

where 1_{n+1} is the unit matrix of order $n+1$.

From F.9 we see that a and D are invertible (check that $D^{-1} = 1_{n+1} - \frac{\vec{b}\,\vec{b}^{\,t}}{1+|b|^2}$). So, the group $O(1, n+1; \mathbb{R})$ splits into four parts according to the signs of a and $\det D$. Actually, these parts are connected components of the group. More precisely, our group as a smooth manifold is diffeomorphic to the product $O(n, \mathbb{R}) \times S^n \times \mathbb{R} \times \mathbb{Z}_2$: to each quadruple $(A, \vec{v}, \tau, \pm 1)$ there correspond the element

$$g = \pm \begin{pmatrix} \cosh\tau & \vec{v}^{\,t} \cdot \sinh\tau \\ \sinh\tau \cdot A\vec{v} & \cosh\tau \cdot A \end{pmatrix} \in O(1, n+1; \mathbb{R}) \tag{F.10}$$

and, conversely, any matrix from $O(1, n+1; \mathbb{R})$ has this form.

Our group acts on the space $\mathbb{R}^{1,n+1}$, preserving the cone

$$C: \; x_0^2 = x_1^2 + \cdots + x_{n+1}^2.$$

It acts also on the projective space associated with $\mathbb{R}^{1,n+1}$. Since scalar matrices act trivially, we have actually the action of the corresponding projective group $PO(1, n+1; \mathbb{R}) = O(1, n+1; \mathbb{R})/\{\pm 1\}$ which is the quotient of $O(1, n+1; \mathbb{R})$ over its center $\{\pm 1_{n+2}\}$. This group has two connected components $PO_{\pm}(1, n+1; \mathbb{R})$ distinguished by the sign of $\det(a^{-1}D)$.

The projectivization of the cone C (with the origin deleted) can be identified with S^n via the coordinates $\alpha_i = \frac{x_i}{x_0}$, $1 \le i \le n+1$ and with $\overline{\mathbb{R}}^{\,n}$ via coordinates $w_j = \frac{x_j}{x_0 - x_{n+1}}$, $1 \le j \le n$.

Conversely, the coordinate on the cone C can be restored up to proportionality by equations:

(F.11)
$$x_0 = \frac{\sum_j w_j^2 + 1}{2}, \quad x_j = w_j \quad \text{for} \quad 1 \le j \le n, \quad x_{n+1} = \frac{\sum_j w_j^2 - 1}{2}.$$

The following fact is well-known and we use it for the matrix definition of conformal group.

THEOREM F.1. *The group $PO(1,\, n+1;\, \mathbb{R})$ acting on S^n (or on $\overline{\mathbb{R}}^n$) coincides with $\overline{Conf}_n$ while its connected subgroup $PO_+(1,\, n+1;\, \mathbb{R})$ coincides with $Conf_n$.*

EXERCISE 25. The inversion $Inv \in \overline{Conf}_n$ in the last realization corresponds to some element of $PO(1,\, n+1;\, \mathbb{R})$, i.e. to a pair of matrices $\pm g \in O(1,\, n+1;\, \mathbb{R})$. Find these matrices.

HINT. Use F.11.

ANSWER.
$$g = diag\ (1,\, 1,\, \ldots,\, 1,\, -1).$$

F.3. In the main text we shall consider in more details the case $n = 2$ and also the cases $n = 3$, $n = 4$. In all these cases the conformal group $Conf_n$ has additional properties which we discuss here.

Case $n = 2$. The group $Conf_2$ is isomorphic to $PO_+(1,\, 3;\, \mathbb{R})$ and also to the Möbius group $PSL(2,\, \mathbb{C})$. The group $\overline{Conf_2}$ is isomorphic to $PO(1,\, 3;\, \mathbb{R})$ and also to the extended Möbius group. Recall that the Möbius group acts on $\overline{\mathbb{C}}$ by so-called fraction-linear (or Möbius) transformations

(F.12)
$$w \to \frac{\alpha\, w + \beta}{\gamma\, w + \delta} \qquad \text{where} \quad \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in SL(2,\, \mathbb{C}).$$

The extended Möbius group besides these transformations contains also the complex conjugation, hence all transformation of the form

(F.13)
$$w \to \frac{\alpha\, \overline{w} + \beta}{\gamma\, \overline{w} + \delta} \qquad \text{where} \quad \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in SL(2,\, \mathbb{C}).$$

Among these transformations there are the so-called **reflections** s which satisfy the equation $s^2 = 1$ and for which the set of fixed points is a circle or a straight line. We denote the set of fixed points by M_s and call it a **mirror**. Conversely, there is a unique reflection with given mirror M; we denote it by s_M.

If the circle M degenerates to a straight line l, the transformation s_M is an ordinary reflection in l. For a unit circle M_0, centered at the origin, the reflection s_{M_0} coincides with the inversion Inv defined by (F.2.2). In general, s_M can be defined as $g \circ Inv \circ g^{-1}$ where $g \in Conf_2$ is any transformation which sends C to M.

EXERCISE 26. Show that all reflections form a single conjugacy class in $\overline{Conf}_2$.

HINT. Show that $\overline{Conf}_2$ acts transitively on $\mathcal{D}$.

EXERCISE 27. Show that the group $\overline{Conf}_2$ is generated by reflections.

HINT. Use the well-known fact that $SL(2, \mathbb{C})$ is generated by elements

$$g(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \, t \in \mathbb{C}, \quad \text{and} \quad s = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{F.14}$$

EXERCISE 28. Show that the conjugacy classes in $Conf_2$ are precisely the level sets $I(g) = const$ for the function

$$I(g) \ := \ \frac{(\operatorname{tr} g)^2}{\det g} - 4 \tag{F.15}$$

with one exception: the set $I(g) = 0$ is the union of two classes: $\{e\}$ and the class of a Jordan block.

EXERCISE 29. Show that all involutions in $Conf_2$ form two conjugacy classes: the unit class and the class which contains a rotation of S^2 on 180° around z-axis.

EXERCISE 30. Show that all involutions in $PO_-(1, 3; \mathbb{R})$ which are not reflections, form a single conjugacy class with a representative acting as the antipodal map on S^2.

We quote two main properties of the group $G = Conf_2$.

PROPOSITION F.1. *For every two triples of different points (z_1, z_2, z_3) and (w_1, w_2, w_3) on $\overline{\mathbb{C}}$ there exists a unique transformation $g \in G$ such that $g(z_i) = w_i, \ i = 1, 2, 3$.*

PROOF. First check it when $w_1 = 0$, $w_2 = 1$, $w_3 = \infty$. The corresponding transformation g_{z_1,z_2,z_3} can be written explicitly:

$$g_{z_1,z_2,z_3}(z) = \frac{z - z_1}{z - z_3} \ : \ \frac{z_2 - z_1}{z_2 - z_3} \tag{F.16}$$

The transformation g which we want is $g = g^{-1}_{w_1,w_2,w_3} \circ g_{z_1,z_2,z_3}$. □

PROPOSITION F.2. *Any circle or a straight line goes under transformations $g \in G$ to a circle or a straight line. (Or else: any disc goes to a disc).*

To prove this statement we use the following

LEMMA F.1. *Let a, c be two real numbers and b be a complex number such that $ac - |b|^2 \ < \ 0$. Then the inequality*

$$a + \bar{b}w + b\bar{w} + cw\bar{w} \ \leq \ 0 \tag{F.17}$$

describes a disc $D \in \mathcal{D}$. More precisely, it is

a) a closed disc with the radius $r = c^{-1}$ and the center $-\frac{b}{c}$, when $c > 0$;

b) a complement of an open disc with the radius $r = -c^{-1}$ and the center $-\frac{b}{c}$, when $c < 0$;

c) a closed half-plane when $c = 0$.

Moreover, any disc $D \in \mathcal{D}$ can be given by an inequality of the form (F.17).

PROOF. It is just a particular case of (F.15). □

The Proposition F.1 follows from Lemma F.1 because the inequality (F.17) goes to the inequality of the same kind under transformations (F.14), hence, under any fractional-linear transformation.

REMARK 4. Note that the set $\overline{Conf_2}\backslash Conf_2$ of conformal transformations of the second kind does not form a group. It is a two-sided coset in $\overline{Conf_2}$ with respect to $Conf_2$. It is worth to know that it possesses both properties listed in Propositions 2 and 3: it acts simply transitively on triples of distinct points in $\overline{\mathbb{C}}$ and preserve circles and discs.

♡

Case $n = 3$. The group $Conf_3 = PSO_0(1, 4; \mathbb{R})$ is isomorphic to the group $PU(1, 1; \mathbb{H})$ which is the quotient of $U(1, 1; \mathbb{H})$ over its center $\{\pm 1_2\}$. The group $U(1, 1; \mathbb{H})$ consists of quaternionic matrices $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ satisfying

$$|a|^2 = |d|^2 = 1 + |b|^2 = 1 + |c|^2, \qquad \bar{a}b = \bar{c}d.$$

Put $a = u \cosh t$, $d = v \cosh t$, where $t \in \mathbb{R}$ and u, v are quaternions of unit norm. Then there exists a quaternion of unit norm w such that $b = w \sinh t$ and $c = \bar{w}u\bar{v} \sinh t$.

If g is not diagonal, the parameters u, v, w and t are defined uniquely. For diagonal matrices we have $t = 0$ and the value of w does not matter. So, our group is a union of $S^3 \times S^3 \times S^3 \times (\mathbb{R}\backslash\{0\})$ and $S^3 \times S^3$.

The group $PU(1, 1; \mathbb{H})$ acts on the unit sphere S^3 by the formula: $u \mapsto (au + b)(cu + d)^{-1}$

Case $n = 4$. The group $Conf_4 = PO_+(1, 5; \mathbb{R})$ is isomorphic to another quaternionic group $PGL(2, \mathbb{H}) = GL(2, \mathbb{H})/\mathbb{R}^\times \cdot 1_2$ which acts on a quaternionic projective space $\mathbb{P}^1(\mathbb{H}) \simeq \overline{\mathbb{H}} \simeq \overline{\mathbb{R}^4} \simeq S^4$. The explicit formula is again: $q \mapsto (aq + b)(cq + d)^{-1}$.

Elaborate!

CHAPTER 6

Definition of Apollonian gasket

6.1. Basic facts

Consider three pairwise tangent discs D^1, D^2, D^3 on S^2. If we delete the interior of these discs from S^2, there remain two curvilinear triangles. Let us inscribe a disc of maximal possible size in each triangle and delete the interior of it. The remaining set consists of 6 triangles. Again, we inscribe a maximal possible disc in each triangle and delete the interior of these discs. We get 18 triangles.

Continuing this procedure, we delete from S^2 the countable set of open discs. The remaining closed set $\mathcal{A}$ is of fractal nature and is called **Apollonian gasket** in honor of the ancient Greek mathematician Apollonius of Perga lived in III-II century BC. Of course, we can replace S^2 by $\overline{\mathbb{R}}^2$ or $\overline{\mathbb{C}}$ and consider the corresponding picture on the extended plane.

According to general practice, we use the term "Apollonian gasket" also for the collection of (open, or closed) discs and the collection of circles which are involved into the construction.

Let us discuss different forms of Apollonian gasket. At first sight, the picture in question looks different for different choices of initial three discs. Nevertheless, all these pictures are in a sense equivalent.

To explain it, consider the group $Conf_2$ of conformal transformations of $\overline{\mathbb{C}}$ given by the formula (F.3.1).

Exercise 31. Show that any two triples of pairwise tangent circles can be transformed one into another by a conformal transformation.

Hint. Show that a triple of pairwise tangent circles is uniquely defined by the triple of tangent points. Then apply the proposition F.1.

So, up to conformal transformation, there is only one class of Apollonian gaskets

Theorem 6.1. *An Apollonian gasket $\mathcal{A}$ is determined by any triple of pairwise tangent discs in it. (In other words, if two Apollonian gaskets have a common triple of pairwise tangent discs, then they coincide).*

The statement looks rather evident and I encourage reader's endeavors to find their own proof. The proof given below based on the special numeration of all discs in a given gasket.

The numeration in question is suggested by the construction of a gasket. Namely, call the initial three discs D^1, D^2, D^3 **discs of level** -1. If we delete from S^2 the union of their interiors, the remaining set is a union of two closed curvilinear triangles. Call it **triangles of level 0** and denote by $T_\pm$. Next, we inscribe in each of these triangles a maximal possible disc, call it **disc of zero level** and denote by $D_\pm$.

After deleting from $T_\pm$ the interior of $D_\pm$, it becomes a union of three triangles. We call it **triangles of first level** and denote by $T_{\pm i_1},\ i_1 = 1, 2, 3$. In each of them we inscribe a maximal possible disc, denoted by $D_{\pm i_1}$, call it a **disc of first level** and continue this procedure.

On the n-th step we consider a triangle $T_{\pm i_1 i_2 \dots i_{n-1}}$, inscribe a maximal possible disc $D_{\pm i_1 i_2 \dots i_{n-1}}$ and delete its interior. The remaining set is a union of three triangles which we label by $T_{\pm i_1 i_2 \dots i_{n-1} i_n},\ i_n = 1, 2, 3$.

We observe, that two different disks of the same level $n \geq 0$ are never tangent to each other.

Thus, we have labelled all triangles (or discs) of level $n \geq 0$ by sequences of the form $\pm i_1 i_2 \dots i_n$ where i_k take values $1, 2, 3$ (see Fig 6.1).

Actually, our numeration scheme is not yet completely determined. We have not precised how we numerate three triangles of n-th level which are contained in a given triangle of $n-1$-st level. There are 3 triangles and 3 possible values of i_n, so there are 6 possible numerations and so far any of them can be used for our purposes.

LEMMA 6.1. *Let D, D', D'', D''' are four pairwise tangent discs on S^2. Then, if three of them belong to some gasket $\mathcal{A}$, so does the fourth.*

PROOF. Assume that D, D', D'' belong to $\mathcal{A}$ and have levels l, m, n respectively. We can assume that $l \leq m \leq n$. To simplify the arguments and make notations uniform, we consider first the case $0 < l < m < n$.

In this case we can suppose that $D = D_{\pm i_1 i_2 \dots i_l}, D' = D_{\pm j_1 j_2 \dots j_m}$, and $D'' = D_{\pm k_1 k_2 \dots k_n}$. By the construction, D'' is a disc inscribed in a triangle $T_{\pm k_1 k_2 \dots k_n}$ which is bounded by arcs of three discs. One of them is $D_{\pm k_1 k_2 \dots k_{n-1}}$ and two other discs have levels, say l' and m', such that $l' \leq m' < n$. (Equality is possible only if $l' = m' = 0$.)

From the construction of $\mathcal{A}$ it is also clear that all discs tangent to D'' except three mentioned above, have level $> n$. But we know that D and D' are tangent to D''. It follows that $l' = l$, $m' = m$ and our three discs are exactly D, D' and $D_{k_1 k_2 \dots k_{n-1}}$. Therefore, the disc $D_{k_1 k_2 \dots k_{n-1}}$ is tangent to D, D', D''. Another disc, which is also tangent to D, D', D'', is $D_{k_1 k_2 \dots k_n k_{n+1}}$ for an appropriate choice of k_{n+1}. We see, that both discs tangent to D, D', D'' belong to $\mathcal{A}$.

In the cases $0 = l < m < n$ or $0 = l = m < n$ the proof is completely analogous. For example, the disc, tangent to D_2 D_3 and D_{+1}, must be either D_+ or D_{+11}. Hence, it belongs to $\mathcal{A}$.

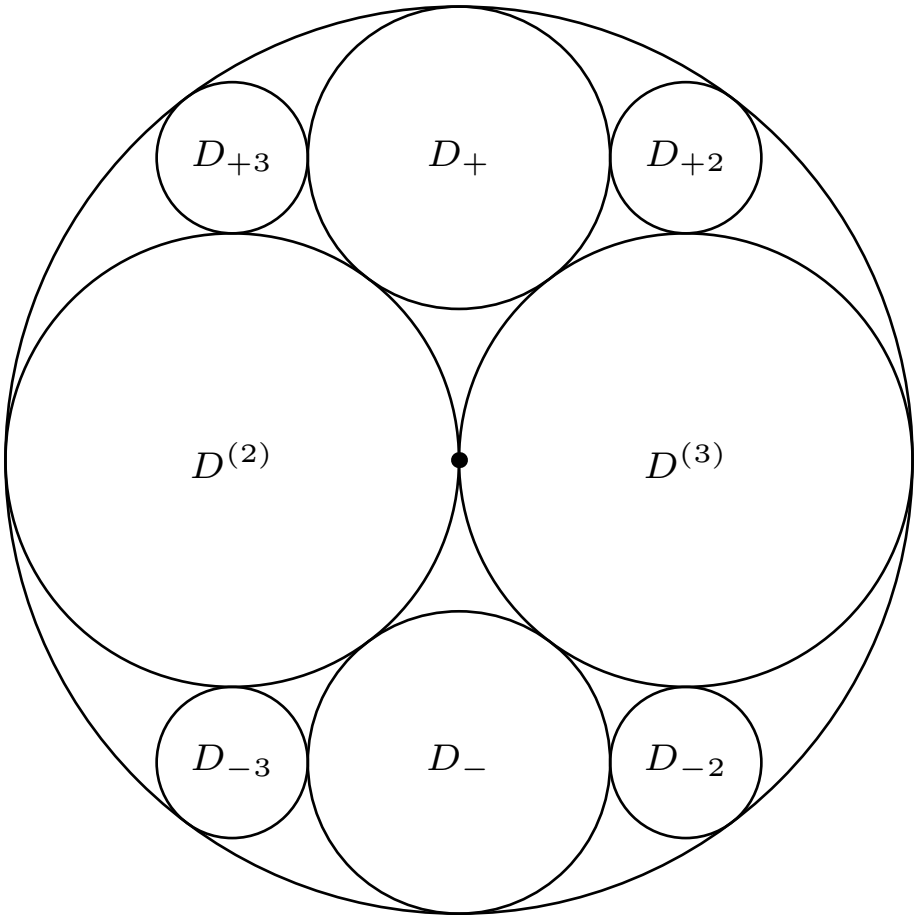

FIGURE 6.1. Numeration of discs in the rectangular gasket

Finally, if $l = m = n = 0$, then D, D', D'' are the three initial discs D^1, D^2, D^3 and D''' coincides with D_+ or D_-, hence, belongs to $\mathcal{A}$. $\square$

PROOF OF THE THEOREM. Let two gaskets $\mathcal{A}$ and $\widetilde{\mathcal{A}}$ have a common triple of pairwise tangent discs D, D', D''. Assume that these disks have level $l \le m \le n$ in $\mathcal{A}$. We want to show that $\mathcal{A} \subset \widetilde{\mathcal{A}}$ using the induction on n.

For $n = 0$ our three discs are just the initial discs D^1, D^2, D^3 for $\mathcal{A}$. According to lemma 2, the discs $D_\pm$ belong to $\widetilde{\mathcal{A}}$ because so do D^1, D^2, D^3.

Use again the induction and suppose that we already know that all discs of level $\le n-1$ in $\mathcal{A}$ belong also to $\widetilde{\mathcal{A}}$. Then any disc of level n, being tangent to three discs of level $\le n-1$, also belongs to $\widetilde{\mathcal{A}}$.

Return to the first induction. Assume that we have proved that if the common discs have level $< n$ in $\mathcal{A}$, then $\mathcal{A} \subset \widetilde{\mathcal{A}}$.

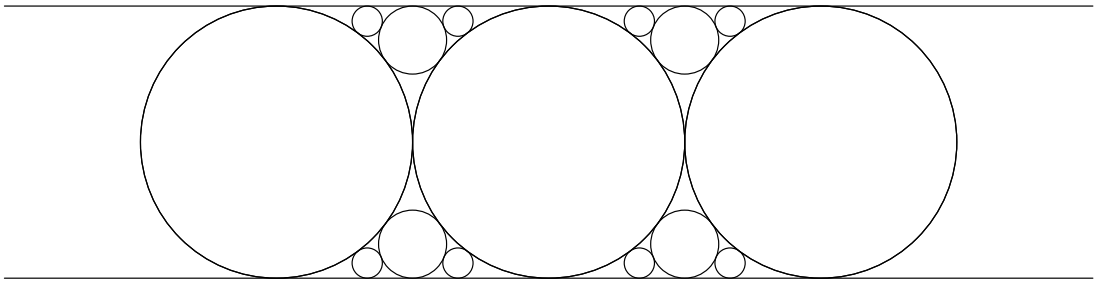

FIGURE 6.2. Band gasket

Let D, D', D'' are common discs of levels $k \le l < n$ respectively. From the proof of Lemma 6.1 we know that among discs, tangent to D, D', D'', there is one which has level $n-1$. Call it D'''. Then D, D', D''' is a common triple of level $\le n-1$ and we are done.

Thus, $\mathcal{A} \subset \widetilde{\mathcal{A}}$. But in the initial data $\mathcal{A}$ and $\widetilde{\mathcal{A}}$ play symmetric roles. Therefore, $\widetilde{\mathcal{A}} \subset \mathcal{A}$ and $\widetilde{\mathcal{A}} = \mathcal{A}$. □

LEMMA 6.2. *The triangle $T_{\pm i_1 i_2 \dots i_n}$ is contained in $T_{\pm j_1 j_2 \dots j_m}$ iff $m < n$, the signs coincide and $i_k = j_k$ for $1 \le k \le m$.*

PROOF. Note, that triangles of the same level can have not more than 3 common points. So, our first triangle is contained only in one of triangles of level m. But it is contained in $T_{\pm i_1 i_2 \dots i_m}$ and in $T_{\pm j_1 j_2 \dots j_m}$. So, we come to the statement of the lemma.

□

There are three most symmetric choices for an initial triple of pairwise tangent circles. The corresponding Apollonian gaskets are shown on Fig. 7, 8 a), 8 b). For future use, we write inside each circle its curvature.

All three gaskets are stereographic projections of a most symmetric gasket on S^2 generated by four pairwise tangent discs of the same size. See Fig. 6.5.

There are some other interesting realizations of Apollonian gasket from which we want to mention two. Their study uses some facts about so-called Fibonacci numbers.

Info G. Fibonacci numbers

The famous Italian mathematician Leonardo from Pisa, often called by a nickname Fibonacci, lived long ago, in 13th century. Among other things he considered the sequence of integers $\{\Phi_k\}$ satisfying the recurrence

$$\Phi_{k+1} = \Phi_k + \Phi_{k-1} \tag{G.1}$$

and the initial condition $\Phi_1 = \Phi_2 = 1$. It looks as follows:

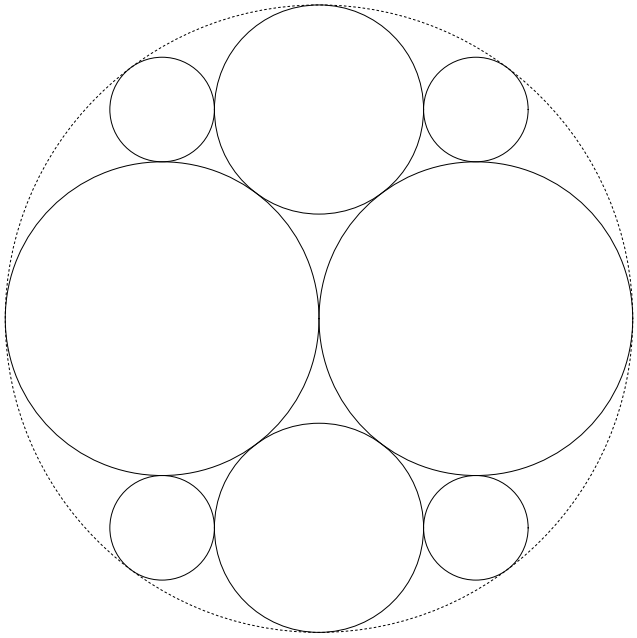

FIGURE 6.3. Rectangular gasket

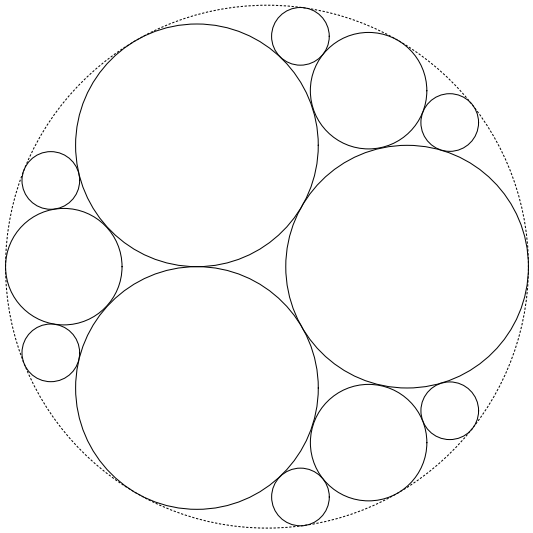

FIGURE 6.4. Triangular gasket

n:	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
Φ_n:	13	-8	5	-3	2	-1	1	0	1	1	2	3	5	8	13

Later these numbers appeared in many algebraic and combinatorial problems and got the name **Fibonacci numbers**. We briefly describe the main facts related to this and similar sequences.

