TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Diplomarbeit

Computer Algebra and Analysis:

Complex Variables Visualized

Ausgeführt am

Research Institute for Symbolic Computation (RISC)

der Johannes Kepler Universität Linz

unter Anleitung von

Univ.-Prof. Dr. Peter Paule

durch

Thomas Ponweiser

Stockfeld 15
4283 Bad Zell

Datum

Unterschrift

Kurzfassung

Ziel dieser Diplomarbeit ist die Visualisierung einiger grundlegender Ergebnisse aus dem Umfeld der Theorie der modularen Gruppe sowie der modularen Funktionen unter Zuhilfenahme der Computer Algebra Software *Mathematica*.

Die Arbeit gliedert sich in drei Teile. Im ersten Kapitel werden für diese Arbeit relevante Begriffe aus der Gruppentheorie zusammengefasst. Weiters werden Möbius Transformationen eingeführt und deren grundlegende geometrische Eigenschaften untersucht.

Das zweite Kapitel ist der Untersuchung der modularen Gruppe aus algebraischer und geometrischer Sicht gewidmet. Der kanonische Fundamentalbereich der modularen Gruppe sowie die daraus abgeleitete Kachelung der oberen Halbebene wird eingeführt. Weiters wird eine generelle Methode zum Auffinden von Fundamentalbereichen für Untergruppen der modularen Gruppe vorgestellt, welche sich zentral auf die Konzepte der 2-dimensionalen hyperbolischen Geometrie stützt.

Im dritten Kapitel geben wir einige konkrete Beispiele, wie die aufgebaute Theorie für die Visualisierung bestimmter mathematischer Sachverhalte angewendet werden kann. Neben der Visualisierung von Graphen modularer Funktionen stellt sich auch der Zusammenhang zwischen modularen Transformationen und Kettenbrüchen als besonders schönes Ergebnis dar.

Abstract

The aim of this diploma thesis is the visualization of some fundamental results in the context of the theory of the modular group and modular functions. For this purpose the computer algebra software *Mathematica* is utilized.

The thesis is structured in three parts. In Chapter 1, we summarize some important basic concepts of group theory which are relevant to this work. Moreover, we introduce Möbius transformations and study their geometric mapping properties.

Chapter 2 is devoted to the study of the modular group from an algebraic and geometric point of view. We introduce the canonical fundamental region which gives rise to the modular tessellation of the upper half-plane. Additionally, we present a general method for finding fundamental regions with respect to subgroups of the modular group based on the concepts of 2-dimensional hyperbolic geometry.

In Chapter 3 we give some concrete examples how the developed results and methods can be exploited for the visualization of certain mathematical results. Besides the visualization of function graphs of modular functions, a particularly nice result is the connection between modular transformations and continued fraction expansions.

Preface

The centerpiece of the present diploma thesis, *Computer Algebra and Analysis: Complex Variables Visualized*, is the modular group. It plays an important role in many areas of mathematics, as for example in number theory due to its connection with partition numbers or continued fractions.

This thesis is structured in three parts. In Chapter 1, we introduce basic notions and definitions which are fundamental for the rest of this work. Firstly, we summarize the most important basic concepts of group theory. Moreover, we introduce *Möbius transformations* and study their connection to *stereographic projection* in detail. Finally, we define the concept of *generalized circles* and *generalized disks*, using an elegant characterization in terms of Hermitian matrices which turns out to be particularly advantageous in the context of Möbius transformations.

Chapter 2 is devoted to the study of the modular group from an algebraic and geometric point of view. Two different and independent algorithms, the T-U algorithm and the T-R algorithm, are presented which both yield group word representations for arbitrary modular transformations in terms of the transformations $T : z \mapsto \frac{1}{z}$, $U : z \mapsto z + 1$ and $R : z \mapsto \frac{1}{z+1}$. Geometric considerations come into play when introducing *fundamental regions* for the action of the modular group on the extended complex plane. A canonical fundamental region is derived which gives rise to the *modular tessellation of the upper-half plane*. Lastly, the basic concepts of 2-dimensional *hyperbolic geometry* are introduced in order to present an alternative and more general method for finding fundamental regions. This method gives rise to so-called *normal polygons* and works as well for subgroups of the modular group.

Finally, Chapter 3 has a clear emphasis on visualization. Firstly, *generalized matrix powers* are introduced as a device for visualizing continuous transitions between given sets and their Möbius-transformed images. Secondly, the relation between the modular transformations, *Ford circles* and *continued fractions* is studied in detail and visually explained using the continued fraction expansion of the irrational number π as an example. Lastly, documenting the important role of the modular group within complex analysis, the most basic

results from the theory of *modular functions* are summarized. Using an adequate color coding, graphs of certain selected modular functions are depicted whose inherent visual aesthetics and symmetry reflect well the beauty of this theory.

Achievements

Complementary to this thesis, the *Mathematica* package *ModularGroup* has been developed. This package, together with some *interactive demonstrations*, may be downloaded from the website of the *Research Institute for Symbolic Computation* (RISC).[1] It contains essentially all algorithms described in this thesis as well as functions for visualization of generalized circles, generalized disks, modular tessellations and more. The main attention has been paid to an efficient and fast implementation of these algorithms, relying in many aspects on compiled functions using machine-precision integer and floating point arithmetic. Note that all figures included in this thesis are based on this package.

