A.A. Kirillov

A Tale of Two Fractals

Birkhäuser

Birkhäuser

A.A. Kirillov

A Tale of Two Fractals

Birkhäuser

A.A. Kirillov
Department of Mathematics
University of Pennsylvania
Philadelphia, PA, USA

ISBN 978-0-8176-8381-8 ISBN 978-0-8176-8382-5 (eBook)
DOI 10.1007/978-0-8176-8382-5
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013933043

Mathematics Subject Classification (2010): 00A09, 26A30, 28A80, 57S

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.birkhauser-science.com)

To Ben and Lisa

Preface

This book is devoted to the phenomenon of fractal sets, or simply *fractals*. Fractals have been known for about more than a century and have been observed in different branches of science. But it is only recently (approximately in the last thirty years) that they have become a subject of mathematical study.

The pioneer of the theory of fractals was Benoit Mandelbrot. His book *Fractals: Form, Chance and Dimension* first appeared in 1977, and a second, enlarged, edition was published in 1982. Since that time, serious articles, surveys, popular papers, and books about fractals have appeared by the dozen (if not by hundreds). Also, in 1993, the specialized journal *Fractals* was published by World Scientific. So, why 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 large number of colorful pictures and amazing examples but no accurate definitions and rigorous results. In contrast, the articles written by professionals are, as a rule, too difficult for beginners and often discuss very special questions, assuming that the motivation and all connections are already known to the reader.

Last, and perhaps the most important reason, is my belief that the endeavor of independent study of the geometry, analysis, and arithmetic of fractals is one of the best ways for a young mathematician to acquire an active and stable knowledge of the basic mathematical tools.

This subject also seems to me to be an excellent opportunity to test your ability to produce creative work in mathematics.[1] I mean here not only solving well-posed problems, but recognizing hidden patterns and formulating new, fruitful problems.

My interest in fractals originates from the lecture course I gave at the University of Pennsylvania in 1995 at the request of our undergraduate students. I repeated the course in 1999, 2003, and 2005. In 2004 and in 2007, I had the opportunity

[1] According to Yu.I. Manin, to create in mathematics is to calculate with excitement.

to present the material in several lectures at the summer school in Dubna, near Moscow, organized for high school seniors and first-year university students who were winners of the Russian Mathematical Olympiad. Both times, I was pleasantly surprised by the activity of the audience and by their quickness in comprehending all of the necessary information.

In this book, we deliberately restrict ourselves to only two examples of fractals: the Sierpiński and Apollonian gaskets. I describe and rigorously formulate several problems that come 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 formulation and understanding) of all the problems requires some preliminary background, which 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, disks, and spheres in $\mathbb{R}^3$. Basic trigonometric formulas. 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 of this is normally contained in the first two or three years of a university mathematics curriculum.

I consider the diversity of the necessary tools and their interconnection a great advantage of this subject, because it is a characteristic feature of modern mathematics.

Let me offer several words about the style of exposition. I tried to avoid two main dangers: being dull by explaining too many details in the most elementary form and being incomprehensible by using very effective but sometimes too abstract modern techniques. It is to the reader to judge the success of this endeavor.

I also tried to communicate an informal knowledge of mathematical tools that distinguish (almost all) professionals from most beginners. Sometimes, one phrase explains more than a long article.[2] So, from time to time, I intentionally use some "high-altitude" notions, explaining each time in the simplest possible words what they mean in the simplest situations.

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

Some additional information is included in the text in the form of sections with the heading “Info.”

I also use “Remarks” as another form of additional information. The end of a remark is indicated by the sign ♡.

The end of a proof (or the absence of proof) is marked by the sign □.

Acknowledgements The author deeply thanks the Erwin Schroödinger International Institute for Mathematical Physics (ESI), where this work was begun; the Max Planck Institute in Bonn (MPI); and the Institute des Hautes Etudes Scientifique (IHES), where it was completed. I also thank the referees for helpful remarks.

Contents

Part I The Sierpiński Gasket

Part II The Apollonian Gasket

Part I
The Sierpiński Gasket

Chapter 1
Definitions and General Properties

1.1 First Appearance and Naive Definition

I will not describe the early manifestations of fractals in the natural sciences (such as investigations of seashore length, cauliflower and snowflake forms); there are enough examples in popular expositions (see, for example, the pioneering book [Man82] or the nice recent book [LGRE00]).

For mathematicians, the simplest and best-known example of a fractal is the famous *Cantor set*. An acquaintance with the Cantor set is a good test to distinguish those who really understand real analysis from those who have merely formally passed a calculus exam. We shall not go into details of this example just yet, but in Sect. 1.2, we shall return to it and show that it is a part of the general theory of self-similar fractals.

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

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

$$\begin{array}{ccccccccccccccccc}
&&&&&&&&1&&&&&&&& \\
&&&&&&&1&&1&&&&&&& \\
&&&&&&1&&2&&1&&&&&& \\
&&&&&1&&3&&3&&1&&&&& \\
&&&&1&&4&&6&&4&&1&&&& \\
&&&1&&5&&10&&10&&5&&1&&& \\
&&1&&6&&15&&20&&15&&6&&1&& \\
&1&&7&&21&&35&&35&&21&&7&&1& \\
..&&\cdots&&\cdots&&\cdots&&\cdots&&\cdots&&\cdots&&\cdots&&..
\end{array}$$

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

A.A. Kirillov, *A Tale of Two Fractals*, DOI 10.1007/978-0-8176-8382-5_1,

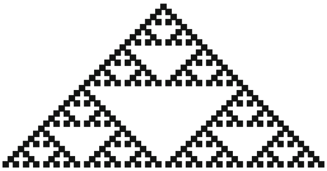


Fig. 1.1 Pascal's triangle mod 2

Now let us replace these numbers by their residues modulo 2. In other words, we put 0 in place of every even number and 1 in place of every odd number. We get the following picture:

$$\begin{array}{c}
1\\
1 \quad 1\\
1 \quad 0 \quad 1\\
1 \quad 1 \quad 1 \quad 1\\
1 \quad 0 \quad 0 \quad 0 \quad 1\\
1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1\\
1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1\\
1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1\\
.. \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad ..
\end{array}$$

How can one describe this picture? Observe that this triangle of size 8 contains three identical triangles of size 4 (left, upper, and right); each of these triangles contains three identical triangles of size 2, consisting of three ones.

The remaining places are occupied by zeros.

Let us try to imagine what happens if we continue our triangle up to 2^N lines for some large 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 Fig. 1.1.

Here the whole triangle contains three triangles of half size that look similar to the whole thing. The space bounded by these triangles is filled by white dots.

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

Another appearance of the same set is related to the following problem of linear algebra. Let E_N be an $N \times N$ 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 a rigorous definition of a limit in this situation.

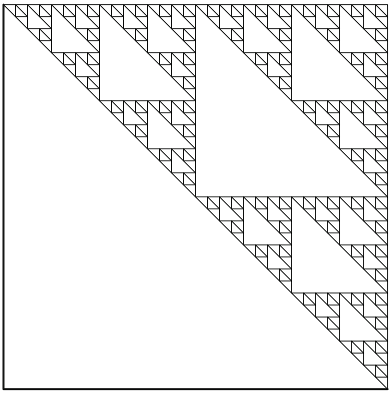


Fig. 1.2 Pascal triangular matrix

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 that establishes the similarity: $E_N A_N = A_N J_N$. It turns out that A_N can be chosen such that it looks like Fig. 1.2.

We leave it to the reader to explain this phenomenon and find the connection of A_N to Pascal's 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 objects of an 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 simpler sets.

Info A. Metric Spaces

We start with some general definitions, which later will be specialized and explained with 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 us to treat uniformly problems that seem completely different.

A.1 Distance and Limit

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 that for every 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 nonnegative real number that 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 $y = (y_1, y_2)$:

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

and the three-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 $d(x_n, a) \to 0$ as $n \to \infty$.

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

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

For example, every 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 the usual distance of Eq. (A.1), the sequence $x_n = \frac{1}{n}$ is a Cauchy sequence, but it 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)–(A.3) above are complete metric spaces according to a 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 A.1. Let (M, d) be a complete metric space and X 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 has been done and what we have to prove, and you will obtain a proof.

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

A.2 Contracting Maps

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 (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., a 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 a 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. To this end, we show that $\{x_n\}$ is a Cauchy sequence. Indeed, let $d(x_0, x_1) = d$. Then, from Eq. (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 every $m < n$ 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 by 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 $d(x, y) = d\big(f(x), f(y)\big) \le \lambda \cdot d(x, y)$. this is possible only if $d(x, y) = 0$; hence $x = y$. □

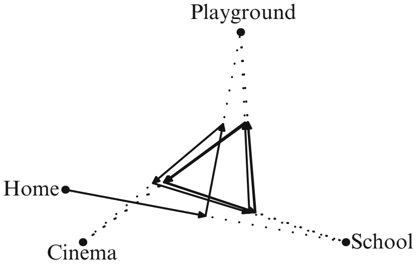


Fig. A.3 Indecisive boy

This theorem solves in particular the following toy problem, which appeared on a mathematical Olympiad for middle-school students.

Problem A.1. A boy came out of his house and went to school. At the halfway point he changed his mind and turned toward a playground. But after walking halfway there, he turned toward a cinema. At the halfway point to the cinema, he decided again to go to school, etc. (See Fig. A.3.)

Where will he end up if he continues to move in this way?

A.3 Compact Sets

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 $S \subset M$ is called an *ε-net* in M if for every $m \in M$, there is a point $s \in S$ such that $d(m, s) < \varepsilon$.

Theorem (Theorem on $\boldsymbol{\varepsilon}$-nets). *A metric space (M, d) is compact iff it is complete and for every $\varepsilon > 0$, there is a finite ε-net in M.*

We give the proof here because it is a good example of mathematical reasoning and because it helps in understanding the nature of compactness.

1. Assume that M is compact. Let us show that M is complete. Consider a Cauchy sequence $\{x_n\}$ in M. We have to show that it has a limit. Since M is compact, the sequence $\{x_n\}$ contains a convergent subsequence $\{y_k\} = \{x_{n_k}\}$. Let a be the limit: $a = \lim_{k\to\infty} y_k$. I claim that a is the limit of $\{x_n\}$. Indeed, since $\{x_n\}$ is a Cauchy sequence, for every $\varepsilon > 0$, there exists a number $N = N(\varepsilon)$ such that $d(x_m, x_n) < \frac{\varepsilon}{2}$ for $m, n > N(\varepsilon)$. Also, there exists a number $K = K(\varepsilon)$ such that $d(y_k, a) < \frac{\varepsilon}{2}$ for $k > K$. So, for $n > \max(N, n_K)$, we have

$$d(x_n, a) \le d(x_n, x_{n_K}) + d(x_{n_K}, a) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

   Therefore, $\lim_{n\to\infty} x_n = a$.

Now we show that M admits a finite ε-net for any $\varepsilon > 0$. Assume the contrary, namely, that for some $\varepsilon > 0$, there is no finite ε-net in M. Then I claim that M contains a sequence $\{x_n\}$ with the property that

$$d(x_m, x_n) \geq \varepsilon \quad \text{for all} \quad m \neq n. \tag{A.6}$$

Indeed, we construct the desired sequence by induction. Choose $x_1 \in M$ arbitrarily. Suppose that points $x_1, x_2, \ldots, x_n$ satisfying Eq. (A.6) are already chosen. Since the finite set $\{x_i\}_{1 \leq i \leq n}$ is not an ε-net, there is a point $x_{n+1} \in M$ such that $d(x_{n+1}, x_i) \geq \varepsilon$ for $1 \leq i \leq n$.

The sequence $\{x_n\}$ satisfying Eq. (A.6) does not contain any convergent subsequence, because every subsequence also satisfies Eq. (A.6), hence is certainly not a Cauchy sequence. A contradiction.

2. Assume that M is complete and that for every $\varepsilon > 0$, there is a finite ε-net in M. Let us show that M is compact, that is, that every sequence $\{x_n\}$ contains a convergent subsequence. It is enough to find in $\{x_n\}$ a Cauchy subsequence. Let S_n be a finite ε-net in M for $\varepsilon = \frac{1}{n}$. Denote the points of S_n by $s_1^{(n)}, s_2^{(n)}, \ldots, s_{|S_n|}^{(n)}$, where $|S_n|$ is the number of points in S_n. Let

$$B_k^n = \left\{x \in M \mid d(x, s_k^{(n)}) \leq \frac{1}{n}\right\}$$

denote the closed ball of radius $\frac{1}{n}$ centered at $s_k^{(n)}$. Since S_n is an ε-net in M, the union of the balls $B_k^{(n)}$, $1 \leq k \leq |S_n|$, covers the whole set M. Put $n = 1$. So at least one of the balls $B_k^{(1)}$, $1 \leq k \leq |S_1|$, contains infinitely many terms of our sequence. Therefore, there exists an infinite subsequence $\{x_k^{(1)}\}$ that is contained in a ball of radius 1.

Now put $n = 2$. At least one of the balls $B_k^{(2)}$ contains infinitely many terms of the subsequence $\{x_k^{(1)}\}$. Hence, there is a subsequence $\{x_k^{(2)}\}$ that is contained in a ball of radius $\frac{1}{2}$. And so on. Consider the diagonal subsequence $\{y_k\} = \{x_k^{(k)}\}$. This is the desired Cauchy subsequence, because its terms, starting with the nth, belong to a ball of radius $\frac{1}{n}$.

Exercise A.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 convergent 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-nets, 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.

◇

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 nonempty compact subsets of M. We want to define a distance between two compact sets so that $\mathbb{K}(M)$ is 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

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

A 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 about how to define the distance between two sets so that axioms 1–3 are satisfied, you will find that the definitions in Eqs. (1.2.2) and (1.2.3) are as simple as possible.

In Fig. 1.3, the first and second terms in Eq. (1.2.3) are the lengths of segments AB and CD respectively.

Exercise 1.1. Prove that the minimum in Eq. (1.2.1) and maximum in Eq. (1.2.2) always exist.

Hint. Use the compactness of sets X and Y.

Exercise 1.2. Compute the distance (a) between the boundary of a square with side 1 and its diagonal; (b) between a unit circle and the disk bounded by this circle.

Answer. (a) $\frac{1+\sqrt{2}}{2}$; (b) 1.

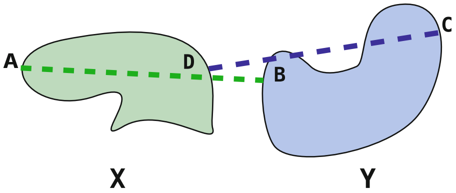


Fig. 1.3 Hausdorff distance

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

Remark 1.1. To make the notion of Hausdorff distance more visual, let us introduce the following definition.

Definition 1.1. Let M be a metric space. A map f of some subset $M' \subset M$ to M is called an *ε-perturbation* if $d(x, f(x)) \leq \varepsilon$ for all $x \in M'$.

Then the statement "the Hausdorff distance between X and Y is equal d" is equivalent to the statement "there exist an ε_1-perturbation $f_1 : X \to Y$ and an ε_2-perturbation $f_2 : Y \to X$ such that $\varepsilon_1 + \varepsilon_2 = d$."

♡

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 that 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 nonempty.)

For the second statement, use the theorem on ε-nets.

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, if M is complete, there is a unique nonempty compact subset $X \subset M$ satisfying $F(X) = X$.*

Definition 1.2. The set X from Theorem 1.2 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 Eq. (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 facts.

Exercise 1.3. Show that the "constant" map f_Y that sends every $X \in \mathbb{K}(M)$ to a fixed $Y \in \mathbb{K}(M)$ is contracting.

Exercise 1.4. Let F_1 and F_2 be two contracting maps of $\mathbb{K}(M)$. Define the map F by

$$F(X) = F_1(X) \bigcup F_2(X).$$

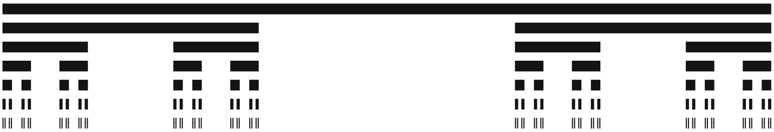


Fig. 1.4 Cantor set (first seven steps in its construction)

Show that F is contracting, using the relation

$$d(X_1 \bigcup X_2,\ Y_1 \bigcup Y_2) \leq \max\big(d(X_1, Y_1), d(X_2, Y_2)\big).$$

Definition 1.3. A set X that is a fixed point for a map Eq. (1.2.5) is called an *inhomogeneous self-similar fractal.*

Example 1. (1) **Cantor set** $C \subset [0,\ 1]$. Here $M = [0,\ 1]$, $f_1(x) = \frac{1}{3}x$, $f_2(x) = \frac{x+2}{3}$. It is instructive to look at 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

$$C_2 = [0,\ 1/3]\cup[2/3,\ 1], \quad C_3{=}[0,\ 1/9]\cup[2/9,\ 1/3]\cup[2/3,\ 7/9]\cup[8/9,\ 1]\ldots.$$

The sequence $\{C_n\}$ is decreasing, $C_{n+1} \subset C_n$, and the limit set is $C = \bigcap_{n\geq 1} C_n$. This construction is shown in Fig. 1.4.

Now put $C_1' = \{0,\ 1\}$. Then

$$C_2' = \{0,\ 1/3,\ 2/3,\ 1\}, \quad C_3' = \{0,\ 1/9,\ 2/9,\ 1/3,\ 2/3,\ 7/9,\ 8/9,\ 1\}, \ldots.$$

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 in this example: if we consider a piece of the Cantor set under a microscope that increases all sizes by a factor of 3^n, we shall see exactly the same picture as that seen with the naked eye.

There is one more sequence $\{C_n''\}$ approximating the Cantor set. It corresponds to the choice $C_1'' = \{0\}$ and admits a simple arithmetic description. Let us write real numbers from $[0,\ 1]$ using base-3 expansions. The notation $0.a_1a_2\ldots a_n$, where the a_i take values $0,\ 1,\ 2$, is used for the number

$$a = \frac{a_1}{3} + \frac{a_2}{3^2} + \cdots + \frac{a_n}{3^n}. \tag{1.2.6}$$

The set C_n'' consists of all expressions (1.2.6) that use only values 0 and 2 for a_i. We obtain the full Cantor set if we also allow infinite three-adic fractions (still with the digits 0 and 2).

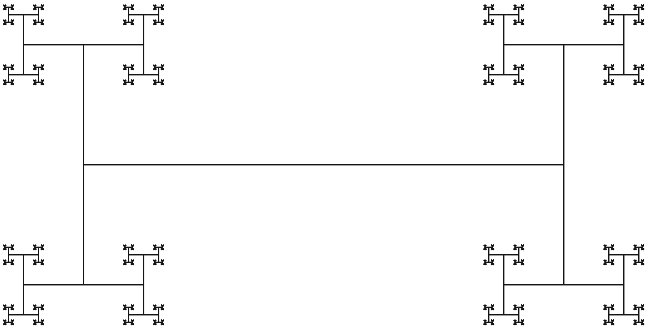


Fig. 1.5 I_α-fractal for $\alpha = 0.5$

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

$$f_1(x, y) = (-\alpha y, \alpha x + 1); \qquad f_2(x, y) = (-\alpha y, \alpha x - 1). \qquad (1.2.7)$$

The corresponding inhomogeneous self-similar fractal is shown in Fig. 1.5, where for typographic convenience the y-axis is horizontal.

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 1.5. 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 1.4. 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 natural choices for the initial set S_0.

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

Second, we let S_0' be the hollow triangle with vertices at $0, 1, \omega$. Then the sequence $S_n' = F^n(S_0')$ is increasing, and $\mathcal{S}$ is the closure of $S_\infty' = \bigcup_{n\ge 0} S_n'$.

Exercise 1.6. How many vertices, edges, and hollow triangles are in S_n'?

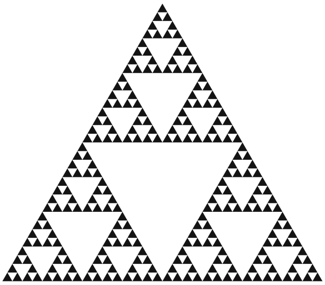


Fig. 1.6 Approximation S_n''

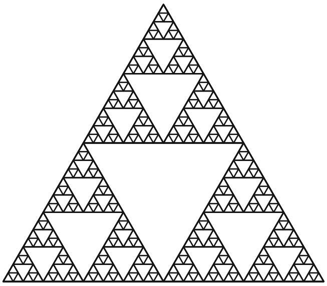


Fig. 1.7 Approximation S_n

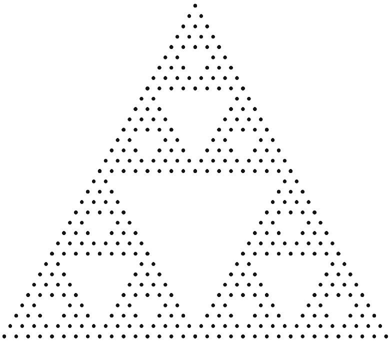


Fig. 1.8 Approximation S_n'

Finally, let S_0 be the set of $0, 1, \omega$. 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 $S_\infty = \bigcup_{n \geq 0} S_n$.

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

Remark 1.2. Looking at Fig. 1.7, you might think that some points of S_n have six neighbors. For example, consider the middle point p in the third row. However, comparing Fig. 1.7 with Fig. 1.8, we see that only four of these six points are genuine neighbors. Note also that if we embed the approximation S_6 in the next approximation S_7, the point p becomes a middle point in the fifth row and will have four neighbors.

Thus, "being a neighbor" is a stronger property than "being at a shortest distance."

♡

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 do we measure the size of a fractal set?

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

Let X be a compact subset of $\mathbb{R}^n$ (to avoid technical complications, we do not consider here more general sets). Then for every $\epsilon > 0$, it admits a finite covering by balls of radius ϵ. (The centers of these balls form an ε-net for X.) Let $N(\epsilon)$ denote the minimal number of balls that cover X.

It is evident that $N(\epsilon)$ grows (more precisely, it is nondecreasing) as 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 purposes, the following proposition will suffice.

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

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

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

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

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

In fact, the first property formally follows from the second one, but we formulated it separately because of its clarity 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 B.1. 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 1.3. There are several variants of this definition. Namely, instead of balls of radius ϵ, one can use arbitrary sets of diameter ε, or when $M = \mathbb{R}^n$, one can use cubes with side length ε.

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

All these variants can lead to a different value of p-measure, but for "nice" examples, including self-similar fractals, they 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:

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

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

Exercise B.2. Show that the Hausdorff dimension of X, when it exists, can be given by the formula

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

Example 2. Let us find the Hausdorff dimensions of the self-similar fractals defined above. In all cases, we assume that not only the Hausdorff dimension but also the Hausdorff measure exists. This is not evident, but the persistent reader can try to prove it by him/herself.

Then we use the following simple arguments to compute it.

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

   It follows from the definition of d-measure that each of these two pieces of C has 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

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

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 arrive at 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 B.1; therefore, $2 \cdot \alpha^d = 1$ and

$$d = \log_\alpha \frac{1}{2} = -\frac{\log 2}{\log \alpha}. \qquad \text{(B.6)}$$

The right-hand side of Eq. (B.6) satisfies the inequality $1 \leq d \leq 2$ for $\alpha \in [\frac{1}{2}, \frac{1}{\sqrt{2}}]$.

Exercise B.3. Prove that Eq. (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 it to the reader to investigate the cases $\alpha = \frac{1}{2}$, $\alpha = \frac{1}{\sqrt{2}}$, and $\alpha \notin [\frac{1}{2}, \frac{1}{\sqrt{2}}]$.

3. Sierpiński gasket. The set $\mathcal{S}$ is the union of three subsets $\frac{1}{2}\mathcal{S}$, $\frac{1}{2}(\mathcal{S} + \omega)$, and $\frac{1}{2}(\mathcal{S}+1)$. So if $\mathcal{S}$ has finite d-measure $\mu_d(\mathcal{S})$, we must have the relation $\mu_d(\mathcal{S}) = 3 \cdot (\frac{1}{2})^d \cdot \mu_d(\mathcal{S})$. It follows that $2^d = 3$ and $d = \log_2 3 \approx 1.5849625$.

Chapter 2
The Laplace Operator on the Sierpiński Gasket

A powerful mathematical method for studying 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 a homogeneous manifold 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, a vast domain of modern mathematics arose that is known as *spectral geometry*. The main subject of this mathematical subfield is the study of spectra of naturally (i.e., geometrically) defined linear operators.

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

In this book, we only briefly mention spectral geometry on fractal sets. Our main goal is the study of harmonic functions on the Sierpiński gasket.

Info C. The Classical Laplace Operator and Harmonic Functions

This section is not necessary for understanding the main text, but it gives motivation for our study of the Laplace operator and harmonic functions on fractal sets.

[1] That is, functions that transform in a prescribed way under the action of the group. Details are explained in textbooks on representation theory.

A.A. Kirillov, *A Tale of Two Fractals*, DOI 10.1007/978-0-8176-8382-5_2,

Here we consider smooth functions and differential operators in some domain $\Omega \subset \mathbb{R}^n$. The reader with some acquaintance with elements of differential geometry on Riemannian manifolds will understand that all our constructions make sense in this general situation.

C.1 Analytic Approach

One of the 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 $\mathbb{R}^n$. More precisely, it is known that every differential operator on $\mathbb{R}^n$ that is invariant under E_n is a polynomial in Δ.

The operator Δ can be expressed as the composition of two other natural operators: the *gradient* and *divergence*:

$$\Delta = \operatorname{div} \circ \operatorname{grad}. \tag{C.1}$$

Here the operator grad acts from the space $C^\infty(\Omega)$ of smooth functions on Ω to the space $\mathrm{Vect}^\infty(\Omega)$ of smooth vector fields on Ω by the formula

$$\operatorname{grad} f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_1} \right). \tag{C.2}$$

The operator div acts from $\mathrm{Vect}^\infty(\Omega)$ to $C^\infty(\Omega)$ by the formula

$$\operatorname{div} v = \frac{\partial v^1}{\partial x_1} + \ldots \frac{\partial v^n}{\partial x_n}. \tag{C.3}$$

There is another, more geometric, definition of the Laplace operator. Take an ε-neighborhood $U_{\varepsilon(x_0)}$ of a point x_0 (i.e., the ball of radius ε centered at the point x_0). Then the integral of f over $U_{\varepsilon(x_0)}$ has the following asymptotic behavior as $\varepsilon \to 0$:

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

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

$$(\Delta f)(x_0) = \lim_{\varepsilon \to 0} \frac{1}{b_n \varepsilon^{n+2}} \int_{U_{\varepsilon(x_0)}} \big(f(x) - f(x_0) \big) d^n x, \tag{C.4}$$

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

Definition C.1. A function satisfying the equation $\Delta f = 0$ is called *harmonic*.

It is known that on any domain $\Omega \subset \mathbb{R}^n$, harmonic functions are characterized by the property

$$\frac{1}{vol(U_{\varepsilon(x_0)})} \int_{U_{\varepsilon(x_0)}} f(x) d^n x = f(x_0), \tag{C.5}$$

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 Ω is a connected domain with boundary (denoted by $\partial\Omega$). Then any nonconstant real harmonic function on Ω attains its maximal value only on the boundary.*

It is also known that for any continuous function φ on the boundary $\partial\Omega$, there exists a unique harmonic function f on Ω that has φ as a boundary value (i.e., such that $f|_{\partial\Omega} = \varphi$).

More precisely, for any point $x \in \Omega$, there exists a probabilistic measure μ_x on $\partial\Omega$ such that $f(x) = \int_{\partial\Omega} \varphi(y) d\mu_x(y)$. The measure μ_x is called *Poisson measure*, and in the case of a smooth boundary, it is given by a density ρ_x) that is a smooth function of $x \in \Omega$ and $y \in \partial\Omega$.

There is a simple physical interpretation of a harmonic function (as a stable heat or charge distribution), and there is a probabilistic interpretation of Poisson measure $\mu_x(A)$ (as a probability of reaching the boundary in a set A starting from x and moving randomly along Ω).

C.2 Algebraic Approach

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

Suppose that 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

$$(v_1, v_2) := \frac{Q_0(v_1 + v_2) - Q_0(v_1) - Q_0(v_2)}{2}. \tag{C.6}$$

In practice, V is usually infinite-dimensional, but the reader can assume that it is finite-dimensional for simplicity. The quadratic form Q_1 can be defined in terms of Q_0 as follows.

Proposition C.1. *There exists a symmetric operator A in V such that*

$$Q_1(v) = (Av, v) \quad \textit{for all} \quad v \in V. \tag{C.7}$$

Remark C.1. Sometimes, the operator A is called a quotient of two forms Q_1 and Q_0. Indeed, every quadratic form Q defines the symmetric bilinear form $\tilde{Q} : V \times V \to \mathbb{R}$ by the formula

$$\tilde{Q}(v_1, v_2) := \frac{Q(v_1 + v_2) - Q(v_1) - Q(v_2)}{2}. \tag{C.8}$$

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

Thus, the operator A can be written as $A = \overline{Q}_0{}^{-1} \circ \overline{Q}_1$.

♡

Now consider the following variational problem: find the extremum of the quadratic form Q_1 under the condition $Q_0 = 1$. Applying the standard theorem about conditional extrema, we get the following result.

Proposition C.2. *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] $Q_0(v) = 1$.

We apply the general algebraic scheme described above to the following situation. Let Ω be a domain in $\mathbb{R}^n$ with a smooth boundary. Denote by V the space of smooth functions on Ω with compact support restricted by some boundary conditions; see below.

There are two natural quadratic forms on V:

$$Q_0(v) = \int_\Omega v^2(x)\, d^n x \quad \text{and} \quad Q_1(v) = \int_\Omega |\text{grad}\, v|^2(x) d^n x, \tag{C.9}$$

where $d^n x$ is the standard measure on $\mathbb{R}^n$ and $|\text{grad}\, v|^2 = \sum_{k=1}^n |\frac{\partial v^k}{\partial x_k}|^2$.

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

$$\int_\Omega (\text{grad}\, v_1, \text{grad}\, v_2) d^n x = \int_\Omega A v_1(x) \cdot v_2(x)\, d^n x. \tag{C.10}$$

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

$$\int_{\partial\Omega} v_1 \partial_\nu v_2\, d^{n-1} y - \int_\Omega \Delta\, v_1(x) \cdot v_2(x)\, d^n x, \tag{C.11}$$

where ∂_ν is the normal derivative and $d^{n-1}y$ is a measure on the boundary $\partial\,\Omega$.