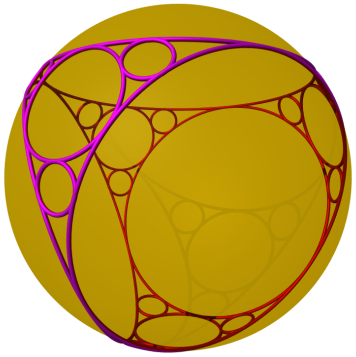

FIGURE 6.5. Spherical gasket

Consider the set V of all two-sided real sequences $\{v_n\}_{n\in\mathbb{Z}}$ satisfying the recurrent relation of type (G.1), i.e. $v_{n+1} = v_n + v_{n-1}$. It is a real vector space where the operations of addition and multiplication by a real number are defined termwise.

The dimension of this space is 2, because any sequence in question is completely determined by two terms v_0, v_1 and these terms can be chosen arbitrarily. We can consider (v_0, v_1) as coordinates in V. So, the series of Fibonacci numbers is a vector in V with coordinates (0, 1). Another known sequence of **Lucas** numbers has coordinates (2, 1).

Let T denote the transformation sending the sequence $\{v_n\}$ to the sequence $\{v_{n+1}\}$ (which also satisfies the same recurrent relation!). It is a linear operator in V. The spectrum of this operator consists of numbers λ, satisfying $\lambda^2 = \lambda + 1$. There are two such numbers: $\phi = \frac{1+\sqrt{5}}{2} \approx 1.618$ and $-\phi^{-1} = \frac{1-\sqrt{5}}{2} = 1 - \phi$. The first of it has a special name **golden ratio** because the rectangle with sides proportional to $\phi : 1$ considered as the most pleasant for human eyes.

For the future use we introduce also the quantities $c = \phi^2 = \frac{3+\sqrt{5}}{2} = \phi+1$ and $\theta = \sqrt{\phi} = \sqrt{\frac{1+\sqrt{5}}{2}}$.

The corresponding eigenvectors of T are geometric progressions $v'_n = \phi^n$ and $v''_n = (-\phi)^{-n}$. Since they are linearly independent, any element of V is a linear combination of these eigenvectors.

In particular, the n-th Fibonacci number can be written as

$$\Phi_n = \alpha \cdot \phi^n + \beta \cdot (-\phi^{-1})^n \quad \text{for appropriate } \alpha \text{ and } \beta.$$

Using the normalization $\Phi_1 = \Phi_2 = 1$, we get $\alpha = -\beta = \frac{1}{\phi+\phi^{-1}} = \frac{1}{\sqrt{5}}$. Thus,

(G.2)

$$\Phi_{2k} = \frac{\phi^{2k} - \phi^{-2k}}{\sqrt{5}} = \frac{c^k - c^{-k}}{\sqrt{5}}; \qquad \Phi_{2k+1} = \frac{\phi^{2k+1} + \phi^{-2k-1}}{\sqrt{5}} = \frac{c^{k+\frac{1}{2}} + c^{-k-\frac{1}{2}}}{\sqrt{5}}.$$

FIGURE 6.6. Quadruples q_1 and $c \cdot q_1$

Conversely,

(G.3)

$$\phi^n = (-1)^n \frac{\Phi_{n+1} + \Phi_{n-1} - \Phi_n\sqrt{5}}{2}; \quad c^n = \frac{\Phi_{2n+1} + \Phi_{2n-1} + \Phi_{2n}\sqrt{5}}{2}.$$

Note also that $\Phi_{-2n} = -\Phi_{2n}$; $\quad \Phi_{-2n-1} = \Phi_{2n+1}$.

It follows that

(G.4)

$$\Phi_n \approx \frac{\phi^n}{\sqrt{5}} \qquad \text{and} \qquad \lim_{n\to\infty} \frac{\Phi_{n+1}}{\Phi_n} = \phi.$$

The Lucas number are given by a more simple expression: $L_n = \phi^n + (-\phi)^{-n}$. They look as follows

n:	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
L_n:	29	18	-11	7	-4	3	-1	2	1	3	4	7	11	18	29

.

♢

6.2. Examples of non-bounded Apollonian tiling

Consider a quadruple q_1 of pairwise tangent discs, one of which is a lower half-plane and other three have the boundary curvatures which form a geometric progression. Then the four curvatures can be written as $0 < x^{-1} < 1 < x$. The number x must satisfy the equation

(6.2.1) $(x + 1 + x^{-1})^2 = 2(x^2 + 1 + x^{-2})$, or $x^2 - 2(x + x^{-1}) + x^{-2} = 1$.

Putting $y := x + x^{-1}$, we obtain $y^2 - 2y - 3 = 0$. So, y is 1 or 3. Only the second value of y gives the real value of x. We have $x = \frac{3+\sqrt{5}}{2} = \frac{2}{3-\sqrt{5}}$ which is the number c which had been introduced in Info G.

The gasket $\mathcal{A}_1$ generated by q_1 has the following property. If we dilate it in ratio c, it goes to its mirror reflection in a vertical line. And if we dilate it in ratio c^2, it goes to one of its horizontal translation. Choosing an appropriate position of $\mathcal{A}_1$, we can arrange, that the mirror in question is an imaginary axis and the translation is an identity – see Fig 11. It means that $\mathcal{A}_1$ is invariant under transformation $w \mapsto -c\bar{w}$. Indeed, $\mathcal{A}_1$ and $-c \cdot \overline{\mathcal{A}}_1$ have a common triple of discs.

The gasket $\mathcal{A}_1$ contains, in particular, a series of discs D_k with boundary curvatures c^k, $k \in \mathbb{Z}$. These discs can be given by inequalities

(6.2.2) $$\left|c^k w + (-1)^k \tfrac{2}{\sqrt{5}} + i\right| \le 1$$

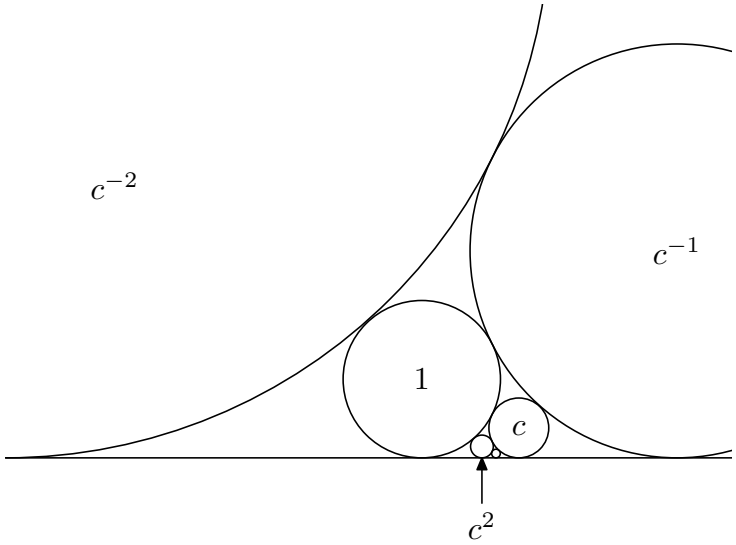

FIGURE 6.7. The gasket $\mathcal{A}_1$

and the corresponding normalized Hermitian matrices are

$$M_k = \begin{pmatrix} \frac{4}{5}c^{-k} & (-1)^k\frac{2}{\sqrt{5}} - i \\ (-1)^k\frac{2}{\sqrt{5}} + i & c^k \end{pmatrix} = \tag{6.2.3}$$
$$\begin{pmatrix} (-\phi)^{-k} & 0 \\ 0 & \phi^k \end{pmatrix} \cdot \begin{pmatrix} \frac{4}{5} & \frac{2}{\sqrt{5}} - i(-1)^k \\ \frac{2}{\sqrt{5}} + i(-1)^k & 1 \end{pmatrix} \cdot \begin{pmatrix} (-\phi)^k & 0 \\ 0 & \phi^k \end{pmatrix}$$

where $\phi := \sqrt{c} \approx 1.618034..$ is the famous "golden ratio".

Each of the relations 6.2.2 and 6.2.3 implies that the dilation $w \to -c \cdot \bar{w}$ send the disc D_n to D_{n-1}, hence, preserve the gasket $\mathcal{A}_1$.

EXERCISE 32. Find a matrix $g \in SL(2, \mathbb{C})$ which transforms the gasket $\mathcal{A}_1$ into the band gasket.

HINT. Find the transformation g which preserves real line and sends the disc D_0 to the parallel line. Show that the images $g \cdot D_k$ will be situated as on Fig. **??**.

Another interesting gasket $\mathcal{A}_2$ with unbounded curvatures can be defined as follows.

Consider a quadruple q_2 whose disks have boundary curvatures forming a geometric progression $(1, \rho, \rho^2, \rho^3)$ where $\rho > 1$. Then the Descartes equation is:

$$(1 + \rho + \rho^2 + \rho^3)^2 = 2(1 + \rho^2 + \rho^4 + \rho^6). \tag{6.2.4}$$

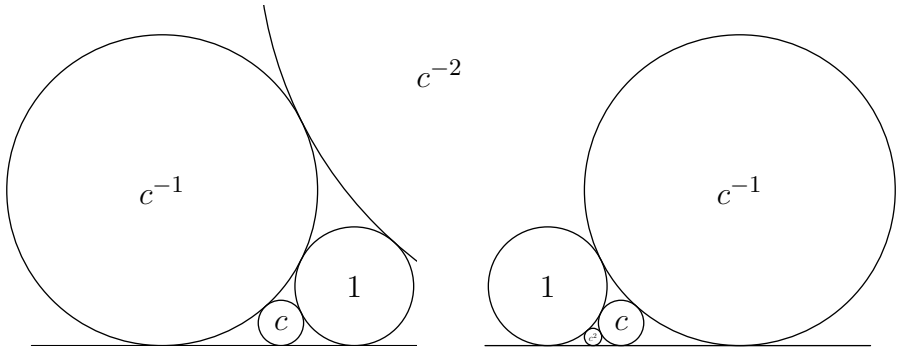

FIGURE 6.8. Gaskets $\mathcal{A}_1$ and $c \cdot \mathcal{A}_1$

Simplifying this equation, write it in the form

$$0 = 1-2\rho-\rho^2-4\rho^3-\rho^4-2\rho^5+\rho^6, \quad \text{or} \quad 4+(\rho+\rho^{-1})+2(\rho^2+\rho^{-2}) = (\rho^3+\rho^{-3}).$$

Introducing $u = \rho + \rho^{-1}$, we get

$$4 + u + 2(u^2 - 2) = (u^3 - 3u), \quad \text{or} \quad u^3 - 2u^2 - 4u = 0.$$

This equation has three solutions: $u = 0,\ 1 - \sqrt{5},\ 1 + \sqrt{5}$. Only the last solution give the real value for ρ and we get

(6.2.5)

$$\rho = \phi + \sqrt{\phi} = \theta^2 + \theta \approx 2.890054...; \quad \rho^{-1} = \phi - \sqrt{\phi} = \theta^2 - \theta \approx 0.346014....$$

The corresponding discs D_k form a spiral, convergent to certain point a when $k \to -\infty$. If we take a for the origin, our spiral will be invariant under multiplication by a complex number λ with $|\lambda| = \rho$. Denote the argument of λ by 2α. Then the corresponding matrices M_k must have the form

$$M_k = \begin{pmatrix} a\rho^k & be^{2ik\alpha} \\ \bar{b}e^{-2ik\alpha} & c\rho^{-k} \end{pmatrix}, \quad ac - |b|^2 = -1. \tag{6.2.6}$$

The condition that discs D_k and D_{k+m} are tangent is $\det(M_k + M_{k+m}) = 0$. This condition actually does not depend on k and leads to the equation

$$\frac{|b|^2}{ac} = \frac{\rho^m + \rho^{-m} + 2}{e^{im\alpha} + e^{-im\alpha} + 2}.$$

Put $s = \frac{1}{2}\log\rho$. Then the right hand side of the equation takes the form

$$\frac{1 + \cosh 2ms}{1 + \cos 2m\alpha} = \left(\frac{\cosh\, ms}{\cos\, m\alpha}\right)^2.$$

We know that D_0 is tangent to D_m for $m = 1, 2, 3$. So, we have

$$\frac{|b|}{\sqrt{ac}} = \frac{\cosh\, s}{|\cos\, \alpha|} = \frac{\cosh\, 2s}{|\cos\, 2\alpha|} = \frac{\cosh\, 3s}{|\cos\, 3\alpha|}. \tag{6.2.7}$$

Since $\cosh 3s = \cosh s\ (2\cosh\ 2s\ -1)$ and $\cos 3\alpha = \cos\alpha\ (2\cos\ 2\alpha\ -1)$, we conclude, comparing the second and last terms in (6.2.7), that $2\cosh\ 2s - 1 = |2\cos\ 2\alpha\ -1|$.

This can happen only if $2\cos 2\alpha - 1 < 0$. Therefore, we get $2\cosh 2s - 1 = 1 - 2\cos 2\alpha$, or $\cosh 2s = 1 - \cos 2\alpha$, which is possible only if $\cos 2\alpha \le 0$.

Using the relation $\cosh 2s = 1 - \cos 2\alpha$, we get, comparing the second and third terms,

$$\cosh s = \pm\frac{\cos\alpha\cdot(1-\cos 2\alpha)}{\cos 2\alpha}.$$

Now, the relation $2\cosh^2 s = \cosh 2s + 1$ gives us the equation:

$$2\left(\frac{\cos\alpha\cdot(1-\cos 2\alpha)}{\cos 2\alpha}\right)^2 = 2 - \cos 2\alpha.$$

Denote $\cos 2\alpha$ by x and write the equation in an algebraic form:

$$\frac{(x+1)(1-x)^2}{x^2} = 2-x, \quad \text{or} \quad (x+1)(1-x)^2 = 2x^2-x^3, \quad \text{or} \quad 2x^3-3x^2-x+1 = 0.$$

It has a solution $x = 1/2$ and this allows us to rewrite it in the simple form $(2x-1)(x^2-x-1) = 0$. So, the other two solutions are ϕ and $-\phi^{-1} = 1-\phi$. Only one of these three solutions is negative: $x = -\phi^{-1}$.

We conclude that $\cos 2\alpha = -\phi^{-1}$, $\cosh 2s = \phi$. Hence, $\rho + \rho^{-1} = 2\phi$ and $\rho = \phi + \sqrt{\phi^2-1} = \theta^2 + \theta$. Also, we get $\frac{|b|}{\sqrt{ac}} = \phi^2$, therefore

$$|b|^2 = \frac{\phi^2}{\sqrt{5}}, \qquad ac = \frac{\phi^{-2}}{\sqrt{5}}. \tag{6.2.8}$$

It follows that we know matrices M_k up to complex conjugation and conjugation by a diagonal matrix. Geometrically, it means that we know the gasket $\mathcal{A}_2$ up to rotation, dilation and reflection in a straight line. In particular, we can put

$$M_k = \frac{1}{\sqrt[4]{5}}\begin{pmatrix}\phi^{-1}\cdot\rho^k & \phi\cdot e^{2ik\alpha}\\ \phi\cdot e^{2ik\alpha} & \phi^{-1}\cdot\rho^{-k}\end{pmatrix} \tag{6.2.9}$$

so that

$$D_0 = \left\{ w \;\middle|\; \left|\ w + 1 + \frac{1}{\sqrt{5}}\ \right| \le \sqrt{\frac{1+2\sqrt{5}}{5}}\ \right\}. \tag{6.2.10}$$

Further, let us compute the number λ which is determined up to complex conjugation. We have

$$2\sin^2\alpha = 1 - \cos 2\alpha = \phi \quad \text{and} \quad 2\cos^2\alpha = 1 + \cos 2\alpha = 1 - \phi^{-1} = \phi^{-2}.$$

Therefore, $\sin^2\alpha = \phi^{-1}$ and $\sin 2\alpha = \pm\theta^{-1}$. So, we have $e^{2i\alpha} = \cos 2\alpha + i\sin 2\alpha = -\phi^{-1} \pm i\theta^{-1}$. Finally,

$$\lambda = \rho e^{2i\alpha} = -(1+\theta^{-1})(1 \mp i\theta). \tag{6.2.11}$$

The corresponding picture is shown on Fig. 6.9.

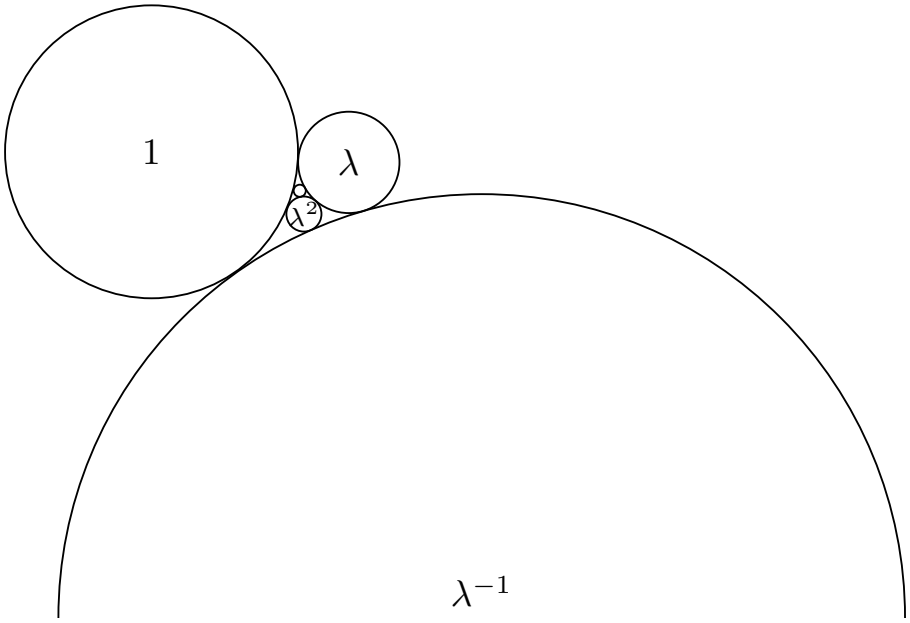

FIGURE 6.9. The gasket $\mathcal{A}_2$

6.3. Two interpretations of the set $\mathcal{D}$

Let $\mathbb{R}^{1,3}$ be the four-dimensional real vector space with coordinates t, x, y, z and with the indefinite scalar product

$$(p_1,\, p_2) = t_1t_2 - x_1x_2 - y_1y_2 - z_1z_2 \tag{6.3.1}$$

The space $\mathbb{R}^{1,3}$ is called the Minkowski space and is the basic object in the Special Relativity Theory. The scalar square $|p|^2 = (p,\, p)$ of a vector $p \in \mathbb{R}^{1,3}$ can be positive, zero or negative. Correspondingly, the vector p is called **time-like**, **light-like** and **space-like** respectively. The time-like vectors are of two kind: the future vectors with $t > 0$ and past vectors with $t < 0$.

The physical meaning of p is an **event** which take place at the moment t of time in the point $(x,\, y,\, z) \in \mathbb{R}^3$.

Physicists call the **whole Lorentz group L** the group of all linear transformations of $\mathbb{R}^{1,3}$ which preserve the scalar product (1.4.1). It splits into four connected components and the component containing the unit is called **proper Lorentz group $\mathbf{L_0}$**. In mathematical papers these groups are denoted by $O(1,\, 3)$ and $SO_+(1,\, 3)$ respectively.

The **Relativity Principle** claims that all physical laws are invariant under the proper Lorentz group.

Algebraically, elements $g \in O(1,\, 3)$ are given by 4×4 real matrices $|g_{i,j}|$ whose rows (columns) are pairwise orthogonal vectors from $\mathbb{R}^{1,3}$, such that

the first row (first column) has the scalar square 1, while all other rows (columns) have the scalar square -1. [1]

We recall in slightly different notations some facts explained in Info F.

An element $g \in O(1, 3)$ belongs to the proper Lorentz group, if two additional conditions are satisfied: $\det g = 1$ and $g_{0,0} > 0$.

Now we show how to use Minkowski space to label the discs on a unit 2-sphere. A disc on S^2 can be defined as an intersection of S^2 with a half-space $H_{u,\tau}$ given by

(6.3.2)

$$H_{u,\tau} = \{v \in \mathbb{R}^3 \mid (u, v) + \tau \leq 0\} \quad \text{where} \quad u \in S^2 \quad \text{and} \quad \tau \in (-1, 1).$$

Instead of the pair $(u, \tau) \in S^2 \times (-1, 1)$ we can use the one space-like vector $p = (t, x, y, z) \in \mathbb{R}^{1,3}$ given by

$$p = \frac{1}{\sqrt{1+\tau^2}} \cdot (\tau, u).$$

Then the half-space in question takes the form

$$H_p = \{v \in \mathbb{R}^3 \mid xv^1 + yv^2 + zv^3 + t \leq 0\}. \tag{6.3.3}$$

It is clear that $H_{p_1} = H_{p_2}$ iff $p_1 = c \cdot p_2$ with $c > 0$. Therefore, we can and will normalize p by the condition $|p|^2 = -1$.

So, the space $\mathcal{D}$ of discs on S^2 is identified with the set P_{-1} of all space-like vectors $p \in \mathbb{R}^{1,3}$ with $|p|^2 = -1$. It is well-known that P_{-1} is a one-sheeted hyperboloid in $\mathbb{R}^4$ and that the group $L_0 \simeq SO_+(1, 3; \mathbb{R})$ acts transitively on it. The stabilizer of the point $(0, 0, 0, 1)$ is isomorphic to the group $SO_+(1, 2; \mathbb{R})$ which is naturally embedded in L_0.

We get the first interpretation of $\mathcal{D}$ as an homogeneous manifold.

Exercise 33. Show that the 3-dimensional hyperboloid in $\mathbb{R}^{1,3}$ defined by the equation $|p|^2 = -1$, is diffeomorphic to $S^2 \times \mathbb{R}$.

Hint. Use the parameters u, τ introduced above.

Our next interpretation of the space $\mathcal{D}$ uses the complex matrix theory. We start with inequality (3.7.61.2.5) and collect the coefficients in the left hand side into a 2×2 matrix $M = \begin{pmatrix} a & b \\ \bar{b} & c \end{pmatrix}$. Recall that we imposed on a, b, c the condition $ac - |b|^2 < 0$. So, M is an Hermitian matrix with $\det M < 0$. Here again we can and will normalize M by the condition $\det M = -1$.

Thus, the set $\mathcal{D}$ is identified with the collection H_{-1} of all Hermitian 2×2 matrices M with $\det M = -1$.