Furthermore it is worth noting that this thesis also contains ideas which have not been found directly in this form in the referenced literature. During the implementation of the enumeration algorithm for modular transformations, it got apparent that the left- or rightmost symbol of the unique T-R group word of any given modular transformation can be read off directly from its corresponding matrix. This observation leads to the T-R algorithm of Section 2.1 and to the alternative proof for the presentation for the modular group in terms of the generators T and R (Theorem 2.13).

For the proof that the region $\mathcal{F} := \{z \in \mathbb{C} \mid |\mathrm{Re}\,(z)| < 1 \wedge |z| > 1\}$ is a fundamental region for the action of the modular group on the upper half-plane, we intentionally take a different and slightly longer track than for example Klein/Fricke [7] or Schoeneberg [14]. We first derive a fundamental region for the action of the homogeneous modular group on $\mathbb{C}^2$ by looking for representative vectors of minimal Euclidean norm. By carrying over the result to the inhomogeneous case, we indeed obtain the fundamental region $\mathcal{F}$ in a very natural and instructive way.

For visualization of objects living on the upper half-plane of $\mathbb{C}$, such as the modular tessellation or graphs of modular functions, we frequently make use of a Möbius transformation which maps the upper half-plane to the unit disk. This allows us to visualize the whole picture rather than just an arbitrary rectangular fragment of it. It turns out that in this context the most natural

[1] http://www.risc.jku.at/

choice for such a Möbius transformation is *not* the well-known Cayley transform, but in fact a map which we introduce in Example 1.43 as the *modified Cayley transform*.

Another idea suggesting itself is to consider the inscribed circle of the canonical fundamental region $\mathcal{F}$ (see Figure 2.2). Indeed, the introduction of so-called *indisks* turns out to be very fruitful in the study of the relation between modular transformations and continued fractions in Section 3.2. It leads to the notion of *indisk-paths*, which in turn give rise to an alternative proof for the presentation of the modular group in terms of the generators T and R (Corollary 3.16).

Outlook

The *Mathematica* package *ModularGroup* may be extended for a systematic treatment of various congruence subgroups of the modular group. The algorithm which has been used for drawing the *normal polygons* in Section 2.4.1 has been one of the latest results of this work. It is still preliminary and may be added to the package at a later stage.

Also the implementation of H. A. Verrill's algorithm for the visualization of fundamental regions of congruence subgroups[2] – which has actually been the starting point for this thesis – is easily possible based on the present *Mathematica* package.

Acknowledgment

I want to thank everybody who supported me in the course of my work on this diploma thesis.

I am particularly grateful to my supervisor *Prof. Peter Paule* for proposing this truly rewarding topic and its neat working title. I very much appreciated his valuable input and guidance as well as our constructive project meetings.

My special thanks also go to *Prof. Günther Karigl* for his prompt and kind answers to my administrative questions and for approving this thesis at the Research Institute for Symbolic Computation of the Johannes Kepler University Linz.

Last but not least I want to thank my family for supporting me. Thanks to my little daughter Lisa for being my sunshine and reminding me of the importance of baby steps. To my beloved wife I want to say thank you for always motivating me. Thank you for being the wind in my sails!

[2] See https://www.math.lsu.edu/∼verrill/

Contents

List of Figures

Chapter 1

Basic notions and definitions

1.1 Groups and algebraic constructions

In this section, we will recapitulate some basic algebraic concepts, most important the notions of *free groups* and *free products*, and the construction of a group in terms of *generators* and *relations*. Additionally, we give the definitions of the *(projective) general* and *special linear groups*. Lastly, *group actions* on sets will be introduced. A reader familiar to this concepts may readily skip this section. Moreover, we will give no rigorous proofs here, as they may be found in many algebra books, as for example in Hungerford [5].

1.1.1 Groups and homomorphisms

If G is a nonempty set, a *binary operation* on G is a function $G \times G \to G$. Commonly used notations for the image of $(a, b) \in G \times G$ under a binary operation are $a \cdot b$ or ab (product notation), $a + b$ (additive notation), $a \circ b$, $a * b$, etc. In this chapter and also later on we will use the *product notation*, $(a, b) \mapsto a \cdot b = ab$, most frequently and we refer to ab as the *product* of a and b.

Definition 1.1 (Groups and monoids)**.** Let G be a nonempty set together with a binary operation on G. If the binary operation satisfies the following three axioms,

(i) *Associativity:* $\forall a, b, c \in G : a(bc) = (ab)c$,

(ii) *Existence of an identity element:* $\exists e \in G \; \forall a \in G : ea = ae = a$,

(iii) *Existence of inverse elements:* $\forall a \in G \; \exists a^{-1} \in G : aa^{-1} = a^{-1}a = e$,

then G is called a *group*. If the axioms (i) and (ii) are satisfied, then G is called a *monoid*. Of course every group is in particular a monoid. The notation $\langle G, \cdot, e\rangle$ will be used for making the identity element and the involved binary operation explicit.