Suppose we restrict v by an appropriate boundary condition that forces the boundary integral in Eq. (C.11) to vanish. Then the operator $-\Delta$ will be exactly the ratio of Q_1 and Q_0.

[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 $V \backslash \{0\}$.

Two special examples are widely known: the *Dirichlet problem*, in which the condition

$$v \big|_{\partial\Omega} = 0 \tag{C.12}$$

is imposed, and the *Neumann problem*, in which the boundary condition is

$$\partial_\nu v \big|_{\partial\Omega} = 0. \tag{C.13}$$

In both cases, $-\Delta$ is a nonnegative self-adjoint operator in $L^2(\Omega, d^n x)$ whose domain of definition consists of differentiable functions v on Ω satisfying boundary conditions and is such that $\Delta v \in L^2(\Omega, d^n x)$ in the sense of generalized functions.

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

We already mentioned that the question, "what can the spectrum of a Laplace operator on a smooth compact manifold be?" has given rise to a whole new domain in mathematics: spectral geometry.

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

$\diamondsuit$

2.1 The Laplace Operator on S_n

In the first version of this book, I wanted to describe in full detail the definition and computation of the spectrum of the Laplace operator on S_N and on $\mathcal{S}$. After that, I learned that such a program had already been realized by several physicists and mathematicians; see, e.g., [Ram84, FS92, MT95]. Therefore, I decided not to repeat the result one more time but instead to concentrate on some different and less-well-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 the Sierpiński gasket $\mathcal{S}$, we consider first the finite approximation S_n of $\mathcal{S}$.

Let us try to follow the algebraic scheme used above. Let S_n be the nth 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 $\frac{3^{n+1}+3}{2}$ points, V_n is a real vector space of dimension $d_n = \frac{3^{n+1}+3}{2}$.

Let us define two quadratic forms on V_n:

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

where the first sum is over all points of $\mathcal{S}_n$, and the second is over all pairs of neighboring points (i.e., points joined by an edge in S_n').

Clearly, these quadratic forms are discrete analogues of the quadratic forms defined by Eq. (C.9) in Info C.

As in the case of the 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

$$Q_1(f) = (\Delta_n f,\ f), \tag{2.1.2}$$

where

$$\big(\Delta_n f\big)(s) = k(s) f(s) - \sum_{s' \leftrightarrow s} f(s'). \tag{2.1.3}$$

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

We introduce two types 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 $\frac{3^n - 3}{2}$. The operator $\Delta_n^{(D)}$ in this space is given by Eq. (2.1.3) for 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 neighboring points to s. The space $V_n^{(N)}$ of functions satisfying this condition again has dimension $\frac{3^n - 3}{2}$. The operator $\Delta_n^{(N)}$ in this space is given by Eq. (2.1.3) for inner points.

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

For illustration and to make things clear, we consider in detail the case $n = 2$.

First let $V = V_2^{(D)}$. It is a 3-dimensional space of functions on S_2 whose values are shown in Fig. 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 the following lemma.

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

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

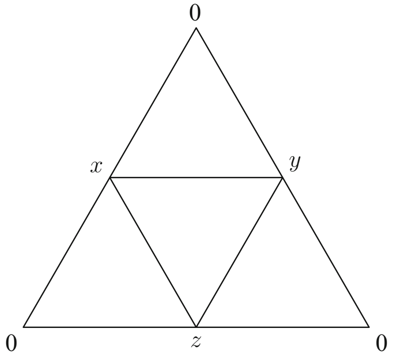


Fig. 2.1 Functions on S_2 with Dirichlet condition

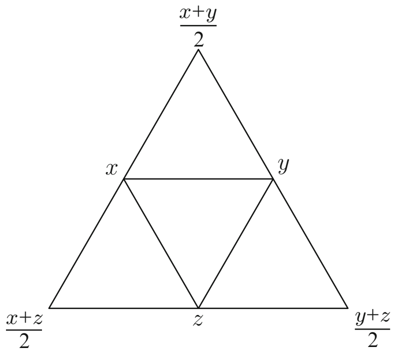


Fig. 2.2 Functions on S_2 with the Neumann condition

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

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 $x = y = z$.

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

Now let $V = V_2^{(N)}$. The values of functions from this space are shown in Fig. 2.2. I leave it to you to check that the operator $\Delta_2^{(N)}$ sends the triple $(x,\ y,\ z)$ to 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.

This means that the 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 if we are to get deep and beautiful results about the spectrum of the Laplace operator. The reader can skip this part and come back to it when he or she is ready to understand the complete set of arguments.

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 to $\mathcal{S}_{n-1}$ is not identically zero. Consider in detail a piece of $\mathcal{S}_n$ around a point where $f \neq 0$. We write the values of f at the corresponding points (values that do not matter are denoted by question marks) as follows:

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

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

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

Adding the 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 the two previous ones, we obtain

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

From Eqs. (2.2.3) and (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 Eq. (2.2.2) gives

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

We arrive at the following alternatives: 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 the following theorem.

Theorem 2.1. *The restriction of any harmonic function on $\mathcal{S}_n$ to $\mathcal{S}_{n-1}$ is also harmonic.*

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

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

Definition 2.1. A function on $\mathcal{S}_\infty$ is called *harmonic* if its restriction to every $\mathcal{S}_n$ is harmonic.

2.3 Eigenfunctions of the Laplace Operator on S_n

Here we consider briefly the spectrum of the operators $\Delta_n^{(D)}$ with the 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 any sequence $\mu_k, \ k = 0, 1, 2, \ldots$ such that $\mu_0 = \mu$ and $P(\mu_k) = \mu_{k-1}$ for $k \geq 1$ a *μ-string*.

We want to extend a function $f \in V_n^{\mu_n}$ in such a way that the extended function belongs to $f \in V_{n+1}^{\mu_{n+1}}$. From Eq. (2.2.6), we know that this 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 S_∞ such that its restriction to $\mathcal{S}_n$ (which can be zero!) belongs to $V_n^{\mu_n}$ for all n.

So, the following problem arises: is such function f on S_∞ uniformly continuous and hence can be extended by continuity to $\mathcal{S}$? When this is the case, we can consider the extended function $\tilde{f}$ as an eigenfunction for the Laplacian on the whole gasket and define the corresponding eigenvalue as the limit of a 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 S_∞.

Chapter 3
Harmonic Functions on the Sierpiński Gasket

In this chapter, we consider in greater detail the harmonic functions on the Sierpiński gasket $\mathcal{S}$. Note that a harmonic function satisfying the Dirichlet boundary condition must be zero, and a harmonic function satisfying the Neumann boundary condition must be a constant. So, we shall consider here harmonic functions with no restrictions on the boundary values.

Recall that the boundary points of $\mathcal{S}$ are $0,\ 1,\ \omega = \frac{1+i\sqrt{3}}{2}$. So the segment $[0,\ 1]$ of the 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 exhibit highly nontrivial analytic, algebraic, 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.*

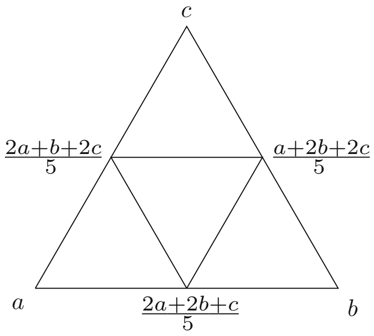
Proof. From linear algebra, we know that if a homogeneous system of linear equations in n variables has only the trivial solution, then the corresponding inhomogeneous system has a unique solution for any right-hand side. It follows that $\dim \mathcal{H}(\mathcal{S}_n) = 3$ for all $n \geq 1$. Therefore, every 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.

Lemma 3.2. *Let $x,\ y,\ z$ be three neighboring points of $\mathcal{S}_m$ that form an equilateral triangle. Put $\alpha = \frac{y+z}{2},\ \beta = \frac{x+z}{2},\ \gamma = \frac{x+y}{2}$. Then $\alpha,\ \beta,\ \gamma$ also form an equilateral*

A.A. Kirillov, *A Tale of Two Fractals*, DOI 10.1007/978-0-8176-8382-5_3,

Fig. 3.1 The ratio 1:2:2



triangle and are neighboring points in $\mathcal{S}_{m+1}$ *(see Fig. 3.1). For every harmonic function* f *on* $\mathcal{S}_{m+1}$*, we have*

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

The informal formulation of this result is that the neighboring points have twice the impact of the opposite one.

Now we can prove the following important result.

Theorem 3.1. *Every harmonic function on* S_∞ *is uniformly continuous, hence has a unique continuous extension to* $\mathcal{S}$*.*

Proof. Let f_{ab}^c be the harmonic function on S_∞ with the boundary values

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

Let us call the quantity

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

the *variation* of the function f on the set X. From the maximum principle, we conclude that

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

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

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

Thus, the function f^c_{ab} belongs to the Hölder class H_β for $\beta = \log_2 \frac{5}{3}$. Therefore, it is uniformly continuous. Hence, it 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} to 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 $\varphi := u^0_{01}$.

Exercise 3.1. Derive from (3.2.1) that

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

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

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

We call functions χ, φ, ψ, ξ *basic functions*. 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 mentioning 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 $\alpha_1(t) = \frac{1+t}{2}$. 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 $\mathcal{H}$. (This follows from the fact that a harmonic function on $\mathcal{S}$ remains harmonic when restricted to a left or right lower subtriangle of $\mathcal{S}$.)

Moreover, we know exactly the eigenvalues of both operators and their eigenfunctions. Indeed, it is easy to check that $1,\ \psi,\ \chi$ are eigenfunctions for A_0 and that $1,\ 1-\xi,\ 1-\varphi$ are eigenfunctions for A_1. The corresponding eigenvalues are $1,\ \frac{3}{5},\ \frac{1}{5}$. (Apply, for example, Lemma 3.2 for the boundary values of χ and for its restriction to the left half of $[0,\ 1]$, and you get $\chi(\frac{t}{2}) = \frac{1}{5}\chi(t)$.)

These relations look more compact in the language of vectors and matrices. Let us introduce the vector functions

$$\overrightarrow{f}(x) = \begin{pmatrix} \psi(x) \\ \chi(x) \\ 1 \end{pmatrix} \quad \text{and} \quad \overrightarrow{g}(x) = \begin{pmatrix} \varphi(x) \\ \xi(x) \\ 1 \end{pmatrix}. \tag{3.2.4}$$

Then the following relations hold:

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

where

$$A_0 = \begin{pmatrix} 3/5 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 3/5 & 0 & 2/5 \\ 0 & 1/5 & 4/5 \\ 0 & 0 & 1 \end{pmatrix}, \quad T = \begin{pmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.2.6}$$

One more useful property of the family of basic functions is that they form an arithmetic progression:

$$\xi - \psi = \psi - \varphi = \varphi - \chi, \tag{3.2.7}$$

as follows from consideration of the boundary values of the corresponding harmonic functions on $\mathcal{S}$.

Exercise 3.2. Using relations (3.2.5)–(3.2.7), complete the table of values of the functions $\chi,\ \varphi,\ \psi,\ \xi$ at the points $k/8,\ k = 0, 1, \ldots, 7, 8$.

Function\Argument	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{3}{4}$	$\frac{7}{8}$	1
χ	0	$\frac{1}{125}$	$\frac{1}{25}$		$\frac{1}{5}$				1
φ	0				$\frac{2}{5}$		$\frac{16}{25}$	$\frac{98}{125}$	1
ψ	0	$\frac{27}{125}$	$\frac{9}{25}$		$\frac{3}{5}$				1
ξ	0				$\frac{4}{5}$		$\frac{24}{25}$	$\frac{124}{125}$	1

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 $\varphi(t) = \frac{\xi(t)+2\chi(t)}{3}$ 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 $\overrightarrow{h}(t) := \begin{pmatrix} \xi(t) \\ \chi(t) \\ 1 \end{pmatrix}$.

From (3.2.5), we derive the following transformation rules for $\overrightarrow{h}$:

$$\overrightarrow{h}\left(\frac{t}{2}\right) = B_0 \overrightarrow{h}(t); \qquad \overrightarrow{h}\left(\frac{1+t}{2}\right) = B_1 \overrightarrow{h}(t), \tag{3.2.8}$$

where

$$B_0 = \begin{pmatrix} 3/5 & 1/5 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \qquad B_1 = \begin{pmatrix} 1/5 & 0 & 4/5 \\ 1/5 & 3/5 & 1/5 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.2.9}$$

Consider now the binary presentations of t and s:

$$t = 0.t_1 t_2 \dots t_k \dots, \qquad s = 0.s_1 s_2 \dots s_k \dots.$$

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

Applying (3.2.8) several times, we get

$$\overrightarrow{h}(t) = B_{t_1} \cdots B_{t_{k-1}} B_0 \overrightarrow{f}(z), \qquad \overrightarrow{h}(s) = B_{t_1} \cdots B_{t_{k-1}} B_1 \overrightarrow{f}(w)$$

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

But

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

while

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

□

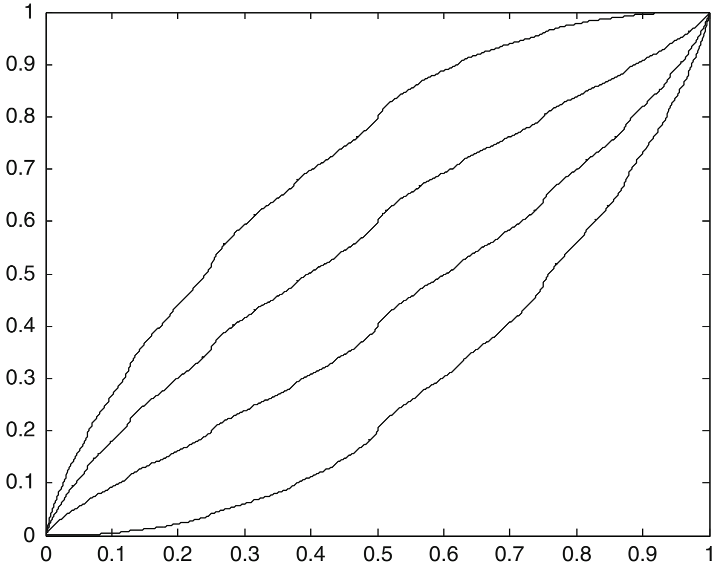


Fig. 3.2 Functions χ, φ, ψ, ξ

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

$$A^{-1}x^{\alpha} \le \psi(x) \le Ax^{\alpha}, \qquad B^{-1}x^{\beta} \le \chi(x) \le Bx^{\beta}$$

$$\text{with } A = \frac{5}{3}, \quad \alpha = \log_2 \frac{5}{3}, \qquad B = 5, \beta = \log_2 5. \tag{3.2.10}$$

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 also have

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

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

As a corollary of Theorem 3.2, we obtain

$$u'(r) = +\infty, \tag{3.2.11}$$

where u is any of 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).

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

It is known, for example, that $\chi'(\frac{1}{3}) = 0$ and $\chi'(\frac{1}{15}) = \infty$. This is proved in the master's thesis of Irina Kalashnikova, an undergraduate and graduate student at the University of Pennsylvania. She also showed that at all rational nondyadic points t, the derivative $\chi'(t)$ takes the value 0 or ∞.

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

Lemma 3.4.

$$\int_0^1 u^c_{a,b}(t)\mathrm{d}t = \frac{3a+3b+c}{7}. \tag{3.2.12}$$

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

$$x = \varphi + \psi - 1 = \chi + \xi - 1; \qquad y = \xi - \psi = \psi - \varphi = \varphi - \chi. \tag{3.2.13}$$

As t runs from 0 to 1, x increases from -1 to 1, while y grows from 0 to $\frac{1}{5}$ at $t = \frac{1}{2}$ 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 given by the following theorem.

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

Exercise 3.3. 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

$$x\left(\frac{t}{2}\right) = \frac{1}{2}x(t) + \frac{3}{10}y(t) - \frac{1}{2}; \qquad y\left(\frac{t}{2}\right) = \frac{1}{10}x(t) + \frac{3}{10}y(t) + \frac{1}{10}$$
$$x\left(\frac{1+t}{2}\right) = \frac{1}{2}x(t) - \frac{3}{10}y(t) + \frac{1}{2}; \quad y\left(\frac{1+t}{2}\right) = -\frac{1}{10}x(t) + \frac{3}{10}y(t) + \frac{1}{10}. \tag{3.2.14}$$

The next two problems are open questions.

Problem 3.2. Compute the following moments:[1]

$$m_n := \int_{-1}^{1} x^n y \mathrm{d}x, \qquad M_n := \int_{-1}^{1} y^n \mathrm{d}x. \tag{3.2.15}$$

Problem 3.3. Compute the Fourier coefficients

$$c_n := \int_{-1}^{1} \mathrm{e}^{-\pi i n x} y \mathrm{d}x. \tag{3.2.16}$$

3.3 Extension and Computation of $\chi(t)$ and $\psi(t)$

There is a method to compute the values of $\chi(t)$ at binary fractions rapidly. Namely, we know that $\chi(t)$ satisfies the relations[2]

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

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

$$\chi(t) := 5^N \chi(2^{-N}|t|) \qquad \text{where } N \text{ is large enough for} \quad 0 \leq 2^{-N}|t| \leq 1. \tag{3.3.2}$$

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

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

It is easy to derive from (3.3.3) the following statement.

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

The relation (3.3.3) allows us not only to compute the values $\chi(k)$ for integer k, but also to formulate the following conjecture.

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

[1] One reviewer computed several moments m_n explicitly. Though they are rational numbers, their lowest terms are rather awkward, and it is difficult to make any conjecture about them in general.

[2] The simplest way to derive these relations is to compare the boundary values of both sides, taking into account that they are harmonic functions. See Theorem 3.6 below.

Table 3.1 Table of values of $\chi(k)$, $3^6\psi(k)$ 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	3,745	−11	9,985	245
2	5	1	1,215	486	35	3,965	5	10,191	206
3	12	2	1,620	405	36	4,200	−2	10,400	209
4	25	1	2,025	405	37	4,429	−11	10,597	197
5	41	1	2,403	378	38	4,625	5	10,755	158
6	60	2	2,700	297	39	4,836	26	10,916	161
7	85	5	2,997	297	40	5,125	1	11,125	209
8	125	1	3,375	378	41	5,417	−23	11,331	206
9	168	−2	3,744	369	42	5,640	−2	11,480	149
10	205	1	7,405	261	43	5,857	17	11,617	137
11	245	5	4,239	234	44	6,125	5	11,775	158
12	300	2	4,500	261	45	6,408	−2	11,936	161
13	361	1	4,761	261	46	6,685	17	12,085	149
14	425	5	4,995	234	47	7,013	53	12,255	170
15	504	14	5,256	261	48	7,500	2	12,500	245
16	625	1	5,625	369	49	7,993	−47	12,745	245
17	749	−11	5,991	366	50	8,345	−11	12,915	170
18	840	−2	6,240	249	51	8,664	14	13,064	149
19	925	5	6,453	213	52	9,025	1	13,225	161
20	1,025	1	6,675	222	53	9,389	−11	13,383	158
21	1,128	−2	6,888	213	54	9,720	14	13,520	137
22	1,225	5	7,065	177	55	10,093	53	13,669	149
23	1,337	17	7,251	186	56	10,625	5	13,875	206
24	1,500	2	7,500	249	57	11,172	−38	14,084	209
25	1,669	−11	7,749	249	58	11,605	1	14,245	161
26	1,805	1	7,935	186	59	12,041	41	14,403	158
27	1,944	14	8,112	177	60	12,600	14	14,600	197
28	2,125	5	8,325	213	61	13,201	1	14,809	209
29	2,321	1	8,547	222	62	13,805	41	15,015	206
30	2,520	14	8,760	213	63	14,532	122	15,260	245
31	2,761	41	9,009	249	64	15,625	1	15,625	365
32	3,125	1	9,375	366	65	16,721	−119	$15,989\frac{2}{3}$	$364\frac{2}{3}$
33	3,492	−38	9,740	365	66	17,460	−38	$16,233\frac{1}{3}$	$243\frac{2}{3}$

A similar approach allows us to compute the values of the extended function ψ at integral points. The key formula is the following analogue of (3.3.3):

$$\left(\Delta_k^2\psi\right)(2^n) = -\frac{1}{3^{n+1}}\chi(k) \quad \text{for} \quad 0 \le k \le 2^n. \tag{3.3.4}$$

In Table 3.1, 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

$$\psi(3+t)+\psi(3-t)=2\psi(3)=\frac{40}{3} \qquad \text{for} \quad |t|\le 1. \qquad (3.3.5)$$

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

$$\psi(48+k)+\psi(48-k)=\frac{25,000}{729}.$$

The same symmetry is observed for φ:

$$\varphi\left(\tfrac{1}{4}+t\right)+\varphi\left(\tfrac{1}{4}+t\right)=2\varphi\left(\tfrac{1}{4}\right) \qquad \text{for} \qquad |t|\le \tfrac{1}{4}. \qquad (3.3.6)$$

All this suggests that we search for minimal "wavelets" such that the graphs of all basic functions can be constructed from affine images of these wavelets.

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

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

It is also of interest to study the 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 that the reader 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 as n goes to ∞.

Info D. Fractional Derivatives and Fractional Integrals

The derivative of order n is defined as the nth iteration of the ordinary derivative. Sometimes, the integral $\int_0^x f(t)\mathrm{d}t$ is called the antiderivative of f, or the derivative of order -1. One can also define the derivative of order $-n$ as the nth iteration of the antiderivative. The explicit form of this operation is

$$f^{(-n)}(x)=\int_0^x \mathrm{d}t_1\int_0^{t_1}\mathrm{d}t_2\cdots\int_0^{t_{n-1}} f(t_n)\mathrm{d}t_n.$$

This iterated integral can be written as the n-dimensional integral

$$\int_{\Delta_x} f(t_n)\mathrm{d}t_1\mathrm{d}t_2\cdots\mathrm{d}t_n,$$

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

$$\int_{\Delta_x} f(t_n)\mathrm{d}t_1\mathrm{d}t_2\cdots\mathrm{d}t_n = \int_0^x \mathrm{vol}\Delta_x(t) f(t)\mathrm{d}t = \int_0^x \frac{(x-t)^{n-1}}{(n-1)!} f(t)\mathrm{d}t. \quad \text{(D.1)}$$

Here $\Delta_x(t)$ is the $(n-1)$-dimensional simplex that 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. Namely, Euler's gamma function $\Gamma(\alpha)$, given by the formula

$$\Gamma(\alpha) = \int_0^\infty x^\alpha \mathrm{e}^{-x} \frac{dx}{x},$$

has the properties

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha), \qquad \Gamma(n+1) = n! \quad \text{for} \quad n \in \mathbb{N}.$$

So it can serve as an interpolation of the factorial function to noninteger values of α. Therefore, we replace n by $-\alpha$ and define an antiderivative of order $-\alpha$, or a *derivative of order* α by the formula

$$f^{(\alpha)}(x) = \int_0^x \frac{(x-t)^{-\alpha-1}}{\Gamma(-\alpha)} f(t)\mathrm{d}t. \quad \text{(D.2)}$$

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

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

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

Hint. Use Euler's beta function, given by

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

and the identity

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

Note also the connection between fractional derivatives and 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 the integral of order α is a convolution with Φ_α.

◇

3.4 Some Arithmetic Properties of Basic Functions

As was shown in Sect. 3.3, the function $\chi(t)$ takes integer values at 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 at integer points.

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

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

We can consider these functions boundary values of harmonic functions defined on the *infinite Sierpiński gasket* bounded by the rays $x \geq 0,\ y = 0$ and $x \geq 0,\ y = \frac{x\sqrt{3}}{2}$ (Fig. 3.3).

We want to study the local behavior of χ in the 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 $k = 2m + 1$.

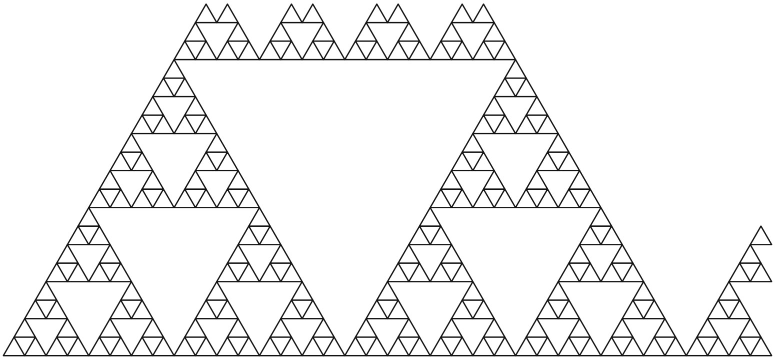


Fig. 3.3 The infinite Sierpiński gasket

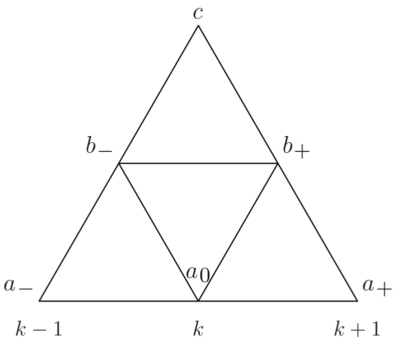


Fig. 3.4 A fragment of the infinite Sierpiński gasket

Theorem 3.6. *For every odd k and every* $\tau \in [0, 1]$, *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 every n and every odd k and odd $l < 2^n$, we have[3]

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

and

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

Here are some particular cases:

(a) $n = 1, k = 2m + 1, l = 1: \quad \chi(4m + 3) \equiv \chi(4m + 1) \mod 11;$
(b) $n = 2, k = 2m + 1, l = 3: \quad \chi(8m + 7) \equiv \chi(8m + 1) \mod 84;$
(c) $k = 1: \quad \chi(2^n + l) \equiv \chi(2^n - l) \mod \big(2\chi(l) + 3^{n+1}\psi(l)\big)$ (here we have not only congruence but in fact equality, since in this case, $2\Delta_1 = 1$).

Proof of the theorem. Consider the triangular piece of the infinite gasket that is based on the segment $[k - 1, k + 1]$. It is shown in Fig. 3.4.

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 the remaining vertices shown in Fig. 3.4 can be uniquely determined from the equations

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

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

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 the 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

$$\begin{aligned}\chi(2^n k + l) - \chi(2^n k - l) &= 5^n \left(\chi\left(k + \tfrac{l}{2^n}\right) - \chi\left(k - \tfrac{l}{2^n}\right)\right) \\ &= 2 \cdot 5^n \Delta_1 \left(2\chi\left(\tfrac{l}{2^n}\right) + 3\psi\left(\tfrac{l}{2^n}\right)\right) = 2 \cdot \Delta_1 \cdot \left(2\chi(l) + 3^{n+1}\psi(l)\right).\end{aligned}$$

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 Function $D(k)$

One more function deserves more detailed study. Let

$$D(k) := \frac{\chi(k+1) - 2\chi(k) + \chi(k-1)}{3}. \tag{3.5.1}$$

Theorem 3.7. *The function $D(k)$ takes integer values at integer points, except $D(0) = \frac{2}{3}$, and has the following properties:*

$$D(2k) = D(k), \qquad D(2k+1) + D(2k-1) = 3D(k) \quad \text{for all } k. \tag{3.5.2}$$

These properties allow us easily to compute the table of values of $D(k)$, which reveals some very interesting behavior. We list here some facts, leaving the proofs to the readers.

1. If $2^a \leq k < 2^{a+1}$, then for $n \geq 0$,

$$D(2^{n+a} + k) = \alpha(k) + \beta(k) \cdot 3^n \tag{3.5.3}$$

for some integers or half-integers $\alpha(k),\ \beta(k)$ and for all $n \geq 0$.

2. The sets $D^{-1}(n)$ seem to have a very special structure. It is clear that it is enough to indicate only the subset $D^{-1}_{\text{odd}}(n)$ of odd numbers from these sets. The following statements, except the first one, are only conjectures, confirmed by numerical computation:

$D^{-1}(\{0, -1, \pm 3, \pm 4, -5, \pm 6, \pm 7, \pm 8, \pm 9, \pm 10, 11, \pm 12\}) = \emptyset$;
$D^{-1}_{\text{odd}}(1) = \{2^n - 3 \mid n \geq 2\}$;
$D^{-1}_{\text{odd}}(2) = \{3\}$;
$D^{-1}_{\text{odd}}(-2) = \{2^n - 3 \mid n \geq 2\}$;
$D^{-1}_{\text{odd}}(5) = \{2^n + 3 \mid n \geq 2\}$;
$D^{-1}_{\text{odd}}(-11) = \{7 \cdot 2^n + 3 \mid n \geq 1\} \bigcup \{11 \cdot 2^n + 3 \mid n \geq 1\}$;
$D^{-1}_{\text{odd}}(14) = \{2^n + 3 \mid n \geq 2\}$.

Hint: Consider possible values of $D(n) \mod 3, \mod 4, \mod 7, \mod 11$.

3.6 The Functions $x(t)$, $y(t)$, and $y(x)$

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

$$x = \varphi + \psi - 1 = \chi + \xi - 1; \qquad y = \xi - \psi = \psi - \varphi = \varphi - \chi. \tag{3.6.1}$$

An alternative definition is $x = u^0_{-1,1}$, $y = u^1_{0,0}$.

As 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:

$$\chi = \frac{x + 1 - 3y}{2}, \quad \varphi = \frac{x + 1 - y}{2}, \quad \psi = \frac{x + 1 + y}{2}, \quad \xi = \frac{x + 1 + 3y}{2}. \tag{3.6.2}$$

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

A disadvantage is the more complicated behavior with respect to A_0 and A_1. Namely, if we introduce the vector function $\overrightarrow{h}(t) = (x(t),\ y(t),\ 1)^t$, then we get the following transformation rules:

$$\overrightarrow{h}\left(\frac{t}{2}\right) = C_0 \overrightarrow{h}(t), \qquad \overrightarrow{h}\left(\frac{1+t}{2}\right) = C_1 \overrightarrow{h}(t), \tag{3.6.3}$$

where

$$C_0 = \frac{1}{10}\begin{pmatrix} 5 & 3 & -5 \\ 1 & 3 & 1 \\ 0 & 0 & 10 \end{pmatrix}, \qquad C_1 = \frac{1}{10}\begin{pmatrix} 5 & -3 & 5 \\ -1 & 3 & 1 \\ 0 & 0 & 10 \end{pmatrix}. \tag{3.6.4}$$

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

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

Often, we will not distinguish between y and $\tilde{y}$ and write simply $y(x)$.[4]

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

Theorem 3.8. *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 nontrivial exercise. In my opinion, the best way to prove the theorem is to show that y is a concave function in x, i.e.,

$$y\left(\frac{x_1 + x_2}{2}\right) > \frac{y(x_1) + y(x_2)}{2}. \tag{3.6.5}$$

Exercise 3.4. Show that the derivative $y'(x)$ satisfies the equations

$$y'\left(x\left(\frac{t}{2}\right)\right) = \frac{3y'(x(t)) + 1}{3y'(x(t)) + 5}, \qquad y'\left(x\left(\frac{1+t}{2}\right)\right) = \frac{3y'(x(t)) - 1}{5 - 3y'(x(t))}. \tag{3.6.6}$$

Hint. Use the relations (3.2.14).