[1] Compare with the properties of the usual orthogonal matrices: all rows (columns) have length 1 and are orthogonal to each other.

EXERCISE 34. Show that the relation between two last interpretations is as follows: to a vector $p = (t, x, y, z) \in \mathbb{R}^{1,3}$ there corresponds the matrix $M = \begin{pmatrix} a & b \\ \bar{b} & c \end{pmatrix}$ with

$$a = t - z, \quad b = x + iy, \quad c = t + z. \tag{6.3.4}$$

HINT. Compare (F.5) and (F.17).

Now we want to describe $\mathcal{D}$ in the second interpretation as an homogeneous space.

We have already seen the action of the group $G = PSL(2, \mathbb{C})$ on $\overline{\mathbb{C}}$ by fractional-linear transformations. Moreover, by Proposition 3, G_2 acts on the set $\mathcal{D}$ of all discs on $\overline{\mathbb{C}}$.

On the other hand, the group $SL(2, \mathbb{C})$ acts on the set H of Hermitian 2×2 matrices by the rule:

$$g: \quad M \mapsto gMg^* \tag{6.3.5}$$

and this action preserves the set H_{-1} of matrices with determinant -1. (Actually, this is a G-action, since the center C of $SL(2, \mathbb{C})$ acts trivially.)

THEOREM 6.2. *There exists a homomorphism $\pi : SL(2, \mathbb{C}) \to L_0 \simeq SO_0(1, 3; \mathbb{R})$, such that the following diagram is commutative:*

$$\begin{array}{ccccc} G & \times & \mathcal{D} & \longrightarrow & \mathcal{D} \\ {\scriptstyle p}\uparrow & & \uparrow\| & & \uparrow\| \\ SL(2, \mathbb{C}) & \times & H_{-1} & \longrightarrow & H_{-1} \\ {\scriptstyle \pi}\downarrow & & \downarrow\| & & \downarrow\| \\ L_0 & \times & P_{-1} & \longrightarrow & P_{-1} \end{array}$$

Where p is the natural projection of $SL(2, \mathbb{C})$ to $PSL(2, \mathbb{C}) \simeq G$ and horizontal arrows denote the actions.

We leave the verification to the reader but give here the explicit formula for the homomorphism π.

EXERCISE 35. Show that the homomorphism π has the form

$$\pi\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \frac{|a|^2+|b|^2+|c|^2+|d|^2}{2} & \mathrm{Re}(a\bar{b} + c\bar{d}) & \mathrm{Im}(\bar{a}b + \bar{c}d) & \frac{|b|^2-|a|^2-|c|^2+|d|^2}{2} \\ \mathrm{Re}(a\bar{c} + b\bar{d}) & \mathrm{Re}(a\bar{d} + b\bar{c}) & \mathrm{Im}(\bar{a}d - \bar{b}c) & \mathrm{Re}(b\bar{d} - a\bar{c}) \\ \mathrm{Im}(a\bar{c} + b\bar{d}) & \mathrm{Im}(a\bar{d} + b\bar{c}) & \mathrm{Re}(\bar{a}d - \bar{b}c) & \mathrm{Im}(b\bar{d} - a\bar{c}) \\ \frac{|c|^2-|a|^2-|b|^2+|d|^2}{2} & \mathrm{Re}(\bar{c}d - \bar{a}b) & \mathrm{Im}(\bar{c}d - \bar{a}b) & \frac{|a|^2-|b|^2-|c|^2+|d|^2}{2} \end{pmatrix}.$$

REMARK 5. The inverse map of $SO^+(1, 3; \mathbb{R}) \to PSL(2, \mathbb{C})$ is well-defined but its lifting to $SL(2, \mathbb{C})$ is defined only up to sign. It is the so-called **spinor representation** of $SO^+(1, 3; \mathbb{R})$.

In particular, all products of the form $2a\bar{a}$, $2a\bar{b}, ...$ etc are well-defined and given in the table:

	$\bar{a}$	$\bar{b}$	$\bar{c}$	$\bar{d}$
$2a$	$g_{00}-g_{03}-g_{30}+g_{33}$	$g_{01}-g_{31}+i(g_{32}-g_{02})$	$g_{10}-g_{13}+i(g_{20}-g_{23})$	$g_{11}+g_{22}+i(g_{21}-g_{12})$
$2b$	$g_{01}-g_{31}+i(g_{02}-g_{32})$	$g_{00}+g_{03}-g_{30}-g_{33}$	$g_{11}-g_{22}+i(g_{12}+g_{21})$	$g_{10}+g_{13}+i(g_{20}+g_{23})$
$2c$	$g_{10}-g_{13}+i(g_{23}-g_{20})$	$g_{11}-g_{22}-i(g_{12}+g_{21})$	$g_{00}-g_{03}+g_{30}-g_{33}$	$g_{01}+g_{31}-i(g_{02}+g_{32})$
$2d$	$g_{11}+g_{22}+i(g_{12}-g_{21})$	$g_{10}+g_{13}-i(g_{20}+g_{23})$	$g_{01}+g_{31}+i(g_{02}+g_{32})$	$g_{00}+g_{03}+g_{30}+g_{33}$

♡

EXERCISE 36. Describe the image under π of the following subgroups of G:

a) $PGL(2, \mathbb{R})$; b) $PSU(2, \mathbb{C})$; c) $PSU(1, 1; \mathbb{C})$.

Hint: Use the fact that the subgroup in question are stabilizers of some geometric objects.

Answers: a) $\pi\big(PGL(2, \mathbb{R})\big) = Stab\,(0, 0, 1, 0) \simeq SO_+(1, 2; \mathbb{R})$;
b) $\pi\big(PSU(2, \mathbb{C})\big) = Stab\,(1, 0, 0, 0) \simeq SO(3, \mathbb{R})$;
c) $\pi\big(PSU(1, 1; \mathbb{C})\big) = Stab\,(0, 0, 0, 1) \simeq SO_+(1, 2; \mathbb{R})$.

An interesting problem is to compare the image under π of the subgroup $SL(2, \mathbb{Z} + i\mathbb{Z})$ with the subgroup $SO_+(1, 3; \mathbb{Z})$.

6.4. Generalized Descartes theorem

Let D_i, $1 \le i \le 4$, be four pairwise tangent discs. Denote by p_i, (resp. M_i) the corresponding space-like vectors with $|p_i|^2 = -1$ (resp. the Hermitian matrices with $\det M_i = -1$).

LEMMA 6.3. *The discs D_1 and D_2 are tangent iff the following equivalent conditions are satisfied:*

a) p_1+p_2 is a future light vector; *b) $(p_1, p_2) = 1$ and p_1+p_2 has positive t-coordinate;* *c) $\det(M_1 + M_2) = 0$ and $\operatorname{tr}(M_1 + M_2) > 0$.*

PROOF. First, we show that the oriented circles $C_i = \partial D_i$, $i = 1, 2$, are negatively (resp. positively) tangent iff $|p_1 \pm p_2|^2 = 0$, or, equivalently, $\det(M_1 \pm M_2) = 0$.

Using the appropriate Möbius transformation, we can assume that the first circle is the real line with a standard orientation. The corresponding vector and matrix are $p_1 = (0, 0, -1, 0)$ and $M_1 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$.

Let C_2 be an oriented circle tangent to C_1. Denote the tangent point by a. Then the transformation $w \mapsto \frac{c}{a-w}$ for a real c preserves C_1. For an appropriate c it sends C_2 to a horizontal line $2i+\mathbb{R}$ with certain orientation. The corresponding vector and matrix are $p_2 = \pm(1, 0, -1, 1)$ and $M_2 = \pm\begin{pmatrix} 2 & -i \\ i & 0 \end{pmatrix}$ where the plus sign corresponds to the standard orientation and the minus sign to the opposite one. We see, that the conditions above are satisfied. Conversely, if these conditions are satisfied, we can make a Möbius

transformation, such that vectors p_1 and p_2 take the form above. Then the corresponding circles are tangent.

The proof of the Lemma follows the same scheme. Note only that for light vectors the sign of the t-coordinates is preserved by G and so is the sign of the trace of M when $\det M = 0$.

□

Come back to the theorem. Consider the the Gram matrix of scalar products for p_i. According to Lemma 6.3, it looks as follows:

$$G_{ij} := (p_i,\, p_j) = 1 - 2\delta_{ij}. \tag{6.4.1}$$

It is well-known that the determinant of the Gram matrix of a system of n vectors in $\mathbb{R}^n$ equals to the square of the determinant composed from coordinates of these vectors. The same is true for pseudo-Euclidean spaces, e.g. for $\mathbb{R}^{1,3}$.

Since $G^2 = 4 \cdot \mathbf{1}$, we have $\det G = 16$. It follows that vectors p_i are linearly independent, hence, form a basis in $\mathbb{R}^{1,3}$.

For any vector $v \in \mathbb{R}^{1,3}$ we define its covariant coordinates v_i and contravariant coordinates v^j with respect to the basis $\{p_i\}$ as follows:

$$v_i = (v,\, p_i); \qquad v = \sum_{j=1}^{4} v^j \cdot p_j. \tag{6.4.2}$$

Let us find the relation between these coordinates. From (6.4.1) and (6.4.2) we have

$$v_i = \left(\sum_{j=1}^{4} v^j \cdot p_j,\, p_i \right) = \sum_{j=1}^{4} G_{ij} v^j = \sum_{j=1}^{4} v^j - 2v^i. \tag{6.4.3}$$

Taking the sum over i we get $\sum_{j=1}^{4} v_i = 4 \sum_{j=1}^{4} v^j - 2 \sum_{j=1}^{4} v^j = 2 \sum_{j=1}^{4} v^j$ and, finally

$$v^j = \frac{1}{2} \sum_{j=1}^{4} v_i - \frac{1}{2} v_j. \tag{6.4.4}$$

From (6.4.3) we also derive the expression for $|v|^2$ in term of coordinates:

$$|v|^2 = \left(\sum_j v^j \right)^2 - 2 \sum_j (v^j)^2 \;=\; \frac{1}{4} \left(\sum_i v_i \right)^2 - \frac{1}{2} \sum_i v_i^2. \tag{6.4.5}$$

It follows that for any light vector v we have

$$\left(\sum_i v_i \right)^2 - 2 \sum_i v_i^2 = 0. \tag{6.4.6}$$

Put, in particular, $v = (1,\, 0,\, 0,\, -1)$. Then $v_i = (v,\, p_i) = t_i + z_i = c_i$ and (6.4.6) gives exactly the statement of Descartes theorem.

Actually the same approach allows to prove more.

THEOREM 6.3. *(Generalized Descartes Theorem) The matrices M_i satisfy the relation*

$$\left(\sum_i M_i\right)^2 - 2\sum_i M_i^2 = -8\cdot\mathbf{1}. \tag{6.4.7}$$

PROOF. Introduce an inner product in the space of 2×2 Hermitian matrices, which correspond to the quadratic form $Q(M) = \det M$. The explicit formula is

$$(M_1,\, M_2) = \frac{\det(M_1+M_2) - \det M_1 - \det M_2}{2}. \tag{6.4.8}$$

In particular, we have $(M,\, 1) = \frac{1}{2}\mathrm{tr}\, M$.

Recall also the Cayley identity which for 2×2 matrices has the form

$$M^2 = M\cdot \mathrm{tr}\, M - \det M\cdot 1. \tag{6.4.9}$$

Let now M_1, M_2, M_3, M_4 be four Hermitian matrices, corresponding to four pairwise tangent discs and normalized by the condition $\det M_i = -1$. Then (6.4.9) takes the form

$$M_i^2 = M_i\cdot \mathrm{tr}\, M_i + 1. \tag{6.4.10}$$

Introduce the notations

$$\Sigma_1 := \sum_{i=1}^{i=4} M_i, \qquad \Sigma_2 := \sum_{i=1}^{i=4} M_i^2.$$

We have seen above that in this case $(M_i,\, M_j) = 1 - 2\delta_{ij}$. In particular, it implies that $(\Sigma_1,\, M_i) = 2$ and $(\Sigma_1,\, \Sigma_1) = 8$. Further, taking the inner product of both sides of (1.5.10) with M_j and making a summation over i, we obtain

$$(\Sigma_2,\, M_j) = \mathrm{tr}\, \Sigma_1. \tag{6.4.11}$$

On the other hand, we have $\Sigma_1^2 = \Sigma_1\cdot \mathrm{tr}\, \Sigma_1 - 8\cdot 1$. Taking inner product with M_j, we get

$$(\Sigma_1^2,\, M_j) = 2\mathrm{tr}\, \Sigma_1 4\mathrm{tr}\, M_j. \tag{6.4.12}$$

Subtracting from (6.4.12) twice (6.4.11), we obtain finally

$$(\Sigma_1^2 - 2\Sigma_2,\, M_j) = -8(1,\, M_j), \quad \text{or} \quad (\Sigma_1^2 - 2\Sigma_2 + 8\cdot 1,\, M_j) = 0.$$

Since M_i form a basis in the space of Hermitian matrices, we get the desired relation (6.4.7). □

The relation (6.4.7) can be considered as the matrix form of the Descartes theorem. It gives us the information not only about radii of tangent discs but also about their configuration.

We mention the following corollary which is useful in computations.

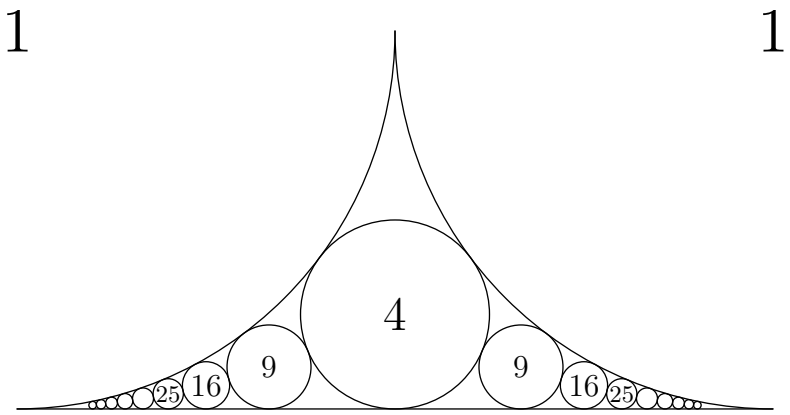

FIGURE 6.10. Quadratic sequences of curvatures

THEOREM 6.4. *Let D_+ and D_- be two tangent discs and let M_+, M_- be corresponding matrices. Suppose, the sequence of discs $\{D_k\}$, $k \in \mathbb{Z}$, has the following property: any D_k is tangent to $D_\pm$ and to $D_{k\pm1}$.*

Then the corresponding sequence of matrices $\{M_k\}$, $k \in \mathbb{Z}$, is quadratic in the parameter k:

(6.4.13)

$$M_k = A\cdot k^2 + B\cdot k + C \quad \textit{where} \quad A = M_+ + M_-,\ B = \frac{M_1 - M_{-1}}{2},\ C = M_0.$$

The illustration of this theorem you can see in pictures on Fig. 6.10 and 7.4.

6.5. Integral solutions to Descartes equation

Here we consider the arithmetic properties of the set of solutions to Descartes equation (5.1.3). Make the following change of variables:

$$\begin{aligned} t &= \frac{c_0 + c_1 + c_2 + c_3}{2}, \quad x = \frac{c_0 + c_1 - c_2 - c_3}{2}, \\ y &= \frac{c_0 - c_1 + c_2 - c_3}{2}, \quad z = \frac{c_0 - c_1 - c_2 + c_3}{2}. \end{aligned} \tag{6.5.1}$$

Then we have

$$t^2 - x^2 - y^2 - z^2 = \frac{(c_0 + c_1 + c_2 + c_3)^2}{2} - (c_0^2 + c_1^2 + c_2^2 + c_3^2)$$

and the equation (5.1.3) becomes

$$t^2 - x^2 - y^2 - z^2 = 0. \tag{6.5.2}$$

In other words, the solutions to (5.1.3) correspond to light vectors in Minsk space.

LEMMA 6.4. *The integral solutions to (5.1.3) correspond to integral light vectors in $\mathbb{R}^{1,3}$ (i.e. light vectors with integral coordinates).*

PROOF. From (5.1.3) it is clear that the sums $c_0 \pm c_1 \pm c_2 \pm c_3$ are always even. So, to any integral solution to (5.1.3) corresponds to a light vector p with integral coordinates. Conversely, from (6.5.2) it follows that the sum $t \pm x \pm y \pm z$ is always even. Therefore, from the equations

$$c_0 = \frac{t+x+y+z}{2}, \quad c_1 = \frac{t+x-y-z}{2},$$
$$c_2 = \frac{t-x+y-z}{2}, \quad c_3 = \frac{t-x-y+z}{2}$$

we deduce that any integral light vector corresponds to an integral solution to (5.1.3). □

Thus, we come to the

PROBLEM 6. Describe the set of integral points on the light cone in $\mathbb{R}^{1,3}$.

The solution for the analogous problem for rational points is well-known. To any rational point $(t,\, x,\, y,\, z)$ of the light cone there corresponds a rational point $\left(\frac{x}{t},\, \frac{y}{t},\, \frac{z}{t}\right)$ of S^2. The stereographic projection sends the point $(\frac{x}{t},\, \frac{y}{t},\, \frac{z}{t}) \in S^2$ to a point $\frac{x+iy}{t-z} \in P^1(\mathbb{Q}[i])$.

Conversely, any $(r+is) \in P^1(\mathbb{Q}[i])$ comes from a rational point

$$\left(\frac{2r}{r^2+s^2+1}, \quad \frac{2s}{r^2+s^2+1}, \quad \frac{r^2+s^2-1}{r^2+s^2+1}\right) \quad \in \quad S^2.$$

Putting $r = \frac{k}{n}$, $s = \frac{m}{n}$, we see that any integral vector on the light cone in $\mathbb{R}^{1,3}$ is proportional (but not necessarily equal) to the vector

$$t = k^2+m^2+n^2, \quad x = 2kn, \quad y = 2mn, \quad z = k^2+m^2-n^2 \tag{6.5.3}$$

with integer k, m, n.

Note, that for any integral light vector p all its multiples np, $n \in \mathbb{Z}$, are also integral light vectors. So, we can restrict ourselves to the study of **primitive** vectors, for which the greatest common divisor of coordinates is equal to 1.

LEMMA 6.5. *Any primitive integral light vector p must have an odd coordinate t and exactly one odd coordinate among x, y, z.*

PROOF. If t is even, then $x^2+y^2+z^2$ is divisible by 4. Since any square has residue 0 or 1 mod 4, it follows that all x, y, z must be even. But then p is not primitive.

If t is odd, then $x^2+y^2+z^2 \equiv 1 \mod 4$. It follows that exactly one of the numbers x, y, z is odd. □

PROBLEM 7. Find a convenient parametrization of all primitive integral light vectors.

For instance, assume that t, z are odd and x, y are even. Is it true that (1.6.3) holds for some relatively prime k, l, m?

Now, consider the subgroup Γ of the Lorentz group G which preserves the set of integral light vectors.

EXERCISE 37. Show that Γ coincides with the group $SO^+(1,3;\mathbb{Z})$ of matrices with integral entries in $SO^+(1,3;\mathbb{R})$.

HINT. Let $g \in \Gamma$. Show that a sum and a difference of any two columns of g is an integer vector and the same property holds for row vectors. Check that coordinates of an integer light vector can not be all odd.

The group Γ acts on the set of all integral light vectors and preserves the subset P of primitive vectors.

EXERCISE 38. a) Find the index of $PSL(2, \mathbb{Z}[i])$ in $PGL(2, \mathbb{Z}[i])$.
b)* What are the images of these subgroups in $O_+(1, 3; \mathbb{R})$?

EXERCISE 39. Show that the homomorphism $\pi : PGL(2, \mathbb{C}) \to SO^+(1, 3; \mathbb{R})$ can be extended to a homomorphism $\overline{\pi} : \overline{G} \to O_+(1, 3; \mathbb{R})$.

HINT. Show that one can take the diagonal matrix diag $(1, 1, -1, 1)$ as the image under $\overline{\pi}$ of the element $s \in G$ acting as complex conjugation.

PROBLEM 8. Describe the Γ-orbits in P.

Info H. Structure of some groups generated by reflections

The theory of groups generated by reflections is a big and very interesting domain in modern mathematics. We consider here only some facts we needed in relation to the Apollonian gaskets.

First, we describe the structure of the so-called **free** group F_n with n generators $x_1, x_2, \ldots, x_n$. This group may be characterized by the following universal property.

For any group G with n generators $y_1, y_2, \ldots, y_n$ there exists a unique homomorphism α of F_n onto G such that $\alpha(x_i) = y_i,\ 1 \le i \le n$.

Let us show that such group exists and is unique up to isomorphism. Indeed, if there are two such groups, F_n with generators $x_1, x_2, \ldots, x_n$ and F'_n with generators $x'_1, x'_2, \ldots, x'_n$, then from the universal property we deduce that there are homomorphisms $\alpha : F_n \to F'_n$ and $\alpha' : F'_n \to F_n$ such that $\alpha(x_i) = x'_i$ and $\alpha'(x'_i) = x_i$. Consider the composition $\alpha' \circ \alpha$. It is a homomorphism of F_n onto itself, preserving the generators. The universal property implies that this homomorphism is identity. The same is true for the composition $\alpha \circ \alpha'$. Hence, F_n and F'_n are isomorphic.

Now, prove the existence. For this end we consider the collection W_n of all words in the alphabet $x_1, x_1^{-1}, \ldots, x_n, x_n^{-1}$ satisfying the condition:

(*) *the letters x_i and x_i^{-1} can not be neighbors*

We denote the length of a word w by $l(w)$. Let $W_n^{(k)}$ be the set of all words of the length k in W_n. It is clear that W_0 contains only the empty word, and W_1 contains $2n$ one-letter words.

EXERCISE 40. Show that $\#(W_n^{(k)}) = 2n(2n-1)^{k-1}$ for $k \geq 1$.

We want to introduce a group structure on W_n. We define the product $w_1 w_2$ of two words w_1, w_2 by induction on the length $l(w_1)$ of the first factor. Namely, if $l(w_1) = 0$, i.e. if w_1 is an empty word, we put $w_1 w_2 := w_2$.

Now assume that the product is defined for $l(w_1) < k$ and consider the case $l(w_1) = k \geq 1$. Let the last letter of w_1 be $x_i^{\epsilon_1}$, $1 \leq i \leq n$, $\epsilon_1 = \pm 1$, and the first letter of w_2 be $x_j^{\epsilon_2}$, $1 \leq j \leq n$, $\epsilon_2 = \pm 1$.

If $i \neq j$ or $i = j, \epsilon_1 + \epsilon_2 \neq 0$, we define the product $w_1 w_2$ just as a juxtaposition (concatenation) of w_1 and w_2. This new word has length $l(w_1) + l(w_2)$ and satisfy the condition (*).

If $i = j$ and $\epsilon_1 + \epsilon_2 = 0$, we denote by $\tilde{w}_1$ (resp. $\tilde{w}_2$) the word obtained from w_1 (resp. w_2) by removing the last (resp. first) letter. Then we put $w_1 w_2 := \tilde{w}_1 \tilde{w}_2$. For example, if $w_1 = x_1$, $w_2 = x_1^{-1} x_2$, we have $\tilde{w}_1 = \emptyset$, $\tilde{w}_2 = x_2$ and $w_1 w_2 = x_2$.

From this definition it easily follows, that always $l(w_1 w_2) \leq l(w_1) + l(w_2)$ and $l(w_1 w_2) \equiv l(w_1) + l(w_2) \mod 2$

To check that W_n is a group with respect to the product defined above, it remains to prove that the operation defined above is associative (induction on the length of the middle factor), admits a unit (empty word) and an inverse element (the same word written back to front with opposite exponents). Traditionally, it is left to a reader.