It is convenient to introduce the following notations operating on subsets of a group G. For $g \in G$ and $A, B \subseteq G$ we define

$$\begin{aligned} A^{-1} &:= \{a^{-1} \mid a \in A\} \subseteq G, \\ AB &:= \{ab \mid a \in A, b \in B\} \subseteq G, \\ gA &:= \{g\}A = \{ga \mid a \in A\} \subseteq G, \\ Ag &:= A\{g\} = \{ag \mid a \in A\} \subseteq G. \end{aligned}$$

Definition 1.2 (Subgroups)**.** Let $\langle G, \cdot, e\rangle$ be a group. A nonempty subset $H \subseteq G$ which is closed under the binary operation $\cdot$ and inversion, i.e. $HH^{-1} = H$, is called a *subgroup* of G, in symbols $H \leq G$. In particular, $\langle H, \cdot, e\rangle$ is itself a group. If $H \leq G$ and $H \neq G$, H is called a *proper subgroup* of G, and we write $H \lneq G$. The subgroup $\{e\} \leq G$ is called the *trivial subgroup*.

Definition 1.3 (Cosets)**.** Let H be as subgroup of a group G and $g \in G$. The set gH is called a *left coset* of H in G and Hg is called a *right coset*.

If H is a subgroup of G, then for every $g \in G$ also gHg^{-1} is a subgroup of G. For $h \in H$, $g \in G$, we call the operation $h \mapsto ghg^{-1}$ *conjugation* of h by g.

Definition 1.4 (Normal subgroups)**.** A subgroup N of a group G which is closed under conjugation by elements of G, i.e. $gNg^{-1} = N$ for all $g \in G$, is called a *normal subgroup* of G, in symbols $N \trianglelefteq G$.

If N is a normal subgroup of G, then we have $gN = Ng$ for every $g \in G$, in other words left and right cosets of N in G coincide. We can therefore define on the set of cosets

$$G/N := \{gN \mid g \in G\}, \tag{1.1}$$

a binary operation by $aNbN := abNN = abN$ for $a, b \in G$. It is easy to see that G/N together with this operation forms a group.

Definition 1.5 (Factor groups)**.** Let N be a normal subgroup of a group G. The set G/N defined in (1.1) together with the binary operation $(aN, bN) \mapsto abN$ is called the *factor group* (or *quotient group*) of G by N.

Definition 1.6 (Homomorphisms)**.** Let G and H be groups (or monoids). A function $f : G \to H$ is a *homomorphism* from G to H, if and only if

$$f(ab) = f(a)f(b) \quad \text{for all } a, b \in G.$$

Depending on f being injective and/or surjective, one distinguishes *monomorphisms* (f is injective), *epimorphisms* (f is surjective) and *isomorphisms* (f is bijective). We call the groups G and H *isomorphic*, in symbols $G \cong H$, if and only if there exists an isomorphism $f : G \to H$. Finally, a homomorphism $f : G \to G$ of G to itself is called *endomorphism*. A bijective endomorphism is called *automorphism*.

Definition 1.7 (Kernel)**.** Let G and H be groups and denote the identity element of H by e_H. For a homomorphism $f : G \to H$, the set

$$\ker f := \{g \in G \mid f(g) = e_H\} \tag{1.2}$$

is called the *kernel* of f.

Lemma 1.8. *Let G and H be groups and $f : G \to H$ be a homomorphism. The kernel of f is a normal subgroup of G and the image $f(G)$ is a subgroup of H:*

$$\ker f \trianglelefteq G \quad \textit{and} \quad f(G) \leq H.$$

Proof. Denote the identity elements of G and H by e_G and e_H respectively. Because f is a homomorphism, we have $f(e_G) = f(e_G e_G) = f(e_G)f(e_G)$. Multiplying this by $f(e_G)^{-1}$ yields $e_H = f(e_G)$. In other words $e_G \in \ker f$ and $\ker f$ is nonempty. To see that $\ker f \leq G$, we need to show that for all $a, b \in \ker f$ also $ab^{-1} \in \ker f$ – but this is easy:

$$f(ab^{-1}) = f(a)f(b^{-1}) = f(a)f(b)^{-1} = e_H {e_H}^{-1} = e_H.$$

To see that $\ker f$ is a normal subgroup of G, let $a \in \ker f$, $g \in G$ and observe

$$f(gag^{-1}) = f(g)f(a)f(g^{-1}) = f(g)e_H f(g)^{-1} = e_H.$$

For this reason $gag^{-1} \in \ker f$ and $\ker f$ is closed under conjugation by elements of G. For the second part of the proof, let $f(a), f(b) \in f(G)$ with $a, b \in G$. Clearly $f(a)f(b)^{-1} = f(a)f(b^{-1}) = f(ab^{-1}) \in f(G)$ and therefore $f(G) \leq H$. □