The relations (3.6.6) allow us to compute the derivative $y'(x)$ explicitly at some points (knowing that the derivative exists).

For example, 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 $y(-1+\varepsilon) > 0$, 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 in Fig. 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, every 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 periodicity begins.

For example, $\frac{5}{6} = 0.11010101\ldots = 0.1(10) = \frac{5}{2(2^2-1)}$.

[4] The reader must nevertheless distinguish between $y(x)$ and $y(x(t))$.

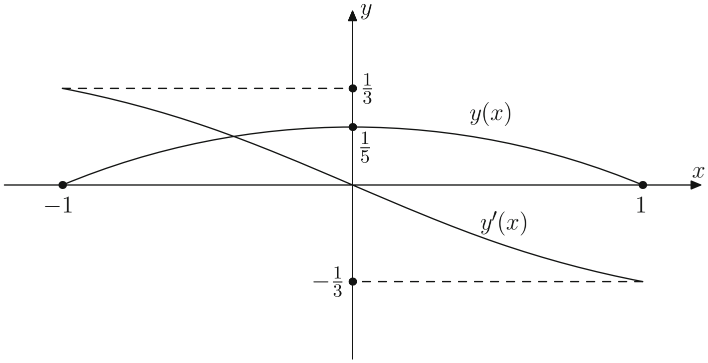


Fig. 3.5 The graphs of the functions $y(x)$ and $y\,'(x)$

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 Sect. 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 (3.6.6).

Exercise 3.5. 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 3.4. Let $\Gamma \subset \mathbb{R}^2$ be the graph of the function $y(x)$. It contains a big subset X of points with rational coordinates. For instance, all the points that correspond to the rational values of the parameter t belong to X.

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

3.7 The Harmonic Image of $\mathcal{S}$

To conclude the first part of the book, we show how the Sierpiński 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 in Fig. 3.6.

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

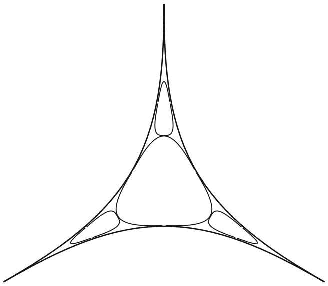


Fig. 3.6 Harmonic image of $\mathcal{S}$

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

3.8 Multidimensional Analogues of $\mathcal{S}$

The Sierpiński gasket has natural analogues in higher dimensions. By definition, the n-dimensional Sierpiński gasket is a self-similar fractal set in $\mathbb{R}^n$ defined by the system of contractions

$$f_i(x) = \frac{x + p_i}{2}, \tag{3.8.1}$$

where the $p_i \in \mathbb{R}^n$, $1 \leq i \leq n + 1$, are not in one hyperplane.

It is not difficult to show that the n-dimensional Sierpiński gasket has Hausdorff dimension $\log_2(n + 1)$ (see Fig. 3.7 for $n = 2$).

Exercise 3.6. Define a projection of the $(2^n - 1)$-dimensional Sierpiński gasket to an n-dimensional plane in such a way that almost all points of the image have a unique preimage.

The theory of harmonic functions on a multidimensional 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 $u^c_{a,b}(t)$.

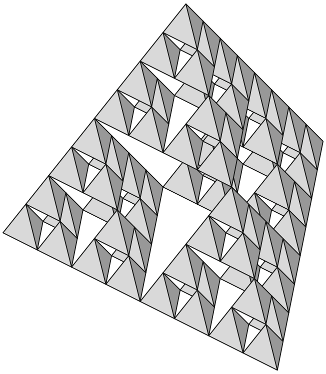


Fig. 3.7 The three-dimensional Sierpiński gasket

We define *basic functions* by

$$\chi(t) = u_{0,1}^{-1}(t), \quad \varphi(t) = u_{0,1}^{0}(t), \quad \psi(t) = u_{0,1}^{n-1}(t), \quad \xi(t) = u_{0,1}^{n}(t), \tag{3.8.2}$$

and the functions $x,\ y$ by

$$x(t) = u_{-1,1}^{0}(t), \qquad y(t) = u_{0,0}^{1}(t). \tag{3.8.3}$$

Then

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

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

Here are the principal relations:

$$\chi(2t) = (n+3)\cdot\chi(t), \qquad \psi(2t) = \frac{n+3}{n+1}\cdot\psi(t); \tag{3.8.4}$$

$$\begin{gathered} \chi(1+\tau) + \chi(1-\tau) = 2 + (n+1)\chi(\tau) \\ \chi(1+\tau) - \chi(1-\tau) = 2\,\tfrac{n+1}{n}\psi(\tau) + \tfrac{(n-1)(n+2)}{n}\chi(\tau); \end{gathered} \tag{3.8.5}$$

$$\begin{gathered} \psi(1+\tau) + \psi(1-\tau) = 2 - \tfrac{n-1}{n+1}\chi(\tau) \\ \psi(1+\tau) - \psi(1-\tau) = \tfrac{2}{n}\,\psi(\tau) + \tfrac{(n-1)(n+2)}{n(n+1)}\chi(\tau). \end{gathered} \tag{3.8.6}$$

These relations allow us to develop the arithmetic theory of basic functions for any integer n, not necessarily positive,[5] analogously to the case $n = 2$.

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

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

Some values of n are of special interest.

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

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

$$\chi(2t) = 3\chi(t), \quad \chi(2^m + k) + \chi(2^m - k) = 2 \cdot 3^m + \chi(k). \tag{3.8.7}$$

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

$$f(k) := \chi(k+1) - 2\chi(k) + \chi(k-1) \quad \text{for any integer} \quad k > 0. \tag{3.8.8}$$

Theorem 3.9. *The function $f(k)$ possesses the following properties:*

$$f(2k) = f(k), \qquad f(2^n + k) + f(2^n - k) = f(k) \quad \textit{for} \quad 0 < k < 2^n. \tag{3.8.9}$$

A 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 it to the reader to consider other negative values for n and discover interesting facts.

Info E. Numerical Systems

E.1. Standard Digital Systems

Most of the real numbers are irrational, so they cannot 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” that contain only finite number of digits.

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

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

$$\mathrm{val}(a) = \sum_{-\infty}^{n} d_{a_k} \cdot b^k. \tag{E.1}$$

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

$$x = \mathrm{val}(a) = \sum_{-\infty}^{n} a_j \cdot b^j. \tag{E.2}$$

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

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

$$\sum_{k \geq 1} (m-1) \cdot m^{-k} = 1. \tag{E.3}$$

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

Motivated by this example, for any numerical system S, we call those numbers that can be written in the form (E.2) with $a_k = 0$ for $k < 0$ *whole numbers*, while *fractional numbers* are those that 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 is denoted by $F(S)$.

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

E.2. Nonstandard Systems

Exercise E.1. 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 every real number can be almost uniquely written in the form (E.2).

Exercise E.2. Introduce a system S with the base $b = 1 + i$ and digits $\{0, 1\}$. Check that here, $W(S) = \mathbb{Z}[i]$, which consists of numbers of the form $a + ib$, $a, b \in \mathbb{Z}$. These are called the *Gaussian integers*. As for $F(S)$, it is a fractal compact

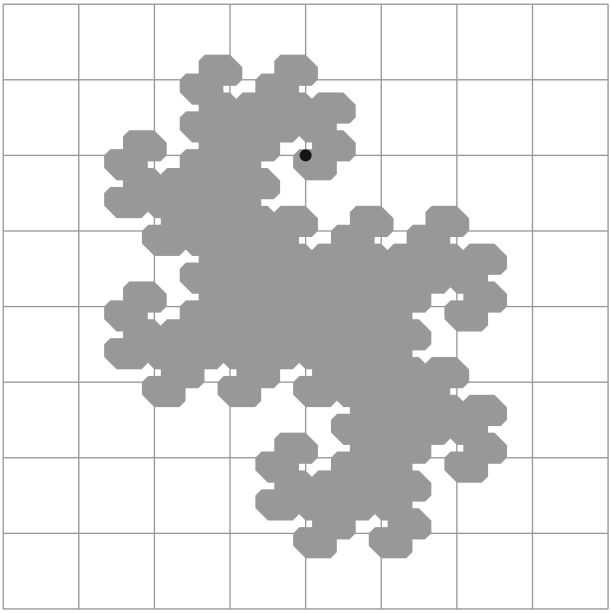


Fig. E.8 The set F

set of dimension 2, determined by the property

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

Here, as always, when an arithmetic operation is applied to a set, it means that it is applied to each element of the set. A picture of this set is shown in Fig. E.8 (taken from the book [Edg90]).

Exercise E.3. Let $\omega = e^{\frac{2\pi i}{3}}$, a cube 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. Continued Fractions

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

$$\mathrm{val}(k) = \cfrac{1}{k_1 + \cfrac{1}{k_2 + \cfrac{1}{k_3 + \cdots + \cfrac{1}{k_n}}}} \tag{E.4}$$

if the sequence k is finite or the limit of the expression (E.4) as $n \to \infty$ 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 continued fraction. As for rational numbers from $(0,\ 1)$, they can be values of two different finite continued fractions: $k = \{k_1, \ldots, k_{n-1}, 1\}$ and $k' = \{k_1, \ldots, k_{n-1} + 1\}$.

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

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

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

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

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

Example. Let $k_n = 2$ for all n. Then $x = \mathrm{val}(k)$ 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 that $x = \sqrt{2} - 1$. So, the square root of 2 is given by an infinite continued fraction,

$$\sqrt{2} = 1 + \cfrac{1}{2 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \ldots}}}},$$

and hence is not a rational number.

This result[6] was known to Pythagoras and kept secret because it undermined his adherents' faith in the power of (rational) numbers.

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

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

First, if the fraction in question is *pure periodic*, i.e., when the number k_n depends only on a residue $n \mod m$ for some m, or *mixed periodic*, when this property holds beginning 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: every 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 continued fraction.

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

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

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

E.4. A General Scheme

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 each 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 $a = \{a_1, a_2 \ldots\}$ we associate the number

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

whenever the infinite product make sense.

Let us explain the relationship to previously described numerical systems.

Let $A_a = \begin{pmatrix} m & 0 \\ a & 1 \end{pmatrix}$, $0 \leq a \leq m-1$. 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

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

This is exactly the standard numerical system.

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 the continued 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 a general identity:

Lemma E.1. *The value of a continued fraction can be computed by the formula*

$$\mathrm{val}(k) = \cfrac{1}{k_1 + \cfrac{1}{k_2 + \cfrac{1}{k_3 + \cdots + \cfrac{1}{k_n}}}} = \frac{(A_{k_1} \cdot A_{k_2} \cdots A_{k_n})_{21}}{(A_{k_1} \cdot A_{k_2} \cdots A_{k_n})_{11}}. \tag{E.5}$$

So, continued fractions form a slight modification of our general scheme.

3.9 Applications of Generalized Numerical Systems

3.9.1 Application to the Sierpiński Gasket

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

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

We also associate the number 0 to the empty sequence.

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

Exercise 3.7. For every infinite sequence a, let us denote by $a^{(n)}$ the sequence of the first n digits of a. Show that the following hold:

(a) The sequence $\mathrm{val}(a^{(n)})$ has a limit as $n \to \infty$. We denote this limit by $\mathrm{val}(a)$.
(b) The point $\mathrm{val}(a)$ belongs to $\mathcal{S}$.
(c) $\mathrm{val}(a) = \mathrm{val}(b)$ iff one sequence can be obtained from the other by substituting the tail of the form $xyyyy\ldots$ by the tail $yxxxx\ldots$.

Exercise 3.8. 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 $\mathcal{S}_n$?

3.9.2 Application to the Question Mark Function

The *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) to 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,[7] the function $\mathbf{?}(\cdot)$ sends a number a represented by the continued fraction

$$a = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_k + \cfrac{1}{\ddots}}}}}$$

to the number

$$\mathbf{?}(a) \;:=\; \sum_{k\geq 1} \frac{(-1)^{k-1}}{2^{a_1+\cdots+a_k-1}} \;=\; \overbrace{0.0\ldots0}^{a_1}\overbrace{1\ldots1}^{a_2}\overbrace{0\ldots0}^{a_3}\ldots.$$

For example,

$$\mathbf{?}\left(\tfrac{3}{7}\right) = \mathbf{?}\left(\frac{1}{2+\frac{1}{3+\frac{1}{\infty}}}\right) = 0.011\overline{0},$$

$$\mathbf{?}\left(\tfrac{\sqrt{2}}{2}\right) = 0.\overline{1100} = \tfrac{4}{5},$$

$$\mathbf{?}\left(\tfrac{e^2-1}{e^2+1}\right) = \sum_{k\geq 0} 2^{-k^2}.$$

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

[7] It would be better to say, by one of the possible definitions (see below).

Part II
The Apollonian Gasket

Introduction

In the second part of the book, we consider another remarkable fractal: the Apollonian gasket $\mathcal{A}$. It consists of circles (or disks) on the two-dimensional sphere and seems rather different from the Sierpiński gasket $\mathcal{S}$. For example, it is not a self-similar fractal (though it can be represented as the union of four subsets homeomorphic to $\mathcal{S}$).

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

Many of facts discussed below are of an elementary geometric nature. However, in modern educational programs, Euclidean geometry occupies a very small place, and we cannot rely on the reader's having acquired the necessary information at school. Therefore, at times, we use more sophisticated tools from algebra and analysis to get the desired geometric 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 4
Circles and Disks on Spheres

Info F. The Conformal Group and Stereographic Projection

F.1 The Conformal Group

The Apollonian gasket, like the Sierpiński gasket, possesses a high degree of symmetry. This symmetry is related to conformal mappings.[1]

For our immediate goals, it is enough to consider only mappings of the extended real plane $\overline{\mathbb{R}}^2$ or the extended complex line $\overline{\mathbb{C}}$. But the main formulas are very similar in all dimensions, and it is natural to speak about conformal mappings of domains in an arbitrary Euclidean space $\mathbb{R}^n$. We will use it later to speak of multidimensional analogues of the Apollonian gasket.

We begin with the following general definition.

Let $\Omega \subset \mathbb{R}^n$ be a domain and let $f : \Omega \to \mathbb{R}^n$ be a smooth map. Then at each point $x \in \Omega$, the derivative Df is a real matrix of size $n \times n$. In terms of coordinates, we have

$$x = (x^1, x^2, \dots, x^n); \quad f = (f^1(x), f^2(x), \dots, f^n(x); \quad Df = \left\| \frac{\partial f^i}{\partial x^j} \right\|. \tag{F.1}$$

Definition F.1. We say that a mapping f is *conformal* if for all $x \in \Omega$, the matrix $Df(x)$ is proportional to an orthogonal matrix $A(x)$.[2]

More precisely, we say that a conformal mapping call f is of the *first kind* if $\det A(x) = 1$, and of the *second kind* if $\det A = -1$.

[1] The notion of conformal mappings makes sense for every Riemannian manifold, but we prefer to keep the exposition on an elementary level.

[2] Recall that a matrix A is said to be *orthogonal* if it satisfies the equation $A^t = A^{-1}$; the corresponding linear operator in $\mathbb{R}^n$ preserves the dot product of vectors.

A.A. Kirillov, *A Tale of Two Fractals*, DOI 10.1007/978-0-8176-8382-5_4,

So infinitesimally, a conformal mapping looks like the composition of a rotation, a dilation and perhaps a reflection. Hence on a small scale, it preserves the geometric form of figures. This explains the name “conformal.”

In the one-dimensional case, this definition is too broad, since it includes all smooth transformations. So in this case, we will use another definition for conformal mappings.

Definition F.2. A function f of one variable is called a *fractional linear transformation*, or *Möbius transformation*, if it has the form

$$f(x) = \frac{ax + c}{bx + d}. \tag{F.2}$$

Such transformations make sense for every field. In particular, we will use real, complex, and quaternionic fractional linear transformations to describe conformal mappings in dimensions $n = 1, 2,\ 3,\ 4$.

In the real case, the transformation (F.2) is by definition a conformal mapping of the extended real line $\overline{\mathbb{R}}$. It belongs to the first or second kind, depending on the sign of $\det \begin{pmatrix} a & c \\ b & d \end{pmatrix} = ad - bc$.

In the complex case, the fractional linear transformations of $\overline{\mathbb{C}}$ (considered as transformations of the extended real plane) are precisely the conformal mappings of the first kind. The conformal mappings of the second kind have the form

$$f(z) = \frac{\alpha \bar{z} + \gamma}{\beta \bar{z} + \delta}. \tag{F.3}$$

In any dimension n, the full conformal group $\overline{G_n}$ of transformations of $\overline{\mathbb{R}}^n$ contains all rotations, translations, dilations, and orthogonal reflections.

There is another remarkable conformal mapping of the second kind that does not belong to any of these types. It is called **inversion** and is denoted by Inv. The formula is

$$\mathrm{Inv}(x) = \frac{x}{|x|^2}. \tag{F.4}$$

Sometimes, Inv is called reflection in the unit sphere. It is justified by the following fact: Inv is conjugate in $\overline{G_n}$ to an ordinary reflection in a hyperplane, and the set of fixed points for Inv, the mirror, is the unit sphere.

It is well known that the group $\overline{G_n}$ is generated by transformations of the above types, including Inv. Moreover, rotations, translations, and Inv already generate $\overline{G_n}$. This property of $\overline{G_n}$ is convenient when we have to show that some geometric structure or object is preserved by conformal mappings. It is enough to check it separately for all rotations and translations and also for Inv.

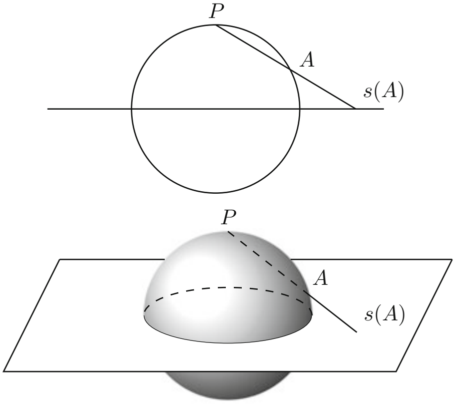


Fig. F.1 Stereographic projection for $n = 1, 2$

The subgroup G_n of conformal mappings of the first kind is characterized by preserving orientation (or by the condition det $Df > 0$). It is a normal subgroup of index 2 in $\overline{G}_n$. As representatives of the two cosets in $\overline{G}_n/G_n$, one can take Id and Inv.[3]

F.2 Stereographic Projection

Here we consider a remarkable example of a conformal map: the stereographic projection s of a sphere S^n to the extended Euclidean space $\overline{\mathbb{R}}^n$.[4]

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 familiar with the subject can start with the case $n = 1$ or $n = 2$, as shown in Fig. F.1. For readers with little or no experience in the terminology of group theory, we recommend first reading Sect. F.4.

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

[3] Here we use the terminology and elementary facts of group theory. The reader can find all necessary information in textbooks on abstract algebra, such as Artin, Michael (1991), *Algebra*, Prentice Hall, ISBN 978-0-89871-510-1.

[4] Until now, we have defined conformal mappings only for domains in extended Euclidean spaces. So strictly speaking, we can not call s a conformal mapping. But using s, we can identify S^n with $\overline{\mathbb{R}}^n$, and conformal mappings of $\overline{\mathbb{R}}^n$ become transformations of S^n. The fact is that they are exactly conformal transformations of S^n as a Riemannian manifold.

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 a 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

$$s(\alpha) = \left(0, \quad \frac{\alpha_1}{1-\alpha_0}, \quad \frac{\alpha_2}{1-\alpha_0}, \quad \dots, \quad \frac{\alpha_n}{1-\alpha_0}\right). \tag{F.5}$$

The inverse map has the form

$$s^{-1}(x) = \left(\frac{|x|^2-1}{|x|^2+1}, \quad \frac{2x_1}{1+|x|^2}, \quad \frac{2x_2}{1+|x|^2}, \quad \dots, \quad \frac{2x_n}{1+|x|^2}\right), \tag{F.6}$$

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

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

Hint. Check that $(1-\alpha_0)X + \alpha_0 P = A$.

So, our map s is geometrically 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 $\alpha_0 = 0$.

Neither the algebraic nor the geometric definition of s makes sense at the point P. We assume additionally that $s(P) = \infty \in \overline{\mathbb{R}}^n$. The map s defined in this way is a bijection between S^n and $\overline{\mathbb{R}}^n$.

The map s transfers the conformal group $\overline{G}_n$ acting on $\overline{\mathbb{R}}^n$ into some group of transformations of the sphere S^n. These are exactly the conformal mappings of S^n as a Riemannian manifold, but we do not discuss this here.

Instead, we show that s sends disks to disks (and circles to circles), which it must do as a conformal mapping.

A general hyperplane in $\mathbb{R}^{n+1}$ is given by the linear equation

$$p_0\alpha_0 + p_1\alpha_1 + \dots + p_n\alpha_n + p_{n+1} = 0. \tag{F.7}$$

This hyperplane divides $\mathbb{R}^{n+1}$ into two half-spaces $H_p^{\pm}$, where the left-hand side of Eq. (F.7) is positive or negative.

Now, fixing $p_0, \dots, p_{n+1}$, the extremal values of the left-hand side of Eq. (F.7) on the unit sphere $\sum_{k=0}^{n} |\alpha_k|^2 = 1$ are $\pm r + p_{n+1}$, where $r = \left(\sum_{k=0}^{n} p_k^2\right)^{\frac{1}{2}}$.

Therefore, the hyperplane H_p intersects the unit sphere along a nontrivial $(n-1)$-sphere when $|p_{n+1}| < r$. We write this condition in the form

$$p_{n+1}^2 - p_0^2 - p_1^2 - \dots - p_n^2 < 0. \tag{F.8}$$

If Eq. (F.8) is satisfied, the intersection of S^n with H_p^- is a disk $\tilde{D}_p \in S^n$, given in coordinates $\alpha_0, \alpha_1, \dots, \alpha_n$ by the linear inequality

$$p_0\alpha_0 + p_1\alpha_1 + \dots + p_n\alpha_n + p_{n+1} \le 0. \tag{F.9}$$

It is natural to consider the vector p an element of the Minkowski space $\mathbb{R}^{1,n}$ (see below) and denote the left-hand side of Eq. (F.8) by $|p|^2$.

Since the inequality (F.9) does not change its meaning after multiplication of p by a positive constant, we can normalize p by the condition[5] $|p|^2 = -1$.

Expressing $\{\alpha_i\}$ in terms of the coordinates $\{x_j\}$ of the point $s(\alpha)$, we get an inequality defining the disk $D_p := s(\tilde{D}_p) \subset \overline{\mathbb{R}}^n$ 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

$$a + (\overrightarrow{p}, \overrightarrow{x}) + c|\overrightarrow{x}|^2 \leq 0, \tag{F.10}$$

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

Now we can use the equation $|p|^2 = ac - |\overrightarrow{p}|^2$ and the normalization $|p|^2 = -1$ to write our inequality as follows:

$$c \cdot \left| x + \frac{\overrightarrow{p}}{c} \right|^2 \leq c^{-1}. \tag{F.11}$$

The last inequality for $c > 0$ describes a ball in $\mathbb{R}^n$ with center $-\frac{\mathbf{p}}{c}$ and radius $\frac{1}{c}$.

If $c < 0$, then Eq. (F.11) describes the complement of a ball with center $-\frac{\mathbf{p}}{c}$ and radius $-\frac{1}{c}$. We agree to associate to this generalized ball the negative curvature c. Its preimage on S^n contains the north pole inside.

Finally, if $c = 0$, then Eq. (F.11) does not make sense, and Eq. (F.10) defines a half-space (the corresponding ball $\tilde{D} \subset S^n$ in this case contains the north pole as a boundary point).

F.3 The Matrix Definition of $\overline{G}_n$

We discuss one more definition of the group $\overline{G}_n$, one that is useful in practical computations. It follows from the matrix realization of the groups $\overline{G}_n$ and G_n. Let $\mathbb{R}^{1,n+1}$ be the $(1, n+1)$-dimensional Minkowski space. It is a Euclidean space of dimension $n+2$ endowed with a bilinear form of signature $(1, n+1)$. In standard coordinates $x = (x^0, x^1, \ldots, x^n, x^{n+1})$, the form looks like

$$(x, y) = x^0y^0 - x^1y^1 - \cdots - x^ny^n - x^{n+1}y^{n+1}. \tag{F.12}$$

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

We use the standard notation $\mathrm{O}(1, n+1; \mathbb{R})$ for the group of invertible linear transformations of $\mathbb{R}^{1,n+1}$ preserving the form (F.12). By $\mathrm{SO}(1, n+1; \mathbb{R})$, we denote the subgroup of matrices with determinant 1.

Elements $g \in \mathrm{O}(1, n+1; \mathbb{R})$ are real matrices of size $n+2$. In the standard basis, they have the block form

$$\left(\begin{array}{|c|c|} a & b \\ \hline c & d \end{array}\right), \tag{F.13}$$

where a is a real number, b is a row $(n+1)$-vector, c is a column $(n+1)$-vector, and d is a square $(n+1) \times (n+1)$ matrix. In order for g to preserve the bilinear form, a, b, c, d must satisfy the conditions

$$a^2 = 1 + bb^t; \quad dd^t = 1_{n+1} + cc^t; \quad ca = bd^t, \tag{F.14}$$

which imply the inequalities

$$a \neq 0, \quad \det d \neq 0. \tag{F.15}$$

Thus, the group $\mathrm{O}(1, n+1; \mathbb{R})$ splits into four separate parts according to the signs of a and $\det d$. It is known that these parts are connected, open in the group $\mathrm{O}(1, n+1; \mathbb{R})$, and closed in the set $\mathrm{Mat}_{n+2}(\mathbb{R})$ of all matrices. The part with $a > 0$, $\det d > 0$ that contains the unit is itself a group, denoted by $\mathrm{SO}_+(1, n+1; \mathbb{R})$; the second component of $\mathrm{SO}(1, n+1; \mathbb{R})$ with $a < 0$, $\det d < 0$ is denoted by $\mathrm{SO}_-(1, n+1; \mathbb{R})$.

Consider now the *projective space* $P^{n+1}(\mathbb{R})$ corresponding to $\mathbb{R}^{1,n+1}$. It is a set, obtained from $\mathbb{R}^{1,n+1}$ by deleting the origin and identifying the proportional vectors.

The ordinary coordinates $\{x^i\}$ of a point $x \in \mathbb{R}^{1,n+1} \backslash \{0\}$ are by definition the homogeneous coordinates of the corresponding point $[x] \in P^n(\mathbb{R})$. They are written as $(x^0 : x^1 : \cdots : x^{n+1})$ to emphasize that only their ratios matter. The linear action of the group $\mathrm{O}(1, n+1; \mathbb{R})$ on the $\mathbb{R}^{1,n+1}$ defines the group $\mathrm{PO}(1, n+1; \mathbb{R})$ of *projective* transformations of $P^{n+1}(\mathbb{R})$.

Introduce the sets

$$\begin{array}{c} U_+ := \{x \in \mathbb{R}^{1,n+1} | (x, x) > 0,\ x^0 > 0\} \text{ and} \\ U_- := \{x \in \mathbb{R}^{1,n+1} | (x, x) > 0,\ x^0 < 0\}. \end{array} \tag{F.16}$$

They are stable under the action of $\mathrm{SO}_+(1, n+1; \mathbb{R})$ and can be interchanged by other components.

Choose the inhomogeneous coordinates $\alpha^i = x^i / x^0$ on $P^n(\mathbb{R})$. In these coordinates, the images of U_+ and U_- both coincide with the interior of the unit ball $B : \sum_{k=1}^{n+1} (\alpha^k)^2 < 1$. The boundary of this ball is the n-dimensional sphere given by the equation

$$\partial B : \sum_{k=1}^{n+1} (\alpha^k)^2 = 1. \tag{F.17}$$

Thus, we get the projective action of $\mathrm{O}(1, n+1; \mathbb{R})$ on the set $S^n \subset P^{n+1}(\mathbb{R})$ (i.e., the action of the corresponding projective group $\mathrm{PO}(1, n+1; \mathbb{R})$). This action is not faithful and has a kernel of order 2. The following theorem is the basis of our matrix definition of the conformal group.

Theorem F.1. *Every conformal mapping of S^n corresponds to two elements of* $\mathrm{PO}(1, n+1; \mathbb{R})$*, while every conformal mapping of the first kind corresponds to the unique element of* $\mathrm{PSO}_+(1, n+1; \mathbb{R})$.

Here we conclude our general survey, and in the next section, we shall consider cases of small dimension.

F.4 Small Dimensions

In the main text, we shall consider mostly the case $n = 2$ and also briefly the cases $n = 3, n = 4$. In all these cases, the conformal group $\overline{G}_n$ has additional properties, which we discuss here.

Case $n = 1$. This case is not important from the point of view of our tiling problem. So I advise the reader to skip this part of the section and come back to it after you have understood the more general situation.

The full group of symmetries for $n = 1$ is the group of fractional linear transformations (F.2) of a real variable x. The basic space is $P^1(\mathbb{R})$, or the circle S^1. The disks are just arcs of the circle, and the Apollonian gasket reduces here to covering a circle by three neighboring arcs.