Let us check that the group W_n has the universal property. Indeed, if G is any group generated by $x_1, x_2, \ldots, x_n$, there is a unique homomorphism $\alpha : W_n \to G$ such that $\alpha(\{x_i\}) = x_i$. (Here $\{x_i\}$ denotes a one letter word). Namely, for a word $w = x_{i_1}^{\epsilon_1} x_{i_2}^{\epsilon_2} \ldots x_{i_k}^{\epsilon_k}$ we must put $\alpha(w) = x_{i_1}^{\epsilon_1} \cdot x_{i_2}^{\epsilon_2} \cdot \cdots \cdot x_{i_k}^{\epsilon_k}$ where the sign "$\cdot$" denotes the multiplication in G. On the other hand, it is easy to check that the so-defined map α is indeed a homomorphism of W_n onto G. We showed the existence of a free group F_n and at the same time proved

PROPOSITION H.1. *Any element of F_n can be uniquely written in the form*

$$g = x_{i_1}^{\epsilon_1} x_{i_2}^{\epsilon_2} \ldots x_{i_k}^{\epsilon_k} \tag{H.1}$$

where the condition (*) is satisfied.

We need also another family of groups Γ_n, $n \geq 1$, which are freely generated by n involutions $s_1, \ldots, s_n$. By definition, the group Γ_n possesses another universal property.

For any group G generated by n involutions $t_1, \dots, t_n$ there exists a unique homomorphism α of Γ_n onto G such that $\alpha(s_i) = t_i,\ 1 \le i \le n$.

The existence and uniqueness (up to isomorphism) of the group Γ_n can be proved in the same way as for F_n. The only difference is that the set W_n now consists of all words in the alphabet $s_1, \dots, s_n$ without repetition of letters.

PROPOSITION H.2. *Any element of Γ_n can be uniquely written in the form*

$$g = s_{i_1} s_{i_2} \cdots s_{i_k},\ k \ge 0, \quad \text{where } i_a \ne i_{a+1} \quad \text{for} \quad 1 \le a \le k-1. \tag{H.2}$$

EXERCISE 41. a) Show that in this case $\#(W_n^{(k)}) = \begin{cases} 1 & \text{for} \quad k = 0 \\ n(n-1)^{k-1} & \text{for} \quad k \ge 1. \end{cases}$

b) Show that Γ_n is isomorphic to F_n/J where F_n is a free group with generators $s_1, \dots, s_n$ and J is the minimal normal subgroup in F_n which contains $s_1^2, \dots, s_n^2$.

THEOREM H.5. *Any non-trivial (i.e., different from e) involution in Γ_n is conjugate to exactly one of generators $s_1, \dots, s_n$.*

PROOF. Let $g \in \Gamma_n$ be an involution. According to Proposition 5, it can be written in the form $g = s_{i_1} s_{i_2} \dots s_{i_n}$. Then $g^{-1} = s_{i_n} s_{i_{n-1}} \dots s_{i_1}$. But $g^{-1} = g$, hence $s_{i_{n-k}} = s_{i_{k+1}}$ for $k = 0, 1, \dots, n-1$.

For $n = 2k$ even, it follows that $k = 0$ and g is an empty word.

For $n = 2k-1$ odd we have $g = w s_{i_k} w^{-1}$ where $w = s_{i_1} \dots s_{i_{k-1}}$. Hence, g is conjugate to s_{i_k}.

Finally, show that s_i is not conjugate to s_j for $i \ne j$. Assume the contrary. Then there is a word w such that $w s_i = s_j w$. Let w_0 be a shortest of such words. From the equation $w_0 s_i = s_j w_0$ we conclude that the first letter of w_0 is s_j and the last letter of w_0 is s_i. Hence, $w_0 = s_j w' s_i$ for some word w'. Then we get $s_j w' = w' s_i$, which is impossible since $l(w') = l(w_0) - 2 < l(w_0)$. □

For small values of n the group Γ_n admits a simpler description. E.g., for $n = 1$ the group Γ_1 is simply a group $\mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$ of order 2.

For $n = 2$ the group Γ_2 is isomorphic to the group $\mathrm{Aff}(1, \mathbb{Z})$ of affine transformations of the integer lattice. It has a matrix realization by matrices of the form $\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$ where $a = \pm 1,\ b \in \mathbb{Z}$. We leave to a reader to check that the matrices $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}$ can be taken as generating involutions s_1, s_2.

For $n = 3$ the group Γ_3 can be realized as a discrete group of transformation acting on the Lobachevsky (=hyperbolic) plane L. Consider e.g.

FIGURE H.11. The action of Γ_3 on L

the Poincaré model of L as the upper half-plane $y > 0$ (see **Info J** below). The three generators of Γ_3 are reflections in three pairwise tangent mirrors. For example, we can take the unit circle M_0 as one of the mirrors and two vertical lines $M_{\pm 1} : x = \pm 1$ as two others. These mirrors bound a triangle T of finite area with 3 infinite vertices. For any word w without repetitions let us denote by T_w the image of T under an element $\gamma \in \Gamma_3$ corresponding to the word w.

It can be proved by induction on $l(w)$ that the triangles T_w are all different, have no common inner points and cover the whole plane.

The case $n = 4$ is more difficult and exactly this case occurs in our study. Moreover, the group Γ_4 arises in two different ways which we discuss in Section 6.1.

CHAPTER 7

Arithmetic properties of Apollonian gaskets

Here we study some arithmetic questions arising when we consider curvatures of discs which constitute an Apollonian gasket.

7.1. The structure of $\overline{\mathbb{Q}}$

We want here to investigate the set $P^1(\mathbb{Q}) = \overline{\mathbb{Q}}$ of rational numbers including the infinite point ∞. It can be called a **rational circle**.

First, think about how to parametrize $\overline{\mathbb{Q}}$. Any number $r \in \overline{\mathbb{Q}}$ can be written in the form $\frac{p}{q}$, where $p,\, q \in \mathbb{Z}$. But the map $\alpha : \mathbb{Z}\times\mathbb{Z} \longrightarrow \overline{\mathbb{Q}}, \quad \alpha(p,\, q) = \frac{p}{q}$ is surjective but by no means injective.

We can impose the condition $\gcd(p,\, q) = 1$, that is p and q are relatively prime, or else the fraction $\frac{p}{q}$ is in lowest terms. Note, however, that the set X of relatively prime pairs $(p,\, q)$ is itself a rather complicated object. The map α, restricted to X, will be "two to one": $\alpha^{-1}(r) = \pm(p,\, q)$. And there is no natural way to choose exactly one representative from every pair $\{(p,\, q), \quad (-p,\, -q)\}$. Though, for all $r = \frac{p}{q} \in \mathbb{Q}$ we can assume $q > 0$. But for $q = 0$ there is no preference between $p = \pm 1$.

REMARK 6. For analytically minded reader, we can say that the situation here is similar to the Riemann surface of the function $f(w) = \sqrt{w}$. The map $z \mapsto w = z^2$ has two preimages for any $w \in \mathbb{C}^\times$, but this double-valued function do not admit any analytic (or even continuous) single-valued branch.

♡

REMARK 7. A remarkable way to label all positive rational numbers was discovered recently by Neil Calkin and Herbert Wilf ("Recounting the rationals", The American Mathematical Monthly, 107 (2000),pp.360-363.) Let $\mathbf{b}(n)$ be the number of partition of an integer $n \geq 0$ into powers of 2, no power of 2 being used more than twice. Than the ratio $r_n = \frac{\mathbf{b}(n)}{\mathbf{b}(n+1)}$ takes any positive rational value exactly once! The initial piece of this numeration is:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\mathbf{b}(n)$	1	1	2	1	3	2	3	1	4	3	5	2	5	3	4	1	5
r_n	1	$\frac{1}{2}$	2	$\frac{1}{3}$	$\frac{3}{2}$	$\frac{2}{3}$	3	$\frac{1}{4}$	$\frac{4}{3}$	$\frac{3}{5}$	$\frac{5}{2}$	$\frac{2}{5}$	$\frac{5}{3}$	$\frac{3}{4}$	4	$\frac{1}{5}$	$\frac{5}{4}$

n	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
$\mathbf{b}(n)$	7	3	8	5	7	2	7	5	8	3	7	4	5	1	6	5	4
r_n	$\frac{4}{7}$	$\frac{7}{3}$	$\frac{3}{8}$	$\frac{8}{5}$	$\frac{5}{7}$	$\frac{7}{2}$	$\frac{2}{7}$	$\frac{7}{5}$	$\frac{5}{8}$	$\frac{8}{3}$	$\frac{3}{7}$	$\frac{7}{4}$	$\frac{4}{5}$	5	$\frac{1}{6}$	$\frac{6}{5}$	$\frac{5}{9}$

It is interesting to compare this numeration with the one giving by Farey series (see below).

♡

Our next step in the study of $\overline{\mathbb{Q}}$ is the introduction of a natural distance between points. In the following we tacitly assume that all rational numbers are written in lowest terms.

Let us call two numbers $r_i = \frac{p_i}{q_i}$, $i = 1, 2$, from $\overline{\mathbb{Q}}$ **friendly** if the following equivalent conditions are satisfied:

$$a)\ |p_1q_2 - p_2q_1| = 1, \qquad b)\ |r_1 - r_2| = \frac{1}{|q_1q_2|}. \tag{7.1.1}$$

It is worth to mention that the friendship relation **is not** an equivalence relation[1]: every integer k is friendly to ∞ but only neighbor integers are friendly to each other.

Note, that the group $PGL(2, \mathbb{Z})$ acts on $\overline{\mathbb{Q}}$ by fraction-linear transformations and this action preserves the friendship relation. We can often use this fact in our study.

LEMMA 7.1. *The group $PSL(2, \mathbb{Z})$ acts simply transitively on the set of all ordered pairs of friendly numbers from $\overline{\mathbb{Q}}$. The group $PGL(2, \mathbb{Z})$ acts transitively but with a non-trivial stabilizer isomorphic to $\mathbb{Z}_2$.*

PROOF. Let $r_i = \frac{p_i}{q_i}$, $i = 1, 2$, be a pair of friendly numbers. Assume for definiteness that $p_1q_2 - p_2q_1 = 1$. We have to show that there is a unique element γ of $PSL(2, \mathbb{Z})$ which sends the standard friendly pair $(\infty, 0)$ to the given pair (r_1, r_2). Let $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be a representative of γ in $SL(2, \mathbb{Z})$. Then we have $\gamma(0) = \frac{b}{d}$, $\quad \gamma(\infty) = \frac{a}{c}$.

The conditions $\gamma(\infty) = r_1$, $\gamma(0) = r_2$ imply $(a, c) = k_1 \cdot (p_1, q_1)$, $\quad (b, d) = k_2 \cdot (p_2, q_2)$. Therefore, $1 = \det g = ad - bc = k_1k_2 \cdot (p_1q_2 - p_2q_1)^{-1} = k_1k_2$ and $k_1 = k_2 = \pm 1$. Hence, $g = \pm \begin{pmatrix} p_1 & p_2 \\ q_1 & q_2 \end{pmatrix}$ is determined up to sign and defines the unique element of $PSL(2, \mathbb{Z})$.

The stabilizer of the pair $(0, \infty)$ in $PGL(2, \mathbb{Z})$ consists of classes of matrices $\begin{pmatrix} 1 & 0 \\ 0 & \pm 1 \end{pmatrix}$. □

EXERCISE 42. Describe all numbers which are friendly

a) to 0; b) to ∞; c) to 1.

[1]As well as in the real life.

We define a **distance** in the set $\overline{\mathbb{Q}}$ by the following way. Given two numbers r' and r'', denote by $d(r', r'')$ the minimal $n \in \mathbb{Z}_+$ for which there exists a chain $r' = r_0, r_1, \ldots, r_{n-1}, r_n = r''$ such that for all k the number r_k is friendly to $r_{k\pm1}$ for $1 \leq k \leq n-1$.

EXERCISE 43. a) Show that $(\overline{\mathbb{Q}}, d)$ is a discrete metric space where the group $PGL(2, \mathbb{Z})$ acts by isometries.

b) Find the stabilizer of the point ∞.

ANSWER. b) The group $\mathrm{Aff}(1, \mathbb{Z})$ of transformations $r \mapsto ar + b,\ a = \pm 1,\ b \in \mathbb{Z}$.

EXERCISE 44. Compute the distances

a) $d(\infty, n)$; b) $d(0, n)$; c) $d(0, \frac{5}{8})$.

ANSWER. a) 1; b) 0 for $n = 0$, 1 for $n = \pm 1$, 2 for $|n| > 1$; c) 4.

EXERCISE 45. . a) Show that for any r', $r'' \in \overline{\mathbb{Q}}$ the distance $d(r', r'')$ is finite.

b) Is the metric space $\overline{\mathbb{Q}}$ bounded?

ANSWER. a) Cf. theorem 6 below; c) No.

Rather interesting and non-trivial problems arise when we consider the geometry of balls and spheres in $\overline{\mathbb{Q}}$. As usual, we define a **ball** with the center a and radius r as the set $B_r(a) = \{b \in \overline{\mathbb{Q}} \mid d(a, b) \leq r\}$. Analogously, a **sphere** is the set $S_r(a) = \{b \in \overline{\mathbb{Q}} \mid d(a, b) = r\}$.

THEOREM 7.1. *The ball $B_n(\infty)$ consists of all rational numbers which can be written as a continuous fraction of length n, i.e. as*

$$r = k_1 + \cfrac{1}{k_2 + \cfrac{1}{k_3 + \cfrac{1}{\ddots\, k_{n-1} + \cfrac{1}{k_n}}}} \cdots \tag{7.1.2}$$

where k_i are arbitrary integers (positive or negative).

PROOF. First of all, let us show, that for any r of the form (7.1.2) the distance $d(\infty, r)$ does not exceed n. We do it by induction on n.

For $n = 1$ it follows from the exercise 44. Assume that the theorem is true for all continuous fractions of length $\leq n-1$ and consider a fraction of length n given by (7.1.2). Denote by r' the number $\frac{1}{r-k_1}$. It is clear that r' is represented by a continuous fraction of length $n-1$, hence, $d(\infty, r') \leq n-1$.

Now, from the invariance of the distance with respect to shifts $r \mapsto r+k$, $k \in \mathbb{Z}$, and with respect to the inversion $r \mapsto r^{-1}$, we have

$$d(\infty, r) = d(\infty, r - k_1) = d(0, r') \le d(0, \infty) + d(\infty, r') \le 1 + (n-1) = n.$$

The first sign $\le$ is just the triangle inequality and the second follows from Exercise 44 a) and from induction hypothesis. □

The structure of spheres is a more delicate question. The "complexity" of a sphere is growing with its radius.

For instance, $S_1(\infty) = \mathbb{Z}$. It is an homogeneous space with respect to the group $\mathrm{Aff}(1, \mathbb{Z})$ which plays the role of the "group of rotations" around the infinite point – see Exercise 44 a).

The sphere $S_2(\infty)$ consist of points $k_1 + \frac{1}{k_2}$ where k_1, $k_2 \in \mathbb{Z}$ and $k_2 \neq 0, \pm 1$. Under the action of $\mathrm{Aff}(1, \mathbb{Z})$ it splits into infinitely many orbits Ω_m, numerated by number $m = |k_2| \ge 2$. The stabilizer of the point $k + \frac{1}{m} \in \Omega_m$ is trivial for $m > 2$ and contains one non-unit element $r \mapsto 2k + 1 - r$ for $m = 2$.

PROBLEM 9. Describe the orbits of $\mathrm{Aff}(1, \mathbb{Z})$ on the sphere $S_k(\infty)$ for $k > 2$.

7.2. Rational parametrization of circles

It is well-known that a circle as a real algebraic manifold is rationally equivalent to a real projective line. It means that one can establish a bijection between a circle and a line, using rational functions with rational coefficients.

E.g. the circle $x^2 + y^2 = 1$ can be identified with a projective line with the parameter t as follows:

$$x = \frac{t^2 - 1}{t^2 + 1}, \qquad y = \frac{2t}{t^2 + 1}; \qquad t = \frac{y}{1 - x} = \frac{1 + x}{y}. \tag{7.2.1}$$

In particular, when t runs through all rational numbers (including ∞), the corresponding points (x, y) run through all rational points[2] of the circle.

From this one can derive the well-known description of primitive integral solutions to the equation $x^2 + y^2 = z^2$. Namely, in every primitive solution exactly one of numbers x, y is even. Assume, it is y; then there are relatively prime numbers a, b such that

$$x = a^2 - b^2, \quad y = 2ab, \quad \pm z = a^2 + b^2 \tag{7.2.2}$$

Analogously, the projectivization of the future light cone in $\mathbb{R}^{1,3}$ is nothing but 2-dimensional sphere which is rationally equivalent to a completed

[2]I.e., points with rational coordinates.

2-dimensional plane. Therefore, all future light vectors (t, x, y, z) with integral non-negative coefficients can be written up to positive proportionality in the form

$$t = k^2 + l^2 + m^2, \quad x = 2km, \quad y = 2lm, \quad z = |k^2 + l^2 - m^2|. \tag{7.2.3}$$

I do not know, if any integral solution can be written exactly in the form (7.2.3) for some integers k, l, m with $\gcd(k, l, m) = 1$.

Next, we take into account that on the real projective line $\overline{\mathbb{R}}$ there is a natural orientation. For our goals it is convenient to define the orientation as a cyclic order for every three distinct points $x_1, x_2, x_3 \in \overline{\mathbb{R}}$. Geometrically, this order means that going from x_1 to x_3 in the positive direction, we pass x_2 on our way. We shall also use the expression "x_2 is between x_1 and x_3". Note, that in this case x_2 **is not** between x_3 and x_1.

EXERCISE 46. a) Show that in case when all x_1, x_2, x_3 are finite (i.e. $\neq \infty$) the statement "x_2 is between x_1 and x_3" is equivalent to the inequality

$$(x_1 - x_2)(x_2 - x_3)(x_3 - x_1) \quad > \quad 0.$$

b) Which of the following are true?
i) 1 is between 0 and ∞;
ii) ∞ is between 0 and 1;
iii) -1 is between 0 and ∞.

Now, we introduce a new operation[3] of "inserting" on $\overline{\mathbb{R}}$. It associates to a ordered pair of rational numbers (r_1, r_2) a third number denoted by $r_1 \downarrow r_2$ so that

$$r_1 \downarrow r_2 := \frac{p_1 + p_2}{q_1 + q_2}, \quad \text{if} \quad r_1 = \frac{p_1}{q_1},\ r_2 = \frac{p_2}{q_2} \tag{7.2.4}$$

where the signs of p_i and q_i are chosen so that $r_1 \downarrow r_2$ is between r_1 and r_2.

EXERCISE 47. Compute the following expressions:
a) $0 \downarrow \infty$; b) $\infty \downarrow 0$; c) $\infty \downarrow -2$; d) $1 \downarrow 2$; e) $2 \downarrow 1$; f) $\frac{1}{2} \downarrow -\frac{1}{3}$.

ANSWER. a) 1; b) -1; c) -3; d) $\frac{3}{2}$; e) ∞; f) -2.

The operation $\downarrow$ has especially nice properties when r_1 and r_2 are friendly numbers. In this case the number $r_1 \downarrow r_2$ is evidently friendly to both r_1 and r_2.

EXERCISE 48. . Show that for friendly numbers r_1, r_2 the number $r_1 \downarrow r_2$ is a unique rational number between r_1 and r_2 (in the sense of the cyclic order described above) which is friendly to both of them.

[3]I learned from R. Borcherds, that this operation is known to mathematicians in England as "English major addition". It is also a subject of one of the standard jokes quoted on Gelfand Seminar.

These consideration lead to a notion of so-called **Farey series.** The standard Farey series F^n of rank n by definition consists of all rational numbers $0 < \frac{p}{q} < 1$ with $1 \le q \le n$ written in increasing order. The number of terms in F^n is equal to $\sum_{k=2}^{n} \varphi(k)$ where $\varphi(k)$ is the Euler totient function which counts the number of integers between 1 and k and prime to k. It is given by the formula

$$\varphi(n) = n \cdot \prod_{p|n}(1 - p^{-1}) \quad \text{where } p \text{ runs through all prime divisors of } n.$$

E.g., the Farey series F^5 contains $\varphi(2)+\varphi(3)+\varphi(4)+\varphi(5) = 1+2+2+4 = 9$ terms:

$$\frac{1}{5},\quad \frac{1}{4},\quad \frac{1}{3},\quad \frac{2}{5},\quad \frac{1}{2},\quad \frac{3}{5},\quad \frac{2}{3},\quad \frac{3}{4},\quad \frac{4}{5}.$$

We refer to [Nev49] for many known facts about standard Farey series, mentioning only some of them here.

EXERCISE 49. Show that neighbor terms of Farey series are friendly numbers.

For our goals we introduce a slightly different definition. Namely, the **modified Farey series** $F^{(n)} \subset \overline{\mathbb{R}}$ is defined as follows.

The series $F^{(0)}$ consists of three numbers: 0, 1 and ∞ with given cyclic order. The series $F^{(n+1)},\ n \ge 1$, is obtained from $F^{(n)}$ by inserting between any consecutive numbers a, b the number $a \downarrow b$. So, the modified Farey series $F^{(n)}$ consists of $3 \cdot 2^n$ cyclic ordered numbers. We denote by $f_k^{(n)}$, $1 - 2^n \le k \le 2^{n+1}$ the k-th member of $F^{(n)}$. So, for any n, we have $f_0^{(n)} = 0$, $\quad f_{2^n}^{(n)} = 1$, $\quad f_{2^{n+1}}^{(n)} = \infty$.

The modified Farey series of rank ≤ 4 are shown below

k:					0				1				2
$f_k^{(0)}$:					$\frac{0}{1}$				$\frac{1}{1}$				$\frac{1}{0}$
k:			-1		0		1		2		3		4
$f_k^{(1)}$:			$-\frac{1}{1}$		$\frac{0}{1}$		$\frac{1}{2}$		$\frac{1}{1}$		$\frac{2}{1}$		$\frac{1}{0}$

k:	-3	-2	-1	0	1	2	3	4	5	6	7	8
$f_k^{(2)}$:	$-\frac{2}{1}$	$-\frac{1}{1}$	$-\frac{1}{2}$	$\frac{0}{1}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{1}$	$\frac{3}{2}$	$\frac{2}{1}$	$\frac{3}{1}$	$\frac{1}{0}$

k:	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$f_k^{(3)}$:	$-\frac{3}{1}$	$-\frac{2}{1}$	$-\frac{3}{2}$	$-\frac{1}{1}$	$-\frac{2}{3}$	$-\frac{1}{2}$	$-\frac{1}{3}$	$\frac{0}{1}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{2}{5}$	$\frac{1}{2}$	$\frac{3}{5}$	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{1}{1}$	$\frac{4}{3}$	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{2}{1}$	$\frac{5}{2}$	$\frac{3}{1}$	$\frac{4}{1}$	$\frac{1}{0}$

To find an explicit formula for numbers $f_k^{(n)}$ is a non-trivial problem. We shall discuss it below.

EXERCISE 50. Show that $f_k^{(n)} = f_{2k}^{(n+1)}$, so that $f_k^{(n)}$ actually depends only on a dyadic number $r = \frac{k}{2^n}$. Sometimes we shall write f_r instead of $f_k^{(n)}$.

FIGURE 7.1. Graph of the function **?**

To simplify the exposition, let us consider the part of $F^{(n)}$ between 0 and 1, i.e. members f_r with r between 0 and 1.

Note, that if we change the procedure and insert between any two numbers a, b not $a \downarrow b$, but the arithmetic mean value $\frac{a+b}{2}$, we obtain on the n-th step the arithmetic progression with $2^n + 1$ terms, starting with 0 and ending by 1. The k-th member of this progression is $a_k^{(n)} = \frac{k}{2^n}$. Or, in the same notations as above, $a_r = r$.