Theorem 1.9 (First isomorphism theorem)**.** *Let G and H be groups and $f : G \to H$ be a homomorphism. Then the quotient group $G/\ker f$ is isomorphic to $f(G)$:*

$$G/\ker f \cong f(G).$$

Proof. Let us again denote the identity element of H by e_H. Define $N := \ker f$ and a map $\varphi : G/N \to f(G)$ by $aN \mapsto f(a)$. Now for arbitrary $a, b \in G$ observe

$$aN = bN \Leftrightarrow ab^{-1} \in N \Leftrightarrow e_H = f(ab^{-1}) = f(a)f(b)^{-1} \Leftrightarrow f(a) = f(b).$$

This shows two things: First ($\Rightarrow$), φ is well defined and second ($\Leftarrow$), φ is injective. Clearly φ is also surjective and by definition we have

$$\varphi(aN)\varphi(bN) = f(a)f(b) = f(ab) = \varphi(abN) \quad \text{for all } a, b \in G.$$

Therefore φ is indeed an isomorphism from G/N to $f(G)$. □

1.1.2 Free monoids

Let Σ be a set of formal symbols, for example $\Sigma = \{a, b, c, \dots\}$. For $n \in \mathbb{N}$, $n \geq 0$, we denote by Σ^n the set of tuples of length n over Σ and we will call these tuples *words of length n over the alphabet Σ*. For simplicity, we will omit parentheses and commas when notating such a word, for example $(a, b, b, a) =: abba \in \Sigma^4$ is a word of length 4. Note the special case for $n = 0$. By definition, Σ^0 contains all words of length 0. Obviously, there is only one such word, the so-called *empty word*, which we denote by ϵ. A second special case occurs for $\Sigma = \emptyset$, where $\Sigma^n = \emptyset$ for all $n > 0$ and $\Sigma^0 = \{\epsilon\}$, because there are no words over the empty alphabet except the empty word. Next, we define

$$\Sigma^\star := \bigcup_{n \geq 0} \Sigma^n. \tag{1.3}$$

as the set of all *words over the alphabet Σ*. It is now just natural to define a binary operation $\cdot$ on $\Sigma^\star$ which is given by concatenation of words:

$$w_1 \cdot w_2 := w_1 w_2, \quad w_1, w_2 \in \Sigma^\star.$$

It is obvious that this operation on $\Sigma^\star$ is associative and has the empty word ϵ as identity element. Therefore, according to Definition 1.1, the algebraic structure $\langle \Sigma^\star, \cdot, \epsilon \rangle$ forms a monoid.

Definition 1.10 (Free monoid)**.** Let Σ be an arbitrary set of symbols (called the alphabet), and $\Sigma^\star$ be defined as in (1.3). Moreover, denote concatenation of words by $\cdot$ and let ϵ be the empty word. Then the algebraic structure $\langle \Sigma^\star, \cdot, \epsilon \rangle$ is called the *free monoid over the alphabet Σ*.

1.1.3 Free groups

In a similar fashion, we can construct also a group from any given formal alphabet Σ. For this purpose we first choose a "disjoint copy" of Σ which we denote by Σ^{-1}. To be precise, Σ^{-1} may be any arbitrary set satisfying $\Sigma \cap \Sigma^{-1} = \emptyset$ and having same cardinality as Σ. In particular we could for example set $\Sigma^{-1} := \Sigma \times \{1\}$. Next, we choose a bijection $f : \Sigma \to \Sigma^{-1}$. We

can extend this bijection between the sets Σ and Σ^{-1} to an involution $\overline{f}$ living on the union $\overline{\Sigma} := \Sigma \cup \Sigma^{-1}$ by defining $\overline{f} := f \cup f^{-1}$ or more verbose:

$$\overline{f}(a) := \begin{cases} f(a) & \text{if } a \in \Sigma \\ f^{-1}(a) & \text{if } a \in \Sigma^{-1} \end{cases}$$

Now we introduce the notation $\overline{f}(a) =: a^{-1}$ for all $a \in \overline{\Sigma}$ and call a^{-1} the *(formal) inverse* of the element a.

Having now defined a formal inverse for each of our symbols in the given alphabet, we can go on and consider the free monoid $\overline{\Sigma}^\star$. If $\sigma_1, \sigma_2, \ldots, \sigma_n$ are symbols of $\overline{\Sigma}$, then we say the word $\sigma_1\sigma_2\ldots\sigma_n$ is *reduced*, if and only if no two subsequent symbols of the word are inverse to each other, that is

$$\sigma_j \neq \sigma_{j+1}^{-1} \quad \text{for all } 1 \leq j < n. \tag{1.4}$$

Clearly every word $w \in \overline{\Sigma}^\star$ can be brought into reduced form by successively "canceling out" adjacent inverse symbols until finally a word is obtained, which satisfies (1.4). We call the result of this procedure the *reduced form of* w. Additionally we define two words $w_1, w_2 \in \overline{\Sigma}^\star$ to be *equivalent* if and only if they have the same reduced form and we write $w_1 \sim w_2$ in this case. If we denote the reduced form of w by $\varphi(w)$, we can write

$$w_1 \sim w_2 \Leftrightarrow \varphi(w_1) = \varphi(w_2).$$

Moreover we observe that if $w_1 \sim w_2$ and $v_1 \sim v_2$, then we also have $w_1v_1 \sim w_2v_2$ because $\varphi(w_1) = \varphi(w_2) =: w$ and $\varphi(v_1) = \varphi(v_2) =: v$ implies