Case $n = 2$. The group G_2 is isomorphic to $\mathrm{PSO}_+(1, 3; \mathbb{R})$ and also to the Möbius group $\mathrm{PSL}(2, \mathbb{C}) = \mathrm{PGL}(2, \mathbb{C})$.[6] The group $\overline{G}_2$ is isomorphic to $\mathrm{PO}_+(1, 3; \mathbb{R})$ and also to the extended Möbius group. Recall that the Möbius group acts on $\overline{\mathbb{C}}$ by fractional linear (or Möbius) transformations

$$w \to \frac{\alpha w + \beta}{\gamma w + \delta}, \quad \text{where} \quad \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in \mathrm{SL}(2, \mathbb{C}). \tag{F.18}$$

Besides these transformations, the extended Möbius group also contains complex conjugation, whence all transformation of the form

$$w \to \frac{\alpha \overline{w} + \beta}{\gamma \overline{w} + \delta} \quad \text{where} \quad \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in \mathrm{SL}(2, \mathbb{C}). \tag{F.19}$$

Among these transformations there are the *reflections* s that 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.

[6] These two groups coincide, because every matrix from $\mathrm{GL}(2, \mathbb{C})$ is proportional to a matrix from $\mathrm{SL}(2, \mathbb{C})$.

If the circle M degenerates to a straight line l, then the transformation s_M is an ordinary reflection in l. For the unit circle M_0 centered at the origin, the reflection s_{M_0} coincides with the inversion Inv defined by Eq. (F.4). In general, s_M can be defined as $g \circ \mathrm{Inv} \circ g^{-1}$, where $g \in G_2$ is any transformation that sends C to M.

Exercise F.2. Show that all reflections form a single conjugacy class in $\overline{G}_2$.

Hint. Show that $\overline{G}_2$ acts transitively on $\mathcal{D}$.

Exercise F.3. Show that the group $\overline{G}_2$ is generated by reflections.

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

$$g(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \ t \in \mathbb{C}, \quad \text{and} \quad s = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{F.20}$$

Exercise F.4. Show that the conjugacy classes in G_2 are precisely the level sets $I(g) = const$ for the function

$$I(g) := \frac{(\operatorname{tr} g)^2}{\det g} - 4 \tag{F.21}$$

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

Exercise F.5. Show that all involutions in G_2 form two conjugacy classes: the unit class and the class that contains a rotation of S^2 through 180° around z-axis.

Exercise F.6. Show that all involutions in $\mathrm{PO}_{-}(1, 3; \mathbb{R})$ that 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_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* $g \in G_2$ *such that* $g(z_i) = w_i$, $i = 1, 2, 3$.

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

$$g_{z_1,z_2,z_3}(z) = \frac{z - z_1}{z - z_3} : \frac{z_2 - z_1}{z_2 - z_3}. \tag{F.22}$$

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

Proposition F.2. *Every circle and every straight line is mapped by transformations* $g \in G_2$ *to a circle or a straight line. (Alternatively, every disk goes to a disk.)*

To prove this statement we use the following lemma.

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

$$a + \bar{b}w + b\bar{w} + cw\bar{w} \le 0 \tag{F.23}$$

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

(a) a closed disk with radius $r = c^{-1}$ and center $-\frac{b}{c}$ when $c > 0$;
(b) the complement of an open disk with radius $r = -c^{-1}$ and center $-\frac{b}{c}$ when $c < 0$;
(c) a closed half-plane when $c = 0$.

Moreover, every disk $D \in \mathcal{D}$ can be given by an inequality of the form (F.23).

Proof. This is just a particular case of Eq. (F.21). □

Proposition F.2 follows from Lemma F.1 because the inequality (F.23) goes to an inequality of the same kind under transformations (F.20), hence under all fractional linear transformations.

Remark F.1. Note that the set $\overline{G}_2\backslash G_2$ of conformal mappings of the second kind does not form a group. It is a two-sided coset in $\overline{G}_2$ with respect to G_2. It is useful to know that it possesses both properties listed in Propositions F.1 and F.2: it acts simply transitively on triples of distinct points in $\overline{\mathbb{C}}$ and preserves circles and disks.

♡

Case $n = 3$. The group $G_3 = \mathrm{PSO}_0(1, 4; \mathbb{R})$ is isomorphic to the group $\mathrm{PU}(1, 1; \mathbb{H})$, which is the quotient of $\mathrm{U}(1, 1; \mathbb{H})$ by its center $\{\pm 1_2\}$. The group $\mathrm{U}(1, 1; \mathbb{H})$ consists of quaternionic[7] matrices $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ satisfying

$$g^* \cdot \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \cdot g = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

which is equivalent to the system of equations

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

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 $c = v\bar{w}u \sinh t$.[8]

[7] The algebra of *quaternions* is a four-dimensional real noncommutative algebra. It can be realized as a subalgebra of $\mathrm{Mat}_2(\mathbb{C})$ or of $\mathrm{Mat}_4(\mathbb{R})$.

[8] Recall that the hyperbolic functions are defined as $\cosh t = \frac{e^t + e^{-t}}{2}$, $\sinh t = \frac{e^t - e^{-t}}{2}$.

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 the union of $S^3 \times S^3 \times S^3 \times (\mathbb{R}\backslash\{0\})$ and $S^3 \times S^3$.

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

Case $n = 4$. The group $G_4 = \mathrm{PO}_{+}(1, 5; \mathbb{R})$ is isomorphic to another quaternionic group, $\mathrm{PGL}(2,\ \mathbb{H}) = \mathrm{GL}(2,\ \mathbb{H})/\mathbb{R}^{\times\cdot}1_2$, which acts on the 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 $q \mapsto (aq + b)(cq + d)^{-1}$.

4.1 Descartes's Theorem on Disks in the Plane

We start with a simple looking geometric problem:

Describe all configurations of four mutually tangent circles in the plane.

Examples of such configurations are shown in Fig. 4.1. We include the cases in which one of the circles degenerates to a straight line (a circle of infinite radius) and in which one of the circles is tangent to others from inside (we shall interpret the later as a circle with negative radius).

There exist some other configurations that we want to exclude. They are shown in Fig. 4.2. Here all four circles have a common point of tangency, finite or infinite. The reason why these configurations are excluded will be clear when we make the formulation of the problem more precise and pass from circles to disks.

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

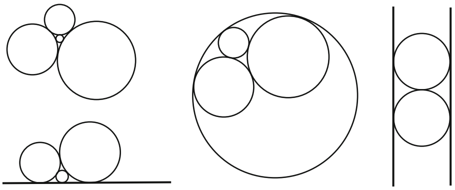


Fig. 4.1 Quadruples of tangent circles

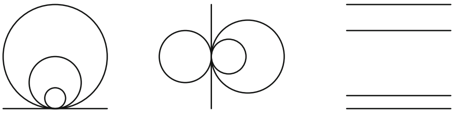


Fig. 4.2 "Wrong quadruples"

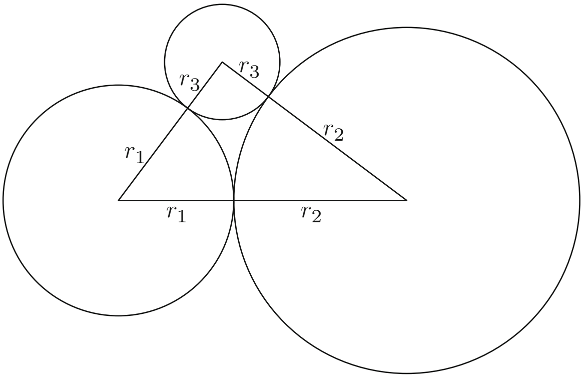


Fig. 4.3 Triple of tangent circles (a)

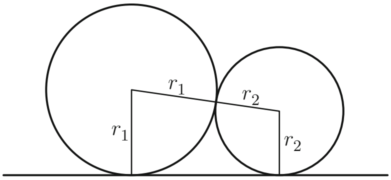


Fig. 4.4 Triple of tangent circles (b)

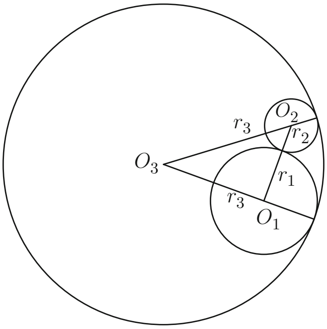


Fig. 4.5 Triple of tangent circles (c)

To approach our problem, we take one step back and consider a triple of mutually tangent circles. There are three kinds of such triples; see Figs. 4.3–4.5.

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

In 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 the 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

$$|O_iO_j| + |O_jO_k| = (r_i + r_j) + (r_j + r_k) \geq r_i + r_k = |O_iO_k|. \qquad (4.1.1)$$

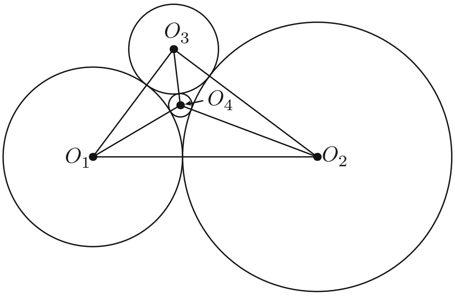


Fig. 4.6 Toward the proof of Descartes's theorem

In 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 general formula that is valid in both cases (a) and (c):

$$|O_i O_j| \;=\; |r_i + r_j|, \tag{4.1.2}$$

provided that we replace r_3 by $-r_3$. Then Eq. (4.1.2) will be satisfied if $r_1+r_2 \leq |r_3|$, or $r_1 + r_2 + r_3 \leq 0$.

In case (b), the center O_3 is situated at infinity. We put $r_3 = \infty$, and Eq. (4.1.2), suitably interpreted, is still satisfied.

If four circles are mutually tangent, then their radii $r_1,\ r_2,\ r_3,\ r_4$ are not arbitrary but must satisfy a particular equation. That equation and some of its consequences were apparently known in ancient Greece, more than two thousand years ago.

More recently, the condition was explicitly given by René Descartes, the famous French mathematician and philosopher of the first half of the seventeenth century. So we call it Descartes's equation.

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

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

The geometric meaning of the quantity c_i is the curvature of the circle of radius r_i.[9]

The equation in question looks as follows:

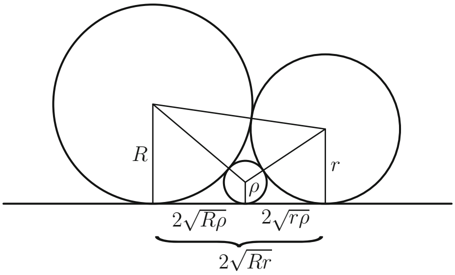
$$(c_1 + c_2 + c_3 + c_4)^2 \;-\; 2(c_1^2 + c_2^2 + c_3^2 + c_4^2) \;=\; 0. \tag{4.1.3}$$

We leave to geometry fans the challenge of recovering the proof of Descartes's theorem using high-school geometry. The following exercise and Fig. 4.6 might help.

Exercise 4.1. Show that the following formula for the area of the triangle $O_1O_2O_3$ above is true for both cases (a) and (c):

[9] The reason why curvatures are better than radii will be explained later, when we develop a group-theoretic approach to the problem. We shall see that the group transformations act linearly in terms of curvatures but not in terms of radii.

Fig. 4.7 Degenerate Descartes's equation



$$S = \sqrt{r_1 r_2 r_3 (r_1 + r_2 + r_3)}. \tag{4.1.4}$$

Note that the expression under the square root sign is always positive.

Hint. Use Heron's formula.

There is a special case of Descartes's theorem that is much easier to prove. Namely, assume that one of the four circles degenerates to a straight line. Let, for example, $c_4 = 0$, so that the relation between the remaining curvatures is

$$(c_1 + c_2 + c_3)^2 - 2(c_1^2 + c_2^2 + c_3^2) = 0. \tag{4.1.5}$$

Fortunately, the left-hand side of Eq. (4.1.5) can be decomposed into simple factors. To this end, we rewrite it in the form of a 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 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

$$\begin{aligned} &-\Big(c_1 - (\sqrt{c_2} + \sqrt{c_3})^2\Big)\Big(c_1 - (\sqrt{c_2} - \sqrt{c_3})^2\Big) \\ &\quad = (\sqrt{c_1} + \sqrt{c_2} + \sqrt{c_3})(-\sqrt{c_1} + \sqrt{c_2} + \sqrt{c_3})(\sqrt{c_1} - \sqrt{c_2} + \sqrt{c_3}) \\ &\qquad \times (\sqrt{c_1} + \sqrt{c_2} - \sqrt{c_3}). \end{aligned}$$

It follows that Eq. (4.1.5) is true iff at least one of the following equations is satisfied:

$$\sqrt{c_1} \pm \sqrt{c_2} \pm \sqrt{c_3} = 0 \quad \text{or} \quad \sqrt{r_2r_3} \pm \sqrt{r_1r_2} \pm \sqrt{r_1r_3} = 0. \tag{4.1.6}$$

The signs in fact depend on the relative sizes of the radii. Thus, for example, 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 Figs. 4.7 and 4.8.

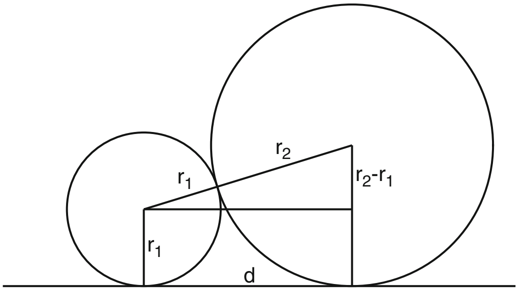


Fig. 4.8 Degenerate triple with $d = 2\sqrt{r_1 r_2}$

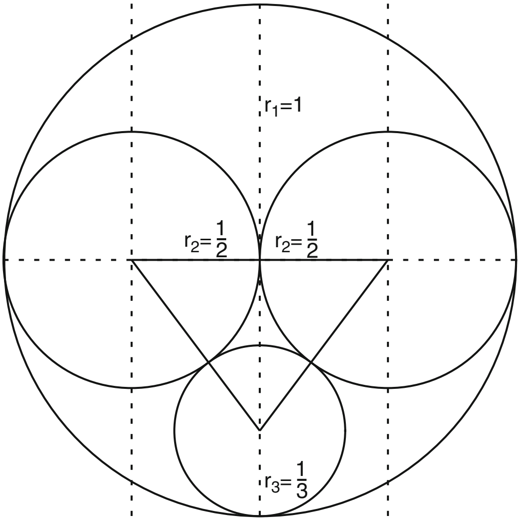


Fig. 4.9 "Violation" of Descartes's equation

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

Namely, we did not take into account the fact that the curvature is a signed quantity: it can be positive or negative. Neglecting this may make the formula (4.1.3) incorrect. Indeed, let us check the equality (4.1.3) in the case shown in Fig. 4.9.

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

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

But if we set 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 at 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 this circle as having negative radius -1.

To make the exposition rigorous, we need either to introduce a notion of orientation for our circles or to consider, instead of circles, the solid disks bounded by them. Readers who are acquainted with the elements of algebraic topology will perhaps prefer the first option. Those who want to remain in the framework of school geometry can simply consider solid disks on the two-dimensional sphere S^2 instead of circles in the plane $\mathbb{R}^2$.

Remark 4.1. These two possibilities are in fact equivalent. Indeed, every disk inherits an orientation from the ambient plane or sphere, and the boundary of an oriented disk 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. 5.3 bounds the domain that is complementary to the unit disk. So we are forced to include domains of this sort in our considerations.

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 the two-dimensional sphere S^2 using stereographic projection (see Info F). Under this identification, the "generalized disks" go the ordinary disks on S^2 that contain the north pole inside. Those disks that 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 disks on the two-dimensional sphere S^2. To each disk $D \in \mathcal{D}$, there corresponds an oriented circle $C = \partial D$.

We can (and will) identify $S^2 \simeq \overline{\mathbb{R}}^2$ with the extended complex plane $\overline{\mathbb{C}}$ and consider our disks and circles subsets of $\overline{\mathbb{C}}$.

Let us say that two disks are **tangent** if they have exactly one common point. In terms of oriented circles, this 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 in Fig. 4.2: they have "positive tangency" and do not correspond to a configuration of four mutually tangent disks.

Let C be an oriented circle of ordinary radius r on $\overline{\mathbb{C}}$. We say that

- C has $c = r^{-1}$ if C bounds an ordinary disk;
- C has curvature $c = -r^{-1}$ if it is the boundary of the complement of a disk;
- C has curvature $c = 0$ if the circle is actually a straight line.

In particular, the outer circle in Fig. 5.3 corresponds to the complement of the open unit disk. Therefore, the curvature of its boundary is -1.

Remark 4.2. Let us look in greater detail at the set of solutions to Eq. (4.1.3).

Note first that if the quadruple (c_1, c_2, c_3, c_4) is a solution, then so is $(-c_1, -c_2, -c_3, -c_4)$.

Further, Eq. (4.1.3) can be written in the form

$$2(c_1 + c_2)(c_3 + c_4) = (c_1 - c_2)^2 + (c_3 - c_4)^2. \tag{4.1.7}$$

We see that either $c_1 + c_2 \geq 0$ and $c_3 + c_4 \geq 0$, or $c_1 + c_2 \leq 0$ and $c_3 + c_4 \leq 0$. Suppose that values are chosen such that $c_1 \geq c_2 \geq c_3 \geq c_4$. Then in the first case, we have $|c_4| \leq c_3 \leq c_2 \leq c_1$, while in the second case, we have $c_4 \leq c_3 \leq c_2 \leq -|c_1|$.

Only in the first case can our solution be interpreted as a set of curvatures of four mutually tangent disks. So only this case will be considered below.[10]

Thus, from now on, we can assume that one of the following situations occurs:

(a) All numbers c_i are positive.
(b) Three numbers are positive, while the fourth is negative and smaller in absolute value than the others.
(c) Three numbers are positive and the fourth is 0.
(d) Two of the c_i are positive and equal, while the other two are equal to zero.

♡

4.2 Proof of Descartes's Theorem for $n = 2$

Here we give a short algebraic proof of Descartes's theorem. We organize this proof in such a way that later, we can prove s stronger and more general theorem by the same method with small modifications.

Let $\mathbb{R}^{1,3}$ be the four-dimensional real vector space with coordinates t, x, y, z and with the indefinite scalar product

$$(p_1, p_2) = t_1 t_2 - x_1 x_2 - y_1 y_2 - z_1 z_2. \tag{4.2.1}$$

The space $\mathbb{R}^{1,3}$ is called Minkowski space and is a basic object in the special theory of relativity. 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 *timelike*, *lightlike*, or *spacelike*. The timelike vectors are of two kinds: future vectors with $t > 0$ and past vectors with $t < 0$.

The physical meaning of p is an *event* that take place at the moment t of time at the point $(x, y, z) \in \mathbb{R}^3$.

[10] In fact, the solutions of the second kind also can be associated with tangent disks, but on the sphere with the opposite orientation.

Physicists call the the group L of all linear transformations of $\mathbb{R}^{1,3}$ that preserve the scalar product (4.2.1) the *full Lorentz group*. It splits into four connected components, and the component containing the unit is called the *proper Lorentz group* L_0. In mathematical papers, these groups are denoted by $\mathrm{O}(1, 3)$ and $\mathrm{SO}_+(1, 3)$, respectively (see the details in Info F).

The *relativity principle* claims that all physical laws are invariant under the proper Lorentz group.

Algebraically, elements $g \in \mathrm{O}(1, 3)$ are given by 4×4 real matrices $|g_{i,j}|$ whose rows (columns) are mutually 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.[11]

An element $g \in \mathrm{O}(1, 3)$ 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 disks on the unit sphere. A disk on S^2 can be defined as the intersection of S^2 with a half-space $H_{u,\tau}$ given by

$$H_{u,\tau} = \{v \in \mathbb{R}^3 \mid (u, v) + \tau \leq 0\}, \quad \text{where} \quad u \in S^2 \quad \text{and} \quad \tau \in (-1, 1). \tag{4.2.2}$$

Instead of the pair $(u, \tau) \in S^2 \times (-1, 1)$, we can use the one spacelike vector $p = (t, x, y, z) \in \mathbb{R}^{1,3}$ given by

$$p = \frac{1}{\sqrt{1 + \tau^2}} \cdot (\tau, u).$$

Namely, the half-space in question has the form

$$H_p = \{v \in \mathbb{R}^3 \mid xv^1 + yv^2 + zv^3 + t \leq 0\}. \tag{4.2.3}$$

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 $|p|^2 = -1$.

The space $\mathcal{D}$ of disks on S^2 is thus identified with the set P_{-1} of all spacelike vectors $p \in \mathbb{R}^{1,3}$ with $|p|^2 = -1$. It is well known that P_{-1} is a hyperboloid of one sheet in $\mathbb{R}^4$ and that the group $L_0 \simeq \mathrm{SO}_+(1, 3; \mathbb{R})$ acts transitively on it. The stabilizer of the point $(0, 0, 0, 1)$ is isomorphic to the group $\mathrm{SO}_+(1, 2; \mathbb{R})$, which is naturally embedded in L_0.

We get our first interpretation of $\mathcal{D}$ as a homogeneous manifold.

Exercise 4.2. Show that the three-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.

[11] Compare with the properties of the usual orthogonal matrices: all rows (columns) have length 1 and are orthogonal to one another.

Our next interpretation of the space $\mathcal{D}$ uses complex matrix theory. We start with inequality (F.23) and collect the coefficients on 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 a 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 $M = -1$.

Exercise 4.3. Show that the two previous interpretations are related 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

$$a = t - z, \quad b = x + iy, \quad c = t + z. \tag{4.2.4}$$

Hint. Compare Eqs. (4.2.3) and (F.23).

Now we want to describe $\mathcal{D}$ in the second interpretation as a homogeneous space.

We have already seen the action of the group $G = \mathrm{PSL}(2,\ \mathbb{C})$ on $\overline{\mathbb{C}}$ by fractional linear transformations. Moreover, by Proposition F.2, G_2 acts on the set $\mathcal{D}$ of all disks on $\overline{\mathbb{C}}$.

On the other hand, the group $\mathrm{SL}(2, \mathbb{C})$ acts on the set H of Hermitian 2×2 matrices by the rule

$$g: \quad M \mapsto gMg^*, \tag{4.2.5}$$

and this action preserves the set H_{-1} of matrices with determinant -1. (Actually, this is a G-action, since the center C of $\mathrm{SL}(2, \mathbb{C})$ acts trivially.)

Theorem 4.1. *There exists a homomorphism* $\pi\colon \mathrm{SL}(2,\ \mathbb{C}) \to L_0 \simeq \mathrm{SO}_0(1,\ 3;\ \mathbb{R})$ *such that the following diagram is commutative:*

$$\begin{tikzcd}
G & \times & \mathcal{D} \arrow[r] & \mathcal{D} \\
\mathrm{SL}(2,\ \mathbb{C}) \arrow[u, "p"] \arrow[d, "\pi"'] & \times & H_{-1} \arrow[u, "\Vert"'] \arrow[d, "\Vert"] \arrow[r] & H_{-1} \arrow[u, "\Vert"'] \arrow[d, "\Vert"] \\
L_0 & \times & P_{-1} \arrow[r] & P_{-1}
\end{tikzcd}$$

where p *is the natural projection of* $\mathrm{SL}(2,\ \mathbb{C})$ *to* $\mathrm{PSL}(2,\ \mathbb{C}) \simeq G$ *and horizontal arrows denote the actions.*

Here we use a convenient way of formulating mathematical statements in the form of commutative diagrams. A diagram consisting of vertices (which denote sets) and arrows (which denote maps) is called *commutative* if the composition of maps along some path joining two vertices depends only on those vertices but not of the choice of path.

We leave the verification to the reader but give here the explicit formula for the homomorphism π.

Exercise 4.4. Show that the homomorphism π has the form

$$\pi\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \frac{|a|^2+|b|^2+|c|^2+|d|^2}{2} & \operatorname{Re}(a\bar{b} + c\bar{d}) & \operatorname{Im}(\bar{a}b + \bar{c}d) & \frac{|b|^2-|a|^2-|c|^2+|d|^2}{2} \\ \operatorname{Re}(a\bar{c} + b\bar{d}) & \operatorname{Re}(a\bar{d} + b\bar{c}) & \operatorname{Im}(\bar{a}d - \bar{b}c) & \operatorname{Re}(b\bar{d} - a\bar{c}) \\ \operatorname{Im}(a\bar{c} + b\bar{d}) & \operatorname{Im}(a\bar{d} + b\bar{c}) & \operatorname{Re}(\bar{a}d - \bar{b}c) & \operatorname{Im}(b\bar{d} - a\bar{c}) \\ \frac{|c|^2-|a|^2-|b|^2+|d|^2}{2} & \operatorname{Re}(\bar{c}d - \bar{a}b) & \operatorname{Im}(\bar{c}d - \bar{a}b) & \frac{|a|^2-|b|^2-|c|^2+|d|^2}{2} \end{pmatrix}.$$

Remark 4.3. The inverse map of $\mathrm{SO}_+(1,\ 3;\ \mathbb{R}) \to \mathrm{PSL}(2,\ \mathbb{C})$ is well defined, but its lifting to $\mathrm{SL}(2,\ \mathbb{C})$ is defined only up to sign. It is called the *spinor representation* of $\mathrm{SO}_+(1,\ 3;\ \mathbb{R})$.

In particular, all products of the form $2a\bar{a}$, $2a\bar{b}, \ldots$, etc., are well defined and given in the following 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 4.5. Describe the image under π of the following subgroups of G:
(a) $\mathrm{PGL}(2,\ \mathbb{R})$; (b) $\mathrm{PSU}(2,\ \mathbb{C})$; (c) $\mathrm{PSU}(1, 1;\ \mathbb{C})$.

Hint. Use the fact that the elements of the subgroup in question are stabilizers of some geometric objects.

Answers:

(a) $\pi\big(\mathrm{PGL}(2,\ \mathbb{R})\big) = \mathrm{Stab}\,(0,\ 0,\ 1,\ 0) \simeq \mathrm{SO}_+(1,\ 2;\ \mathbb{R})$;
(b) $\pi\big(\mathrm{PSU}(2,\ \mathbb{C})\big) = \mathrm{Stab}\,(1,\ 0,\ 0,\ 0) \simeq \mathrm{SO}(3,\ \mathbb{R})$;
(c) $\pi\big(\mathrm{PSU}(1, 1;\ \mathbb{C})\big) = \mathrm{Stab}\,(0,\ 0,\ 0,\ 1) \simeq \mathrm{SO}_+(1,\ 2;\ \mathbb{R})$.

An interesting problem is to compare the image under π of the subgroup $\mathrm{SL}(2,\ \mathbb{Z} + i\mathbb{Z})$ with the subgroup $\mathrm{SO}_+(1,\ 3;\ \mathbb{Z})$.

4.2.1 Generalized Descartes's Theorem

We use the terminology of Minkowski space introduced at the beginning of this section.

Let D_i, $1 \le i \le 4$, be four mutually tangent disks. Denote by p_i, (resp. M_i) the corresponding spacelike vectors with $|p_i|^2 = -1$ (resp. the Hermitian matrices with $\det M_i = -1$).

Lemma 4.1. *The disks 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* $tr\,(M_1 + M_2) > 0$

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 the 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 the horizontal line $2i + \mathbb{R}$ with a 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 find a Möbius transformation such that the 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-coordinate is preserved by G and so is the sign of the trace of M when $\det M = 0$. □

We now return to the theorem. Consider the *Gram matrix* of scalar products for p_i. According to Lemma 4.1, it looks as follows:

$$G_{ij} := (p_i, p_j) = 1 - 2\delta_{ij}. \tag{4.2.6}$$

It is well known that the determinant of the Gram matrix of a system of n vectors in $\mathbb{R}^n$ equals the square of the determinant consisting of the coordinates of these vectors. The same is true up to sign 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 the vectors p_i are linearly independent, and hence they 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:

$$v_i = (v, p_i); \quad v = \sum_{j=1}^{4} v^j \cdot p_j. \tag{4.2.7}$$

Let us find the relationship between these coordinates. From Eqs. (4.2.6) and (4.2.7), we have

$$v_i = \left(\sum_{j=1}^{4} v^j \cdot p_j,\ p_i\right) = \sum_{j=1}^{4} G_{ij} v^j = \sum_{j=1}^{4} v^j - 2v^i. \tag{4.2.8}$$

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,

$$v^j = \frac{1}{2}\sum_{j=1}^{4} v_i - \frac{1}{2} v_j. \tag{4.2.9}$$

From Eq. (4.2.8) we also derive an expression for $|v|^2$ in term of coordinates:

$$|v|^2 = \left(\sum_j v^j\right)^2 - 2\sum_j (v^j)^2 = \frac{1}{4}\left(\sum_i v_i\right)^2 - \frac{1}{2}\sum_i v_i^2. \tag{4.2.10}$$

It follows that for every light vector v, we have

$$\left(\sum_i v_i\right)^2 - 2\sum_i v_i^2 = 0. \tag{4.2.11}$$

Put, in particular, $v = (1, 0, 0, -1)$. Then $v_i = (v, p_i) = t_i + z_i = c_i$, and Eq. (4.2.11) gives exactly the statement of Descartes's theorem.

In fact, the same approach allows us to prove more.

Theorem 4.2 (Generalized Descartes's theorem). *The matrices M_i satisfy the relation*

$$\left(\sum_i M_i\right)^2 - 2\sum_i M_i^2 = -8 \cdot \mathbf{1}. \tag{4.2.12}$$

Proof. Introduce an inner product on the space of 2×2 Hermitian matrices that corresponds to the quadratic form $Q(M) = \det M$. The explicit formula is

$$(M_1, M_2) = \frac{\det(M_1 + M_2) - \det M_1 - \det M_2}{2}. \tag{4.2.13}$$

In particular, we have $(M, 1) = \frac{1}{2}\operatorname{tr} M$.