Now we are prepared to define a remarkable function first introduced by Hermann Minkowski. He called it a "question mark function" and denoted it by $?(x)$, see **Info E** in Part I.

THEOREM (Minkowski Theorem). *There exists a unique continuous and strictly increasing function* $?\colon [0,1] \to [0,1]$ *such that*

(7.2.5)
$$?\,(a \downarrow b) \;=\; \frac{?(a) + ?(b)}{2} \quad \textit{for all friendly rational numbers} \quad a, b \in [0,1].$$

SKETCH OF THE PROOF. The formula 7.2.5 and induction over n imply that if the desired function exists, it must have the property $?\big(f_k^{(n)}\big) = a_k^{(n)}$. It follows that $?\big(f_r\big) = r$ for all $r \in \mathbb{Z}[\frac{1}{2}] \bigcap [0,\,1]$.

On the other hand, we can define **?** on $\mathbb{Z}[\frac{1}{2}]$ by the formula $?(f_r) = r$. Since both sets $\{f_k^{(n)}\}$ and $\{a_k^{(n)}\}$ are dense in $[0,\,1]$, the function can be extended uniquely as a monotone function from $[0,\,1]$ to $[0,\,1]$. E.g., we can put

(7.2.6)
$$?(x) = \lim_{n\to\infty} ?(x_n)$$

where $\{x_n\}$ is a monotone sequence of rational numbers converging to x. □

The inverse function p to the question mark function solves the problem of computing $f_k^{(n)}$ posed above, since for any dyadic $r \in [0,\,1]$ we have $f_r = p(r)$.

On the set $\mathbb{Z}[\frac{1}{2}] \bigcap [0,\,1]$ of binary fractions the function $p(x)$ can be computed step by step using the property

(7.2.7)
$$p\left(\frac{2k+1}{2^{n+1}}\right) = p\left(\frac{k}{2^n}\right) \downarrow\; p\left(\frac{k+1}{2^n}\right)$$

which follows immediately from 7.2.5 and repeat the construction of the modified Farey series.

THEOREM 7.2. *The function* $p := ?^{-1}$ *has the following properties.*

1. a) $p(1-x) = 1 - p(x)$; *b)* $p(\frac{x}{2}) = \frac{p(x)}{1+p(x)}$; *c)* $p(\frac{1+x}{2}) = \frac{1}{2-p(x)}$.

2. $(p)'(\frac{k}{2^n}) = \infty$ *for any* n *and* $0 \leq k \leq 2^n$.

3. For any rational non-dyadic number $r \in [0,1]$ *the value* $p(r)$ *is a quadratic irrationality, i.e. has a form* $r_1 + \sqrt{r}_2$ *for some rational* r_1, r_2.

4. The following remarkable formula takes place:

(7.2.8)

$$p\left(\underbrace{0.0\ldots00}_{k_1}\underbrace{11\ldots11}_{l_1}\ldots\underbrace{00\ldots00}_{k_n}\underbrace{11\ldots11}_{l_n}\ldots\right) = \cfrac{1}{k_1+\cfrac{1}{l_1+\cfrac{1}{\ddots\, k_n+\cfrac{1}{l_n+\cfrac{1}{\ddots}}}}}\ldots$$

where in the left hand side the binary system is used while in the right hand side we use so-called continuous fraction. The formula 7.2.8 works also for finite binary fractions.[4]

Proof. (Sketch of) The relations 1 a) - c) can be derived from the following useful fact

Lemma 7.2. *Let* $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL(2,\mathbb{Z})$. *Then the transformation of* $\overline{\mathbb{Q}}$ *given by*

$$r \mapsto g \cdot r := \frac{ar+b}{cr+d}$$

commutes with the insertion operation $\downarrow$, *i.e.*

$$(g \cdot r_1) \downarrow (g \cdot r_2) \;=\; g \cdot (r_1 \downarrow r_2). \tag{7.2.9}$$

We leave the proof of this claim to the readers and make only two useful remarks, each of which can be a base for a proof.

1. The transformations in question send friendly pairs to friendly pairs.
2. The group $GL(2,\mathbb{Z})$ is generated by 2 elements:

$$g_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad g_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Now we prove the relation 1 a). Consider the following diagram

$$\begin{array}{ccc} [0,\,1] & \xrightarrow{x\mapsto 1-x} & [0,\,1] \\ {\scriptstyle p}\big\downarrow & & \big\downarrow{\scriptstyle p} \\ [0,\,1] & \xrightarrow{x\mapsto 1-x} & [0,\,1] \end{array} \tag{7.2.10}$$

The relation 1a) claims that it is commutative. To check it choose a point $x \in [0,\,1]$ which is a dyadic fraction $r = \frac{k}{2^n} = a_r$.

[4]Guess about the form of the right hand side of the formula in this case.

Then the vertical arrow sends it to $p(a_r) = f_r$ and the horizontal arrow sends this number f_r to f_{1-r} (check it, looking on a table above!)

On the other hand, the horizontal arrow sends r to $1 - r = a_{1-r}$ and then the vertical arrow sens a_{1-r} to f_{1-r}. Thus, for any number of the form $\frac{k}{2^n}$ the relation 1a) holds. By continuity, it holds everywhere.

Consider the relation 1b). It is equivalent to the commutativity of the diagram

$$\begin{array}{ccc} [0,\,1] & \xrightarrow{x\,\mapsto\,x/2} & [0,\,\frac{1}{2}] \\ {\scriptstyle p}\big\downarrow & & \big\downarrow{\scriptstyle p} \\ [0,\,1] & \xrightarrow{x\,\mapsto\,\frac{x}{1+x}} & [0,\,\frac{1}{2}] \end{array} \tag{7.2.11}$$

Here again we start with an element $r = a_r \in [0,\,1]$. The horizontal arrow sends it to $a_{r/2}$ and then the vertical arrow transforms it to $f_{r/2}$.

On the other hand, the vertical arrow sends a_r to f_r and we have to show that the horizontal arrow transforms it to $f_{r/2}$. I.e., we want to check the equality $\frac{f_r}{1+f_r} = f_{r/2}$. For this we observe that the transformation $x \mapsto \frac{x}{1+x}$ maps the segment $[0,\,1]$ to the segment $[0,\,\frac{1}{2}]$. Since it belongs to $PGL(2,\,\mathbb{Z})$, it transforms the Farey series into its part, sending f_0 and f_1 to f_0 and $f_{\frac{1}{2}}$ respectively. Then, by induction on n, we check that it sends $f_{\frac{2k}{2^n}}$ to $f_{\frac{k}{2^n}}$.

The relation 1c) can be proved in the same way using the diagram

$$\begin{array}{ccc} [0,\,1] & \xrightarrow{x\,\mapsto\,\frac{1+x}{2}} & [\frac{1}{2},\,1] \\ {\scriptstyle p}\big\downarrow & & \big\downarrow{\scriptstyle p} \\ [0,\,1] & \xrightarrow{x\,\mapsto\,\frac{1}{2-x}} & [\frac{1}{2},\,1] \end{array} \tag{7.2.12}$$

The point is that affine transformations respect halfsums while the transformations from $PGL(2,\,\mathbb{Z})$ respect insert operation. I recommend to the reader to formulate and prove some other properties of **?** and p using other diagrams.

It is also useful to extend the definition of **?** and p to the whole set $\overline{\mathbb{R}}$ by the formulae:

$$p\left(\frac{1}{x}\right) = \frac{1}{p(x)}; \qquad p(-x) = -p(x). \tag{7.2.13}$$

The property 2 we verify only at the point $x = 0$. The general case $x = \frac{k}{2^n}$ can be done similarly, or reduces to the case $x = 0$ by 1 a) – 1 c).

We have $p(0) = 0$, $p(\frac{1}{2^n}) = \frac{1}{n+1}$. So, if $\frac{1}{2^n} \le \Delta x \le \frac{1}{2^{n-1}}$, we have $\frac{1}{n+1} \le \Delta p \le \frac{1}{n}$.

Therefore, $\frac{2^{n-1}}{n+1} \le \frac{\Delta p}{\Delta x} \le \frac{2^n}{n}$ for $\frac{1}{2^n} \le \Delta x \le \frac{1}{2^{n-1}}$ and $p'(0) = +\infty$.

The statement 3 follows from the formula (2.2.8). As for this formula, it can be proved for finite fractions by induction, using the Farey series.

Note, that in the last section of Part I we used (2.2.8) as a definition of the question mark function. □

REMARK 8. Let us interpret the function $p := ?^{-1}$ as a distribution function for a probability measure μ on $[0,1]$: the measure of an interval $[a,\, b]$ is equal to $p(b) - p(a)$. This measure is a weak limit of the sequence of discrete measures μ_n, $n \geq 1$, concentrated on the subset $F^{(n)}$ so that the point $f_k^{(n)}$ has the mass $\frac{1}{2^n}$ for $1 \leq k \leq 2^n$.

It is clear that the support of μ is the whole segment [0, 1] (i.e. measure of any interval $(a,\, b) \subset [0,\, 1]$ is positive). While for an ordinary Farey series the measure defined in a similar way is uniform, in our case it is far from it. The detailed study of this measure is a very promising subject (see, e.g. [de Rha59]).

♡

EXERCISE 51. 28. Find the values of $?(x)$ and $?'(x)$ at the point $x = \frac{1}{3}$.

HINT. Using the relation $\frac{1}{2} - \frac{1}{4} + \frac{1}{8} - \frac{1}{16} + \frac{1}{32} - \frac{1}{64} + \cdots = \frac{1}{3}$, show that

$$?\left(\frac{1}{3} - \frac{1}{3\cdot 4^n}\right) = \frac{\Phi_{2n-1}}{\Phi_{2n+1}}, \qquad ?\left(\frac{1}{3} + \frac{2}{3\cdot 4^n}\right) = \frac{\Phi_{2n}}{\Phi_{2n+2}}$$

where Φ_n is the n-th Fibonacci number given by the formula

$$\Phi_n = \frac{\phi^n - (-\phi)^{-n}}{\phi + \phi^{-1}}$$

where $\phi = \frac{\sqrt{5}+1}{2} \approx 1.618...$ is so-called "golden ratio".[5]

ANSWER. $?\left(\frac{1}{3}\right) = \frac{3-\sqrt{5}}{2}$; $?'\left(\frac{1}{3}\right) = 0$.

PROBLEM 10. Is it true that $?'(x) = 0$ for all rational numbers except $a_k^{(n)}$?

We can sum up the content of this section: there is a monotone parametrization of all rational numbers in [0, 1] by more simple set of all binary fractions in the same interval.

If we remove the restriction $r \in [0,\, 1]$, we get a parametrization of $\overline{\mathbb{Q}}$ by $\overline{\mathbb{Z}[\frac{1}{2}]}$ which preserves the cyclic order on the circle introduced above.

REMARK 9. There is a interesting geometric interpretation of Farey series and of Minkowski question function. It was discovered by George de Rahm [Rh].

Consider a square $[-1,\, 1]\times[-1,\, 1] \subset \mathbb{R}^2$. Let us split every side in 3 equal parts and join the neighbor splitting points. We get an octagon with equal angles but different sides. Repeat this procedure: split every side of the octagon into 3 equal parts and join the neighbor splitting points. The result will be a convex polygon with 16 sides which is contained in the octagon.

[5] See **Info G**.

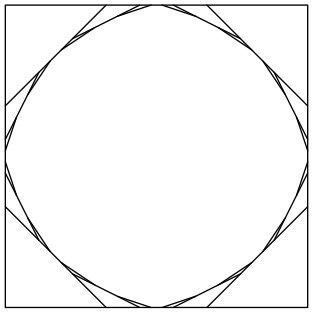

FIGURE 7.2. The de Rham curve

Proceeding in this way, we get a nested series of convex polygons Π_n, $n \geq 1$ with 2^{n+1} sides. The intersection of all these polygons is a convex domain D bounded by a C^2-smooth curve C (see Fig. 7.2). Note the following facts:

a) The centers of sides of every Π_n belong to C. Let us numerate those of them which belong to the upper half of C by numbers $r_k = \frac{k}{2^n}$, $-2^n \leq k \leq 2^n$

b) Let the upper half of C is given by the equation $y = f(x)$, $|x| \leq 1$. Let x_k be the x-coordinate of r_k. Then $f'(x_k) = f_{r_k}$, the member of the n-th Farey series.

♡

7.3. Nice parametrizations of discs tangent to a given disc

Let A be an Apollonian gasket. Choose a disc $D \in A$ corresponding to an Hermitian matrix M and consider those discs in A which are tangent to D.

The tangent points form a countable subset $T \subset \partial D$. We show later that one can parametrize points of T by rational numbers (including ∞) so that the natural cyclic order on T, as a part of ∂D, corresponds to the cyclic order on $\overline{\mathbb{Q}}$, as a part of $\overline{\mathbb{R}}$.

Let D_r be the disc tangent to D at the point $t_r \in T$ and let M_r be the corresponding Hermitian matrix.

We say that a parametrization $r \to t_r$ of T by $\overline{\mathbb{Q}}$ is **nice** if it has the following properties:

1. If $r = \frac{p}{q}$ in simple terms, then

$M_r = Ap^2 + 2Bpq + Cq^2 - M$ where A, B, C are fixed Hermitian matrices.

2. The disc D_r is tangent to $D_{r'}$ iff $r = \frac{p}{q}$ and $r' = \frac{p'}{q'}$ are friendly, i.e. iff $|pq' - p'q| = 1$.

Of course, the condition 1. and 2. are very strong and contain all the information about tangent discs. Therefore the next result is rather important.

THEOREM 7.3. *Nice parametrizations exist and have an additional property:*

Let v_0, v_1, v_2, v_3 be vectors in $\mathbb{R}^{1,3}$ corresponding to matrices $A+C$, B, $A-C$, M. Then the Gram matrix of their scalar products has the form

$$G = ||(v_i, v_j)|| = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{7.3.1}$$

First Main Example: Band gasket. Let $D = \{w \in \mathbb{C} \mid \text{Im } w \leq 0\}, D_\infty = \{w \in \mathbb{C} \mid \text{Im } w \geq 1\}$. Let D_0, D_1 be the discs of unit diameter, tangent to D at points $0, 1$ and to D_∞ at points i, $i+1$.

Then $\partial D = \overline{\mathbb{R}}$, $T = \overline{\mathbb{Q}}$. The tautological parametrization of T is nice with

$$M = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}, \quad M_{\frac{p}{q}} = \begin{pmatrix} 2p^2 & -2pq - i \\ -2pq + i & 2q^2 \end{pmatrix}, \quad D_{\frac{p}{q}} : \left| w - \frac{2pq+i}{2q^2} \right| \leq \frac{1}{2q^2}.$$

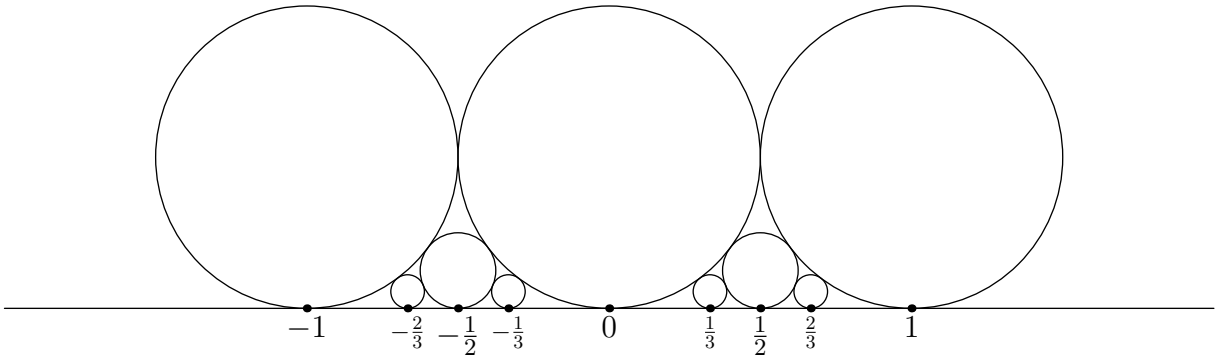

FIGURE 7.3. Nice parametrization of a line in the Band gasket

Second Main Example: Rectangular gasket. Let $D = \{w \in \mathbb{C} \mid |w| \geq 1\}$ be a complement to the open unit disc, D_0 is given by the condition $|w - \frac{1}{2}| \leq \frac{1}{2}$, D_∞ by the condition $|w + \frac{1}{2}| \leq \frac{1}{2}$ and D_1 by the condition $|w - \frac{2i}{3}| \leq \frac{1}{3}$.

Here ∂D is the unit circle and a nice parametrization is $t_r = \frac{p+iq}{p-iq}$ so that

$$M = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad M_r = \begin{pmatrix} p^2+q^2-1 & -(p+iq)^2 \\ -(p-iq)^2 & p^2+q^2+1 \end{pmatrix},$$

$$D_{\frac{p}{q}} : \left| w - \frac{(p+iq)^2}{p^2+q^2+1} \right| \leq \frac{1}{p^2+q^2+1}.$$

PROOF OF THE THEOREM 7.3. Let D_0, D_1, D_∞ be any three discs from A which are tangent to D and to each other. We associate the labels 0, 1 and ∞ to the corresponding tangent points in ∂D.

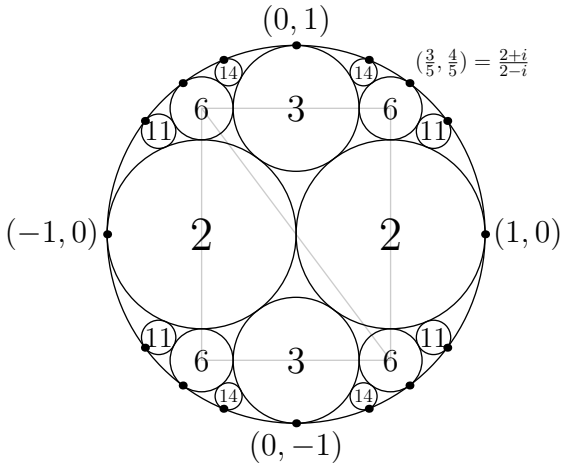

FIGURE 7.4. Nice parametrization of the outer circle in the Rectangular gasket

Then, assuming that theorem is true and parametrization is nice, we can compute A, B, C from the equations

$$M_\infty = A - M, \quad M_0 = C - M, \quad M_1 = A + 2B + C - M.$$

We get

$$A = M + M_\infty, \quad C = M + M_0, \quad B = \frac{1}{2}(M_1 - M - M_0 - M_\infty).$$

Then, using the property of matrices M_0, M_1, M_∞ and M, we can check the relation (2.3.1). From there, the statement 2 of the theorem easy follows if we define M_r using statement 1. □

Practically, the nice parametrization can be defined step by step. Assume that the discs D_{r_1} and D_{r_2} corresponding to friendly rational numbers r_1 and r_2 are already defined and are tangent to D and to each other. Then we associate to $r = r_1 \downarrow r_2$ the disc tangent to D_{r_1}, D_{r_2} and D.

Actually, there are two such discs and two possible values of $r = r_1 \downarrow r_2$; the right choice is uniquely determined by the cyclic order.

COROLLARY. *The boundary curvature of the disc tangent to D at the point $r = \frac{p}{q}$ (in the simplest form) is given by a quadratic polynomial in p, q:*

$$c(p, q) = \big(c_\infty + c\big) \cdot p^2 + \big(c_1 - c_0 - c_\infty - c\big) \cdot pq + \big(c_0 + c\big) \cdot q^2 - c \tag{7.3.2}$$

where c_i is the boundary curvature of the disc D_i.

In particular, if four pairwise tangent discs in an Apollonian gasket A have integral boundary curvatures, then all discs from A have this property.

EXERCISE 52. 29. For the triangular Apollonian gasket find the curvatures of discs tangent to the outer disc.

Answer: $c(p,q) = \frac{2(p^2-pq+q^2)}{\sqrt{3}} + 1$.

EXERCISE 53. 30. Describe the canonical parametrization for the outer circle of the triangular gasket.

HINT. Label by 0, 1, ∞ the tangent points corresponding to three maximal inner discs.

7.4. Integral Apollonian gaskets

There are many models of Apollonian gasket for which the curvatures of all circles are integers. We call them **integral** gaskets. For each such gasket we can choose the quadruple of discs such that corresponding boundary curvatures form an integral quadruple ($c_1 \geq c_2 \geq c_3 \geq c_4$) with minimal c_1. Call it **basic quadruple**.

LEMMA 7.3. *For a basic quadruple we have*

$$c_4 \leq 0, \qquad |c_4| < c_3 < \left(1 + \frac{2}{\sqrt{3}}\right)|c_4| \approx 2.1547... \cdot |c_4|.$$

PROOF. Let D_i, $1 \leq i \leq 4$, be a quadruple of pairwise tangent discs with curvatures c_i, $1 \leq i \leq 4$,. The first inequality was already proved (see Remark in the end of 5.1).

Consider now the Descartes equation (5.1.3) as a quadratic equation in c_1 with given c_2, c_3, c_4. Then we get

$$c_1 = c_2 + c_3 + c_4 \pm 2\sqrt{c_2c_3 + c_3c_4 + c_4c_2}. \tag{7.4.1}$$

Since the initial quadruple is basic, we have to choose minus sign in (7.4.1) (otherwise we could replace c_1 by smaller quantity).

The inequality $c_1 \geq c_2$ together with (2.4.1) gives $c_3+c_4 \geq 2\sqrt{c_2c_3 + c_3c_4 + c_4c_2}$, or $(c_3 - c_4)^2 \geq 4c_2(c_3 + c_4) \geq (c_3 + c_4)^2$. It can be true only when $c_4 \leq 0$.

Finally, for non-positive c_4 we have $(c_3 - c_4)^2 \geq 4c_2(c_3 + c_4) \geq 4c_3(c_3 + c_4)$, or $3c_3^2 + 6c_3c_4 + c_4^2 \leq 4c_4^2$. It gives $\sqrt{3}(c_3 + c_4) \leq -2c_4$, hence $c_3 \leq \frac{2+\sqrt{3}}{\sqrt{3}}|c_4|$. □

Here is the list of basic quadruples of small sizes, generating non-isomorphic gaskets in the order of increasing $|c_4|$:

$c_4 = 0$ (1, 1, 0, 0);
$c_4 = -1$ (3, 2, 2, −1);
$c_4 = -2$ (7, 6, 3, −2);
$c_4 = -3$ (13, 12, 4, −3), (8, 8, 5, −3);
$c_4 = -4$ (21, 20, 5, −4), (9, 9, 8, −4);
$c_4 = -5$ (31, 30, 6, −5), (18, 18, 7, −5);
$c_4 = -6$ (43, 42, 7, −6), (15, 14, 11, −6), (19, 15, 10, −6);
$c_4 = -7$ (57, 56, 8, −7), (20, 17, 12, −7), (32, 32, 9, −7);
$c_4 = -8$ (73, 72, 9, −8), (24, 21, 13, −8), (25, 25, 12, −8);
$c_4 = -9$ (91, 90, 10, −9), (50, 50, 11, −9), (22, 19, 18, −9), (27, 26, 14, −9);

$c_4 = -10$
(111, 110, 11, −10), (62, 60, 12, −10), (39, 35, 14, −10), (27, 23, 18, −10);
$c_4 = -11$
(133, 132, 12, −11), (72, 72, 13, −11), (37, 36, 16, −11), (28, 24, 21, −11).
Three general formulae:
$c_4 = -km$ $\quad (k^2 + km + m^2,\ k(k+m),\ m(k+m),\ -km)$
$c_4 = 1 - 2k$ $\quad (2k^2,\ 2k^2,\ 2k+1,\ 1-2k)$
$c_4 = -4k$ $\quad \big((2k+1)^2,\ (2k+1)^2,\ 4(k+1),\ -4k\big)$
Many other interesting facts about integral gaskets the reader can find in [G].