$$\varphi(w_1v_1) = \varphi(wv) = \varphi(w_2v_2).$$

Thus we see that the equivalence relation $\sim$ is compatible with the operation of word concatenation. Therefore we can consider also the set of equivalence classes $\overline{\Sigma}^\star/\sim$ as a monoid under the operation of word concatenation. Obviously the set of reduced words is a system of representatives for $\overline{\Sigma}^\star/\sim$ and we agree to denote an equivalence class $[w]_\sim$ simply by the reduced word w. $\overline{\Sigma}^\star/\sim$ is not just a monoid, but in fact a group, as the inverse of a word $\sigma_1\sigma_2\ldots\sigma_{n-1}\sigma_n$ is trivially given by $\sigma_n^{-1}\sigma_{n-1}^{-1}\ldots\sigma_2^{-1}\sigma_1^{-1}$. We call a reduced word of $\overline{\Sigma}^\star$ (resp. its corresponding equivalence class in $\overline{\Sigma}^\star/\sim$) a *group word over the alphabet* Σ.

Definition 1.11 (Free group)**.** Let Σ be an arbitrary set of formal symbols and define $\overline{\Sigma} := \Sigma \cup \Sigma^{-1}$ as above. On the free monoid $\overline{\Sigma}^\star$ define an equivalence relation $\sim$ by identifying words with the same reduced form. Then the algebraic structure

$$\Sigma_\sim := \langle \overline{\Sigma}^\star/\sim, \cdot, \epsilon \rangle \tag{1.5}$$

is called the *free group over the alphabet* Σ.

Remark 1.12. In the exceptional case when $\Sigma = \emptyset$, we obtain the trivial group by this construction: If $\Sigma = \emptyset$ then we can also choose $\Sigma^{-1} = \emptyset$ (Σ and Σ^{-1} are then disjoint as required). It follows that also $\overline{\Sigma}$ is empty and we end up with the trivial monoid $\overline{\Sigma}^\star = \{\epsilon\}$, which corresponds to the trivial group.

Example 1.13. Let Σ be the singleton set $\{a\}$. Now the free group over Σ consists precisely of the following group words:

$$\Sigma_\sim = \{\epsilon, a, aa, aaa, \ldots, a^{-1}, a^{-1}a^{-1}, \ldots\}.$$

If we now introduce the notation

$$\epsilon =: a^0, \quad \underbrace{aa\ldots a}_{k \text{ times}} =: a^k, \quad \underbrace{a^{-1}a^{-1}\ldots a^{-1}}_{k \text{ times}} =: a^{-k},$$

we have $a^n \cdot a^m = a^{n+m}$ for all $n, m \in \mathbb{Z}$ and it is thus evident that the free group over any one-element alphabet is isomorphic to the group $\langle \mathbb{Z}, +, 0\rangle$.

Note that this is the only case (besides the trivial one, when $\Sigma = \emptyset$), where the free group is commutative. In fact, the free group construction always yields a group with "richest possible" structure in the following sense:

Theorem 1.14. *Let Σ be an arbitrary set and G be an arbitrary group. If any mapping $\varphi : \Sigma \to G$ is given, then we can always extend φ in a unique way to a homomorphism $\overline{\varphi} : \Sigma_\sim \to G$.*

In fact, the property described in Theorem 1.14 completely characterizes the free group up to isomorphism, as stated in the next Theorem:

Theorem 1.15 (Universal mapping property)**.** *Let Σ be an arbitrary set and F be a group together with an injective map $\iota : \Sigma \to F$. If F has the property that for any map $\varphi : \Sigma \to G$, where G is an arbitrary group, there is a homomorphism $\overline{\varphi} : F \to G$ such that $\varphi = \overline{\varphi} \circ \iota$, then F is (isomorphic to) the free group over the alphabet Σ.*

1.1.4 Generators and relations

Definition 1.16. Let $\langle G, \cdot, 1\rangle$ be a group and $\Sigma \subseteq G$ a subset. We say G is *generated by* the elements of Σ, if every element $g \in G$ can be written as a product of elements from Σ in the following sense:

$$\forall g \in G\ \exists n \in \mathbb{N},\ (\sigma_j) \in \Sigma^n,\ (k_j) \in \mathbb{Z}^n : \quad g = \sigma_1^{k_1}\sigma_2^{k_2}\ldots\sigma_n^{k_n}. \tag{1.6}$$

Moreover in this case we call the elements of Σ *generators* of G.