Recall also the Cayley identity, which for 2×2 matrices has the form

$$M^2 = M \cdot \operatorname{tr} M - \det M \cdot 1. \tag{4.2.14}$$

Now let M_1, M_2, M_3, M_4 be four Hermitian matrices, corresponding to four mutually tangent disks and normalized by the condition $\det M_i = -1$. Then Eq. (4.2.14) takes the form

$$M_i^2 = M_i \cdot \operatorname{tr} M_i + 1. \tag{4.2.15}$$

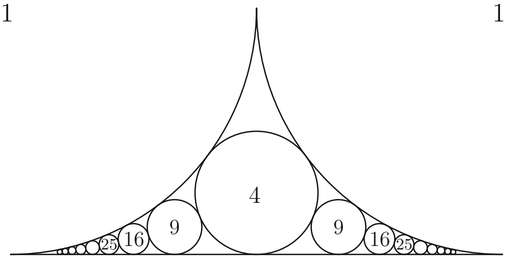


Fig. 4.10 Quadratic sequences of curvatures

Introduce the notation

$$\Sigma_1 := \sum_{i=1}^{i=4} M_i, \quad \Sigma_2 := \sum_{i=1}^{i=4} M_i^2.$$

We have seen above that in this case, $(M_i, M_j) = 1 - 2\delta_{ij}$. In particular, this implies that $(\Sigma_1, M_i) = 2$ and $(\Sigma_1, \Sigma_1) = 8$. Further, taking the inner product of both sides of Eq. (4.2.15) with M_j and summing over i, we obtain

$$(\Sigma_2, M_j) = \operatorname{tr} \Sigma_1. \tag{4.2.16}$$

On the other hand, we have $\Sigma_1^2 = \Sigma_1 \cdot \operatorname{tr} \Sigma_1 - 8 \cdot 1$. Taking the inner product with M_j, we get

$$(\Sigma_1^2, M_j) = 2\operatorname{tr} \Sigma_1 - 4\operatorname{tr} M_j. \tag{4.2.17}$$

Subtracting Eq. (4.2.16) from Eq. (4.2.17) twice, we finally obtain

$$(\Sigma_1^2 - 2\Sigma_2, M_j) = -8(1, M_j), \quad \text{or} \quad (\Sigma_1^2 - 2\Sigma_2 + 8 \cdot 1, M_j) = 0.$$

Since the M_i form a basis in the space of Hermitian matrices, we get the desired relation (4.2.12). □

The relation (4.2.12) can be considered the matrix form of Descartes's theorem. It gives us information not only about radii of tangent disks but also about their configuration.

We mention the following corollary, which is useful in computations.

Theorem 4.3 (Quadratic series of curvatures). *Let D_+ and D_- be two tangent disks and let M_+, M_- be the corresponding matrices. Suppos, the sequence of disks $\{D_k\}$, $k \in \mathbb{Z}$, has the following property: every 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:

$$M_k = A \cdot k^2 + B \cdot k + C, \quad \textit{where} \quad A = M_+ + M_-, \ B = \frac{M_1 - M_{-1}}{2}, \ C = M_0. \tag{4.2.18}$$

An illustration of this theorem can be seen in Figs. 4.10 and 6.4.

Chapter 5
Definition of the Apollonian Gasket

5.1 Basic Facts

Consider all possible finite or countable configurations of disks on S^2 such that no two disks have a common interior point. We call such a configuration a *tiling* of S^2 by disks. A naive approach to Apollonian gaskets is based on the belief that starting with some "natural" initial configuration of disks and inserting the maximal possible number of disks in all available gaps, we arrive eventually at the same picture (up to conformal mapping). In this form, the statement is wrong and becomes true only if we restrict the initial configurations rather severely.

Here we give a rigorous definition of the Apollonian gasket and prove its uniqueness, assuming that initially we have four mutually tangent disks $D^1,\ D^2,\ D^3,\ D^4$ on S^2.

The further construction of the gasket A goes as follows. At the first step, we delete the interiors of disks D^i, $1 \leq i \leq 4$, from S^2. Then four closed curvilinear triangles remain. Call them *triangles of level 1* and denote them by T_i, $i = 1,\ 2,\ 3,\ 4$, so that T_i has no common point with D^i. The union of these triangles we denote by A' and consider it the first approximation to A.

Next, we inscribe in each triangle T_i a largest possible disk, denote it by D_i, and call it a *disk of level 1*. Then we delete the interiors of disks of level 1 from A' and obtain as remainder the union of 12 closed triangles. The part of the remainder that belongs to T_i consists of three triangles. Each of them is bounded by D_i and two of the initial disks, say D^j and D^k. Note that the indices $i,\ j,\ k$ are distinct, and let m be the fourth index. Then we denote the corresponding triangle by T_{im}. In total, the remainder consists of 12 triangles T_{im}, $i \neq m$. We denote it by A'' and consider it a second approximation to A.

At the third step, we inscribe a disk D_{im} of maximal possible size in each triangle T_{im} and delete the interior of the disk. The remaining set consists of 36 triangles, which we number $T_{im_1m_2}$, $i \neq m_1$, $m_1 \neq m_2$. The union $\bigcup_{im_1m_2} T_{im_1m_2}$ is the third approximation A''', and so on.

A.A. Kirillov, *A Tale of Two Fractals*, DOI 10.1007/978-0-8176-8382-5_5,
© Springer Science+Business Media New York 2013

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

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

Let us discuss different forms of the Apollonian gasket. At first glance, the pictures in question look different for different choices of the initial four disks. Nevertheless, all these pictures are in a sense equivalent.

To understand this, consider the group G of conformal mappings of $\overline{\mathbb{C}}$ given by the formula (F.18).

Exercise 5.1. Show that any two quadruples of mutually tangent circles can be transformed one into another by a conformal mapping.

Hint. Show that a triple of mutually tangent circles is uniquely defined by the triple of tangent points and apply Proposition F.1. Then show that there are exactly two quadruples that contain the given triple and that these two quadruples can be transformed into each other by a conformal mapping.

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

Theorem 5.1. *An Apollonian gasket $\mathcal{A}$ is determined by any triple of mutually tangent disks in it. (In other words, if two Apollonian gaskets have a common triple of mutually tangent disks, then they coincide).*

The statement looks rather evident, and I encourage readers to endeavor to find their own proof. The proof given below is rather long and is based on the special numeration of all disks in a given gasket.

The numeration in question is suggested by the construction of a gasket. Namely, call the initial four disks D^1, D^2, D^3, D^4 *disks of level* 0. If we delete from S^2 the union of their interiors, the remaining set is a union of four closed curvilinear triangles. Above, we called them triangles of level 1 and denoted them by T_i, $i = 1, 2, 3, 4$, so that T_i has no common point with D^i. Next, we inscribed in each of these triangles a maximal possible disk, called it a disk of level 1, and denoted it by D_i.

After deleting from T_i the interior of D_i, it becomes a union of three triangles. We call them *triangles of the level 2* and denote them by T_{ij}, $j \neq i$. In each of them we inscribe a maximal possible disk denoted by D_{ij}, call it a *disk of level 2*, and continue this procedure.

At the nth step, we consider a triangle $T_{i_1 i_2 \dots i_{n-1}}$, inscribe a maximal possible disk $D_{i_1 i_2 \dots i_{n-1}}$, and delete its interior. The remaining set is a union of three *triangles of level n*, which we label with $T_{i_1 i_2 \dots i_{n-1} i_n}$, $i_n \neq i_{n-1}$. The maximal disk in a triangle of level n is called a *disk of level n* and has the same label as the ambient triangle.

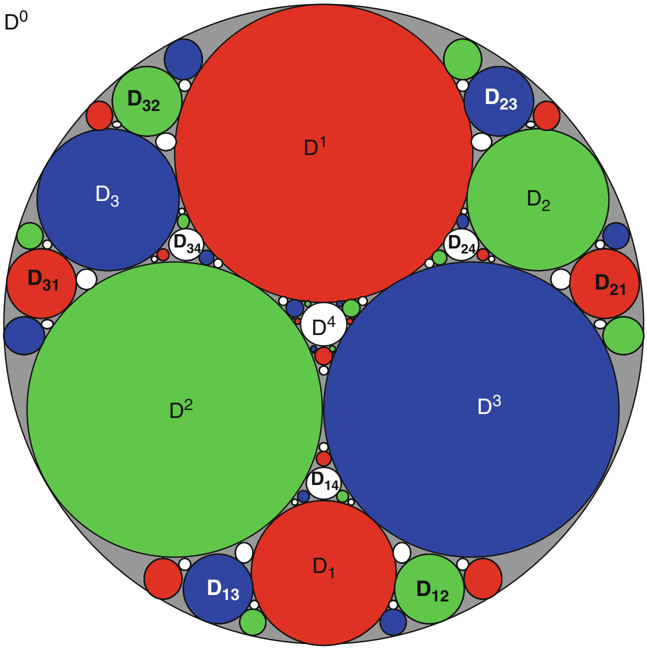


Fig. 5.1 Numeration of disks in the triangular gasket

We observe that two different disks of the same level $n \geq 1$ are never tangent to each other.

Thus, we have labeled all triangles (or disks) of level $n \geq 1$ by sequences of the form $i_1 i_2 \ldots i_n$, where i_k take values $1, 2, 3, 4$ and $i_k \neq i_{k+1}$ (see Fig. 5.1). Besides these, there are only four initial disks, labeled by D^i, $1 \leq i \leq 4$.

Lemma 5.1. *Let D, D', D'', D''' be four mutually tangent disks on S^2. Then if three of them belong to some gasket $\mathcal{A}$, then so does the fourth.*

Proof. Assume that D, D', D'' belong to $\mathcal{A}$ and have levels m, m', m'', respectively. We can assume that $m \leq m' \leq m''$. Since the disks of the same level $m \geq 1$ cannot be tangent, we have to consider the following four cases:

$$(1)\ 0 < m < m' < m''; \qquad (2)\ 0 = m < m' < m'';$$
$$(3)\ 0 = m = m' < m''; \qquad (4)\ 0 = m = m' = m''.$$

In the first case, we can suppose that $D = D_{i_1 i_2 \ldots i_m}$, $D' = D_{j_1 j_2 \ldots j_{m'}}$, and $D'' = D_{k_1 k_2 \ldots k_{m''}}$. By construction, D'' is a disk inscribed in a triangle $T_{k_1 k_2 \ldots k_{m''}}$ that is bounded by arcs of three disks. One of them is $D_{k_1 k_2 \ldots k_{m''-1}}$ of level $m'' - 1$, and the other two disks have levels, say p and q, such that $p \leq q < m'' - 1$. (Equality is possible only if $p = 0$.)

From the construction of $\mathcal{A}$, it is also clear that all disks tangent to D'' except the three mentioned above have level $> m''$. But we know that D and D' are tangent to D''. It follows that $m = p$, $m' = q$, and our three disks are exactly $D,\ D'$, and $D_{k_1k_2\ldots k_{m''-1}}$. Therefore, the disk $D_{k_1k_2\ldots k_{m''-1}}$ is tangent to $D,\ D',\ D''$. Another disk that is also tangent to $D,\ D',\ D''$ is $D_{k_1k_2\ldots k_{m''}k_{m''-1}}$ of level $m'' + 1$. We see that both disks tangent to $D,\ D',\ D''$ belong to $\mathcal{A}$. □

In the other cases, the proof is completely analogous, but simpler. For example, the disk tangent to $D^1\ D^2$, and D_{34} must be either D_3 or D_{343}. Hence it belongs to $\mathcal{A}$. □

Proof of the theorem. Let two gaskets $\mathcal{A}$ and $\tilde{\mathcal{A}}$ have a common triple of mutually tangent disks $D,\ D',\ D''$. Assume that these disks have level $l \le m \le n$ in $\mathcal{A}$. We want to show that $\mathcal{A} \subset \tilde{\mathcal{A}}$ using induction on n.

For $n = 0$, our three disks are just the initial disks, say D^1, D^2, and D^3, for $\mathcal{A}$. According to Lemma 5.1, the disks D^4 and D_4 belong to $\tilde{\mathcal{A}}$, because so do $D^1,\ D^2,\ D^3$.

Again use induction and suppose that we already know that all disks of level $\le n - 1$ in $\mathcal{A}$ belong also to $\tilde{\mathcal{A}}$. Then every disk of level n, being tangent to three disks of level $\le n - 1$, also belongs to $\tilde{\mathcal{A}}$.

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

Let $D,\ D',\ D''$ be common disks of levels $k \le l < n$ respectively. From the proof of Lemma 5.1, we know that among the disks tangent to $D,\ D',\ D''$, there is one that 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 $\mathcal{A} \subset \tilde{\mathcal{A}}$. But in the initial data $\mathcal{A}$ and $\tilde{\mathcal{A}}$ play symmetric roles. Therefore, $\tilde{\mathcal{A}} \subset \mathcal{A}$ and $\tilde{\mathcal{A}} = \mathcal{A}$. □

Lemma 5.2. *The triangle $T_{i_1i_2\ldots i_n}$ is contained in $T_{j_1j_2\ldots j_m}$ iff $m \le n$ and $i_k = j_k$ for $1 \le k \le m$.*

Proof. Note that triangles of the same level cannot have more than three common points. So our first triangle is contained in only one triangle of level m. But it is contained in $T_{i_1i_2\ldots i_m}$ and in $T_{j_1j_2\ldots j_m}$. So we come to the statement of the lemma. □

5.2 Examples of Gaskets

There are three maximally symmetric choices for an initial triple of mutually tangent circles. The corresponding Apollonian gaskets are shown in Figs. 5.2–5.4. We call them the *band gasket*, the *rectangular gasket*, and the *triangular gasket*.

All three gaskets are stereographic projections of a maximally symmetric gasket on S^2 generated by four mutually tangent disks of the same size. See Fig. 5.5.

There are some other interesting realizations of Apollonian gaskets, of which we want to mention two. Their study uses some facts about the Fibonacci numbers.

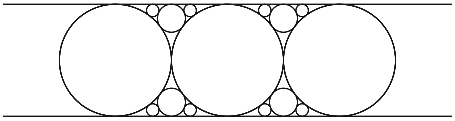


Fig. 5.2 Band gasket

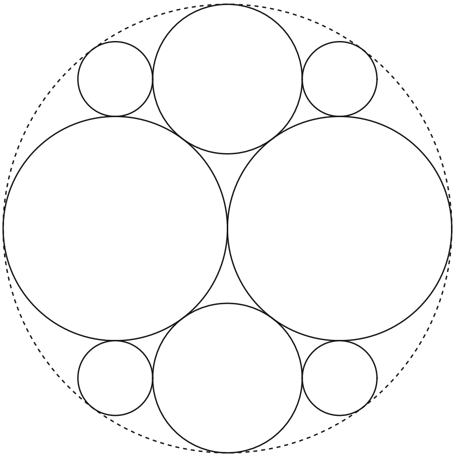


Fig. 5.3 Rectangular gasket

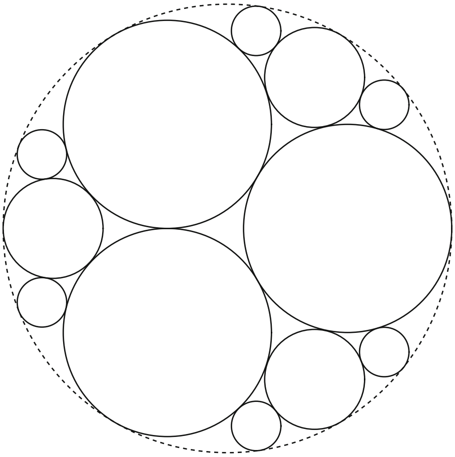


Fig. 5.4 Triangular gasket

Info G. The Fibonacci Numbers

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

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

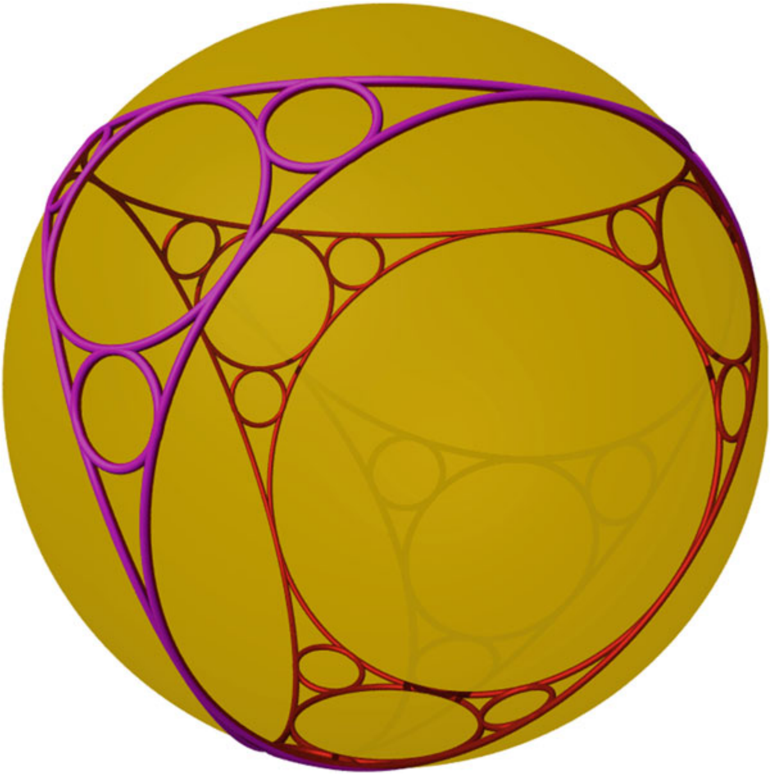


Fig. 5.5 Spherical gasket

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

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 were given the name *Fibonacci numbers*. We briefly describe the main facts related to this and similar sequences.

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

The dimension of this space is 2, because every 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) coordinates in V. So the series of Fibonacci numbers is a vector in V with coordinates $(0, 1)$. Another well-known sequence, called the **Lucas** numbers, has coordinates $(2, 1)$.

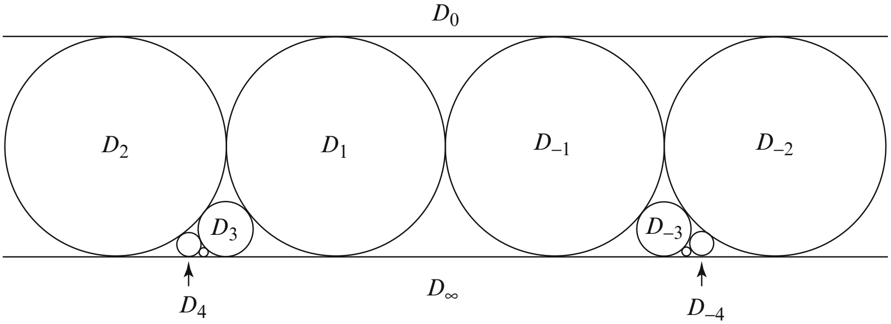


Fig. 5.6 The image of $\mathcal{A}_1$ in the band gasket

Let T denote the transformation sending the sequence $\{v_n\}$ to the sequence $\{v_{n+1}\}$ (which also satisfies the same recurrence 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 these numbers has a special name, the *golden ratio*, because the rectangle with sides in proportion $\phi : 1$ is considered the most pleasant to the human eye.

For 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, every element of V is a linear combination of these eigenvectors.

In particular, the nth 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,

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

Conversely,

$$\phi^n = \frac{\Phi_{n+1} + \Phi_{n-1} + \Phi_n\sqrt{5}}{2}; \quad c^n = \frac{\Phi_{2n+1} + \Phi_{2n-1} + \Phi_{2n}\sqrt{5}}{2}. \tag{G.3}$$

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

It follows that

$$\Phi_n \approx \frac{\phi^n}{\sqrt{5}} \qquad \text{and} \qquad \lim_{n\to\infty} \frac{\Phi_{n+1}}{\Phi_n} = \phi. \tag{G.4}$$

The Lucas number are given by a simpler 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

◇

5.3 Examples of Unbounded Apollonian Tilings

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

$$(x + 1 + x^{-1})^2 = 2(x^2 + 1 + x^{-2}), \quad \text{or} \quad x^2 - 2(x + x^{-1}) + x^{-2} = 1. \tag{5.3.1}$$

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 a real value of x. We have $x = \frac{3+\sqrt{5}}{2} = \frac{2}{3-\sqrt{5}}$, which is the number c introduced in Info G.

The gasket $\mathcal{A}_1$ generated by q_1 has the following property. If we dilate it by the factor c, it goes to its mirror reflection in a vertical line. And if we dilate by c^2, it goes to one of its horizontal translations. Choosing an appropriate position of $\mathcal{A}_1$, we can arrange the mirror in question to be the imaginary axis and the translation to be the identity; see Fig. 5.8. This means that $\mathcal{A}_1$ is invariant under the transformation $w \mapsto -c\bar{w}$. Indeed, $\mathcal{A}_1$ and $-c \cdot \overline{\mathcal{A}_1}$ have a common triple of disks.

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

$$\left| c^k w + (-1)^k \tfrac{2}{\sqrt{5}} + i \right| \leq 1, \tag{5.3.2}$$

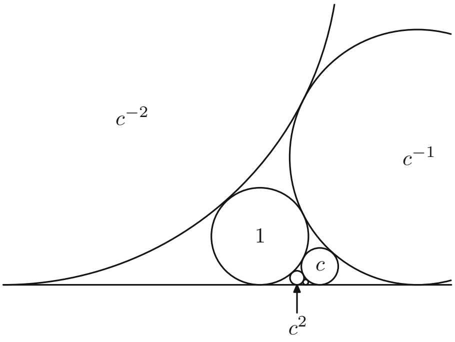


Fig. 5.7 The gasket $\mathcal{A}_1$

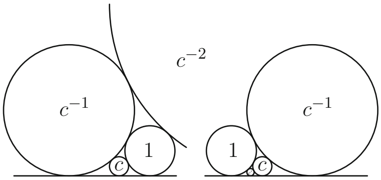


Fig. 5.8 Gaskets $\mathcal{A}_1$ and $c \cdot \mathcal{A}_1$

and the corresponding normalized Hermitian matrices (see Sect. 4.3) are

$$\begin{aligned} M_k &= \begin{pmatrix} \frac{4}{5}c^{-k} & (-1)^k\frac{2}{\sqrt{5}} - i \\ (-1)^k\frac{2}{\sqrt{5}} + i & c^k \end{pmatrix} \\ &= \begin{pmatrix} (-\phi)^{-k} & 0 \\ 0 & \phi^k \end{pmatrix} \cdot \begin{pmatrix} \frac{4}{5} & \frac{2}{\sqrt{5}} - i(-1)^k \\ \frac{2}{\sqrt{5}} + i(-1)^k & 1 \end{pmatrix} \cdot \begin{pmatrix} (-\phi)^k & 0 \\ 0 & \phi^k \end{pmatrix}, \end{aligned} \tag{5.3.3}$$

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

Each of the relations (5.3.2) and (5.3.3) implies that the dilation $w \to -c \cdot \bar{w}$ sends the disk D_n to D_{n-1} and hence preserves the gasket $\mathcal{A}_1$ (Fig. 5.8).

Exercise 5.2. Find a matrix $g \in \mathrm{SL}(2, \mathbb{C})$ that transforms the gasket $\mathcal{A}_1$ into the band gasket.

Hint. Find the transformation g that preserves the real line and sends the disk D_0 to a parallel line. Show how the images will be situated (see Fig. 5.6).

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 Descartes's equation is

$$(1 + \rho + \rho^2 + \rho^3)^2 = 2(1 + \rho^2 + \rho^4 + \rho^6). \tag{5.3.4}$$

Simplify this equation by writing 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 gives a real value for ρ, and we get

$$\rho = \phi+\sqrt{\phi} = \theta^2+\theta \approx 2.890054\ldots; \quad \rho^{-1} = \phi-\sqrt{\phi} = \theta^2-\theta \approx 0.346014\ldots. \tag{5.3.5}$$

The corresponding disks D_k form a spiral that converges to certain point a as $k \to -\infty$. If we take for a 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

$$M_k = \begin{pmatrix} a\rho^k & b\mathrm{e}^{2ik\alpha} \\ \bar{b}\mathrm{e}^{-2ik\alpha} & c\rho^{-k} \end{pmatrix}, \quad ac - |b|^2 = -1. \tag{5.3.6}$$

The condition that the disks 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 it leads to the equation

$$\frac{|b|^2}{ac} = \frac{\rho^m + \rho^{-m} + 2}{\mathrm{e}^{im\alpha} + \mathrm{e}^{-im\alpha} + 2}.$$

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

$$\frac{|b|}{\sqrt{ac}} = \frac{\cosh s}{|\cos\alpha|} = \frac{\cosh 2s}{|\cos 2\alpha|} = \frac{\cosh 3s}{|\cos 3\alpha|}. \tag{5.3.7}$$

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 Eq. (5.3.7), 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 $\cos 2\alpha \leq 0$.

Using the relation $\cosh 2s = 1 - \cos 2\alpha$, we get, comparing the second and third terms,

$$\cosh s = \pm\frac{\cos\alpha \cdot (1 - \cos 2\alpha)}{\cos 2\alpha}.$$

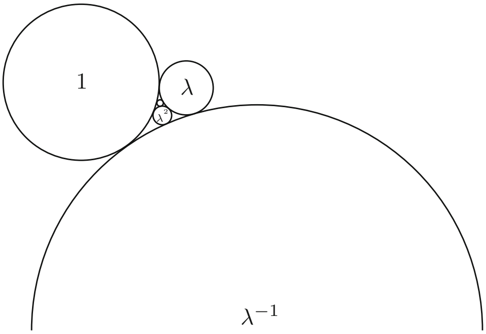
Now the relation $2\cosh^2 s = \cosh 2s + 1$ gives us the equation

$$2\left(\frac{\cos\alpha \cdot (1 - \cos 2\alpha)}{\cos 2\alpha}\right)^2 = 2 - \cos 2\alpha.$$

Denote $\cos 2\alpha$ by x and write the equation in the algebraic form

$$\frac{(x+1)(1-x)^2}{x^2} = 2 - x, \quad \text{or} \quad (x+1)(1-x)^2 = 2x^2 - x^3,$$
$$\text{or} \quad 2x^3 - 3x^2 - x + 1 = 0.$$

Fig. 5.9 The gasket $\mathcal{A}_2$



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 $\frac{|b|}{\sqrt{ac}} = \phi^2$, and therefore

$$|b|^2 = \frac{\phi^2}{\sqrt{5}}, \qquad ac = \frac{\phi^{-2}}{\sqrt{5}}. \tag{5.3.8}$$

It follows that we know the matrices M_k up to complex conjugation and conjugation by a diagonal matrix. Geometrically, this means that we know the gasket $\mathcal{A}_2$ up to rotation, dilation, and reflection in a straight line. In particular, we can put

$$M_k = \frac{1}{\sqrt[4]{5}} \begin{pmatrix} \phi^{-1} \cdot \rho^k & \phi \cdot e^{2ik\alpha} \\ \phi \cdot e^{2ik\alpha} & \phi^{-1} \cdot \rho^{-k} \end{pmatrix}, \tag{5.3.9}$$

so that

$$D_0 = \left\{ w \,\middle|\, \left| w + 1 + \frac{1}{\sqrt{5}} \right| \leq \sqrt{\frac{1 + 2\sqrt{5}}{5}} \right\}. \tag{5.3.10}$$

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,

$$\lambda = \rho e^{2i\alpha} = -(1 + \theta^{-1})(1 \mp i\theta). \tag{5.3.11}$$

The corresponding picture is shown as Fig. 5.9.

5.4 Integral Solutions to Descartes's Equation

Here we consider the arithmetic properties of the set of solutions to Descartes's equation (4.1.3). Make the following change of variables:

$$\begin{aligned} t &= \frac{c_0 + c_1 + c_2 + c_3}{2}, \quad x = \frac{c_0 + c_1 - c_2 - c_3}{2}, \\ y &= \frac{c_0 - c_1 + c_2 - c_3}{2}, \quad z = \frac{c_0 - c_1 - c_2 + c_3}{2}. \end{aligned} \tag{5.4.1}$$

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 Eq. (4.1.3) becomes

$$t^2 - x^2 - y^2 - z^2 = 0. \tag{5.4.2}$$

In other words, the solutions to Eq. (4.1.3) correspond to light vectors in Minkowski space.

Lemma 5.3. *The integral solutions to Eq. (4.1.3) correspond to integral light vectors in $\mathbb{R}^{1,3}$ (i.e., light vectors with integral coordinates).*

Proof. From Eq. (4.1.3), it is clear that the sums $c_0 \pm c_1 \pm c_2 \pm c_3$ are always even. So every integral solution to Eq. (4.1.3) corresponds to a light vector p with integral coordinates. Conversely, from Eq. (5.4.2), it follows that the sum $t \pm x \pm y \pm z$ is always even. Therefore, from the equations

$$\begin{aligned} c_0 &= \frac{t + x + y + z}{2}, \quad c_1 = \frac{t + x - y - z}{2}, \\ c_2 &= \frac{t - x + y - z}{2}, \quad c_3 = \frac{t - x - y + z}{2}, \end{aligned}$$

we deduce that every integral light vector corresponds to an integral solution to Eq. (4.1.3). □

Thus, we arrive at the following problem.

Problem 5.1. Describe the set of integral points on the light cone in $\mathbb{R}^{1,3}$.

The solution to the analogous problem for rational points is well known. To every 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 the point $\frac{x+iy}{t-z} \in P^1(\mathbb{Q}[i])$.

Conversely, every $(r + is) \in P^1(\mathbb{Q}[i])$ 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 every integral vector on the light cone in $\mathbb{R}^{1,3}$ is proportional (but not necessarily equal) to the vector

$$t = k^2 + m^2 + n^2, \quad x = 2kn, \quad y = 2mn, \quad z = k^2 + m^2 - n^2, \qquad (5.4.3)$$

with integer k, m, n.

Note that for every 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, namely vectors whose greatest common divisor of their coordinates is equal to 1.

Lemma 5.4. *Every 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 every 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 5.2. Find a convenient parameterization of all primitive integral light vectors.

For instance, assume that t, z are odd and x, y are even. Is it true that Eq. (5.4.3) holds for some relatively prime k, m, n?

Now consider the subgroup Γ of the Lorentz group G that preserves the set of integral light vectors.

Exercise 5.3. Show that Γ coincides with the group $\mathrm{SO}_+(1,3;\mathbb{Z})$ of matrices with integral entries in $\mathrm{SO}_+(1,3;\mathbb{R})$.

Hint. Let $g \in \Gamma$. Show that the sum and difference of any two columns of g is an integer vector and that the same property holds for row vectors. Check that the coordinates of an integer light vector cannot be all odd.

The group Γ acts on the set of all integral light vectors and preserves the subset P of primitive vectors.

Exercise 5.4. (a) Find the index of $\mathrm{PSL}(2, \mathbb{Z}[i])$ in $\mathrm{PGL}(2, \mathbb{Z}[i])$.
(b)* What are the images of these subgroups in $\mathrm{O}_+(1, 3; \mathbb{R})$?