Info I. Möbius inversion formula

In number-theoretic computations the so-called Möbius inversion formula is frequently used. We explain here how it works.

Suppose, we have a partially ordered set X with the property: for any element $x \in X$ there are only finitely many elements which are less than x. Let now f be any real or complex-valued function on X. Define a new function F by the formula

$$F(x) = \sum_{y \le x} f(y). \tag{I.1}$$

Proposition I.1. *There exists a unique function $\widetilde{\mu}$ on $X \times X$ with the properties:*
1. $\widetilde{\mu}(x, y) = 0$ unless $x < y$ and $\mu(x, x) = 1$
2. $\widetilde{\mu}(x, x) = 1$
3. If the functions f and F are related by (I.1), then

$$F(x) = \sum_{y \le x} \widetilde{\mu}(x, y) F(y). \tag{I.2}$$

In many applications the set X is a semi-group of non-negative elements in some partially ordered abelian group G and the order relation is translation-invariant: $x < y$ is equivalent to $a + x < a + y$ for any $a \in G$. In this case μ is also translation-invariant: $\widetilde{\mu}(a+x, a+y) = \widetilde{\mu}(x, y)$, hence, can be written in the form $\mu(y-x)$ where μ is a function on G which is zero outside X. The inversion formula takes the form

$$F(x) = \sum_{y \le x} \mu(x - y) F(y) \qquad \text{(Möbius inversion formula)}. \tag{I.3}$$

We leave the proofs for the interested reader and consider only some examples which we need in our book.

Example 1. Let $G = \mathbb{Z}$ with the standard order. Then the formula (I.1) takes the form $F(n) = \sum_{m \le n} f(m)$ and the inversion formula is $f(n) =$

$F(n) - F(n-1)$. We see that in this case the proposition 5 is true and the function μ is given by

$$\mu(n) = \begin{cases} 1 & \text{if} \quad n = 0 \\ -1 & \text{if} \quad n = 1 \\ 0 & \text{otherwise} \end{cases}$$

Example 2. $G = G_1 \times G_2$ and the order on G is the product of orders on G_1 and on G_2, i.e.

$$(g_1,\, g_2) > (0,\, 0) \quad \Leftrightarrow \quad g_1 > 0 \quad \& \quad g_2 > 0.$$

Here the μ-function for G is simply the product of μ-functions for G_1 and G_2.

Note, that if G_1 and G_2 are ordered groups, $G = G_1 \times G_2$ is only partially ordered.

Example 3. $G = \mathbb{Q}^\times$ is the multiplicative group of non-zero rational numbers. The partial order is defined as follows: $r_1 \le r_2$ iff the number $\frac{r_2}{r_1}$ is an integer. So, in this case $X = \mathbb{Z}_+$ with the order relation $m < n$ iff $m|n$ (m is a divisor of n).

It is easy to see that this partially ordered group is the direct sum of the countable number of copies of $\mathbb{Z}$ with a usual order. Indeed, any element of G can be uniquely written in the form $r = \prod_{k\ge 1} p_k^{n_k}$ where p_k is the k-th prime number, $n_k \in \mathbb{Z}$ and only finite number of n_k are non-zero. The number r is an integer iff all n_k are non-negative.

Therefore, the function μ is the product of infinitely many functions from example 1. The exact definition is:

DEFINITION I.1. $\mu(n) = \begin{cases} 1 & \text{if} \quad n = 1 \\ (-1)^k & \text{if } n \text{ is a product of } k \text{ distinct primes} \\ 0 & \text{otherwise} \end{cases}$

The equation (I.3) in this case is the classical Möbius inversion formula

$$F(n) = \sum_{d|n} \mu(d) F\left(\frac{n}{d}\right). \tag{I.4}$$

As an application, we derive here the formula for the Euler φ-function.

Let us classify the numbers $k \le n$ according to the value of $d = \gcd(k,\, n)$. It is clear that $\gcd(\frac{k}{d},\, \frac{n}{d}) = 1$. It follows that the number of those k for which $\gcd(k,\, n) = d$ is equal to $\varphi(\frac{n}{d})$. We have obtained the identity

$$n = \sum_{d|n} \varphi\left(\frac{n}{d}\right).$$

Applying the Möbius inversion formula, we get

$$\varphi(n) = \sum_{d|n} \mu(d) \cdot \frac{n}{d}, \qquad \text{or} \qquad \frac{\varphi(n)}{n} = \sum_{d|n} \frac{\mu(d)}{d}. \tag{I.5}$$

♢

7.4.1. Some computations. The well-known unsolved problem is to compute the Hausdorff dimension of the Apollonian gasket and the Hausdorff measure of its different modifications (e.g. spherical or triangular gaskets). Though we know the answer for the first question with high degree of accuracy: in [?M]) it is shown that the Hausdorff dimension of the Apollonian gasket is $d = 1.308535???...$, we have no idea of the nature of this number. For example, is it irrational? Can it be expressed in terms of some logarithms as for the Cantor set or Sierpiński gasket? Has it any interesting arithmetic properties?

Another interesting problem is to compute the total area of the discs in some Apollonian gasket, which are tangent to a given disk D, e.g., to the outer disc in rectangular or triangular gasket.

We start, however, with a slightly easier problem. Consider the First Main Example of the band gasket above. We want to compute the total area of the discs in the Band gasket, which are tangent to the real axis at the points of the segment [0, 1]. More natural question, which has a simpler answer is to compute the area of the part of the unit square with vertices 0, 1, $1 + i$, i, covered by the discs, tangent to the lower side of the square.

We know that the diameter of the disc with tangent point $\frac{m}{n} \in [0, 1]$ is $\frac{1}{2n^2}$. Hence, its area is $\frac{\pi}{4n^4}$. There are $\varphi(n)$ of discs of this size. So, for the area in question we have an expression

$$A \quad = \quad \frac{\pi}{4} \cdot \sum_{n\geq 1} \frac{\varphi(n)}{n^4}. \tag{7.4.6}$$

This number can be expressed through the values of the Riemann ζ-function at points 3 and 4.

Let us use the formula for $\varphi(n)$ obtained in Info I. The formula (I.5) takes the form

$$A \quad = \quad \frac{\pi}{4} \cdot \sum_{n\geq 1} \sum_{d|n} \mu(d) \frac{d}{n^3}.$$

We denote $\frac{n}{d}$ by m and make the summation on d and m. We get

$$A \quad = \quad \frac{\pi}{4} \cdot \sum_{d\geq 1} \sum_{m\geq 1} \frac{\mu(d)}{m^3 d^4} \quad = \quad \frac{\pi}{4} \cdot \sum_{m\geq 1} \frac{1}{m^3} \cdot \sum_{d\geq 1} \frac{\mu(d)}{d^4}.$$

The sum $\sum_{m\geq 1} \frac{1}{m^3}$ is, by definition, the value $\zeta(3)$. On the other hand, the sum $\sum_{d\geq 1} \frac{\mu(d)}{d^4}$ can be written as

$$\sum_{k\geq 0} (-1)^k \cdot \sum_{1\leq i_1<i_2<\cdots<i_k} (p_{i_1} p_{i_2} \cdots p_{i_k})^{-4} = \prod_{i\geq 1} \left(1 - \frac{1}{p_i^4}\right) = \frac{1}{\sum_{n\geq 1} \frac{1}{n^4}} = \frac{1}{\zeta(4)}.$$

Finally, we get

$$A \;=\; \frac{\pi}{4} \cdot \frac{\zeta(3)}{\zeta(4)} \;=\; \frac{45\zeta(3)}{2\pi^3} \approx 0.76.$$

The total area of discs tangent to the outer disc of the rectangular gasket is equal

$$\frac{\pi}{2}\cdot\sum_{\gcd(p,q)=1}\frac{1}{(p^2+q^2+1)^2}$$

It can be expressed in terms of the ζ-function related to the Gauss field $\mathbb{Q}[i]$.

EXERCISE 54. Let Σ_m denote the sum $\sum_{\mathbb{Z}^2\setminus\{(0,0)\}}\frac{1}{(k^2+l^2)^m}$. Show that

$$\sum_{\gcd\,(p,q)=1}\frac{1}{(p^2+q^2)^m}=\frac{\Sigma_m}{\zeta(2m)}. \tag{7.4.7}$$

and

$$\sum_{\gcd\,(p,q)=1}\frac{1}{(p^2+q^2+1)^2}=\sum_{m=1}^{\infty}(-1)^{m-1}\frac{m\cdot\Sigma_{m+1}}{\zeta(2m+2)}. \tag{7.4.8}$$

CHAPTER 8

Geometric and group-theoretic approach

Info J. Hyperbolic (Lobachevsky) plane L

A hyperbolic space satisfies all axioms of Euclidean space except the famous 5-th postulate about uniqueness of parallel lines. Such a space exists in all dimensions, but here we consider only the case $n = 2$. We collect here some information about the 2-dimensional hyperbolic space, a.k.a. Lobachevsky plane L.

There are three most convenient models of L.

J.1. The first Poincaré model

Let $\mathbb{C}$ be the complex plane with a complex coordinate $z = x + iy$. Denote by H the upper half-plane of $\mathbb{C}$ given by the condition $\operatorname{Im} z > 0$. The first Poincaré model identifies L, as a set, with H. The group $\overline{G}$ of conformal transformations of both kinds (see Info F) acts on H and is, by definition, the full group of symmetries of L. So, according to F. Klein philosophy, So, by definition, geometric properties of L are those which are invariant under the group $\overline{G}$.

In particular, the distance $d(z_1,\, z_2)$ between two points $z_1,\, z_2 \in H$ must be $\overline{G}$-invariant. It turns out that this condition defines the distance uniquely up to scale.

To find the explicit formula for the distance, we can procede as follows. To any pair $p = (z_1,\, z_2)$ there corresponds a quadruple $q(p) = (z_1,\, z_2,\, \bar{z}_1,\, \bar{z}_2)$. The correspondence $p \to q(p)$ is clearly invariant under the action of $PSL(2,\, \mathbb{R})$.

On the other hand, it is well-known that for any quadruple $q = (z_1,\, z_2,\, z_3,\, z_4)$ of points in $\mathbb{C}$ the so-called **cross-ratio** $\lambda(q) := \frac{z_2 - z_3}{z_1 - z_3} : \frac{z_2 - z_4}{z_1 - z_4}$ does not change under fractional-linear transformations from $PSL(2,\, \mathbb{C})$.

Introduce the quantity

$$\Delta(p) := \lambda\big(q(p)\big) = \frac{z_2 - \bar{z}_1}{z_1 - \bar{z}_1} : \frac{z_2 - \bar{z}_2}{z_1 - \bar{z}_2} = \frac{|z_1 - \bar{z}_2|^2}{4 \operatorname{Im} z_1 \operatorname{Im} z_2}. \tag{J.1}$$

This function on the set of pairs of points in H is positive, symmetric and invariant with respect to full group $\overline{G}$. Let us clarify how it is related to the desired distance. For this end we restrict our consideration to the subset T of H consisting of points ie^{τ}, $\tau \in \mathbb{R}$. This subset is invariant under dilations $z \mapsto e^t z$ and admits a natural dilation-invariant distance $d(\tau_1,\, \tau_2) = |\tau_1 - \tau_2|$.

Compare this distance with the restriction of Δ to $T \times T$.

$$\Delta(ie^{\tau_1},\, ie^{\tau_2}) = \frac{(e^{\tau_1}+e^{\tau_2})^2}{4e^{\tau_1+\tau_2}} = \frac{1}{4}\left(e^{\tau_1-\tau_2}+2+e^{\tau_2-\tau_1}\right) = \cosh^2\left(\frac{\tau_1-\tau_2}{2}\right).$$

We come to the relation

$$\Delta(z_1,\, z_2) = \cosh^2\left(\frac{d(z_1,\, z_2)}{2}\right) = \frac{\cosh\left(d(z_1,\, z_2)\right)+1}{2}. \tag{J.2}$$

It holds on $T \times T$ and both sides are G-invariant.

Exercise 55. 31. Show that $G \cdot (T \times T) = H \times H$. More precisely any pair of points $(z_1,\, z_2)$ can be obtained by a transformation $g \in G$ from a pair $(i,\, ie^{\tau})$ for an appropriate $\tau \in \mathbb{R}$.

It follows from the exercise that the relation (J.3) holds everywhere. The simple computation leads to the final formula

$$\cosh\, d(z_1,\, z_2) = 2\Delta\,(z_1,\, z_2) - 1 = \frac{(x_1-x_2)^2+y_1^2+y_2^2}{2y_1y_2} \tag{J.3}$$

It is well-known that the area of a domain $\Omega \subset L$ and the length of a curve $C \subset L$ are given by integrals[1]

$$\text{area}\,(\Omega) = \int_\Omega \frac{dx \wedge dy}{y^2}, \qquad \text{length}\,(C) = \int_C \frac{\sqrt{(dx)^2+(dy)^2}}{y}. \tag{J.4}$$

Exercise 56. 32. Show that the geodesics, i.e. the shortest curves, are half-circles orthogonal to the real axis (including vertical rays).

Hint. Use the fact that any two points p, q on L define a unique geodesic. So, this geodesic must be invariant under any transformation $g \in G$ which preserves or permutes these two points. Apply this to the the points $p = ir$, $q = ir^{-1}$ and transformations $s:\, z \mapsto -\bar{z}$, $\quad t:\, z \mapsto -z^{-1}$.

There is a remarkable relation between the area of a triangle with geodesic sides and its angles:

$$\text{area}\,(ABC) = \pi - A - B - C. \tag{J.5}$$

Exercise 57. 33. Check the formula (J.4) for a triangle with 3 zero angles given by inequalities $\quad -a \le x \le a, \quad x^2+y^2 \ge a^2$.

Exercise 58. 34. Show that the set of points $B_r(a) = \{z \in L \;\big|\; d(z,\, a) \le r\}$ (Lobachevsky disc) in the first Poincaré model is just an ordinary disc with the center a' and the radius r'. Express a' and r' in terms of a and r.

Answer. $a' = \operatorname{Re} a + i\cosh r \cdot \operatorname{Im} a, \qquad r' = \sinh\, r \cdot \operatorname{Im} a\ .$

[1]Actually, the first integrand here is the unique (up to a scalar factor) differential 2-form which is invariant under the action of G. It is covariant under $\overline{G}$: a conformal transformation of second kind changes the sign of the form. The second integrand is the square root of the unique (also up to a scalar factor) $\overline{G}$-invariant quadratic differential form (i.e. metric) on L.

EXERCISE 59. 35. Consider an Euclidean disc $D:\ (x-a)^2+(y-b)^2 \le r^2$ on H. Find its diameter d and the area A in the sense of the hyperbolic geometry.

ANSWER. $d = \log\ \frac{b+r}{b-r}$; $\qquad A = 2\pi\left(\frac{b}{\sqrt{b^2-r^2}} - 1\right) = 4\pi \sinh^2\left(\frac{d}{4}\right).$

J.2. The second Poincaré model.

Sometimes another variant of the Poincaré model is more convenient. Namely, a Möbius transformation $h:\ w \mapsto \frac{w-i}{w+i}$, sends the real line to the unit circle and the upper half-plane H to the interior D^0 of the unit disc $D:\ x^2+y^2 \le 1$. All we said above about H can be repeated for D^0 *mutatis mutandis*.

Thus, the group $\overline{G}$, acting on the upper half-plane, is replaced by the group $\overline{G}' = h \cdot \overline{G} \cdot h^{-1}$ acting on D^0. The connected component of unit in $\overline{G}$ is the group $h \cdot PSL(2,\, \mathbb{R}) \cdot h^{-1} = PSU(1, 1;\, \mathbb{C})$.

To a pair $p' = (w_1,\, w_2) \in D^0 \times D^0$ we associate in a $\overline{G}'$-invariant way the quadruple $q'(p') = (w_1,\, w_2,\, \bar{w}_1^{-1},\, \bar{w}_2^{-1})$. Introduce the function

$$\Delta'(p) := \lambda\big(q'(p')\big) = \frac{|1 - w_1\bar{w}_2|^2}{(1-|w_1|^2)(1-|w_2|^2)} \tag{J.6}$$

The subgroup of dilation of H given by matrices $g_\tau = \begin{pmatrix} e^{\tau/2} & 0 \\ 0 & e^{-\tau/2} \end{pmatrix}$ comes to the subgroup of matrices $g'_\tau = h \cdot g_\tau \cdot h^{-1} = \begin{pmatrix} \cosh\,\tau/2 & \sinh\,\tau/2 \\ \sinh\,\tau/2 & \cosh\,\tau/2 \end{pmatrix}$. This subgroup preserves the interval $h \cdot T = T' = (-1,\, 1) \subset D^0$. Introduce the local parameter t on T' so that $x = \tanh\,\frac{t}{2}$. Then the transformation g'_τ takes a simple form $t \mapsto t + \tau$. Therefore, the invariant distance on T is $d(t_1,\, t_2) = |t_1 - t_2|$. On the other hand,

$$\Delta'\big(\tanh\,\frac{t_1}{2},\, \tanh\,\frac{t_2}{2}\big) = \frac{(1 - \tanh\,\frac{t_1}{2}\,\tanh\,\frac{t_2}{2})^2}{(1-\tanh^2\,\frac{t_1}{2})(1-\tanh^2\,\frac{t_2}{2})} = \cosh^2\left(\frac{t_1-t_2}{2}\right).$$

Then (J.2) and (J.3) take the form

$$\Delta'(w_1,\, w_2) = \cosh^2\left(\frac{d'(w_1,\, w_2)}{2}\right) = \frac{\cosh\big(d'(w_1,\, w_2)\big) + 1}{2} \tag{J.7}$$

$$\cosh\,d(w_1,\, w_2) = \frac{|1 - w_1\bar{w}_2|^2 + |w_1 - w_2|^2}{(1-|w_1|^2)(1-|w_2|^2)}. \tag{J.8}$$

The formula (J.4) is replaced by

(J.9)

$$\text{area}\,(\Omega) = \int_\Omega \frac{4\ dx \wedge dy}{(1-x^2-y^2)^2}, \qquad \text{length}\,(C) = \int_C \frac{2\sqrt{(dx)^2 + (dy)^2}}{1-x^2-y^2}.$$

The geodesics are arcs of circles orthogonal to ∂D (including the diameters of the disc). The formula (**??**J.1.5) remains to be true.

EXERCISE 60. 36. Show that the set of points $\{z \in L \mid d(z, a) \leq r\}$ (Lobachevsky disc) in the second variant of the Poincaré model is an ordinary disc with the center a' and radius r'. Express a' and r' in terms of a and r.

ANSWER. $a' = \frac{a}{\sqrt{a^2+1}\cosh r}$, $\qquad r' = \frac{(a+a^{-1})}{a\tanh r + a^{-1}\coth r}$.

EXERCISE 61. 37. Find the diameter d and the area A of a disc $D_r(a, b)$: $(x-a)^2 + (y-b)^2 \leq r^2$ in D.

ANSWER. $d = \log \frac{b+r}{b-r}$; $\qquad A = 2\pi\left(\frac{b}{\sqrt{b^2-r^2}} - 1\right) = 4\pi \sinh^2\left(\frac{d}{4}\right)$.

J.2. The Klein model. .

The extended Möbius group $\overline{G}$ is isomorphic to $PO(2,\, 1,\, \mathbb{R}) \subset PGL(3,\, \mathbb{R})$ (see Info F). Therefore, there is one more realization of the hyperbolic plane L. It is the so-called **Klein model** which we describe now.

The group $O(2, 1, \mathbb{R})$ acts on the real vector space $\mathbb{R}^{2,1}$ with coordinates X, Y, Z preserving the cone $X^2 + Y^2 = Z^2$. Consider the real projective plane $P := P^2(\mathbb{R})$ with homogeneous coordinates $(X : Y : Z)$ and local coordinates $x = \frac{X}{Z}$, $y = \frac{Y}{Z}$. The corresponding projective action of $PO(2,\, 1,\, \mathbb{R})$ on P preserves the circle $x^2 + y^2 = 1$ and the open disk $D^0 : x^2 + y^2 < 1$. It is the Klein model of L.

The explicit formula of the group action is

(J.10)

$$x \mapsto \frac{a'x + b'y + c'}{ax + by + c}, \quad y \mapsto \frac{a''x + b''y + c''}{ax + by + c} \quad \text{where} \quad g = \begin{pmatrix} a' & b' & c' \\ a'' & b'' & c'' \\ a & b & c \end{pmatrix}$$

belongs to $O(2,\, 1,\, \mathbb{R}) \subset GL(3,\, \mathbb{R})$.

We know that $g \in O(2,\, 1,\, \mathbb{R})$ iff $g^t I g = I$ where $I = \operatorname{diag}(1, 1, -1)$, or, in full details:

(J.11)

$$(a')^2 + (a'')^2 = a^2 + 1, \quad (b')^2 + (b'')^2 = b^2 + 1, \quad (c')^2 + (c'')^2 = c^2 - 1,$$
$$a'b' + a''b'' = ab, \qquad b'c' + b''c'' = bc, \qquad c'a' + c''a'' = ca.$$

EXERCISE 62. 38. a) Show that the group $O(2,\, 1,\, \mathbb{R})$ has four connected components characterized by the signs of $\det g$ and c.

b) Show that $PO(2,\, 1,\, \mathbb{R})$ has two connected components: $PSO_+(2,\, 1,\, \mathbb{R})$ and $PSO_-(2,\, 1,\, \mathbb{R})$ distinguished by the sign of $a'b'' - a''b'$.

Note, that the Klein model uses the same set D^0 and the same abstract group $\overline{G} \simeq PO(2,\, 1;\, \mathbb{R})$, as the second Poincaré model, but the group actions are different.

More precisely, there exist a smooth map $f : D^0 \to D^0$ and a homomorphism $\alpha : \overline{G} \to PO(2,\, 1,\, \mathbb{R})$ such that the following diagram is commutative: To describe the homomorphism α, consider first the connected component

of unit $G \subset \overline{G}$ which we identify with the group $PSU(1,\,1;\,\mathbb{C})$. The restriction of α to this subgroup induces the homomorphism $\widetilde{\alpha} : SU(1,\,1;\,\mathbb{C}) \to SO_+(2,\,1;\,\mathbb{R})$ which has the form:

(J.12)

$$g = \begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix} \quad \to \quad \widetilde{\alpha}(g) = \begin{pmatrix} \mathrm{Re}(a^2+b^2) & -\mathrm{Im}(a^2+b^2) & 2\,\mathrm{Re}\,(ab) \\ \mathrm{Im}(a^2+b^2) & \mathrm{Re}(a^2-b^2) & -2\,\mathrm{Im}\,(ab) \\ 2\,\mathrm{Re}\,(\overline{a}b) & -2\,\mathrm{Im}\,(\overline{a}b) & |a|^2+|b|^2 \end{pmatrix}.$$

The second connected component of $\overline{G}$ is a two-sided G-coset $c\cdot G = G\cdot c$ where c acts as the complex conjugation on D^0. From the relation $c\cdot g\cdot c = \overline{g}$ we derive that $\alpha(c) = diag(-1,\,1,\,-1) \in SO_-(2,\,1;\,\mathbb{R})$, i.e. $\alpha(c)$ acts on D^0 by the rule: $x \mapsto x$, $y \mapsto -y$.

Therefore, the horizontal diameter of D^0 is the set of fixed point of an involution $\alpha(c)$, hence, is a geodesic in the Klein model. Of course, the same is true for all other diameters.

The remarkable property of Klein model is that all geodesics are the ordinary straight lines. Indeed, the projective transformations send lines to lines (in contrast with conformal transformations which sends circles to circles).