Let us first consider the case when some $g \in G$ can be generated in two different ways, for example

$$g = \sigma_1^{k_1}\sigma_2^{k_2}\dots\sigma_n^{k_n} = \tau_1^{\ell_1}\tau_2^{\ell_2}\dots\tau_m^{\ell_m}$$

with $\sigma_j, \tau_j \in \Sigma$ and $k_j, \ell_j \in \mathbb{Z}$. This immediately gives

$$\sigma_1^{k_1}\sigma_2^{k_e}\dots\sigma_n^{k_n} \cdot \tau_m^{-\ell_m}\tau_{m-1}^{-\ell_{m-1}}\dots\tau_1^{-\ell_1} = 1. \tag{1.7}$$

We call the group word occurring on the left hand side of (1.7) a *relation* on G. The set of all such possible relations on G forms a normal subgroup of $\Sigma_\sim$. To see this, let $\varphi : \Sigma \to G$ be the canonical embedding. According to Theorem 1.14 there exists a unique extension of φ to a homomorphism $\overline{\varphi} : \Sigma_\sim \to G$. In fact this homomorphism just evaluates group words over Σ in the obvious way to concrete group elements in G. By definition, the set of all relations on G is now given by $N := \ker(\overline{\varphi})$ which is, by Lemma 1.8, a normal subgroup of $\Sigma_\sim$. It now follows from the first isomorphism theorem (Theorem 1.9) that $G \cong \Sigma_\sim/N$. In the case when the product representation in (1.6) is unique for all $g \in G$, then N just consists of the identity element and G is isomorphic to $\Sigma_\sim$. In this case G is said to be *relation-free*, which is where the terminology "free group" actually comes from.

Summing up, we see that every group can be described by supplying a set of generators Σ and a the set of relations $N \trianglelefteq \Sigma_\sim$. Of course it is also sufficient to supply just a subset R of relations from which the other relations can be derived. This leads to the following definition:

Definition 1.17 (Presentation of a group)**.** Let Σ be an arbitrary set and $R \subseteq \Sigma_\sim$ a set of group words over the alphabet Σ. A group G is said to be the *group defined by* the *generators* $\sigma \in \Sigma$ *and relations* R, if G is obtained by the following construction:

1. Let N be the normal subgroup of $\Sigma_\sim$ generated by the relations R:

$$N := \bigcap_{R \subseteq N' \trianglelefteq \Sigma_\sim} N'.$$

2. Define $G := \langle \Sigma \mid R \rangle := \Sigma_\sim/N$.

Finally $\langle \Sigma \mid R \rangle$ is called a *presentation* of G.

For easier notation we may omit braces when listing Σ and R in a presentation of G. Moreover we may write out relations more explicitly, for example instead of $G = \langle \{a\} \mid \{a^2\} \rangle$ we may write $G = \langle a \mid a^2 = 1 \rangle$.

Remark 1.18. The presentation $\langle \Sigma \mid R \rangle$ of a group G is by far not unique as the normal group N can be generated by many different subsets $R \subseteq N$. Even worse, the *word problem*, i.e. the question, if a given group word $w \in \Sigma_\sim$ is in N when R is given, is in general not decidable, which means that there is no general algorithm for deciding, if two given group words represent the same element of G.

Example 1.19. The group $\langle t, r \mid t^2 = r^3 = 1 \rangle$ consists of all group words of the form

$$r^{k_1} t r^{k_2} t \dots t r^{k_n}, \quad \text{with } k_1, k_n \in \{0, \pm 1\} \text{ and } k_2, \dots, k_{n-1} \in \{\pm 1\}.$$

We conclude this section with the statement, that among all groups, which satisfy a given set of relations R, the group constructed as above is in a certain sense the largest possible one.

Theorem 1.20. *Let $\langle \Sigma \mid R \rangle$ be a presentation of a group G. If H is any group generated by Σ and if H satisfies all relations R, then there is an epimorphism $G \to H$.*

1.1.5 Free products

Let $\langle G_i, \cdot, e_i \rangle_{i \in I}$ a family of pairwise disjoint groups, i.e. $G_i \cap G_j = \emptyset$ for $i \neq j$. We define the *free product* of the family of groups $(G_i)_{i \in I}$ by a construction in terms of generators and relations (see Definition 1.17) as follows: As the set of generators we choose the alphabet $\Sigma = \bigcup_{i \in I} G_i$ and as the set of relations we take all the relations coming from any of the groups G_i. Note that a priori the neutral elements $e_i \in G_i$ are all different symbols, but the final factorization by the normal subgroup N automatically identifies them with each other, as N includes all the e_i as trivial relations.

Definition 1.21 (Free product)**.** Let $(G_i)_{i \in I}$ be a family of disjoint groups. For every $i \in I$, let $\overline{\varphi}_i$ be the unique extension of the identity map on G_i to a homomorphism $G_{i\sim} \to G_i$ according to Theorem 1.14.[1] The *free product* of the family $(G_i)_{i \in I}$ is defined as

$$\prod_{i \in I}^{\star} G_i := \langle \Sigma \mid R \rangle, \quad \text{where } \Sigma = \bigcup_{i \in I} G_i \text{ and } R = \bigcup_{i \in I} \ker \overline{\varphi}_i.$$

If only a small finite number of groups is involved, for example only the two groups G and H, we will write $G * H$ for their free product.

[1] In other words, $\overline{\varphi}_i$ is the map which evaluates group words over the alphabet G_i to concrete elements of G_i in the obvious way.

Example 1.22. The free product of the groups $G = \langle t \mid t^2 = 1 \rangle$ and $H = \langle r \mid r^3 = 1 \rangle$ is the group $G * H = \langle t, r \mid t^2 = r^3 = 1 \rangle$ which we have already seen in Example 1.19.