Exercise 5.5. Show that the homomorphism $\pi\colon \mathrm{PGL}(2, \mathbb{C}) \to \mathrm{SO}_+(1, 3; \mathbb{R})$ can be extended to a homomorphism $\overline{\pi}\colon \overline{G} \to \mathrm{O}_+(1, 3; \mathbb{R})$.

Hint. Show that one can take the diagonal matrix $\mathrm{diag}(1, 1, -1, 1)$ as the image under $\overline{\pi}$ of the element $s \in G$ acting as complex conjugation.

Problem 5.3. Describe the Γ-orbits in P.

Info H. Structure of Groups Freely Generated by Reflections

The theory of groups generated by reflections is a large and very interesting domain in modern mathematics. We consider here only some facts that we need in relation to the Apollonian gaskets.

First, we describe the structure of the *free* group F_n on n generators $x_1, x_2, \dots, x_n$. This group may be characterized by the following universal property.

> *For every group G with n generators $y_1, y_2, \dots, y_n$, 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 a group exists and is unique up to isomorphism. Indeed, if there are two such groups, F_n with generators $x_1, x_2, \dots, x_n$ and F'_n with generators $x'_1, x'_2, \dots, x'_n$, 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 the identity. The same is true for the composition $\alpha \circ \alpha'$. Hence F_n and F'_n are isomorphic.

Now we prove the existence. For this, we consider the collection W_n of all words in the alphabet $x_1, x_1^{-1}, \dots, x_n, x_n^{-1}$ satisfying the following condition:

$(*)$ *The letters x_i and x_i^{-1} cannot be neighbors.*

We denote the length of a word w by $l(w)$. Let $W_n^{(k)}$ be the set of all words of 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 H.1. Show that $\#(W_n^{(k)}) = 2n(2n-1)^{k-1}$ for $k \ge 1$.

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 the 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 $l(w_1) = k \ge 1$. Let the last letter of w_1 be $x_i^{\varepsilon_1}$, $1 \le i \le n$, $\varepsilon_1 = \pm 1$, and let the first letter of w_2 be $x_j^{\varepsilon_2}$, $1 \le j \le n$, $\varepsilon_2 = \pm 1$.

If $i \ne j$ or $i = j, \varepsilon_1 + \varepsilon_2 \ne 0$, we define the product $w_1 w_2$ simply as the juxtaposition (concatenation) of w_1 and w_2. This new word has length $l(w_1) + l(w_2)$ and satisfies condition $(*)$.

If $i = j$ and $\varepsilon_1 + \varepsilon_2 = 0$, 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 we always have $l(w_1 w_2) \le l(w_1) + l(w_2)$ 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) and that it admits a unit (empty word) and an inverse element (the same word written back to front with opposite exponents). As is traditional, this checking is left to the reader.

Let us check that the group W_n has the universal property. Indeed, if G is any group generated by $y_1, y_2, \dots, y_n$, there is a unique homomorphism $\alpha\colon W_n \to G$ such that $\alpha(\{x_i\}) = y_i$. (Here $\{x_i\}$ denotes a one-letter word.) Namely, for a word $w = x_{i_1}^{\varepsilon_1} x_{i_2}^{\varepsilon_2} \dots x_{i_k}^{\varepsilon_k}$, we must put $\alpha(w) = y_{i_1}^{\varepsilon_1} \cdot y_{i_2}^{\varepsilon_2} \cdot \dots \cdot y_{i_k}^{\varepsilon_k}$, 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 have established the existence of a free group F_n and at the same time proved the following proposition.

Proposition H.1. *Every element of F_n can be uniquely written in the form*

$$g = x_{i_1}^{\varepsilon_1} x_{i_2}^{\varepsilon_2} \dots x_{i_k}^{\varepsilon_k}, \tag{H.1}$$

where the condition $(*)$ is satisfied.

We need also another family of groups Γ_n, $n \geq 1$, which are freely generated by n involutions $s_1, \dots, s_n$. By definition, the group Γ_n possesses another universal property.

For every group G generated by n involutions $t_1, \dots, t_n$, there exists a unique homomorphism α of Γ_n onto G such that $\alpha(s_i) = t_i$, $1 \leq i \leq n$.

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*

$$g = s_{i_1} s_{i_2} \dots s_{i_k},\ k \geq 0, \quad \textit{where } i_a \neq i_{a+1} \quad \textit{for} \quad 1 \leq a \leq k-1. \tag{H.2}$$

Exercise H.2. (a) Show that in this case,

$$\#(W_n^{(k)}) = \begin{cases} 1 & \text{for } k = 0, \\ n(n-1)^{k-1} & \text{for } k \geq 1. \end{cases}$$

(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 that contains $s_1^2, \dots, s_n^2$.

Theorem H.1. *Every nontrivial (i.e., different from e) involution in Γ_n is conjugate to exactly one of the generators $s_1, \dots, s_n$.*

Proof. Let $g \in \Gamma_n$ be an involution. According to Proposition H.2, 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 $g^{-1} = g$, whence $s_{i_{n-k}} = s_{i_{k+1}}$ for $k = 0, 1, \dots, n-1$.

For $n = 2k$ even, it follows that $k = 0$ and g is the empty word.

For $n = 2k - 1$ odd, we have $g = ws_{i_k}w^{-1}$, where $w = s_{i_1} \dots s_{i_{k-1}}$. Hence, g is conjugate to s_{i_k}.

Finally, we show that s_i is not conjugate to s_j for $i \neq j$. Assume the contrary. Then there is a word w such that $ws_i = s_j w$. Let w_0 be a shortest such word. 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. For example, for $n = 1$, the group Γ_1 is simply the 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 it to the 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., the Poincaré model of L as the upper half-plane $y > 0$ (see **Info H** below). The three generators of Γ_3 are reflections in three mutually 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 the two others. These mirrors bound a triangle T of finite area with three infinite vertices. For every 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 distinct, 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 Sects. 7.1 and 7.2.

Chapter 6
Arithmetic Properties of Apollonian Gaskets

Here we study some arithmetic questions arising when we consider the curvatures of disks which constitute an Apollonian gasket.

6.1 The Structure of $\overline{\mathbb{Q}}$

Here we want 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 parameterize $\overline{\mathbb{Q}}$. Every number $r \in \overline{\mathbb{Q}}$ can be written in the form $\frac{p}{q}$, where $p, q \in \mathbb{Z}$. But the map $\alpha : \mathbb{Z} \times \mathbb{Z} \longrightarrow \overline{\mathbb{Q}}, \alpha(p, q) = \frac{p}{q}$, is surjective but by no means injective.

We can impose the condition $\gcd(p, q) = 1$, that is, that p and q be relatively prime, or in other words, that the fraction $\frac{p}{q}$ be 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 $\{(p, q), (-p, -q)\}$. However, 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 6.1. For the analytically minded reader, we can say that the construction 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 each $w \in \mathbb{C}^\times$, but this double-valued function does not admit an analytic (or even continuous) single-valued branch.

♡

Remark 6.2. A remarkable way to label all positive rational numbers was discovered recently by Neil Calkin and Herbert Wilf ("Recounting the Rationals," *American Mathematical Monthly* 107 (2000), pp. 360–363). Let $\mathbf{b}(n)$ be the number of partitions of an integer $n \geq 0$ into powers of 2, no power of 2 being used more

A.A. Kirillov, *A Tale of Two Fractals*, DOI 10.1007/978-0-8176-8382-5_6,

than twice. Than the ratio $r_n = \frac{\mathbf{b}(n)}{\mathbf{b}(n+1)}$ takes every positive rational value exactly once! The initial piece of this numeration is given in the following table:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$\mathbf{b}(n)$	1	1	2	1	3	2	3	1	4	3	5	2	5	3	4	1	5	4
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	$\frac{1}{5}$	$\frac{5}{4}$	$\frac{4}{7}$

n	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	$\frac{7}{3}$	$\frac{3}{8}$	$\frac{8}{5}$	$\frac{5}{7}$	$\frac{7}{2}$	$\frac{2}{7}$	$\frac{7}{5}$	$\frac{5}{8}$	$\frac{8}{3}$	$\frac{3}{7}$	$\frac{7}{4}$	$\frac{4}{5}$	5	$\frac{1}{6}$	$\frac{6}{5}$	$\frac{5}{9}$	

It is of interest to compare this numeration with the one given 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:

$$(a)\ \ |p_1q_2 - p_2q_1| = 1, \qquad (b)\ \ |r_1 - r_2| = \frac{1}{|q_1q_2|}. \tag{6.1.1}$$

It is worth mentioning that the friendship relation **is not** an equivalence relation:[1] every integer k is friendly to ∞, but only neighboring integers are friendly to each other.

Note that the group PGL(2, $\mathbb{Z}$) acts on $\overline{\mathbb{Q}}$ by fractional linear transformations and that this action preserves the friendship relation. We can often use this fact in our study.

Lemma 6.1. *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 nontrivial 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, $\mathbb{Z}$)

[1] Just as in the real life.

that 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 $\mathrm{SL}(2, \mathbb{Z})$. Then we have $\gamma(0) = \frac{b}{d}$, $\gamma(\infty) = \frac{a}{c}$.

The conditions $\gamma(\infty) = r_1$, $\gamma(0) = r_2$ imply $(a, c) = k_1 \cdot (p_1, q_1)$, $(b, d) = k_2 \cdot (p_2, q_2)$. 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 $\mathrm{PSL}(2, \mathbb{Z})$.

The stabilizer of the pair $(0, \infty)$ in $\mathrm{PGL}(2, \mathbb{Z})$ consists of classes of matrices $\begin{pmatrix} 1 & 0 \\ 0 & \pm 1 \end{pmatrix}$. □

Exercise 6.1. Describe all numbers that are (a) friendly to 0; (b) friendly to ∞; (c) friendly to 1.

We define a **distance** in the set $\overline{\mathbb{Q}}$ in 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 $r' = r_0, r_1, \dots, r_{n-1}, r_n = r''$ such that for all k, the number r_k is friendly to $r_{k\pm 1}$ for $1 \leq k \leq n-1$.

Exercise 6.2. (a) Show that $(\overline{\mathbb{Q}}, d)$ is a discrete metric space on which the group $\mathrm{PGL}(2, \mathbb{Z})$ 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 6.3. 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 6.4. (a) Show that for every $r', r'' \in \overline{\mathbb{Q}}$, the distance $d(r', r'')$ is finite.
(b) Is the metric space $\overline{\mathbb{Q}}$ bounded?

Answer. (a) See Theorem 6.1 below; (c) No.

Rather interesting and nontrivial problems arise when we consider the geometry of balls and spheres in $\overline{\mathbb{Q}}$. As usual, we define a **ball** with 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 6.1. *The ball $B_n(\infty)$ consists of all rational numbers that can be written as a continued fraction of length n, i.e., as*

$$r = k_1 + \cfrac{1}{k_2 + \cfrac{1}{k_3 + \cfrac{1}{\ddots\, k_{n-1} + \cfrac{1}{k_n}}}} \tag{6.1.2}$$

where k_i are arbitrary integers (positive or negative).

proof. First of all, let us show that for every r of the form (6.1.2), the distance $d(\infty, r)$ does not exceed n. We shall do this by induction on n.

For $n = 1$, the result follows from Exercise 6.3. Assume that the theorem is true for all continued fractions of length $\leq n-1$ and consider a continued fraction of length n given by Eq. (6.1.2). Denote by r' the number $\frac{1}{r-k_1}$. It is clear that r' is represented by a continued fraction of length $n-1$, whence $d(\infty, r') \leq n-1$. 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

$$d(\infty, r) = d(\infty, r - k_1) = d(0, r') \leq d(0, \infty) + d(\infty, r') \leq 1 + (n-1) = n.$$

The first sign $\leq$ is just the triangle inequality, and the second follows from Exercise 6.3(a) and from the induction hypothesis. □

The structure of spheres is a more delicate question. The “complexity” of a sphere grows with its radius.

For instance, consider $S_1(\infty) = \mathbb{Z}$. It is a 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 6.3(a).

The sphere $S_2(\infty)$ consists 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, enumerated by the number $m = |k_2| \geq 2$. The stabilizer of the point $k + \frac{1}{m} \in \Omega_m$ is trivial for $m > 2$ and contains one nonunit element $r \mapsto 2k + 1 - r$ for $m = 2$.

Problem 6.1. Describe the orbits of $\mathrm{Aff}(1, \mathbb{Z})$ on the sphere $S_k(\infty)$ for $k > 2$.

6.2 Rational Parameterization of Circles

It is well known that a circle as a real algebraic manifold is rationally equivalent to a real projective line. This means that one can establish a bijection between a circle and a projective line using rational functions with rational coefficients.

For example, the circle $x^2 + y^2 = 1$ can be identified with a projective line with parameter t as follows:

$$x = \frac{t^2 - 1}{t^2 + 1}, \qquad y = \frac{2t}{t^2 + 1}; \qquad t = \frac{y}{1 - x} = \frac{1 + x}{y}. \tag{6.2.1}$$

In particular, as t runs through all rational numbers (including ∞), the corresponding points (x, y) run through all rational points[2] of the circle.

From this fact 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

[2] That is, points with rational coordinates.

one of the numbers x, y is even. Assume that it is y; then there are relatively prime numbers a, b such that

$$x = a^2 - b^2, \quad y = 2ab, \quad \pm z = a^2 + b^2. \tag{6.2.2}$$

Analogously, the projectivization of the future light cone in $\mathbb{R}^{1,3}$ is nothing but the two-dimensional sphere, which is rationally equivalent to a completed two-dimensional plane. Therefore, all future light vectors (t, x, y, z) with integral nonnegative coefficients can be written, up to positive proportionality, in the form

$$t = k^2 + l^2 + m^2, \quad x = 2km, \quad y = 2lm, \quad z = |k^2 + l^2 - m^2|. \tag{6.2.3}$$

I do not know whether any integral solution can be written exactly in the form (6.2.3) for some integers k, l, m with $\gcd(k, l, m) = 1$.[3]

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 6.5. (a) Show that in the case that 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

$$(x_1 - x_2)(x_2 - x_3)(x_3 - x_1) > 0.$$

(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 ∞.

We now introduce a new operation[4] of "inserting" on $\overline{\mathbb{R}}$. It associates to an ordered pair of rational numbers (r_1, r_2) a third number, denoted by $r_1 \downarrow r_2$, such that

$$r_1 \downarrow r_2 := \frac{p_1 + p_2}{q_1 + q_2}, \quad \text{if} \quad r_1 = \frac{p_1}{q_1},\ r_2 = \frac{p_2}{q_2}, \tag{6.2.4}$$

where the signs of p_i and q_i are chosen such that $r_1 \downarrow r_2$ is between r_1 and r_2.

Exercise 6.6. 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.

[3] As one of the reviewers pointed out, the quadruple $(7, 2, 3, 6)$ is a counterexample, since 7 is not a sum of three squares.

[4] I learned from R. Borcherds that this operation is known to mathematicians in England as "English major addition." It is also the subject of one of the standard jokes quoted in Gelfand's seminar.

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 6.7. Show that for friendly numbers r_1, r_2, the number $r_1 \downarrow r_2$ is the unique rational number between r_1 and r_2 (in the sense of the cyclic order described above) that is friendly to both of them.

These considerations lead to the notion of **Farey series**. The standard Farey series F^n of rank n by definition consists of all rational numbers $0 < \frac{p}{q} < 1$ with $1 \leq q \leq n$ 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 that are relatively prime to k. It is given by the formula

$$\varphi(n) = n \cdot \prod_{p|n} \left(1 - p^{-1}\right), \quad \text{where } p \text{ runs through all prime divisors of } n.$$

For example, 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 6.8. Show that neighboring 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 $F^{(n+1)}$, $n \geq 1$, is obtained from $F^{(n)}$ by inserting between every pair of 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 $f_k^{(n)}$, $1 - 2^n \leq k \leq 2^{n+1}$, the kth member of $F^{(n)}$. In particular, for every $n \geq 0$, we have $f_0^{(n)} = 0$, $f_{2^n}^{(n)} = 1$, $f_{2^{n+1}}^{(n)} = \infty$.

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 the numbers $f_k^{(n)}$ is a nontrivial problem. We shall discuss it below.

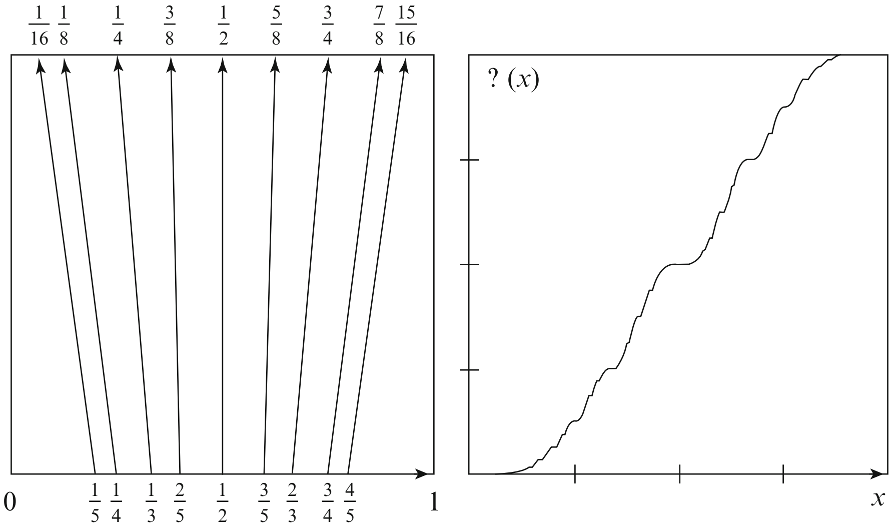


Fig. 6.1 Graph of the function **?**

Exercise 6.9. Show that $f_k^{(n)} = f_{2k}^{(n+1)}$, so that $f_k^{(n)}$ actually depends only on the dyadic number $r = \frac{k}{2^n}$. Sometimes, we shall write f_r instead of $f_k^{(n)}$.

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 at the nth step, the arithmetic progression with $2^n + 1$ terms starting with 0 and ending by 1. The kth member of this progression is $a_k^{(n)} = \frac{k}{2^n}$, or in the same notation as above, $a_r = r$ (Fig. 6.1).

Now we are prepared to define a remarkable function first introduced by Hermann Minkowski. He called it the "question mark function" and denoted it by $\mathbf{?}(x)$; see **Info E** in Part I.

Theorem (Minkowski's theorem). *There exists a unique continuous and strictly increasing function* $\mathbf{?} : [0,1] \to [0,1]$ *such that*

$$\mathbf{?}\,(a \downarrow b) = \frac{\mathbf{?}(a) + \mathbf{?}(b)}{2} \quad \textit{for all friendly rational numbers} \quad a, b \in [0,1]. \tag{6.2.5}$$

Sketch of proof. The formula (6.2.5) and induction over n imply that if the desired function exists, it must have the property $\mathbf{?}\big(f_k^{(n)}\big) = a_k^{(n)}$. It follows that $\mathbf{?}\big(f_r\big) = r$ 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 $\mathbf{?}(f_r) = r$. 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 $[0,\ 1]$. For example, we can put

$$\mathbf{?}(x) = \lim_{n\to\infty} \mathbf{?}(x_n), \tag{6.2.6}$$

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 every 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

$$p\left(\frac{2k+1}{2^{n+1}}\right) = p\left(\frac{k}{2^n}\right) \downarrow p\left(\frac{k+1}{2^n}\right), \tag{6.2.7}$$

which follows immediately from Eq. (6.2.5), and repeating the construction of the modified Farey series.

Theorem 6.2. *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 every n and* $0 \le k \le 2^n$.
3. *For every rational nondyadic 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. *We have the following remarkable formula:*

$$p\left(\underbrace{0.0\ldots00}_{k_1}\underbrace{11\ldots11}_{l_1}\ldots\underbrace{00\ldots00}_{k_n}\underbrace{11\ldots11}_{l_n}\ldots\right) = \cfrac{1}{k_1+\cfrac{1}{l_1+\cfrac{1}{\ddots\, k_n+\cfrac{1}{l_n+\cfrac{1}{\ddots}}}}\ldots} \tag{6.2.8}$$

where on the left-hand side, the binary system is used, while on the right-hand side, we use a continued fraction. The formula (6.2.8) *works also for finite binary fractions.*[5]

Sketch of proof. The relations 1(a)–(c) can be derived from the following useful fact.

[5] Guess about the form of the right-hand side of the formula in this case.

Lemma 6.2. *Let* $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}(2, \mathbb{Z})$. *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.,*

$$(g \cdot r_1) \downarrow (g \cdot r_2) \;=\; g \cdot (r_1 \downarrow r_2). \tag{6.2.9}$$

We leave the proof of this claim to the reader and make only two useful remarks, each of which could serve as the basis for a proof.

1. The transformations in question send friendly pairs to friendly pairs.
2. The group $\mathrm{GL}(2, \mathbb{Z})$ is generated by two 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 relation 1(a). Consider the following diagram:

$$\begin{tikzcd}
[0,\ 1] \arrow[r, "x \mapsto 1-x"] \arrow[d, "p"'] & {[0,\ 1]} \arrow[d, "p"] \\
{[0,\ 1]} \arrow[r, "x \mapsto 1-x"] & {[0,\ 1]}
\end{tikzcd} \tag{6.2.10}$$

Relation 1(a) claims that it is commutative. To check this, choose a point $x \in [0,\ 1]$ that is a dyadic fraction $r = \frac{k}{2^n} = a_r$. Then the vertical arrow sends this number to $p(a_r) = f_r$, and the horizontal arrow sends f_r to f_{1-r} (check this, consulting the table above).

On the other hand, the horizontal arrow sends r to $1 - r = a_{1-r}$, and then the vertical arrow sends a_{1-r} to f_{1-r}. Thus for every number of the form $\frac{k}{2^n}$, relation 1(a) holds. By continuity, it holds everywhere.

Consider relation 1(b). It is equivalent to the commutativity of the diagram

$$\begin{tikzcd}
[0,\ 1] \arrow[r, "x \mapsto x/2"] \arrow[d, "p"'] & {[0,\ \frac{1}{2}]} \arrow[d, "p"] \\
{[0,\ 1]} \arrow[r, "x \mapsto \frac{x}{1+x}"] & {[0,\ \frac{1}{2}]}
\end{tikzcd} \tag{6.2.11}$$

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 $f_{r/2}$. That is, 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 $\mathrm{PGL}(2,\ \mathbb{Z})$, it transforms the part of Farey series between f_0 and f_1 to the par between f_0 and $f_{1/2}$. Then by induction on n, we check that it sends $f_{\frac{2k}{2^n}}$ to $f_{\frac{k}{2^n}}$.

The relation 1(c) can be proved in the same way using the diagram

$$\begin{tikzcd}
{[0,\ 1]} \arrow[r, "x \mapsto \frac{1+x}{2}"] \arrow[d, "p"'] & {[\frac{1}{2},\ 1]} \arrow[d, "p"] \\
{[0,\ 1]} \arrow[r, "x \mapsto \frac{1}{2-x}"] & {[\frac{1}{2},\ 1]}
\end{tikzcd} \tag{6.2.12}$$

The point is that affine transformations respect half-sums, while the transformations from $\mathrm{PGL}(2,\ \mathbb{Z})$ respect the insertion operation.

I recommend that the reader 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 formulas

$$p\left(\frac{1}{x}\right) = \frac{1}{p(x)}; \qquad p(-x) = -p(x). \tag{6.2.13}$$

We shall verify property 2 only at the point $x = 0$. The general case $x = \frac{k}{2^n}$ can be done similarly, or it can be reduced 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 $\frac{1}{2^n} \leq \Delta x \leq \frac{1}{2^{n-1}}$, we have $\frac{1}{n+1} \leq \Delta p \leq \frac{1}{n}$.

Therefore, $\frac{2^{n-1}}{n+1} \leq \frac{\Delta p}{\Delta x} \leq \frac{2^n}{n}$ for $\frac{1}{2^n} \leq \Delta x \leq \frac{1}{2^{n-1}}$ and $p'(0) = +\infty$.

Statement 3 follows from the formula (6.2.8). 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 Eq. (6.2.8) as a definition of the question mark function.

□

Remark 6.3. Let us interpret the function $p := \mathbf{?}^{-1}$ 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[6] 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., the measure of every 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]).

♡

[6] We say that a measure μ on $[0,\ 1]$ is a weak limit of the sequence of measures μ_n if for every continuous function f on $[0,\ 1]$, we have $\lim_{n\to\infty} \int_0^1 f(x)\mathrm{d}\mu_n(x) = \int_0^1 f(x)\mathrm{d}\mu(x)$.

Exercise 6.10. Find the values of $\mathbf{?}(x)$ and $\mathbf{?}'(x)$ 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 nth Fibonacci number, given by the formula

$$\Phi_n = \frac{\phi^n - (-\phi)^{-n}}{\phi + \phi^{-1}},$$

where $\phi = \frac{\sqrt{5}+1}{2} \approx 1.618\ldots$ is the golden ratio.[7]

Answer. $\mathbf{?}\left(\frac{1}{3}\right) = \frac{3-\sqrt{5}}{2}$; $\quad \mathbf{?}'\left(\frac{1}{3}\right) = 0.$

Problem 6.2. Is it true that $\mathbf{?}'(x) = 0$ for all rational numbers except $a_k^{(n)}$?

We can sum up the content of this section as follows: there is a monotone parameterization of all rational numbers in [0, 1] by the simpler set of all binary fractions in the same interval.

If we remove the restriction $r \in [0, 1]$, we get a parameterization of $\overline{\mathbb{Q}}$ by $\overline{\mathbb{Z}[\frac{1}{2}]}$ that preserves the cyclic order on the circle introduced above.

Remark 6.4. There is an interesting geometric interpretation of the Farey series and of the Minkowski question mark function. It was discovered by George de Rham [de Rha59].

Consider the square $[-1, 1] \times [-1, 1] \subset \mathbb{R}^2$. Let us divide every side into three equal parts and join neighboring division points. We get an octagon with equal angles but unequal sides. Repeat this procedure: divide every side of the octagon into three equal parts and join the neighboring division points. The result will be a convex polygon with 16 sides that is contained in the octagon. 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. 6.2). Note the following facts:

(a) The center of each side of every Π_n belongs to C. Let us enumerate those that belong to the upper half of C by the numbers $r_k = \frac{k}{2^n}$, $-2^n \leq k \leq 2^n$.
(b) Let the upper half of C be 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 nth Farey series.

♡

[7] See **Info G**.

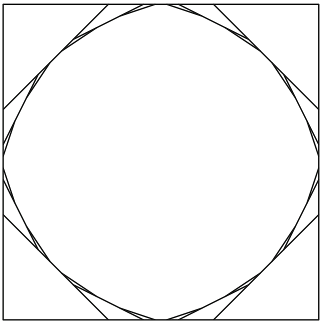


Fig. 6.2 The de Rham curve

6.3 Nice Parameterizations of Disks Tangent to a Given Disk

Let A be an Apollonian gasket. Choose a disk $D \in A$ corresponding to a Hermitian matrix M and consider those disks in A that are tangent to D.

The tangent points form a countable subset $T \subset \partial D$. We shall show later that one can parameterize points of T by rational numbers (including ∞) in such a way 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 disk tangent to D at the point $t_r \in T$ and let M_r be the corresponding Hermitian matrix.

We say that a parameterization $r \to t_r$ of T by $\overline{\mathbb{Q}}$ is *nice* if it has the following properties:

1. If $r = \frac{p}{q}$ in lowest terms, then

   $M_r = Ap^2 + 2Bpq + Cq^2 - M$, where A, B, C are fixed Hermitian matrices.

2. The disk 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, conditions 1 and 2 are very strong and contain all the information about tangent disks. Therefore, the next result is rather important.

Theorem 6.3. *Nice parameterizations exist and have the following 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*

$$G = \|(v_i, v_j)\| = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \qquad (6.3.1)$$

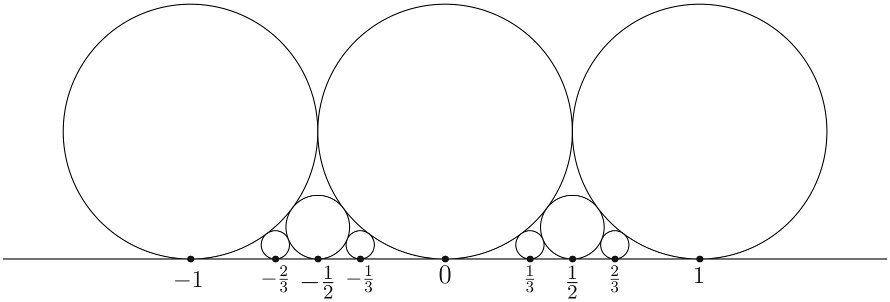


Fig. 6.3 Nice parameterization of a line in the band gasket

First main example: band gasket. Let $D = \{w \in \mathbb{C} \mid \operatorname{Im} w \leq 0\}$, $D_\infty = \{w \in \mathbb{C} \mid \operatorname{Im} w \geq 1\}$. Let $D_0,\ D_1$ be the disks of unit diameter, tangent to D at points $0,\ 1$ and to D_∞ at points $i,\ i+1$ (Fig. 6.3).

Then $\partial D = \overline{\mathbb{R}},\ T = \overline{\mathbb{Q}}$. The tautological parameterization of T is nice, with

$$M = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix},\ M_{\frac{p}{q}} = \begin{pmatrix} 2p^2 & -2pq - i \\ -2pq + i & 2q^2 \end{pmatrix},\ D_{\frac{p}{q}} : \left| w - \frac{2pq + i}{2q^2} \right| \leq \frac{1}{2q^2}.$$

Second main example: rectangular gasket. Let $D = \{w \in \mathbb{C} \mid |w| \geq 1\}$ be the complement to the open unit disk, and let D_0 be 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 parameterization 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 Theorem 6.3. Let $D_0,\ D_1,\ D_\infty$ be any three disks from A that are tangent to D and to each other. We associate the labels 0, 1 and ∞ to the corresponding tangent points in ∂D (Fig. 6.4).