To compute the map f we use the following particular cases of (J.3.3):

$$\widetilde{\alpha} : \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix} \quad \to \quad \begin{pmatrix} \cos\,2\theta & -\sin\,2\theta & 0 \\ \sin\,2\theta & \cos\,2\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$\widetilde{\alpha} : \begin{pmatrix} \cosh\,t & \sinh\,t \\ \sinh\,t & \cosh\,t \end{pmatrix} \quad \to \quad \begin{pmatrix} \cosh\,2t & 0 & \sinh\,2t \\ 0 & 1 & 0 \\ \sinh\,2t & 0 & \cosh\,2t \end{pmatrix}.$$

We see, that rotation to the angle 2θ in the Poincaré model corresponds to the same rotation in the Klein model.

On the contrary, the motion along the diameter

$$x \mapsto \frac{x\cosh\,t + \sinh\,t}{x\sinh\,t + \cosh\,t}, \quad \text{or, if} \quad x = \tanh\,\tau, \quad \tau \to \tau + t$$

goes to the motion

$$x \mapsto \frac{x\cosh\,2t + \sinh\,2t}{x\sinh\,2t + \cosh\,2t}, \quad \text{or} \quad \tau \to \tau + 2t$$

with doubled speed.

We conclude, that in polar coordinates $(r,\,\alpha)$ in the Poincaré model and $(\rho,\,\theta)$ in the Klein model, the diffeomorphism f take the form

(J.13) $$f(r,\,\alpha) = (\rho,\,\theta) \quad \text{where} \quad \theta = \alpha, \qquad \rho = \tanh\,\left(2\,\tanh^{-1}(r)\right).$$

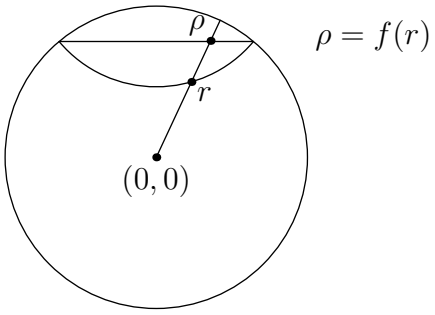

FIGURE J.1. Diffeomorphism f

EXERCISE 63. 39. Show that the relation (**??**J.3.4) between r and ρ can be written also in the following forms:

$$\text{(J.14)} \qquad a) \quad \frac{1+\rho}{1-\rho} = \left(\frac{1+r}{1-r}\right)^2; \qquad b) \quad \rho = \frac{2}{r+r^{-1}}.$$

Another interesting geometric fact is that the diffeomorphism f "straightens" arcs of circles orthogonal to the boundary, sendig them into corresponding chords (see Fig. 21).

The Klein model has two disadvantages: more complicated formula for the distance between two points and non-conformness (the angles between curves are not equal to euclidean angles on the model).

The area form and the length element for the Klein model in polar coordinates (ρ, θ) look like

(J.15)

$$\text{area}\,(\Omega) = \int_\Omega \frac{\rho^2 d\rho \wedge d\theta}{(1-\rho^2)\sqrt{1+\rho^2}}, \quad \text{length}\,(C) = \frac{1}{2}\int_C \frac{\sqrt{(d\rho)^2 + \rho^2(1-\rho^2)(d\theta)^2}}{1-\rho^2}.$$

EXERCISE 64. 40. Prove that the Klein and the second Poincaré models are related geometrically as follows.

Let s be the restriction of the stereographic projection to the open southern hemisphere S^2_-. It sends S^2_- onto the open horizontal disc D bounded by the equator. Let p be the vertical projection of S^2_- to D.

Then the map $s \circ p^{-1} : D \to D$ is the isomorphism between Klein and Poincaré models.

♢

8.1. Action of the group G and Apollonian gaskets

Here we consider in more details the action of the Möbius group G and extended Möbius group $\overline{G}$ in connection with Apollonian gaskets.

If we apply a (extended) Möbius transformation to a given Apollonian gasket $\mathcal{A}$, we obtain another gasket $\mathcal{A}'$. Moreover, we know that any Apollonian gasket can be obtained in this way from one fixed gasket. So, the set

of all possible Apollonian gaskets form an homogeneous space with G (or $\overline{G}$) as a group of motions.

Let $\mathrm{Aut}\,(\mathcal{A})$ (resp. $\overline{\mathrm{Aut}}\,(\mathcal{A})$) denote the subgroup of G (resp. of $\overline{G}$) consisting of transformations which preserve the gasket $\mathcal{A}$.

THEOREM 8.1. *The subgroups* $\mathrm{Aut}\,\mathcal{A} \subset G$ *and* $\overline{\mathrm{Aut}}\,(\mathcal{A}) \subset \overline{G}$ *are discrete.*

PROOF. Let D_1, D_2, D_3 are three pairwise tangent discs in $\mathcal{A}$. Choose three interior points $w_1 \in D_1$, $w_2 \in D_2$, $w_3 \in D_3$. Afterwards, choose a neighborhood of unit $U \subset G$ which is small enough so that for any $g \in U$ we have: $g \cdot w_1 \in D_1$, $g \cdot w_2 \in D_2$, $g \cdot w_3 \in D_3$. On the other hand, if $g \in \mathrm{Aut}\,\mathcal{A}$, then it must send discs D_1, D_2, D_3 to some other discs of $\mathcal{A}$. Hence, an element $g \in U \bigcap \mathrm{Aut}(\mathcal{A})$ preserves D_1, D_2, D_3, hence, their tangent points, and so must be identity. This proves discreteness of $\mathrm{Aut}\,(\mathcal{A})$ in G.

The other statement can be proved in the same way considering four pairwise tangent discs. □

We want to describe the algebraic structure of the groups $\mathrm{Aut}(\mathcal{A})$ and $\overline{\mathrm{Aut}}(\mathcal{A})$. Fix one special gasket, e.g. the strip gasket shown on Fig. 6. We denote it by $\mathcal{A}_0$. Besides, we denote by D_1, D_2, D_3, D_4 correspondingly the half-plane $\mathrm{Im}\, w \geq 1$, the half-plane $\mathrm{Im}\, w \leq -1$, the disc $|w-1| \leq 1$ and the disc $|w+1| \leq 1$. We call it a **original quadruple** in $\mathcal{A}_0$ and denote it q_0.

First of all we want to describe the subgroup of $\overline{G}$ which preserves the basic quadruple.

THEOREM 8.2. *The group* $\overline{G}$ *acts simply transitively on the set of all ordered quadruples. The stabilizer in* $\overline{G}$ *of the original unordered quadruple is contained in* $\overline{\mathrm{Aut}}(\mathcal{A}_0)$ *and is isomorphic to* S_4*: all permutations of discs in the quadruple are possible.*

PROOF. Let $Q' = (D_1', D_2', D_3', D_4')$ be any ordered quadruple. There exists a unique element $g \in G$ which transforms the ordered triple $T_0 = (D_1, D_2, D_3)$ into the triple $T' = (D_1', D_2', D_3')$ (since an ordered triple is completely characterized by the ordered triple of tangent points).

The disc $g(D_4)$ is one of the two discs which are tangent to D_1', D_2', D_3'. These two discs are intertwined by a unique element of $\overline{G}$ preserving D_1', D_2', D_3'. Namely, by the reflection s in the mirror orthogonal to D_1', D_2', D_3'. (It is evident for the initial triple (D_1, D_2, D_3), hence is true for any triple.) Thus, exactly one of elements g and $s \circ g$ transforms q_0 into q'.

It remains to check that the stabilizer of q_0 in $\overline{G}$ is isomorphic to S_4. We already know that any permutation s of discs in q_0 can be achieved by an element $g \in \overline{G}_0$, since there is a $g \in \overline{G}$ which sends (D_1, D_2, D_3, D_4) to $(D_{s(1)}, D_{s(2)}, D_{s(3)}, D_{s(4)})$. Assume that $g \neq e$ belongs to the stabilizer of the ordered quadruple q_0 in $\overline{G}$. Then g can not be in G (it has at least six fixed points).

Recall that the set $\overline{G} \backslash G$ of antiholomorphic transformations, being not a group, still acts simply transitively on the set of ordered triples of distinct

FIGURE 8.2. Basic reflections

points. The stabilizer of an ordered triple is the reflection in the mirror passing through 3 points in question. (It is an easy exercise). Therefore, it can not have 6 fixed points which are not all on the same circle. (For the original quadruple these points are 0, ∞ and $\pm 1 \pm i$.) □

There are four quadruples q_i, $1 \leq i \leq 4$, which have with q_0 a common triple $T_i = Q_0 \backslash \{D_i\}$. Denote by D_i' the disc in q_i which is not in q_0 and by s_i a reflection which sends D_i to D_i' and preserves all other discs from q_0. See Fig.8.1

THEOREM 8.3. *The group generated by reflections s_i, $1 \leq i \leq 4$, is isomorphic to the group Γ_4 introduced in Info H.*

SCHEME OF THE PROOF. First of all we recall (see Info H) that we have labelled elements of the group Γ_4 by words in the alphabet {1, 2, 3, 4} which do not contain any digit twice in a row. We call such words **reduced**.

Recall also that $l(w)$ denotes the length of a word w and $W^{(k)}$ denotes the set of all reduced words of length k. Thus, the set $W^{(0)}$ contains only an empty word $\emptyset$, the set $W^{(1)}$ contains four words $\{i\}$, where $i = 1, 2, 3, 4$, the set W_2 contains six words $\{ij\}$, $i \neq j$, etc.

Evidently, we have an action of Γ_4 on the gasket $\mathcal{A}_0$: the generators act as reflections $\{s_i\}$. Let $D_i(\gamma)$ denote the image of the disc D_i under the action of the element $\gamma \in \Gamma_4$. The idea of the proof is to show that all discs $D_i(\gamma)$ are different.

First, we observe that $D_i(\gamma_1) \neq d_j(\gamma_2)$ for $i \neq j$. It follows from the fact that we can color all discs from $\mathcal{A}_0$ in four colors so that in any quadruple of pairwise tangent discs all four colors occur. Indeed, the set $S^2 \backslash q_0$ consists of four triangles bounded by three discs of different color. So, we can for a new disc, inscribed in each triangle, use the complementary color. In this new picture again all quadruples contains four discs of different color and we can continue the coloring.

The action of Γ_4 preserves the coloring, since the generators have this property.

Now, we can define a new numeration of discs in $\mathcal{A}_0$. Namely, let us consider all finite non-empty words in the alphabet {1, 2, 3, 4} without repeating digits. To a one-digit word $\{i\}$ we associate the disc $D_i' = s_i D_i \in q_0$. In general, we associate to a word $\{i_1 i_2 \dots i_k\}$ the disc $s_{i_1} s_{i_2} \cdots s_{i_k} D_{i_1}$.

It is enough to check that $D_i(\gamma) \neq D_i$ for $\gamma \neq e$.

We leave it as a (non-trivial) exercise. One way is to compare the numeration of discs in Section 1.2 with labelling of elements of Γ_4 above. Another way is to see, how the numeration changes when we replace the quadruple q_0 by $q_i := s_i \cdot q_0$. □

We continue to study the action of $\overline{G}$ on discs.

FIGURE 8.3. Stabilizer of a pair of discs

EXERCISE 65. 41. a) Find all transformations $g \in \overline{G}$ which preserve the unordered triple D_1, D_2, D_3.

b) Same question about unordered quadruple D_1, D_2, D_3, D_4.

Hint: a) Consider the triple of tangent points: $1 \pm i$ and ∞.

b) Find which solutions to a) preserve the disc D_4.

From the Exercise 41 we derive

THEOREM 8.4. *a) The stabilizer $S \subset G$ of any unordered triple of pairwise tangent discs in $\mathcal{A}$ is contained in* $\mathrm{Aut}(\mathcal{A})$ *and is isomorphic to S_3: all permutations of the triple are possible.*

b) The stabilizer $\overline{S} \subset \overline{G}$ of any unordered triple in $\mathcal{A}$ is contained in $\overline{\mathrm{Aut}}(\mathcal{A})$ *and is isomorphic to $S_3 \times S_2$; the central element, generating S_2 is the reflection in the mirror, orthogonal to ∂D_1, ∂D_2, ∂D_3.*

c) The stabilizer in G of any unordered quadruple of pairwise tangent discs in $\mathcal{A}$ is contained in $\mathrm{Aut}(\mathcal{A})$ *and is isomorphic to A_4: all even permutations of the quadruple are possible;*

d) The stabilizer in $\overline{G}$ of any unordered quadruple in $\mathcal{A}$ is contained in $\overline{\mathrm{Aut}}(\mathcal{A})$ *and is isomorphic to S_4: all permutations of the quadruple are possible.*

e) The group $\overline{\mathrm{Aut}}(\mathcal{A})$ *acts simply transitively on the set of ordered quadruples in $\mathcal{A}$. With respect to* $\mathrm{Aut}(\mathcal{A})$ *the ordered quadruples form two orbits.*

For an ordered triple $\widetilde{T}$ the stabilizer in G is trivial, so an element $g \in G$ is completely determined by the ordered triple $g \cdot \widetilde{T}$. By the same reason, an element $g \in \overline{G}$ is completely determined by the ordered quadruple $g \cdot \widetilde{q}$.

Now consider all pairs of tangent discs in $\mathcal{A}_0$. They form an homogeneous set with respect to the group $\overline{\mathrm{Aut}}(\mathcal{A})$. The stabilizer of $\{D_1, D_2\}$ coincides with the group $\mathrm{Aff}(1, \mathbb{Z})$ which is isomorphic to $\mathbb{Z}_2 * \mathbb{Z}_2$. Indeed the stabilizer in question consists of transformations $w \to \pm w + k,\ k \in \mathbb{Z}$ and is freely generated by reflections $s_1(w) = -w,\ s_2(w) = 1 - w$.

Finally, consider the stabilizer in $\mathrm{Aut}(\mathcal{A})$ of the disc $D_1 \in \mathcal{A}_0$. It is convenient to replace the gasket $\mathcal{A}_0$ by $\frac{1}{2}(\mathcal{A}_0 + 1 - i)$, so that D_1 becomes the upper half-plane and the tangent points of D_1 with D_3 and D_4 will be 0 and 1 – see Fig. 8.3

Then the stabilizer of this new D_1 in G is a subgroup of $PSL(2, \mathbb{R})$ which stabilizes the upper half plane. We leave to the reader to check that it coincides with $PSL(2, \mathbb{Z}) \subset G$. The stabilizer in $\overline{G}$ is obtained by adding the reflection $s_0(w) = -\bar{w}$.

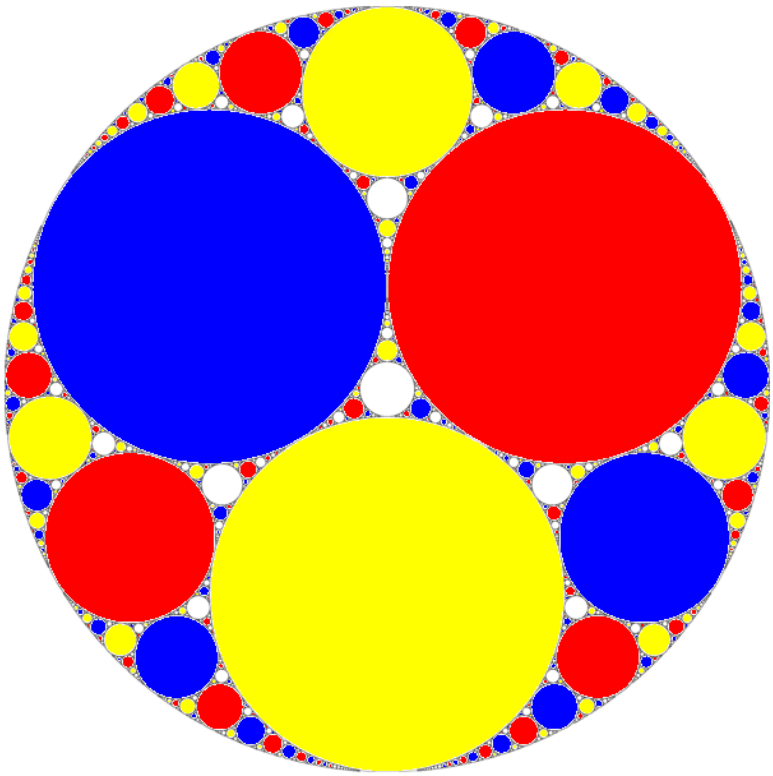

FIGURE 8.4. Orbits of Γ_4

8.2. Action of the group Γ_4 on a Apollonian gasket

Let q_0 be the original quadruple (see the text before Theorem 10). Denote by s_i, $1 \le i \le 4$, the reflection, preserving three discs from q_0, excepting D_i.

THEOREM 8.5. *a) The group, generated by s_1, s_2, s_3, s_4 is isomorphic to Γ_4. The action of this group on discs has four orbits, each of which contains one of the initial discs D_1, D_2, D_3, D_4.*

b) The stabilizer of D_1 is generated by reflections s_2, s_3, s_4 and is isomorphic to Γ_3. The action of this group on discs, tangent to D_1 has three orbits, each of which contains one of the discs D_2, D_3, D_4.

c) The stabilizer of D_1, D_2 is generated by reflections s_3, s_4 and is isomorphic to Γ_2. The action of this group on discs, tangent to both D_1, D_2 has two orbits, each of which contains one of the discs D_3, D_4.

We omit the proof based on the results of previous sections but give here the illustration where discs of four different Γ_4-orbits have different colors

There is another group generated by reflection which acts on an Apollonian gasket. Namely, let h_{ij} be the reflection in the mirror which passes

through the tangent point t_{ij} of D_i and D_j and is orthogonal to two other initial discs. It is clear, that this reflection interchanges D_i and D_j and preserves two other initial discs. Let H be the group generated by six reflections h_{ij} We leave to the reader to check that H is finite and isomorphic to the permutation group S_4.

THEOREM 8.6. *The full group Aut (A) of fractional-linear transformation of an Apollonian gasket A is a semidirect product $H \ltimes \Gamma_4$ of the subgroup H and the normal subgroup Γ_4.*

SCHEME OF PROOF. By definition, H permutes the initial discs, hence, conjugation with $h \in H$ make the corresponding permutation of generators s_i. It follows, that action of H normalizes the action of Γ_4.

Further, from Theorem 13 we conclude that Γ_4 can transform any unordered quadruple q to the initial quadruple q_0 (also considered as unordered). Since H permutes the four discs of q_0, using the group $H \ltimes \Gamma_4$, we can transform any ordered quadruple q in A to the ordered quadruple q_0.

Let now $\gamma \in G$ be any transformation of A. It sends the initial ordered quadruple q_0 to some ordered quadruple q. There exists an element $\gamma' \in H \ltimes \Gamma_4$ which sends q back to q_0. The composition $\gamma' \circ \gamma$ preserves q_0, hence is an identity. Therefore $\gamma = (\gamma')^{-1}$ belongs to $H \ltimes \Gamma_4$ and we are done. □

EXERCISE 66. 45. Let $\mathcal{M}$ be the collection of all mirrors for $\mathcal{A}_0$. Is it an homogeneous space for Γ_4, for $\mathrm{Aut}(\mathcal{A})$ and for $\overline{\mathrm{Aut}}(\mathcal{A})$?

Here we construct a group of transformations of quite a different kind. Let s_i, $i = 0, 1, 2, 3$, denote linear transformations of $\mathbb{R}^4$ which send a point $c = (c_0, c_1, c_2, c_3)$ to the point $c' = (c'_0, c'_1, c'_2, c'_3)$ where

$$c'_k = \begin{cases} c_k & \text{if} \quad k \neq i \\ 2\sum_{j \neq i} c_j - c_i & \text{if} \quad k = i. \end{cases} . \tag{8.2.1}$$

LEMMA 8.1. *The transformations s_i preserve the quadratic form*

$$Q(c) = \frac{(c_0 + c_1 + c_2 + c_3)^2}{2} - (c_0^2 + c_1^2 + c_2^2 + c_3^2),$$

hence, send a solution to the Descartes equation to a solution.

PROOF. The hyperplane M_i given by the equation $c_i = \sum_{j \neq i} c_j$ is invariant under s_i, since for the points of this hyperplane we have $c'_i = 2c_i - c_i = c_i$. Hence, s_i is a reflection in M_i in the direction of the i-th coordinate axis.

From (5.1.3) we see, first, that the Descartes equation has the form $Q(c) = 0$ and, second, that the coordinate c_i of a solution c satisfies to a quadratic equation $c_i^2 + pc_i + q = 0$ where $p = -2\sum_{j \neq i} c_j$. Therefore, the second solution c'_i to this equation satisfies $c'_i + c_i = -p$ (Vieta theorem). Thus, we get another solution to the Descartes equation if we replace c_i by c'_i leaving all other coordinates unchanged. □

Recall that we have defined above the change of coordinates (6.5.1) which sends integral solutions to Descartes equation to integral light vectors in the Minkowski space $\mathbb{R}^{1,3}$ with the coordinates t, x, y, z. So, we can consider the transformations s_i acting in $\mathbb{R}^{1,3}$. Lemma 8.2 implies that they belong to the pseudo-orthogonal group $O(1,\,3;\,\mathbb{R})$. Actually, one can prove a more precise statement.

EXERCISE 67. 46. Show that s_i acts in $\mathbb{R}^{1,3}$ as a reflection:

$$s_i(v) = v - \frac{2(v,\xi_i)}{(\xi_i,\xi_i)}\xi_i \tag{8.2.2}$$

where ξ_i, $0 \le i \le 3$, are the column vectors of the matrix $\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$ which reduces $Q(c)$ to the diagonal form.

HINT. Check that the transformations (6.5.1) in the space $\mathbb{R}^4$ with coordinates $(c_0,\,c_1,\,c_2,\,c_3)$ are reflections.

Let Γ_4 be the free product of four copies of the group $\mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$.

LEMMA 8.2. *The group Γ_4 is isomorphic to the semi-direct product $\mathbb{Z}_2 \ltimes F_3$ where F_3 is a free group with 3 generators and the non-trivial element of $\mathbb{Z}_2$ acts on F_3 by the outer automorphism inverting all generators.*

PROOF. Indeed, let

$$\Gamma_4 = \langle s_0, s_1, s_2, s_3 \mid s_i^2 = 1 \rangle.$$

Introduce the new generators: $s := s_0$ and $\tau_i := s_0 s_i$, $i = 1, 2, 3$. Then $s^2 = 1$, $s\tau_i s = \tau_i^{-1}$ and we have only show that τ_i are free generators. The proof can be obtained from the explicit realization of Γ_4 given above. . □

We define the homomorphism $\Phi : \mathfrak{g}_4^* \to O(1,\,3;\,\mathbb{R})$ by $\Phi(s_i) = s^i$, $i = 0, 1, 2, 3$.

THEOREM 8.7. *Φ is an isomorphism of Γ_4 to some discrete subgroup $\widetilde{\Gamma}_4$ in $O^+(1, 3;\,\mathbb{R})$.*

The generators of $\widetilde{\Gamma}_4$ are

$$\Phi(\tau_1) = \begin{pmatrix} 5 & -4 & 2 & 2 \\ 4 & -3 & 2 & 2 \\ 2 & -2 & 1 & 0 \\ 2 & -2 & 0 & 1 \end{pmatrix}, \qquad \Phi(\tau_2) = \begin{pmatrix} 5 & 2 & -4 & 2 \\ 2 & 1 & -2 & 0 \\ 4 & 2 & -3 & 2 \\ 2 & 0 & -2 & 1 \end{pmatrix},$$

$$\Phi(\tau_3) = \begin{pmatrix} 5 & 2 & 2 & -4 \\ 2 & 1 & 0 & -2 \\ 2 & 0 & 1 & -2 \\ 4 & 2 & 2 & -3 \end{pmatrix}; \qquad \Phi(s) = \begin{pmatrix} 2 & -1 & -1 & -1 \\ 1 & 0 & -1 & -1 \\ 1 & -1 & 0 & -1 \\ 1 & -1 & -1 & 0 \end{pmatrix}.$$

The first three matrices are unipotent with Jordan block structure $(3, 1)$. It would be interesting to give a direct geometric proof of the discreteness of the group $\widetilde{\Gamma}_4$ (see e.g. [CH65]).