Example 1.23. Let F_n and F_m be the free groups generated by n and m elements respectively. Then the free product $F_n * F_m = F_{n+m}$ is the free group generated by $n + m$ elements.

1.1.6 Basic linear groups

In this section we give the definitions of some important basic groups which will be needed throughout this document. We start with the general and special linear group.

Definition 1.24 (General linear group)**.** Let F be a field and $n > 0$. The group of invertible n-by-n matrices over F is called *general linear group* and is denoted by

$$\mathrm{GL}_n(F) := \{M \in F^{n\times n} \mid \det M \neq 0\}. \tag{1.8}$$

Definition 1.25 (Special linear group)**.** Let R be a commutative ring with 1 and $n > 0$. The group of n-by-n matrices over R having determinant 1, is called *special linear group* and is denoted by

$$\mathrm{SL}_n(R) := \{M \in R^{n\times n} \mid \det M = 1\}. \tag{1.9}$$

Of course in the case, when R is contained in the field F, then the special linear group $\mathrm{SL}_n(R)$ is a subgroup of the general linear group $\mathrm{GL}_n(F)$.

For both, the general and special linear group, one can construct the *projective groups* by identifying matrices which differ by a scalar multiple. For this purpose we define the *center* of a group G in the usual way:

$$\mathrm{Z}(G) := \{z \in G \mid zg = gz \quad \forall g \in G\}. \tag{1.10}$$

If G is a matrix group, $\mathrm{Z}(G)$ consists precisely of all scalar multiples of the identity matrix within G, which is exactly what we need for this construction.

Definition 1.26 (Projective linear groups)**.** As above, let F be a field, R a ring with 1 and $n > 0$. The *projective general linear group* and the *projective special linear group* are defined as

$$\mathrm{PGL}_n(F) := \mathrm{GL}_n(F)/\mathrm{Z}(\mathrm{GL}_n(F)), \tag{1.11}$$

$$\mathrm{PSL}_n(R) := \mathrm{SL}_n(R)\,/\mathrm{Z}(\mathrm{SL}_n(R)). \tag{1.12}$$

Example 1.27. Consider a matrix $M \in \mathrm{SL}_2(\mathbb{R})$. The equivalence class $[M]_\sim \in \mathrm{PSL}_2(\mathbb{R})$ consists precisely of the two matrices M and $-M$, whereas the equivalence class $[M]_\sim \in \mathrm{PGL}_2(\mathbb{R})$ is just the set $\{\lambda M \mid \lambda \in \mathbb{R} \setminus \{0\}\}$. For a clearer distinction of these equivalence classes let us introduce the notation

$$\begin{aligned} {}_{\pm}M &:= [M]_\sim \in \mathrm{PSL}_2(\mathbb{R}) \quad \text{and} \\ {}_{\lambda}M &:= [M]_\sim \in \mathrm{PGL}_2(\mathbb{R}). \end{aligned}$$

Clearly ${}_{\lambda}M = {}_{\lambda}(-M)$, which allows us to define an injective homomorphism ι which embeds $\mathrm{PSL}_2(\mathbb{R})$ into $\mathrm{PGL}_2(\mathbb{R})$:

$$\iota : \left\{ \begin{array}{rcl} \mathrm{PSL}_2(\mathbb{R}) & \to & \mathrm{PGL}_2(\mathbb{R}) \\ {}_{\pm}M & \mapsto & {}_{\lambda}M \end{array} \right.$$

Because $\det(\lambda M) = \lambda^2 \det(M) = \lambda^2 > 0$, the image of ι consists of equivalence classes of matrices with positive determinant only, whereas $\mathrm{PGL}_2(\mathbb{R})$ also contains equivalence classes of matrices with negative determinant. Therefore the embedding ι is not surjective and $\mathrm{PSL}_2(\mathbb{R})$ isomorphic to a proper subgroup of $\mathrm{PGL}_2(\mathbb{R})$:

$$\mathrm{PSL}_2(\mathbb{R}) \cong \iota(\mathrm{PSL}_2(\mathbb{R})) \lneq \mathrm{PGL}_2(\mathbb{R}).$$

In contrast to that, if we switch to $\mathrm{PSL}_2(\mathbb{C})$ and $\mathrm{PGL}_2(\mathbb{C})$ respectively and define ι analogously, we see that for every equivalence class $[M]_\sim \in \mathrm{PGL}_2(\mathbb{C})$, the matrix $M' := \frac{1}{\sqrt{\det M}} M$ is in $\mathrm{SL}_2(\mathbb{C})$ and satisfies $\iota([M']_\sim) = [M]_\sim$. Therefore ι is now also surjective and thus we have

$$\mathrm{PSL}_2(\mathbb{C}) \cong \mathrm{PGL}_2(\mathbb{C}).$$

More generally, for an arbitrary field F the groups $\mathrm{PSL}_n(F)$ and $\mathrm{PGL}_n(F)$ are isomorphic, if and only if F is closed under taking the n-th root, otherwise $\mathrm{PSL}_n(F)$ is isomorphic to a proper subgroup of $\mathrm{PGL}_n(F)$.