Then, assuming that the theorem is true and the parameterization 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

$$A = M + M_\infty, \quad C = M + M_0, \quad B = \frac{1}{2}(M_1 - M - M_0 - M_\infty).$$

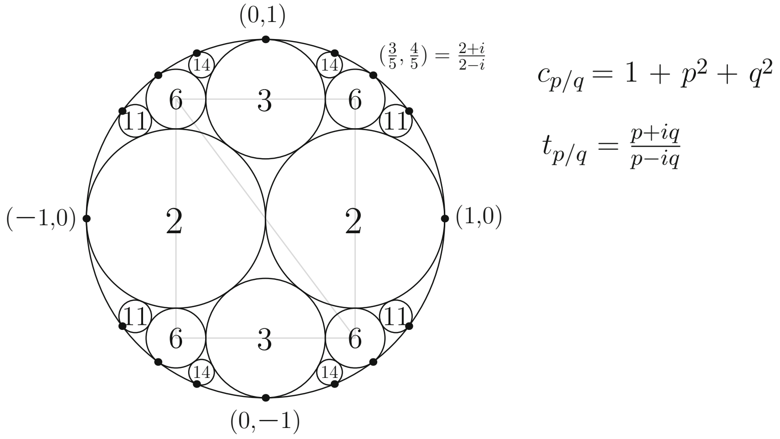

(0,1)
$(\frac{3}{5},\frac{4}{5}) = \frac{2+i}{2-i}$
14
14
6
3
6
11
11
(−1,0)
2
2
(1,0)
11
11
6
3
6
14
14
(0,−1)
$c_{p/q} = 1 + p^2 + q^2$
$t_{p/q} = \frac{p+iq}{p-iq}$

Fig. 6.4 Nice parameterization of the outer circle in the rectangular gasket

Then, using the property of matrices M_0, M_1, M_∞, and M, we can check relation (6.3.1). From there, statement 2 of the theorem follows easily if we define M_r using statement 1. □

Practically, a nice parameterization can be defined step by step. Assume that the disks 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 disk tangent to D_{r_1}, D_{r_2}, and D.

In fact, there are two such disks 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 disk tangent to D at the point $r = \frac{p}{q}$ (in lowest terms) is given by a quadratic polynomial in p, q:*

$$c(p,q) = \big(c_\infty + c\big)\cdot p^2 + \big(c_1 - c_0 - c_\infty - c\big)\cdot pq + \big(c_0 + c\big)\cdot q^2 - c, \quad (6.3.2)$$

where c_i is the boundary curvature of the disk D_i.

In particular, if four mutually tangent disks in an Apollonian gasket A have integral boundary curvatures, then all disks from A have this property.

Exercise 6.11. For the triangular Apollonian gasket, find the curvatures of the disks tangent to the outer disk.

Answer. $c(p,q) = \frac{2(p^2-pq+q^2)}{\sqrt{3}} + 1.$

Exercise 6.12. Describe the canonical parameterization for the outer circle of the triangular gasket.

Hint. Label by $0,\ 1,\ \infty$ the tangent points corresponding to the three maximal inner disks.

6.4 Integral Apollonian Gaskets

6.4.1 Basic Quadruples

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 disks 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 a *basic quadruple*.

Lemma 6.3. *For a basic quadruple, we have*

$$c_4 \leq 0, \qquad |c_4| < c_3 < \left(1 + \frac{2}{\sqrt{3}}\right)|c_4| \approx 2.1547\ldots \cdot |c_4|.$$

Proof. Let D_i, $1 \leq i \leq 4$, be a quadruple of mutually tangent disks with curvatures c_i, $1 \leq i \leq 4$. The first inequality has already been proved (see Remark 4.2).

Consider now Descartes's equation (4.1.3) as a quadratic equation in c_1 with given c_2, c_3, c_4. Then we get

$$c_1 = c_2 + c_3 + c_4 \pm 2\sqrt{c_2c_3 + c_3c_4 + c_4c_2}. \tag{6.4.1}$$

Since the initial quadruple is basic, we have to choose the minus sign in Eq. (6.4.1) (otherwise, we could replace c_1 by a smaller quantity).

The inequality $c_1 \geq c_2$ together with Eq. (6.4.1) gives $c_3 + c_4 \geq 2\sqrt{c_2c_3 + c_3c_4 + c_4c_2}$, or $(c_3 - c_4)^2 \geq 4c_2(c_3 + c_4) \geq (c_3 + c_4)^2$. This can be true only when $c_4 \leq 0$.

Finally, for nonpositive c_4, we have $(c_3 - c_4)^2 \geq 4c_2(c_3 + c_4) \geq 4c_3(c_3 + c_4)$, or $3c_3^2 + 6c_3c_4 + c_4^2 \leq 4c_4^2$. This gives $\sqrt{3}(c_3 + c_4) \leq -2c_4$, hence $c_3 \leq \frac{2+\sqrt{3}}{\sqrt{3}}|c_4|$. □

Here is a list of basic quadruples of small sizes generating nonisomorphic gaskets in order of increasing $|c_4|$:

$c_4 = 0$ $(1, 1, 0, 0)$;
$c_4 = -1$ $(3, 2, 2, -1)$;
$c_4 = -2$ $(7, 6, 3, -2)$;
$c_4 = -3$ $(13, 12, 4, -3)$, $(8, 8, 5, -3)$;
$c_4 = -4$ $(21, 20, 5, -4)$, $(9, 9, 8, -4)$;
$c_4 = -5$ $(31, 30, 6, -5)$, $(18, 18, 7, -5)$;
$c_4 = -6$ $(43, 42, 7, -6)$, $(15, 14, 11, -6)$, $(19, 15, 10, -6)$;
$c_4 = -7$ $(57, 56, 8, -7)$, $(20, 17, 12, -7)$, $(32, 32, 9, -7)$;
$c_4 = -8$ $(73, 72, 9, -8)$, $(24, 21, 13, -8)$, $(25, 25, 12, -8)$;
$c_4 = -9$ $(91, 90, 10, -9)$, $(50, 50, 11, -9)$, $(22, 19, 18, -9)$,
$(27, 26, 14, -9)$;
$c_4 = -10$
$(111, 110, 11, -10)$, $(39, 35, 14, -10)$, $(27, 23, 18, -10)$;

$c_4 = -11$
(133, 132, 12, −11), (72, 72, 13, −11), (37, 36, 16, −11), (28, 24, 21, −11).
Three general formulas:

$$\begin{array}{ll} c_4 = -km & (k^2 + km + m^2,\ k(k+m),\ m(k+m),\ -km) \\ c_4 = 1 - 2k & (2k^2,\ 2k^2,\ 2k+1,\ 1-2k) \\ c_4 = -4k & \big((2k+1)^2,\ (2k+1)^2,\ 4(k+1),\ -4k\big) \end{array}$$

The reader can find many other interesting facts about integral gaskets in [G03]. However, a description of all basic quadruples is still unknown.

Info I. The Möbius Inversion Formula

In number-theoretic computations, the *Möbius inversion formula* is frequently used. We explain here how it works.

Suppose we have a partially ordered set X with the property that for every element $x \in X$, there are only finitely many elements that are less than x. Let now f be any real- or complex-valued function on X. Define a new function F by the formula

$$F(x) = \sum_{y \leq x} f(y). \tag{I.1}$$

Proposition I.1. *There exists a unique function $\tilde{\mu}$ on $X \times X$ with the following properties:*

1. $\tilde{\mu}(x, y) = 0$ *unless* $y \leq x$.
2. $\tilde{\mu}(x, x) = 1$.
3. *If the functions f and F are related by Eq. (I.1), then*

$$f(x) = \sum_{y \leq x} \tilde{\mu}(x, y) F(y). \tag{I.2}$$

In many applications, the set X is a semigroup of nonnegative elements in some partially ordered abelian group G, and the order relation is translation-invariant: $y < x$ is equivalent to $a + y < a + x$ for every $a \in G$. In this case, μ is also translation-invariant, $\tilde{\mu}(a + x, a + y) = \tilde{\mu}(x, y)$, and hence it can be written in the form $\mu(x - y)$, where μ is a function on G that is zero outside X. The inversion formula takes the form

$$f(x) = \sum_{y \leq x} \mu(x - y) F(y) \qquad \text{(Möbius inversion formula).} \tag{I.3}$$

We leave the proofs for the interested reader and consider only some examples that 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, Proposition I.1 is true and the function μ is given by

$$\mu(n) = \begin{cases} 1 & \text{if } n = 0, \\ -1 & \text{if } n = 1, \\ 0 & \text{otherwise.} \end{cases}$$

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.,

$$(g_1, g_2) > (0, 0) \quad \Leftrightarrow \quad g_1 > 0 \quad \text{and} \quad g_2 > 0.$$

Here the μ-function for G is simply the product of the μ-functions for G_1 and G_2.

Note that if G_1 and G_2 are ordered groups, the $G = G_1 \times G_2$ is only partially ordered.

Example 3. $G = \mathbb{Q}^\times$ is the multiplicative group of nonzero rational numbers. The partial order is defined as follows: $r_1 \le r_2$ if the number $\frac{r_2}{r_1}$ is an integer. So in this case, $X = \mathbb{Z}_+$ with the order relation $m < n$ if $m \mid n$ (m is a divisor of n).

It is easy to see that this partially ordered group is the direct sum of a countable number of copies of $\mathbb{Z}$ with the usual order. Indeed, every element of G can be uniquely written in the form $r = \prod_{k \ge 1} p_k^{n_k}$, where p_k is the kth prime number, $n_k \in \mathbb{Z}$, and only finitely many of n_k are nonzero. The number r is an integer iff all n_k are nonnegative.

Therefore, the function μ is the product of infinitely many functions from Example 1. The exact definition is as follows.

Definition I.1.

$$\mu(n) = \begin{cases} 1\&\text{if } n = 1, \\ (-1)^k & \text{if } n \text{ is a product of } k \text{ distinct primes,} \\ 0 & \text{otherwise.} \end{cases}$$

Equation (I.3) in this case is the classical Möbius inversion formula

$$f(n) = \sum_{d \mid n} \mu(d) F\left(\frac{n}{d}\right). \tag{I.4}$$

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

$$n = \sum_{d \mid n} \varphi\left(\frac{n}{d}\right).$$

Applying the Möbius inversion formula, we get

$$\varphi(n) = \sum_{d|n} \mu(d) \cdot \frac{n}{d}, \qquad \text{or} \qquad \frac{\varphi(n)}{n} = \sum_{d|n} \frac{\mu(d)}{d}. \tag{I.5}$$

◇

6.4.2 Some Computations

A 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). We know the answer to the first question to a high degree of accuracy: in [MC03], it is shown that the Hausdorff dimension of the Apollonian gasket is $d = 1.308535???\ldots$. However, we have no idea about 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 disks in some Apollonian gasket that are tangent to a given disk D, e.g., to the outer disk in the 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 disks in the band gasket that are tangent to the real axis at the rational points of the segment $[0, 1]$. A more natural question, one with 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 disks tangent to the lower side of the square.

We know that the diameter of the disk with tangent point $\frac{m}{n} \in [0, 1]$ is $\frac{1}{n^2}$. Hence its area is $\frac{\pi}{4n^4}$. There are $\varphi(n)$ disks of this size. So for the area in question, we have the expression

$$A \quad = \quad \frac{\pi}{4} \cdot \sum_{n \geq 1} \frac{\varphi(n)}{n^4}. \tag{6.4.2}$$

This number can be expressed through the values of the Riemann ζ-function at the points 3 and 4.

Let us use the formula for $\varphi(n)$ obtained in Info I. The formula (6.4.2) takes the form

$$A \quad = \quad \frac{\pi}{4} \cdot \sum_{n \geq 1} \sum_{d|n} \frac{\mu(d)}{dn^3}.$$

We denote $\frac{n}{d}$ by m and sum over 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

$$A=\frac{\pi}{4}\cdot\frac{\zeta(3)}{\zeta(4)}=\frac{45\zeta(3)}{2\pi^3}\approx 0.872284.$$

The total area of the disks tangent to the outer disk of the rectangular gasket is equal to

$$\frac{\pi}{2}\cdot\sum_{\gcd(p,q)=1}\frac{1}{(p^2+q^2+1)^2}.$$

It can be expressed in terms of the ζ-function related to the Gaussian field $\mathbb{Q}(i)$.

Exercise 6.13. Let Σ_m denote the sum $\sum_{\mathbb{Z}^2\setminus\{(0,0)\}}\frac{1}{(k^2+l^2)^m}$. Show that

$$\sum_{\gcd(p,q)=1}\frac{1}{(p^2+q^2)^m}=\frac{\Sigma_m}{\zeta(2m)} \tag{6.4.2}$$

and

$$\sum_{\gcd(p,q)=1}\frac{1}{(p^2+q^2+1)^2}=\sum_{m=1}^{\infty}(-1)^{m-1}\frac{m\cdot\Sigma_{m+1}}{\zeta(2m+2)}. \tag{6.4.3}$$

Chapter 7
Geometric and Group-Theoretic Approach

Info J. The Hyperbolic (Lobachevsky) Plane L

A hyperbolic space satisfies all of the axioms of Euclidean space except for the famous fifth postulate about the 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 two-dimensional hyperbolic space, also known as the Lobachevsky plane L. We introduce three convenient models of L.

J.1 The First Poincaré Model

Let $\mathbb{C}$ be the complex plane with a complex coordinate $z = x + iy$, $x, y \in \mathbb{R}$ and i the imaginary unit. 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 mappings of both kinds (see Info F) acts on H and is, by definition, the full group of symmetries of L. So according to Felix Klein's philosophy, the geometric properties of L are those that 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 an explicit formula for the distance, we can proceed as follows. To every 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 $\mathrm{PSL}(2, \mathbb{R})$.

On the other hand, it is well known that for every 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 $\mathrm{PSL}(2, \mathbb{C})$.

We introduce the quantity

$$\Delta(p) := \lambda\bigl(q(p)\bigr) = \frac{z_2 - \bar{z}_1}{z_1 - \bar{z}_1} : \frac{z_2 - \bar{z}_2}{z_1 - \bar{z}_2} = \frac{|z_1 - \bar{z}_2|^2}{4 \operatorname{Im} z_1 \operatorname{Im} z_2}. \tag{J.1}$$

A.A. Kirillov, *A Tale of Two Fractals*, DOI 10.1007/978-0-8176-8382-5_7,

This function on the set of pairs of points in H is positive, symmetric, and invariant with respect to the full group $\overline{G}$. Let us clarify how it is related to the desired distance. To this end, we restrict our consideration to the subset T of H consisting of points $z(\tau) = i\mathrm{e}^{\tau}$, $\tau \in \mathbb{R}$. This subset is invariant under dilations $z(\tau) \mapsto \mathrm{e}^{t} z(\tau) = z(t + \tau)$ and admits a natural dilation-invariant distance $d(z(\tau_1), z(\tau_2)) = |\tau_1 - \tau_2|$.

Compare this distance with the restriction of Δ to $T \times T$:

$$\Delta(z(\tau_1), z(\tau_2)) = \frac{(\mathrm{e}^{\tau_1} + \mathrm{e}^{\tau_2})^2}{4\mathrm{e}^{\tau_1+\tau_2}} = \frac{1}{4}\,(\mathrm{e}^{\tau_1-\tau_2} + 2 + \mathrm{e}^{\tau_2-\tau_1}) = \cosh^2\left(\frac{\tau_1 - \tau_2}{2}\right).$$

We come to the relation

$$\Delta(z_1, z_2) = \cosh^2\left(\frac{d(z_1, z_2)}{2}\right) = \frac{\cosh\big(d(z_1, z_2)\big) + 1}{2}. \tag{J.2}$$

It holds on $T \times T$, and both sides are G-invariant.

Exercise J.1. Show that $G \cdot (T \times T) = H \times H$. More precisely, every pair of points (z_1, z_2) can be obtained by a transformation $g \in G$ from a pair $(i, i\mathrm{e}^{\tau})$ for an appropriate $\tau \in \mathbb{R}$.

It follows from the exercise that the relation (J.3) holds everywhere. A 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]

$$\text{area}\,(\Omega) = \int_{\Omega} \frac{\mathrm{d}x \wedge \mathrm{d}y}{y^2}, \qquad \text{length}\,(C) = \int_{C} \frac{\sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2}}{y}. \tag{J.4}$$

Exercise J.2. 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. Hence this geodesic must be invariant under every transformation $g \in G$ that preserves or permutes these two points. Apply this to the points $p = ir$, $q = ir^{-1}$ and transformations $s : z \mapsto -\bar{z}$, $t : z \mapsto -z^{-1}$.

[1] Indeed, the first integrand here is the unique (up to a scalar factor) differential 2-form that is invariant under the action of G. It is covariant under $\overline{G}$: a conformal mapping 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.

There is a remarkable relation between the area of a triangle with geodesic sides and its angles:

$$\operatorname{area}(ABC) = \pi - A - B - C. \tag{J.5}$$

Exercise J.3. Check formula (J.5) for a triangle with three zero angles given by the inequalities $-a \le x \le a$, $x^2 + y^2 \ge a^2$.

Exercise J.4. Show that the set of points $B_r(a) = \{z \in L \mid d(z, a) \le r\}$ (Lobachevsky disk) in the first Poincaré model is just an ordinary disk with center a' and radius r'. Express a' and r' in terms of a and r.

Answer. $a' = \operatorname{Re} a + i \cosh r \cdot \operatorname{Im} a, \qquad r' = \sinh r \cdot \operatorname{Im} a$.

Exercise J.5. Consider the Euclidean disk $D : (x - a)^2 + (y - b)^2 \le r^2$ on H. Find its diameter d and area A in the sense of 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 disk $D : x^2 + y^2 \le 1$. 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 the identity in $\overline{G}$ is the group $h \cdot \mathrm{PSL}(2, \mathbb{R}) \cdot h^{-1} = \mathrm{PSU}(1, 1; \mathbb{C})$.

To a pair $p' = (w_1, w_2) \in D^0 \times D^0$ we associate in a $\overline{G}'$-invariant way the quadruple $q'(p') = (w_1, w_2, \bar{w}_1^{-1}, \bar{w}_2^{-1})$. We 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 dilations of H in $\overline{G}$ given by matrices $g_\tau = \begin{pmatrix} e^{\tau/2} & 0 \\ 0 & e^{-\tau/2} \end{pmatrix}$ becomes the subgroup of matrices $g'_\tau = h \cdot g_\tau \cdot h^{-1} = \begin{pmatrix} \cosh \tau/2 & \sinh \tau/2 \\ \sinh \tau/2 & \cosh \tau/2 \end{pmatrix}$ in $\overline{G}'$. This subgroup preserves the interval $h \cdot T = T' = (-1, 1) \subset D^0$. Let us now introduce the local parameter t on T' so that $x = \tanh \frac{t}{2}$. Then the transformation g'_τ takes the simple form t to $t + \tau$. Therefore, the invariant distance on T is $d(t_1, t_2) = |t_1 - t_2|$. On the other hand,

$$\Delta' \left(\tanh \frac{t_1}{2}, \tanh \frac{t_2}{2} \right) = \frac{(1 - \tanh \frac{t_1}{2} \tanh \frac{t_2}{2})^2}{(1 - \tanh^2 \frac{t_1}{2})(1 - \tanh^2 \frac{t_2}{2})} = \cosh^2 \left(\frac{t_1 - t_2}{2} \right).$$

Then Eqs. (J.2) and (J.3) take the form

$$\Delta'(w_1, w_2) = \cosh^2\left(\frac{d'(w_1, w_2)}{2}\right) = \frac{\cosh\left(d'(w_1, w_2)\right) + 1}{2}, \tag{J.7}$$

$$\cosh\, d(w_1, w_2) = \frac{|1 - w_1\bar{w}_2|^2 + |w_1 - w_2|^2}{(1 - |w_1|^2)(1 - |w_2|^2)}. \tag{J.8}$$

Formula (J.4) is replaced by

$$\text{area}\,(\Omega) = \int_\Omega \frac{4\,\mathrm{d}x \wedge \mathrm{d}y}{(1 - x^2 - y^2)^2}, \qquad \text{length}\,(C) = \int_C \frac{2\sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2}}{1 - x^2 - y^2}. \tag{J.9}$$

The geodesics are arcs of circles orthogonal to ∂D (including the diameters of the disk). Formula (J.5) remains true.

Exercise J.6. Show that the set of points $\{z \in L \mid d(z, a) \leq r\}$ (Lobachevsky disk) in the second variant of the Poincaré model is an ordinary disk with center a' and radius r'. Express a' and r' in terms of a and r.

Answer. $a' = \frac{2a}{1+|a|^2+(1-|a|^2)\cosh r}$; $\qquad r' = \frac{(1-|a|^2)\sinh r}{1+|a|^2+(1-|a|^2)\cosh r}$.

Exercise J.7. Find the diameter d and area A of the disk $D_r(a, b) : (x - a)^2 + (y - b)^2 \leq r^2$ in D.

Answer. $d = \log \frac{(1+r)^2-a^2-b^2}{(1-r)^2-a^2-b^2}$; $\qquad A = 4\pi \sinh^2(\frac{d}{4})$.

J.3 The Klein Model

The extended Möbius group $\overline{G}$ is isomorphic to $\mathrm{PO}(2, 1, \mathbb{R}) \subset \mathrm{PGL}(3, \mathbb{R})$ (see Info F). Therefore, there is one more realization of the hyperbolic plane L. It is the *Klein model*, which we describe now.

The group $\mathrm{O}(2, 1, \mathbb{R})$ 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 $\mathrm{PO}(2, 1, \mathbb{R})$ on P preserves the circle $x^2 + y^2 = 1$ and the open disk $D^0 : x^2 + y^2 < 1$. This is the Klein model of L.

An explicit formula for the group action is

$$x \mapsto \frac{a'x + b'y + c'}{ax + by + c}, \quad y \mapsto \frac{a''x + b''y + c''}{ax + by + c}, \tag{J.10}$$

where

$$g = \begin{pmatrix} a' & b' & c' \\ a'' & b'' & c'' \\ a & b & c \end{pmatrix} \quad \text{belongs to} \quad \mathrm{O}(2,\ 1,\ \mathbb{R}) \subset \mathrm{GL}(3,\ \mathbb{R}).$$

We know that $g \in \mathrm{O}(2,\ 1,\ \mathbb{R})$ iff $g^t I g = I$, where $I = \mathrm{diag}\,(1, 1, -1)$, or in full detail,

$$\begin{gathered} (a')^2 + (a'')^2 = a^2 + 1, \quad (b')^2 + (b'')^2 = b^2 + 1, \quad (c')^2 + (c'')^2 = c^2 - 1, \\ a'b' + a''b'' = ab, \qquad b'c' + b''c'' = bc, \qquad c'a' + c''a'' = ca. \end{gathered} \tag{J.11}$$

Exercise J.8. (a) Show that the group $\mathrm{O}(2,\ 1,\ \mathbb{R})$ has four connected components characterized by the signs of $\det g$ and c.

(b) Show that $\mathrm{PO}(2,\ 1,\ \mathbb{R})$ has two connected components: $\mathrm{PSO}_+(2,\ 1,\ \mathbb{R})$ and $\mathrm{PSO}_-(2,\ 1,\ \mathbb{R})$ 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 $\overline{G} \simeq \mathrm{PO}(2,\ 1;\ \mathbb{R})$ as the second Poincaré model, but the group actions are different. More precisely, there exist a smooth map $f\colon D^0 \to D^0$ and a homomorphism $\alpha\colon \overline{G} \to \mathrm{PO}(2,\ 1,\ \mathbb{R})$ such that the following diagram is commutative:

$$\begin{tikzcd} \overline{G} \arrow[d, "\alpha"'] & \times & D^0 \arrow[r, "\text{conformal action}"] \arrow[d, "f"] & D^0 \arrow[d, "f"] \\ \mathrm{PO}(2,\ 1,\ \mathbb{R}) & \times & D^0 \arrow[r, "\text{projective action}"] & D^0 \end{tikzcd}$$

To describe the homomorphism α, consider first the connected component of the identity $G \subset \overline{G}$, which we identify with the group $\mathrm{PSU}(1,\ 1;\ \mathbb{C})$. The restriction of α to this subgroup induces the homomorphism $\tilde{\alpha}\colon \mathrm{SU}(1,\ 1;\ \mathbb{C}) \to \mathrm{SO}_+(2,\ 1;\ \mathbb{R})$, which has the form

$$g = \begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix} \quad \to \quad \tilde{\alpha}(g) = \begin{pmatrix} \mathrm{Re}(a^2 + b^2) & -\mathrm{Im}(a^2 + b^2) & 2\,\mathrm{Re}\,(ab) \\ \mathrm{Im}(a^2 + b^2) & \mathrm{Re}(a^2 - b^2) & -2\,\mathrm{Im}\,(ab) \\ 2\,\mathrm{Re}\,(\overline{a}b) & -2\,\mathrm{Im}\,(\overline{a}b) & |a|^2 + |b|^2 \end{pmatrix}. \tag{J.12}$$

The second connected component of $\overline{G}$ is a two-sided G-coset $c \cdot G = G \cdot c$, where c acts as complex conjugation on D^0. From the relation $c \cdot g \cdot c = \overline{g}$, we derive that $\alpha(c) = \mathrm{diag}(-1,\ 1,\ -1) \in \mathrm{SO}_-(2,\ 1;\ \mathbb{R})$, 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 points of an involution $\alpha(c)$ and 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 ordinary straight lines. Indeed, the projective transformations send lines to lines (in contrast to conformal mappings, which send circles to circles). In this model, the violation of the fifth postulate is most transparent.

To compute the map f, we use the following particular cases of Eq. (J.12):

$$\tilde{\alpha} : \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix} \quad \to \quad \begin{pmatrix} \cos 2\theta & -\sin 2\theta & 0 \\ \sin 2\theta & \cos 2\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$\tilde{\alpha} : \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix} \quad \to \quad \begin{pmatrix} \cosh 2t & 0 & \sinh 2t \\ 0 & 1 & 0 \\ \sinh 2t & 0 & \cosh 2t \end{pmatrix}.$$

We see that rotation through the angle 2θ in the Poincaré model corresponds to the same rotation in the Klein model.

In contrast, 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, α) in the Poincaré model and (ρ, θ) in the Klein model, the diffeomorphism f takes the form

$$f(r, \alpha) = (\rho, \theta) \quad \text{where} \quad \theta = \alpha, \qquad \rho = \tanh\big(2 \tanh^{-1}(r)\big). \tag{J.13}$$

Exercise J.9. Show that the relation between r and ρ in Eq. (J.13) can be written also in the following forms:

$$(a) \quad \frac{1+\rho}{1-\rho} = \left(\frac{1+r}{1-r}\right)^2; \qquad (b) \quad \rho = \frac{2}{r + r^{-1}}. \tag{J.14}$$

Another interesting geometric fact is that the diffeomorphism f "straightens" arcs of circles orthogonal to the boundary, sending them into corresponding chords (see Fig. J.1).

The Klein model has two disadvantages: a more complicated formula for the distance between two points and nonconformality (the angles between curves are not equal to Euclidean angles in the model).

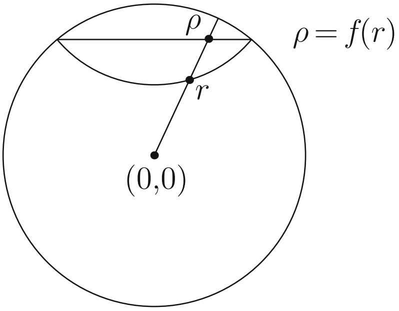


Fig. J.1 The diffeomorphism f

The area form and the length element for the Klein model in polar coordinates (ρ, θ) look like

$$\operatorname{area}(\Omega) = \int_\Omega \frac{\rho^2 \mathrm{d}\rho \wedge \mathrm{d}\theta}{(1-\rho^2)\sqrt{1+\rho^2}}, \tag{J.15}$$

$$\operatorname{length}(C) = \frac{1}{2}\int_C \frac{\sqrt{(\mathrm{d}\rho)^2 + \rho^2(1-\rho^2)(\mathrm{d}\theta)^2}}{1-\rho^2}.$$

Exercise J.10. 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 disk D bounded by the equator. Let p be the vertical projection of S^2_- to D.

Then the map $s \circ p^{-1}\colon D \to D$ is an isomorphism between the Klein and Poincaré models.

$\diamond$

7.1 The Möbius Group and Apollonian Gaskets

Here we consider in greater detail the action of the Möbius group G and extended Möbius group $\overline{G}$ in connection with Apollonian gaskets.

If we apply an (extended) Möbius transformation to a given Apollonian gasket $\mathcal{A}$, we obtain another gasket $\mathcal{A}'$. Moreover, we know that every Apollonian gasket can be obtained in this way from one fixed gasket. So, the set of all possible Apollonian gaskets forms a homogeneous space with G (or $\overline{G}$) as a group of motions.

Let $\operatorname{Aut}(\mathcal{A})$ (resp. $\overline{\operatorname{Aut}}(\mathcal{A})$) denote the subgroup of G (resp. of $\overline{G}$) consisting of transformations that preserve the gasket $\mathcal{A}$.

Theorem 7.1. *The subgroups* $\operatorname{Aut}\mathcal{A} \subset G$ *and* $\overline{\operatorname{Aut}}(\mathcal{A}) \subset \overline{G}$ *are discrete.*

Proof. Let D_1, D_2, D_3 be three mutually tangent disks in $\mathcal{A}$. Choose three interior points $w_1 \in D_1$, $w_2 \in D_2$, $w_3 \in D_3$. Afterward, choose a neighborhood of identity element $U \subset G$ that is small enough so that for every $g \in U$ 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 $g \in \operatorname{Aut} \mathcal{A}$, then it must send disks D_1, D_2, D_3 to some other disks of $\mathcal{A}$. Hence, an element $g \in U \bigcap \operatorname{Aut}(\mathcal{A})$ preserves D_1, D_2, D_3, hence their tangent points, and so must be the identity. This proves the discreteness of Aut $(\mathcal{A})$ in G.

The other statement can be proved in the same way by considering four mutually tangent disks. □

We want to describe the algebraic structure of the groups $\operatorname{Aut}(\mathcal{A})$ and $\overline{\operatorname{Aut}}(\mathcal{A})$. Fix one special gasket, e.g., the band gasket shown in Fig. 5.2. We denote it by $\mathcal{A}_0$. We denote respectively by D_1, D_2, D_3, D_4 the half-plane $\operatorname{Im} w \geq 1$, the half-plane $\operatorname{Im} w \leq -1$, the disk $|w-1| \leq 1$, and the disk $|w+1| \leq 1$. We call these the *original quadruple* in $\mathcal{A}_0$ and denote it by q_0.