CHAPTER 9

Many-dimensional Apollonian gaskets

9.1. General approach

Consider the analogue of the Descartes disc problem: find the relation between curvatures of $n+2$ pairwise tangent balls in $\mathbb{R}^n$.

Here again, it is better to extend $\mathbb{R}^n$, adding one infinite point ∞. The resulting space $\overline{\mathbb{R}}^n$ is topologically equivalent to the unit sphere S^n in the vector space R^{n+1} with coordinates $\alpha_1, \ldots, \alpha_{n+1}$ given by the equation $\sum_{k=1}^{n+1} \alpha_k^2 = 1$.

Let $\mathcal{B}_n$ be the set of all balls in $\overline{\mathbb{R}^n}$. We introduce several parametrizations of $\mathcal{B}_n$. It is instructive to compare this general result with the case $n = 2$ studied in the previous sections.

First parametrization. Let $\mathbb{R}^{1,n+1}$ be $(n+2)$-dimensional real vector space with coordinates $(p^0, \ldots, p^{n+1})$, endowed with the quadratic form

$$|p|^2 := \big(p^0\big)^2 - \big(p_1\big)^2 - \big(p_2\big)^2 - \cdots - \big(p_{n+1}\big)^2. \tag{9.1.1}$$

To any vector $p \in \mathbb{R}^{1,n+1}$ with $|p|^2 < 0$ we associate a half-space $H_p \subset \mathbb{R}^{n+1}$ defined by the condition

$$H_p := \left\{ \alpha \in \mathbb{R}^{n+1} \;\middle|\; p^0 + \sum_{k=1}^{n+1} p^k \alpha_k \;\le\; 0 \right\}. \tag{9.1.2}$$

Exercise 68. 47. Show that the intersection $S^n \bigcap H_p$ is:

• for $|p|^2 > 0$ – empty;

• for $|p|^2 = 0$ – either the whole sphere, or a single point (which one?);

• for $|p|^2 < 0$ – a closed ball which we denote by B_p.

Hint. Consider in $\mathbb{R}^{n+1}$ the projection on a line, orthogonal to H_p.

It is clear that for $c > 0$ the half-spaces H_p and H_{cp} coincide, hence, $B_p = B_{cp}$. So, we can and will normalize p by the condition $|p|^2 = -1$. Thus, the set $\mathcal{B}_n$ of all balls in S^n is parametrized by the points of the hyperboloid $|p|^2 = -1$ in $\mathbb{R}^{1,n+1}$.

Second parametrization. Define the stereographic projection $s : S^n \to \overline{\mathbb{R}}^n$ as in **Info F.** This map gives a bijection of S^n onto $\overline{\mathbb{R}}^n$ and sends balls to balls.

The inequality from (9.1.2) goes to the inequality

$$a + (b,\, x) + c(x, x) < 0 \tag{9.1.3}$$

where $x = (x_1, \dots, x_n)$, $b = (p^1, \dots, p^n)$, $a = p^0 - p^{n+1}$, $c = p^0 + p^{n+1}$ and the condition $ac - |b|^2 <$ is satisfied. We normalize, as we did before, the vector $(p^0, \dots, p^{n+1})$, or the triple (a, b, c), by the condition $|p|^2 = ac - |b|^2 = -1$.

We live to a reader to find a proof of the following

LEMMA 9.1. *Two balls B_{p_1} and B_{p_2} are tangent iff $|p_1 + p_2|^2 = 0$.*

EXERCISE 69. 48. Assume that ∂B_{p_1} and ∂B_{p_2} contain a common point x. Find the angle between the radii of B_{p_1} and B_{p_2} at x.

HINT. Use the fact that the answer practically does not depend on the dimension n: only the intersection of the whole picture with the plane passing through the centers of balls and the tangent point matters.

ANSWER.

$$\cos\alpha = -(p_1, p_2). \tag{9.1.4}$$

Let now B_{p_k}, $k = 1, 2, \dots, n+2$, be pairwise tangent balls in $\overline{\mathbb{R}^n}$. Then, exactly as in section 1.4, we see that[1]

$$\big(p_i, p_j\big) = 1 - 2\delta_{i,j}.$$

So, the eigenvalues of the Gram matrix $G_{i,j} = \big(p_i, p_j\big)$ are 2 with multiplicity $n+1$ and $-n$ with multiplicity 1. Therefore, the Gram matrix is non-singular and the vectors p_k, $1 \le k \le n+2$, form a basis in $\mathbb{R}^{1,n+1}$.

Further, we introduce for any vector $v \in \mathbb{R}^{1,n+1}$ two kind of coordinates: the covariant coordinates $v_k = (v, p_k)$ and contravariant coordinates v^k by the condition $v = \sum v^k p_k$.

The relations between two kind of coordinates are derived exactly as we did in section 1.4 for 2-dimensional case. They are:

$$v_j = \sum_i v^i - 2v^j, \qquad v^i = \frac{1}{2n}\sum_j v_j - \frac{1}{2}v_i.$$

The quadratic form in these coordinates is expressed like

$$|v|^2 \;= \big(\sum_i v^i\big)^2 - 2\sum_i (v^i)^2 \;=\; \frac{1}{2n}\left(\big(\sum_i v_i\big)^2 - n\sum_i (v_i)^2\right).$$

Put now $v = (1, -1, 0, \dots, 0, 0)$; then $v_k = \big(v, p_k\big) = p_k^{n+1} + p_k^0 = c_k$. Recall that c_k is the curvature of the ball B_{p_k}. Since $|v|^2 = 0$, we get

$$\big(\sum_k c_k\big)^2 = n \cdot \sum_k c_k^2, \tag{9.1.5}$$

which is the n-dimensional analogue of the Descartes equation.

[1]It follows also from (9.1.4) since for externally tangent balls $\cos\alpha = \cos\pi = -1$.

EXERCISE 70. 49.* Prove the n-dimensional analogue of the generalized Descartes equation:

$$\Sigma_1^2 = n \cdot \Sigma_2 - 2n^2 \cdot 1 \tag{9.1.6}$$

where

$$\Sigma_1 = \sum_{i=0}^{n+1} M_i, \qquad \Sigma_2 = \sum_{i=0}^{n+1} M_i^2 \tag{9.1.7}$$

and M_i, $0 \le i \le n+1$, are matrices corresponding to $n+2$ pairwise tangent balls in $\mathbb{R}^n$.

Let $\{B_k^0\}_{1\le k\le n}$ be a set of pairwise tangent balls in $\mathbb{R}^n$. We want to describe all sequences $\{B_j\}_{j\in\mathbb{Z}}$ of balls in $\overline{\mathbb{R}}^n$ which have the property: B_j is tangent to $B_{j\pm1}$ and to all $\{B_k^0\}_{1\le k\le n}$. Let d_k be the curvature of B_k^0 and c_j be the curvature of B_j.

From (5) we have two equations:

$$(c_j + c_{j\pm1} + d_1 + \cdots + d_n)^2 = n \cdot \left(c_j^2 + c_{j\pm1}^2 + d_1^2 + \cdots + d_n^2\right).$$

Subtracting one from another, we get

$$2c_j + c_{j+1} + c_{j-1} + 2d_1 + \cdots + 2d_n \;=\; n(c_{j+1} + c_{j-1}),$$

or

$$(n-1)(c_{j+1} + c_{j-1}) - 2c_j \;=\; 2(d_1 + \cdots + d_n).$$

It is a inhomogeneous recurrent equation for the sequence $\{c_j\}$. Subtracting two such equation for consequent j's, we get an homogeneous recurrent equation:

$$(n-1)c_{j+1} - (n+1)c_j + (n+1)c_{j-1} - (n-1)c_{j-2} = 0. \tag{9.1.8}$$

The corresponding characteristic equation is

$$(n-1)\lambda^3 - (n+1)\lambda^2 + (n+1)\lambda - (n-1) = 0 \tag{9.1.9}$$

with roots $\lambda_0 = 1,\ \lambda_{\pm1} = \frac{1\pm\sqrt{n(2-n)}}{n-1}$. Note the different structure of these roots and, consequently, different behavior of the series $\{c_j\}$ in cases $n = 2,\ n = 3$ and $n > 3$.

When $n = 2$, the characteristic equation has a triple root $\lambda = 1$. It follows that the corresponding sequence $\{c_j\}$ is quadratic in j. Indeed, the left hand side of (4.1.9) is for $n = 2$ exactly the third difference of the sequence $\{c_j\}$.

For $n = 3$, the characteristic equation has roots $\lambda_k = -e^{\frac{2k\pi i}{3}}$, $k = -1,\ 0,\ 1$, i.e. three 6-th roots from unit which are not cubic roots. Therefore, the sequence $\{c_j\}$ is 6-periodic. Moreover, not only curvatures but the balls themselves form a 6-periodic sequence. This fact was known already in ancient Greece (see [Sod36] for the details).

There is one more circumstance. Since only 3 from 6 possible sixth roots are used, the sequence $\{c_j\}$ is not only 6-periodic, but has an additional property: $c_j + c_{j+3}$ is independent on j.

We leave to the reader to formulate the corresponding geometric property of the ball sequence.

EXERCISE 71. 50. Let B_1, B_2, B_3 be three unit balls in $\mathbb{R}^3$ which are pairwise tangent. Find 6 balls which are tangent to all B_k, $k = 1, 2, 3$.

HINT. The corresponding curvatures are 0, 0, 3, 6, 6, 3.

For $n > 3$ the situation is quite different. The characteristic equation has one real root $\lambda_0 = 1$ and two complex roots $\lambda_{\pm 1} = \frac{1 \pm i\sqrt{n^2-2n}}{n-1}$ of absolute value 1. Write them in the form $\lambda_{\pm 1} = e^{\pm i\alpha}$. Then $\cos\alpha = \frac{1}{n-1}$.

PROPOSITION 9.1. *All integral solutions to the equation* $\cos\frac{2\pi}{m} = \frac{1}{n}$ *have the form:* $m = n = 1$, $\quad m = 2$, $n = -1$, $\quad m = 3$, $n = -2$, $\quad m = 6$, $n = 2$.

It follows that the sequence of balls $\{B_j\}_{j\in\mathbb{Z}}$ in $\mathbb{R}^3$ has a quasiperiodic character and self-intersects infinitely many times.

From the recurrent relation

$$c_{j+1} = \frac{2}{n-1}c_j - c_{j-1} \tag{9.1.10}$$

we conclude also that for $n > 3$ the curvatures cannot be integers for all j.

9.2. 3-dimensional Apollonian gasket

As we saw above, the case $n = 3$ is exceptional. From any integral solution $(c_1, \ldots, c_5)$ to Descartes equation we can made five new solutions; namely, i-th transformation s_i replace c_i by $\sum_{j\neq i} c_j - c_i$ and preserves all other c_j. The transformations s_i satisfy as before the relations $s_i^2 = \mathrm{Id}$, but moreover, they satisfy the relations $(s_i s_j)^3 = \mathrm{Id}$ for $i \neq j$. Hence, any pair (s_i, s_j), $i \neq j$, generates a group isomorphic to S_3, the Weyl group for $\mathbb{A}_2$.

Still more interesting is that any three reflections (s_i, s_j, s_k) generate the affine Weyl group for $\mathbb{A}_2$ which is a semi-direct product $S_3 \ltimes \mathbb{Z}^2$.

PROPOSITION 9.2. *For any three pairwise tangent balls the set of balls tangent to all three can be parametrized by a circle* $T = \mathbb{R}/2\pi\mathbb{Z}$, *so that the balls* B_α *and* B_β *are tangent iff* $|\alpha - \beta| = \frac{\pi}{3} \mod \mathbb{Z}$.

PROPOSITION 9.3. *For any two pairwise tangent balls the set of balls tangent to both of them can be parametrized by a sphere* S^2, *or, better by* $\overline{\mathbb{R}^2}$ *so that the balls* B_α *and* B_β *are tangent iff* $|\alpha - \beta| = 1$.

We leave to the reader to prove the propositions and relate their statements to the structure of subgroups $\langle s_i, s_j\rangle$ and $\langle s_i, s_j, s_k\rangle$.

PROBLEM 11. Determine the structures of the group $\Gamma = \langle s_1, s_2, s_3, s_4, s_5\rangle$ and its subgroup $\langle s_i, s_j, s_k, s_l\rangle$.

A lot of useful information about this problem can be found in the book [EGM]. See also [C] as a very interesting introduction to the theory of quadratic forms.

The notion of a nice parametrization can be generalized to the 3-dimensional case. Consider the algebraic number field $K = \mathbb{Q}[\epsilon]$ where $\epsilon = e^{\frac{2\pi i}{3}}$ is a cubic root of unit. A general element of K has the form $k = \alpha\epsilon + \beta\bar{\epsilon}$ where α, $\beta \in \mathbb{Q}$ and bar means the complex conjugation. Note that

$$||k||_K^2 = |k|^2 = k\bar{k} = \alpha^2 - \alpha\beta + \beta^2. \tag{9.2.1}$$

Denote by E the set of all complex numbers of the form $a\epsilon + b\bar{\epsilon}$ where a, $b \in \mathbb{Z}$. It is the set of integers in the algebraic number field K. There are six invertible integers with norm 1: ± 1, $\pm\epsilon$, $\pm\bar{\epsilon}$. They are called **units** of the field K. It is well-known that any element of E can be uniquely (modulo units) written as a product of prime numbers. As for prime numbers, they include all rational (i.e. ordinary) primes of the form $p = 3m - 1$ and also the numbers $k = a\epsilon + b\bar{\epsilon}$ for which $|k|^2 = a^2 - ab + b^2$ is equal to 3 or to a rational prime of the form $3m + 1$.

It follows that any element $k \in K$ can be uniquely (modulo units) written as a fraction $\frac{p}{q}$ where p, $q \in E$ have no common factors (except units). It can be also written as $k = \frac{l\epsilon + m\bar{\epsilon}}{n}$ where l, m, n are ordinary integers with $gcd\,(l,\, m,\, n) = 1$.

Definition 9.1. Let D be a 3-ball in an integral 3-dimensional Apollonian gasket $\mathcal{A}$. A parametrization of ∂D by the points of $\overline{\mathbb{R}}^2$ is called **nice** if the tangent points for D and other balls in $\mathcal{A}$ correspond exactly to elements of $\overline{K} \subset \overline{\mathbb{R}}^2$.

Let $D_k \in \mathcal{A}$ be the ball tangent to D which corresponds to the point $k = \frac{p}{q} \in \overline{K}$.

Theorem 9.1. *Nice parametrizations exist and have the properties:*
a) Let $K \ni k = \frac{p}{q}$. The curvature c_k of the ball D_k has the form

$$c_k = \alpha|p|^2 + \beta p\bar{q} + \bar{\beta}\bar{p}q + \gamma|q|^2 + \delta \tag{9.2.2}$$

where α, $\gamma \in \mathbb{R}$, $\beta \in \mathbb{C}$.

b) There is a coordinate system $(x_1,\, x_2,\, x_3)$ in the ambient space $\mathbb{R}^3$ such that

$$x_i = \frac{\alpha_i|p|^2 + \beta_i p\bar{q} + \bar{\beta}_i\bar{p}q + \gamma_i|q|^2 + \delta_i}{\alpha|p|^2 + \beta p\bar{q} + \bar{\beta}\bar{p}q + \gamma|q|^2 + \delta}. \tag{9.2.3}$$

c) Let $k_i = \frac{p_i}{q_1}$, $i = 1,\, 2$. The balls D_{k_1} and D_{k_2} are tangent iff

$$|k_1 - k_2| \,=\, \frac{1}{|q_1 q_2|}. \tag{9.2.4}$$

We leave to the reader the proof of the theorem and developing of the matrix variant of the theory.

In conclusion, we illustrate theorem 16 by two examples of nice parametrizations in a 3-dimensional Apollonian gasket.

We associate with a ball in $\mathbb{R}^3$ with a center $x+iy+jz$ and radius r the Hermitian matrix $\begin{pmatrix} a & b \\ \bar{b} & c \end{pmatrix}$ where $c=\frac{1}{r},\ b=\frac{x+iy+jz}{r},\ \bar{b}=\frac{x-iy-jz}{r},\ a=\frac{x^2+y^2+z^2-r^2}{r}$.

Our gasket $\mathcal{A}$ is the analog of the band plane gasket. It contains two half-spaces: $z\geq 1$ and $-z\geq 1$ corresponding to matrices $M_{\pm}=\begin{pmatrix} 2 & \mp j \\ \pm j & 0 \end{pmatrix}$; further, it contains infinitely many unit balls corresponding to matrices $\begin{pmatrix} |v|^2-1 & v \\ -\bar{v} & 1 \end{pmatrix}$ where v runs through the lattice $V\subset\mathbb{C}$ generated by 2ϵ and $2\bar{\epsilon}$.

The first example is the parametrization of all balls tangent to the plane $z=1$ by the elements of $\bar{K}$. Namely, to $k=\frac{p}{q}\in\bar{K}$ we associate the matrix

$$M_k=\begin{pmatrix} 4|p|^2+|q|^2-2 & 2p\bar{q}+(1-|q|^2)j \\ 2\bar{p}q-(1-|q|^2)j & |q|^2 \end{pmatrix}. \tag{9.2.5}$$

The corresponding ball is tangent to the plane at the point $t_k=-2\frac{p}{q}+(1-\frac{1}{|q|^2})j$ and has the radius $r=\frac{1}{|q|^2}$.

The second example is the parametrization of all balls tangent to the unit ball corresponding to the matrix $M=\left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. Here we have

$$M_k=\begin{pmatrix} |p|^2+|q|^2+1 & 2p\bar{q}+(|p|^2-|q|^2)j \\ 2\bar{p}q+(|q|^2-|p|^2)j & |p|^2+|q|^2-1 \end{pmatrix}. \tag{9.2.6}$$

The corresponding ball is tangent to the unit ball at the point $t_k=\frac{-2p\bar{q}+(|q|^2-|p|^2)j}{|p|^2+|q|^2}$ and has the radius $r=\frac{1}{|p|^2+|q|^2-1}$.

Bibliography

A. Popular books, lectures and surveys

[Bar88] M. Barnsley, *Fractals Everywhere*, Academic Press Inc., 1988.

[LGRE00] N. Lesmoir-Gordon, W. Rood, and R. Edney, *Fractal Geometry*, Icon Books, UK, Totem books, USA, 2000.

[Sod36] F. Soddy, *The kiss precise*, Nature **7 (June 20)** (1936), 1021.

[Ste03] K. Stephenson, *Circle packing: A mathematical tale*, Notices of the AMS **50** (2003), 1376–1388.

[Str99] Robert S. Stricharts, *Analysis on Fractals*, Notices of the AMS **46** (1999), 1999–1208.

B. Books

[Bea83] A.F. Beardon, *The Geometry of discrete groups*, Graduate Texts in Mathematics, vol. 91, Springer, 1983.

[Con97] John H. Conway, *The sensual (quadratic) form*, with the assistance of Francis Y. C. Fung, Carus Mathematical Monographs, vol. 26, MAA, Washington, DC, 1997.

[Cox69] H.S.M. Coxeter, *Introduction to Geometry*, Wiley & Sons, 1969.

[CH91] R. Courant and D. Hilbert, *Methods of mathematical Physics, vol. 2*, Wiley, 1991.

[CH65] Moser W.O.J. Coxeter H.M.S., *Generators and relations for discrete groups, 2d ed.*, Ergebnisse der Mathematik und ihrer Grenzgebiete. Reihe, Gruppentheorie. n.F., vol. 14, Springer-Verlag, Berlin ; New York, 1965.

[Edg90] Edgar, *Measure Topology and Fractal Geometry*, GTM, Springer-Verlag Inc., 1990.

[EGM98] J. Elstrodt, F. Grunewald, and J. Mennike, *Groups Acting on Hyperbolic Space*, Springer, 1998; Russian transl. in MCCME, Moscow, 2003.

[Kig01] Jun Kigami, *Analysis on fractals*, Cambridge Tracts in Mathematics, vol. 143, Cambridge University Press, Cambridge, 2001.

[Man82] B. Mandelbrot, *The fractal geometry of Nature*, Freeman, San Francisco, 1982.

[Thu97] W. Thurston, *Three-Dimensional Geometry and Topology*, Princeton Mathematical Series, vol. 35, Princeton University Press, Princeton, NJ, 1997.

C. Research papers

[AS97] D. Aharonov and K. Stephenson, *Geometric sequences of discs in the Apollonian packing*, Algebra i Analiz **9** (1997), 104-140; English transl., St. Petersburg Math. J. **9** (1998), 509-545.

[Bar92] M.T. Barlow, *Harmonic analysis on fractal sets,* Seminar Bourbaki, Exp. 755, Astérisque **206** (1992), 345-368.

[Bro85] R. Brooks, *The spectral geometry of the Apollonian packing*, Comm.Pure Appl. Math. **38** (1985), 358-366.

[BL04] A. F. Beardon and L. Lorentzen, *Continued fractions and restrained sequences of Möbius maps*, Rocky Mountain J. Math. **34** (2004), 441–466.

[DK] R. L. Dobrushin and S. Kusuoka, *Statistical mechanics and fractals*, Lecture Notes in Math, vol. 1567, Springer, Berlin, 1993, pp. 39–98.

[FS92] M. Fukushima and T. Shima, *On a spectral analysis for the Sierpiński gasket*, Potential Anal. **1** (1992), 1-35.

[GLMCL$^+$03] R. L. Graham, J. C. Lagarias, Mallows C. L., A. Wilks, and C. Yan, *Apollonian Packings: Number Theory*, Journal of Number Theory **100** (2003), 1-45.

[KS43] E. Kasner and F. Supnik, *The Apollonian packing of circles*, Proc. Nat. Acad. Sci. USA **29** (1943), 378-384.

[MC03] MacMullen C.T., *Hausdorff dimension and conformal dynamics III: Computation of dimension*, Preprint (2003).

[MT95] Leonid Malozemov and Alexander Teplyaev, *Pure point spectrum of the Laplacians on fractal graphs*, J. Funct. Anal. **129** (1995), 390-405.

[Nev49] E.H. Neville, *The structure of Farey series*, Proc. of the London Math. Soc., ser. 2 **51** (1949), 132-144.

[Ram84] R. Rammal, *Spectrum of harmonic exitations on fractals*, J. Physique **45** (1984), 191-206.

[de Rha59] G. de Rham, *Sur les courbes limites de polygones obtenus par trisection*, Enseignement Math **(2) 5** (1959), 29–43 (French).

[de Rha56] ———, *Sur une courbe plane*, J. Math. Pures Appl. **(9) 35** (1956), 25-42 (French).

[de Rha47] ———, *Un peu de mathmatiques propos d'une courbe plane*, Elemente der Math. **2** (1947), 73–76, 89–97 (French).

[de Rha56] ———, *Sur quelques courbes definies par des equations fonctionnelles*, Univ. e Politec. Torino. Rend. Sem. Mat. **16** (1956/1957), 101–113 (French).

[Sal43] R. Salem, *On some singular monotonic functions which are strictly increasing*, Trans. Amer. Math. Soc. **53** (1943), 427–439.

[Str00] Robert S. Strichartz, *Taylor approximations on Sierpinski gasket type fractals*, J. Funct. Anal. **174** (2000), 76–127.

[TAV00] Teplyaev A. V., *Gradients on fractals*, Jour. Funct. Anal. **174** (2000), 128–154.

D. Web sites

http://en.wikipedia.org/wiki/Fractal
http://classes.yale.edu/fractals/
http://www.faqs.org/faqs/fractal-faq/