First of all, we want to describe the subgroup of $\overline{G}$ that preserves the basic quadruple.

Theorem 7.2. *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* $\overline{\operatorname{Aut}}(\mathcal{A}_0)$ *and is isomorphic to S_4: all permutations of disks 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 $g \in G$ that 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 disk $g(D_4)$ is one of the two disks that are tangent to D'_1, D'_2, D'_3. These two disks 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. (This is obvious for the initial triple (D_1, D_2, D_3), and hence is true for every triple.) Thus, exactly one of the 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 every permutation s of disks in q_0 can be achieved by an element $g \in \overline{G}_0$, since there is a $g \in \overline{G}$ that 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 cannot be in G (it has at least six fixed points).

Recall that the set $\overline{G} \backslash G$ of antiholomorphic transformations, not being a group, still acts simply transitively on the set of ordered triples of distinct points. The stabilizer of an ordered triple is the reflection in the mirror passing through three points in question. (It is an easy exercise.) Therefore, it cannot have six fixed points that 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, $1 \leq i \leq 4$, that have with q_0 a common triple $T_i = Q_0 \backslash \{D_i\}$. Denote by D'_i the disk in q_i that is not in q_0 and by s_i the reflection that sends D_i to D'_i and preserves all other disks from q_0. See Fig. 7.1.

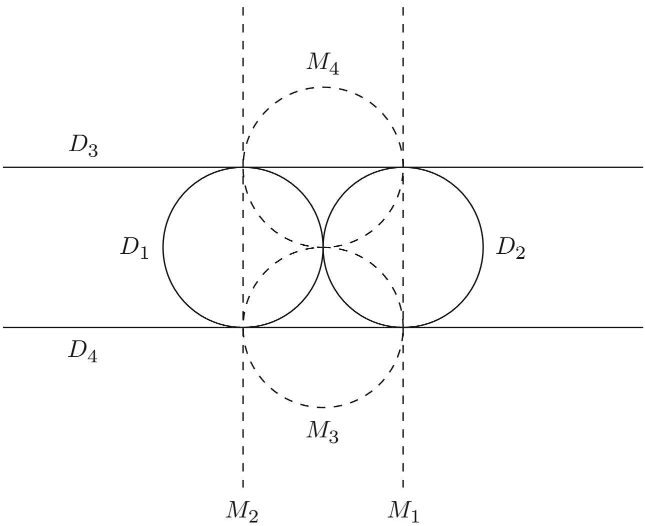


Fig. 7.1 Basic reflections: $s_i(D_j) = D_j$ for $i \neq j$. Here s_i is the reflection in M_i

Theorem 7.3. *The group generated by reflections s_i, $1 \leq i \leq 4$, is isomorphic to the group Γ_4 introduced in Info H.*

Outline of the proof. First of all, we recall (see Info H) that we have labeled elements of the group Γ_4 with words in the alphabet $\{1, 2, 3, 4\}$ that 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 the empty word $\emptyset$; the set $W^{(1)}$ contains four words $\{i\}$, where $i = 1, 2, 3, 4$,; the set $W^{(2)}$ contains 12 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 disk D_i under the action of the element $\gamma \in \Gamma_4$. The idea of the proof is to show that all disks $D_i(\gamma)$ are distinct.

First, we observe that $D_i(\gamma_1) \neq D_j(\gamma_2)$ for $i \neq j$. This follows from the fact that we can color all disks from $\mathcal{A}_0$ in four colors, so that all four colors occur in every quadruple of mutually tangent disks. Indeed, the set $S^2 \backslash q_0$ consists of four triangles bounded by three disks of different colors. So, for a new disk inscribed in each triangle, we can use the complementary color. In this new picture, again all quadruples contain four disks of different colors 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 disks in $\mathcal{A}_0$. Namely, let us consider all finite nonempty words in the alphabet $\{1, 2, 3, 4\}$ without repeating digits. To a one-digit word $\{i\}$ we associate the disk $D'_i = s_i D_i \in q_i$. In general, we associate to a word $\{i_1 i_2 \ldots i_k\}$ the disk $s_{i_1} s_{i_2} \cdots s_{i_k} D_{i_1}$.

It is enough to check that $D_i(\gamma) \neq D_i$ for $\gamma \neq e$. We leave this as a (nontrivial) exercise. One way is to compare the numeration of disks in Sect. 5.1 with labeling of elements of Γ_4 above. Another way is to see how the numeration changes when we replace the quadruple q_0 by $q_i := \underline{s_i} \cdot q_0$. □

We continue to study the action of $\overline{G}$ on disks.

Exercise 7.1. (a) Find all transformations $g \in \overline{G}$ that preserve the unordered triple D_1, D_2, D_3.
(b) Same question about the 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 disk D_4.

From Exercise 7.1, we derive the following result.

Theorem 7.4. *(a) The stabilizer $S \subset G$ of any unordered triple of mutually tangent disks 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 $\partial D_1, \partial D_2, \partial D_3$.*
(c) The stabilizer in G of every unordered quadruple of mutually tangent disks 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 every 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 $\tilde{T}$, the stabilizer in G is trivial, so an element $g \in G$ is completely determined by the ordered triple $g \cdot \tilde{T}$. For the same reason, an element $g \in \overline{G}$ is completely determined by the ordered quadruple $g \cdot \tilde{q}$.

Now consider all pairs of tangent disks in $\mathcal{A}_0$. They form a 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 $\Gamma_2 = \mathbb{Z}_2 * \mathbb{Z}_2$. 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 disk $D_1 \in \mathcal{A}_0$. 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. 7.2.

Then the stabilizer of this new D_1 in G is a subgroup of $\mathrm{PSL}(2, \mathbb{C})$ that stabilizes the upper half-plane. We leave it to the reader to check that it coincides with $\mathrm{PSL}(2, \mathbb{R}) \subset G$. The stabilizer in $\overline{G}$ is obtained by adding the reflection $s_0(w) = -\bar{w}$.

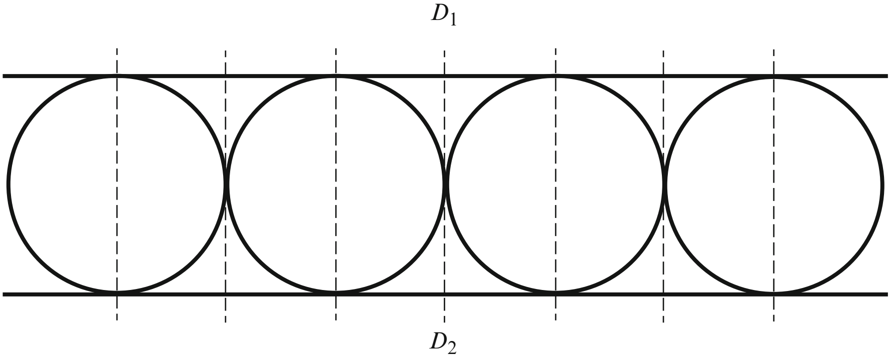


Fig. 7.2 The stabilizer of a pair (D_1, D_2)

The three generators of Γ_4 that preserve D_1 are

$$s_1 : w \mapsto -2 - \bar{w},\ s_2 : w \mapsto 2 - \bar{w},\ s_3 : w \mapsto \bar{w}^{-1}. \tag{7.1.1}$$

So the stabilizer of D_1 in Aut$\mathcal{A}$ is generated by

$$a_1 = s_0 s_1 : w \mapsto w-2,\ a_2 = s_0 s_2 : w \mapsto w+2,\ a_3 = s_0 s_3 : w \mapsto -w^{-1}. \tag{7.1.2}$$

It is not difficult to see that the matrices $\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ generate a subgroup of SL$(2, \mathbb{Z})$ that consists of matrices with two even and two odd elements.

7.2 Action of the Group Γ_4 on an Apollonian Gasket

Let q_0 be the original quadruple (see the text before Theorem 7.2). Denote by s_i, $1 \le i \le 4$, the reflection preserving three disks from q_0, excepting D_i.

Theorem 7.5. *(a) The group generated by s_1, s_2, s_3, s_4 is isomorphic to Γ_4. The action of this group on disks has four orbits, each of which contains one of the initial disks 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 disks tangent to D_1 has three orbits, each of which contains one of the disks 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 disks tangent to both D_1, D_2 has two orbits, each of which contains one of the disks D_3, D_4.

We omit the proof based on the results of previous sections but give here an illustration in which disks of four different Γ_4-orbits have different colors (Fig. 7.3).

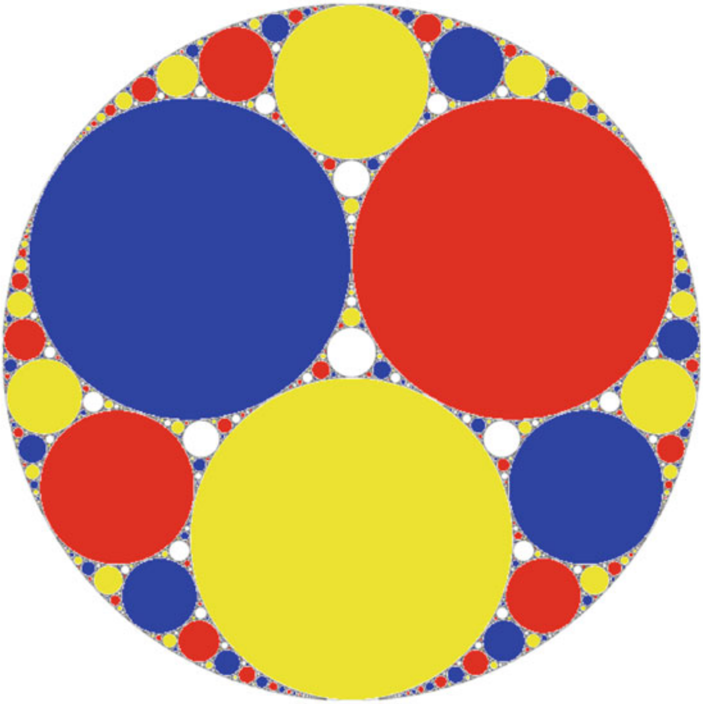


Fig. 7.3 Orbits of Γ_4

There is another group generated by reflections that acts on an Apollonian gasket. Namely, let h_{ij} be the reflection in the mirror that passes through the tangent point t_{ij} of D_i and D_j and is orthogonal to two other initial disks. It is clear that this reflection interchanges D_i and D_j and preserves two other initial disks. Let H be the group generated by the six reflections h_{ij}. We leave it to the reader to check that H is finite and isomorphic to the permutation group S_4.

Theorem 7.6. *The full group* $\mathrm{Aut}(A)$ *of fractional linear transformations of an Apollonian gasket A is the semidirect product $H \ltimes \Gamma_4$ of the subgroup H and the normal subgroup* Γ_4.

Outline of proof. By definition, H permutes the initial disks, and hence conjugation with $h \in H$ yields the corresponding permutation of generators s_i. It follows that the action of H normalizes the action of Γ_4.

Further, from Theorem 7.4, we conclude that Γ_4 can transform any unordered quadruple q to the initial quadruple q_0 (also considered unordered). Since H permutes the four disks 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 $\gamma' \in H \ltimes \Gamma_4$ that sends q back to q_0. The composition $\gamma' \circ \gamma$ preserves q_0, hence is the identity. Therefore, $\gamma = (\gamma')^{-1}$ belongs to $H \ltimes \Gamma_4$, and we are done. □

Exercise 7.2. Let $\mathcal{M}$ be the collection of all mirrors for $\mathcal{A}_0$. Is it a 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 $\mathbb{R}^4$ that 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

$$c'_k = \begin{cases} c_k & \text{if } k \neq i, \\ 2\sum_{j \neq i} c_j - c_i & \text{if } k = i. \end{cases} \qquad (7.2.1)$$

Lemma 7.1. *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 of Descartes's equation to another 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 ith coordinate axis.

From Eq. (4.1.3), we see first that Descartes's equation has the form $Q(c) = 0$, and second that the coordinate c_i of a solution c satisfies the 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 $c'_i + c_i = -p$ (Viète's theorem). Thus, we get another solution to Descartes's equation if we replace c_i by c'_i, leaving all other coordinates unchanged. □

Recall that we have defined above the change of coordinates (5.4.1) that sends integral solutions to Descartes's equation to integral light vectors in Minkowski space $\mathbb{R}^{1,3}$ with coordinates t, x, y, z. So we can consider the transformations s_i acting on $\mathbb{R}^{1,3}$. Lemma 7.2 implies that they belong to the pseudoorthogonal group $\mathrm{O}(1, 3; \mathbb{R})$. In fact, one can prove a more precise statement.

Exercise 7.3. Show that s_i acts on $\mathbb{R}^{1,3}$ as a reflection:

$$s_i(v) = v - \frac{2(v, \xi_i)}{(\xi_i, \xi_i)}\xi_i, \qquad (7.2.2)$$

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 diagonal form.

Hint. Check that the transformations (5.4.1) 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 7.2. *The group Γ_4 is isomorphic to the semidirect product $\mathbb{Z}_2 \ltimes F_3$, where F_3 is the free group on three generators, and the nontrivial 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 to show that the τ_i are free generators. The proof can be obtained from the explicit realization of Γ_4 given above. □

We define the homomorphism $\Phi \colon \Gamma_4 \to \mathrm{O}(1,\ 3;\ \mathbb{R})$ by $\Phi(s_i) = s^i$, $i = 0, 1, 2, 3$.

Theorem 7.7. *Φ is an isomorphism of Γ_4 to some discrete subgroup $\tilde{\Gamma}_4$ in* $\mathrm{O}_+(1, 3; \mathbb{R})$.

The generators of $\tilde{\Gamma}_4$ 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 $\tilde{\Gamma}_4$ (see, e.g., [CH65]).

Chapter 8
MultiDimensional Apollonian Gaskets

8.1 General Approach

Consider the analogue of the Descartes disk problem: find the relationship between curvatures of $n + 2$ mutually 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 parameterizations of $\mathcal{B}_n$. It is instructive to compare this general result with the case $n = 2$ studied in the previous sections.

First Parameterization. Let $\mathbb{R}^{1,n+1}$ be the $(n + 2)$-dimensional real vector space with coordinates $(p^0, \ldots, p^{n+1})$, endowed with the quadratic form

$$|p|^2 := \left(p^0\right)^2 - \left(p_1\right)^2 - \left(p_2\right)^2 - \cdots - \left(p_{n+1}\right)^2. \tag{8.1.1}$$

To every 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

$$H_p := \left\{ \alpha \in \mathbb{R}^{n+1} \;\middle|\; p^0 + \sum_{k=1}^{n+1} p^k \alpha_k \leq 0 \right\}. \tag{8.1.2}$$

Exercise 8.1. 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 onto a line orthogonal to H_p.

A.A. Kirillov, *A Tale of Two Fractals*, DOI 10.1007/978-0-8176-8382-5_8,

It is clear that for $c > 0$, the half-spaces H_p and H_{cp} coincide, whence $B_p = B_{cp}$. 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 parameterized by the points of the hyperboloid $|p|^2 = -1$ in $\mathbb{R}^{1,n+1}$.

Second Parameterization. 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 (8.1.2) goes to the inequality

$$a + (b,\, x) + c(x,x) < 0, \tag{8.1.3}$$

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 leave it to the reader to find a proof of the following lemma.

Lemma 8.1. *Two balls B_{p_1} and B_{p_2} are tangent iff $|p_1 + p_2|^2 = 0$.*

Exercise 8.2. 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 essentially 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.

$$\cos\alpha = -(p_1,\, p_2). \tag{8.1.4}$$

Let now $B_{p_k}, k = 1,\, 2,\, \dots,\, n+2$, be mutually tangent balls in $\overline{\mathbb{R}^n}$. Then, exactly as in Sect. 4.2, we see that[1]

$$\big(p_i,\, p_j\big) = 1 - 2\delta_{i,j}.$$

Thus 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 nonsingular, and the vectors p_k, $1 \le k \le n+2$, form a basis in $\mathbb{R}^{1,n+1}$.

Further, we introduce for each vector $v \in \mathbb{R}^{1,n+1}$, two kinds of coordinates: covariant coordinates $v_k = (v,\, p_k)$ and contravariant coordinates v^k by the condition $v = \sum v^k p_k$.

The relations between the two kinds of coordinates are derived exactly as we did in Sect. 1.4 for the two-dimensional case. They are

$$v_j = \sum_i v^i - 2v^j, \qquad v^i = \frac{1}{2n}\sum_j v_j - \frac{1}{2}v_i.$$

[1] This follows also from (8.1.4), since for externally tangent balls, $\cos\alpha = \cos\pi = -1$.

The quadratic form in these coordinates is expressed as

$$|v|^2 = \left(\sum_i v^i\right)^2 - 2\sum_i (v^i)^2 = \frac{1}{2n}\left(\left(\sum_i v_i\right)^2 - n\sum_i (v_i)^2\right).$$

Put now $v = (1, -1, 0, \ldots, 0, 0)$; 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

$$\left(\sum_k c_k\right)^2 = n \cdot \sum_k c_k^2, \tag{8.1.5}$$

which is the n-dimensional analogue of Descartes's equation.

Exercise 8.3. * Prove the n-dimensional analogue of the generalized Descartes equation:

$$\Sigma_1^2 = n \cdot \Sigma_2 - 2n^2 \cdot 1, \tag{8.1.6}$$

where

$$\Sigma_1 = \sum_{i=0}^{n+1} M_i, \qquad \Sigma_2 = \sum_{i=0}^{n+1} M_i^2, \tag{8.1.7}$$

and M_i, $0 \le i \le n+1$, are matrices corresponding to $n+2$ mutually tangent balls in $\mathbb{R}^n$.

Let $\{B_k^0\}_{1 \le k \le n}$ be a set of mutually tangent balls in $\mathbb{R}^n$. We want to describe all sequences $\{B_j\}_{j \in \mathbb{Z}}$ of balls in $\overline{\mathbb{R}}^n$ that have the property that B_j is tangent to $B_{j\pm 1}$ and to all $\{B_k^0\}_{1 \le k \le n}$. Let d_k be the curvature of B_k^0 and let c_j be the curvature of B_j. From (8.1.5) we have two equations:

$$(c_j + c_{j\pm 1} + d_1 + \cdots + d_n)^2 = n \cdot \big(c_j^2 + c_{j\pm 1}^2 + d_1^2 + \cdots + d_n^2\big).$$

Subtracting one from the other, 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 an inhomogeneous recurrence equation for the sequence $\{c_j\}$. Subtracting two such equations for successive j's, we get the homogeneous recurrence equation

$$(n-1)c_{j+1} - (n+1)c_j + (n+1)c_{j-1} - (n-1)c_{j-2} = 0. \tag{8.1.8}$$

The corresponding characteristic equation is

$$(n-1)\lambda^3 - (n+1)\lambda^2 + (n+1)\lambda - (n-1) = 0 \tag{8.1.9}$$

with roots $\lambda_0 = 1$, $\lambda_{\pm 1} = \frac{1 \pm \sqrt{n(2-n)}}{n-1}$. Note the different structures of these roots and consequently the different behaviors 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, for $n = 2$, the left-hand side of (8.1.8) is exactly the third difference of the sequence $\{c_j\}$.

For $n = 3$, the characteristic equation has roots 1 and $\frac{1 \pm \sqrt{-3}}{2}$, i.e., two sixth roots of unity that are not cube roots. Therefore, the sequence $\{c_j\}$ is 6-periodic. Moreover, not only the curvatures but the balls themselves form a 6-periodic sequence. This fact was known already in ancient Greece (see [Sod36] for details).

There is one more circumstance that we would like to mention. Since only three out of the six possible sixth roots of unity have been used, the sequence $\{c_j\}$ not only is 6-periodic, but has an additional property: $c_j + c_{j+3}$ is independent of j.

We leave to the reader to formulate the corresponding geometric property of the ball sequence.

Exercise 8.4. Let B_1, B_2, B_3 be three unit balls in $\mathbb{R}^3$ that are mutually tangent. Find six balls that are tangent to all of 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 $\lambda_{\pm 1} = \mathrm{e}^{\pm i\alpha}$. Then $\cos\alpha = \frac{1}{n-1}$.

Proposition 8.1. *All integral solutions to the equation* $\cos\frac{2\pi}{m} = \frac{1}{n}$ *have the form* $m = n = 1$*;* $m = 2, n = -1$*;* $m = 3, n = -2$*;*$m = 6, n = 2$*.*

It follows that for $n > 3$, the sequence of balls $\{B_j\}_{j \in \mathbb{Z}}$ in $\mathbb{R}^n$ tangent to n given balls has a quasiperiodic character and self-intersects infinitely many times.

From the recurrence relation

$$c_{j+1} = \frac{2}{n-1} c_j - c_{j-1}, \tag{8.1.10}$$

we conclude also that for $n > 3$, the curvatures cannot be integers for all j.

8.2 The Three-Dimensional Apollonian Gasket

As we saw above, the case $n = 3$ is exceptional. From any integral solution $(c_1, \ldots, c_5)$ to Descartes's equation, we can make five new solutions; namely, the ith transformation s_i replaces 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, every 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 the semidirect product $S_3 \ltimes \mathbb{Z}^2$.

Proposition 8.2. *For any three mutually tangent balls, the set of balls tangent to all three can be parameterized by the 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 8.3. *For any two mutually tangent balls, the set of balls tangent to both of them can be parameterized 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 these propositions and relate their statements to the structure of the subgroups $\langle s_i, s_j \rangle$ and $\langle s_i, s_j, s_k \rangle$.

Problem 8.1. 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 great deal of useful information about this problem can be found in the book [EGM98]. See also [Con97] as a very interesting introduction to the theory of quadratic forms.

The notion of a nice parameterization can be generalized to the three-dimensional case. Consider the algebraic number field $K = \mathbb{Q}[\varepsilon]$, where $\varepsilon = \mathrm{e}^{\frac{2\pi i}{3}}$ is a cube root of unity. A general element of K has the form $k = \alpha\varepsilon + \beta\bar{\varepsilon}$, where $\alpha, \beta \in \mathbb{Q}$, and bar means complex conjugation. Note that

$$||k||_K^2 = |k|^2 = k\bar{k} = \alpha^2 - \alpha\beta + \beta^2. \tag{8.2.1}$$

Denote by E the set of all complex numbers of the form $a\varepsilon + b\bar{\varepsilon}$, 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, $\pm\varepsilon$, $\pm\bar{\varepsilon}$. They are called *units* of the ring E. It is well known that every element of E can be uniquely (modulo units) written as a product of primes. As for the primes, they include all rational (i.e., ordinary) primes of the form $p = 3m - 1$ and also the numbers $k = a\varepsilon + b\bar{\varepsilon}$ 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 every 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 $k = \frac{l\varepsilon + m\bar{\varepsilon}}{n}$, where l, m, n are ordinary integers with $\gcd(l, m, n) = 1$.

Definition 8.1. Let D be a 3-ball in an integral three-dimensional Apollonian gasket $\mathcal{A}$. A parameterization 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 that corresponds to the point $k = \frac{p}{q} \in \overline{K}$.

Theorem 8.1. *Nice parameterizations exist and have the following properties:*

(a) Let $K \ni k = \frac{p}{q}$. The curvature c_k of the ball D_k has the form

$$c_k = \alpha|p|^2 + \beta p\bar{q} + \bar{\beta}\bar{p}q + \gamma|q|^2 + \delta, \tag{8.2.2}$$

where $\alpha,\ \gamma,\ \delta \in \mathbb{R}$, $\beta \in \mathbb{C}$.

(b) There is a coordinate system $(x_1,\ x_2,\ x_3)$ *in the ambient space* $\mathbb{R}^3$ *such that*

$$x_i = \frac{\alpha_i|p|^2 + \beta_i p\bar{q} + \bar{\beta}_i\bar{p}q + \gamma_i|q|^2 + \delta_i}{\alpha|p|^2 + \beta p\bar{q} + \bar{\beta}\bar{p}q + \gamma|q|^2 + \delta}. \tag{8.2.3}$$

(c) Let $k_i = \frac{p_i}{q_1}$, $i = 1,\ 2$. *The balls* D_{k_1} *and* D_{k_2} *are tangent iff*

$$|k_1 - k_2| = \frac{1}{|q_1q_2|}. \tag{8.2.4}$$

We leave to the reader the proof of the theorem and development of the matrix variant of the theory.

In conclusion, we illustrate Theorem 8.1 by two examples of nice parameterizations for a three-dimensional Apollonian gasket.

We associate to a ball in $\mathbb{R}^3$ with 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 analogue 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 $2\bar{\varepsilon}$.

Our first example is the parameterization 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

$$M_k = \begin{pmatrix} 4|p|^2 + |q|^2 - 2 & 2p\bar{q} + (1 - |q|^2)j \\ 2\bar{p}q - (1 - |q|^2)j & |q|^2 \end{pmatrix}. \tag{8.2.5}$$

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 radius $r = \frac{1}{|q|^2}$.

Our second example is the parameterization 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

$$M_k = \begin{pmatrix} |p|^2 + |q|^2 + 1 & 2p\bar{q} + (|p|^2 - |q|^2)j \\ 2\bar{p}q + (|q|^2 - |p|^2)j & |p|^2 + |q|^2 - 1 \end{pmatrix}. \tag{8.2.6}$$

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, Boston, 1988)

[LGRE00] N. Lesmoir-Gordon, W. Rood, R. Edney, *Fractal Geometry* (Icon Books, UK/Totem books, USA, 2000)

[Sod36] F. Soddy, The kiss precise. Nature **7**, 1021 (1936)

[Ste03] K. Stephenson, Circle packing: a mathematical tale. Not. Am. Math. Soc. **50**, 1376–1388 (2003)

[Str99] R.S. Stricharts, Analysis on fractals. Not. Am. Math. Soc. **46**, 1999–1208 (1999)

B Books

[Bea83] A.F. Beardon, in *The Geometry of Discrete Groups*. Graduate Texts in Mathematics, vol. 91 (Springer, Berlin, 1983)

[Con97] J.H. Conway, in *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, New York, 1969)

[CH91] R. Courant, D. Hilbert, *Methods of Mathematical Physics*, vol. 2 (Wiley, New York, 1991)

[Edg90] Edgar, in *Measure Topology and Fractal Geometry*. GTM (Springer, Berlin, 1990)

[EGM98] J. Elstrodt, F. Grunewald, J. Mennike, in *Groups Acting on Hyperbolic Space* (Springer, Berlin, 1998). Russian translation in MCCME, Moscow, 2003

[EGM98] J. Kigami, in *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)

[CH65] W.O.J. Moser, H.M.S. Coxeter , in *Generators and Relations for Discrete Groups*, 2nd edn. Ergebnisse der Mathematik und ihrer Grenzgebiete. Reihe, Gruppentheorie. n.F., vol. 14 (Springer, Berlin, 1965)

[Thu97] W. Thurston, in *Three-Dimensional Geometry and Topology*. Princeton Mathematical Series, vol. 35 (Princeton University Press, Princeton, 1997)

A.A. Kirillov, *A Tale of Two Fractals*, DOI 10.1007/978-0-8176-8382-5,

C Research Papers

[AS97] D. Aharonov, K. Stephenson, Geometric sequences of discs in the Apollonian packing. Algebra i Analiz **9**, 104–140 (1997). English translation, St. Petersburg Math. J. **9**, 509–545 (1998)

[Bar92] M.T. Barlow, Harmonic analysis on fractal sets. Seminar Bourbaki, Exp. 755, Astérisque **206**, 345–368 (1992)

[BL04] A.F. Beardon, L. Lorentzen, Continued fractions and restrained sequences of Möbius maps. Rocky Mountain J. Math. **34**, 441–466 (2004)

[Bro85] R. Brooks, The spectral geometry of the Apollonian packing. Comm. Pure Appl. Math. **38**, 358–366 (1985)

[DK] R.L. Dobrushin, S. Kusuoka, in *Statistical Mechanics and Fractals*. Lecture Notes in Mathematics, vol. 1567 (Springer, Berlin, 1993), pp. 39–98

[FS92] M. Fukushima, T. Shima, On a spectral analysis for the Sierpiński gasket. Potential Anal. **1**, 1–35 (1992)

[G03] R.L. Graham, J.C. Lagarias, C.L. Mallows, A. Wilks, C. Yan, Apollonian packings: number theory. J. Num. Theor. **100**, 1–45 (2003)

[KS43] E. Kasner, F. Supnik, The Apollonian packing of circles. Proc. Nat. Acad. Sci. USA **29**, 378–384 (1943)

[MC03] C.T. MacMullen, Hausdorff dimension and conformal dynamics III: computation of dimension. Amer. J. Math. **120**(4), 691–721 (1988)

[MT95] L. Malozemov, A. Teplyaev, Pure point spectrum of the Laplacians on fractal graphs. J. Funct. Anal. **129**, 390–405 (1995)

[Nev49] E.H. Neville, The structure of Farey series. Proc. London Math. Soc. **51**(2), 132–144 (1949)

[Ram84] R. Rammal, Spectrum of harmonic excitations on fractals. J. Physique **45**, 191–206 (1984)

[de Rha59] G. de Rham, Sur les courbes limites de polygones obtenus par trisection. Enseignement Math. **5**(2) , 29–43 (1959) (French)

[de Rha56] G. de Rham, Sur une courbe plane. J. Math. Pures Appl. (**9**) **35**, 25–42 (1956) (French)

[de Rha47] G. de Rham, Un peu de mathmatiques propos d'une courbe plane. Elemente der Math. **2**, 73–76, 89–97 (1947) (French)

[de Rha56] G. de Rham, Sur quelques courbes definies par des equations fonctionnelles. Univ. e Politec. Torino. Rend. Sem. Mat. **16**, 101–113 (1956/1957) (French)

[Sal43] R. Salem, On some singular monotonic functions which are strictly increasing. Trans. Am. Math. Soc. **53**, 427–439 (1943)

[Str00] Robert S. Strichartz, Taylor approximations on Sierpinski gasket type fractals. J. Funct. Anal. **174**, 76–127 (2000)

[TAV00] A.V. Teplyaev, Gradients on fractals. J Funct. Anal. **174**, 128–154 (2000)

D Web Sites

http://en.wikipedia.org/wiki/Fractal
http://classes.yale.edu/fractals/
http://www.faqs.org/faqs/fractal-faq/

Index

A.A. Kirillov, *A Tale of Two Fractals*, DOI 10.1007/978-0-8176-8382-5,