1.1.7 The action of a group on a set

Definition 1.28 (Group action)**.** Let $\langle G, \cdot, 1\rangle$ be a group and $\mathcal{S}$ be an arbitrary set. A function $G \times \mathcal{S} \to \mathcal{S}$, $(g, x) \mapsto gx$ is called a *group action* of G on the set $\mathcal{S}$ (alternatively we say G *acts on the set* $\mathcal{S}$), if and only if

$$\forall x \in \mathcal{S},\ g_1, g_2 \in G: \quad 1x = x \quad \text{and} \quad (g_1 g_2)x = g_1(g_2 x). \tag{1.13}$$

Theorem 1.29. *Let G be a group acting on the set $\mathcal{S}$.*

(i) The relation $\sim$ on $\mathcal{S}$ defined by

$$x_1 \sim x_2 \quad :\Leftrightarrow \quad \exists g \in G : x_1 = gx_2 \tag{1.14}$$

is an equivalence relation.

(ii) For all $x \in \mathcal{S}$ the set $G_x := \{g \in G \mid gx = x\}$ forms a subgroup of G.

Definition 1.30 (Orbit, stabilizer)**.** For $x \in \mathcal{S}$, the set

$$Gx := \{gx \in \mathcal{S} \mid g \in G\},$$

which is identical to the equivalence class $[x]_\sim \in \mathcal{S}/\sim$ of the relation $\sim$ defined in (1.14), is called the *orbit* of x under G. The group

$$G_x := \{g \in G \mid gx = x\} \leq G$$

is called the *stabilizer* of x.

Theorem 1.31. *The action of a group G on a set $\mathcal{S}$ induces a homomorphism from G to the group of permutations on $\mathcal{S}$ which is given by $g \mapsto (x \mapsto gx)$.*

Corollary 1.32 (Cayley)**.** *The action of a group G on itself induces a monomorphism from G to the group of permutations of its elements. Hence every group is isomorphic to a permutation group. In particular every finite group G is isomorphic to a subgroup of the symmetric group S_n with $n = |G|$.*

1.2 Möbius transformations

In the following we define the group of Möbius transformations and collect some useful basic properties.

Definition 1.33. A non-constant rational function $\varphi \in \mathbb{C}(z)$ of the form

$$\varphi(z) = \frac{az + b}{cz + d}, \quad a, b, c, d \in \mathbb{C}, \quad ad - bc \neq 0$$

is called *Möbius transformation.*

Remark 1.34. The condition $ad - bc \neq 0$ just ensures that φ is in fact non-constant.

Theorem 1.35. *The set of Möbius transformations forms a group under the action of function composition and can be identified with the projective general linear group* $\mathrm{PGL}_2(\mathbb{C})$ *or the projective special linear group* $\mathrm{PSL}_2(\mathbb{C})$.

Proof. Let φ and ψ be Möbius transformations with

$$\varphi(z) = \frac{az+b}{cz+d}, \quad \psi(z) = \frac{ez+f}{gz+h}.$$

First we make the trivial observation that composing those two transformations again yields a rational function of the desired form:

$$\varphi \circ \psi(z) = \frac{a\frac{ez+f}{gz+h} + b}{c\frac{ez+f}{gz+h} + d} = \frac{aez + af + bgz + bh}{cez + cf + dgz + dh} = \frac{(ae+bg)z + (af+bh)}{(ce+dg)z + (cf+dh)} \tag{1.15}$$

Having a closer look on the resulting coefficients one might notice that they relate to the following matrix product:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix} \tag{1.16}$$

This motivates the definition of a mapping π between matrices in $\mathrm{GL}_2(\mathbb{C})$ and Möbius transformations:

$$\pi : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \frac{az+b}{cz+d} \tag{1.17}$$

Note that the domain of π is $\mathrm{GL}_2(\mathbb{C})$, i.e. the set of 2-by-2 matrices with nonzero determinant. This is perfectly consistent with the condition $ad-bc \neq 0$ we have for Möbius transformations. For this reason π is a well-defined function from $\mathrm{GL}_2(\mathbb{C})$ to the set of Möbius transformations.

But π is not just a function, it is in fact a homomorphism as we see from (1.15) and (1.16). Trivially π is also surjective, which carries over the group structure of $\mathrm{GL}_2(\mathbb{C})$ to the set of Möbius transformations. The kernel of π comprises of all multiples of the identity matrix. Therefore, by the first isomorphism theorem (Theorem 1.9), the group of Möbius transformations is isomorphic to $\mathrm{GL}_2(\mathbb{C})/\ker\pi \cong \mathrm{PGL}_2(\mathbb{C})$ and we have seen in Example 1.27 that $\mathrm{PGL}_2(\mathbb{C}) \cong \mathrm{PSL}_2(\mathbb{C})$. □

Remark 1.36. We note that the nature of Möbius transformations is threefold: Firstly, as in Definition 1.33, we can regard a Möbius transformation φ as purely algebraic object, namely as rational function, i.e. the (formal) quotient of two polynomials in $\mathbb{C}[z]$:

$$\varphi_{\mathrm{alg}} = \frac{az+b}{cz+d} \in \mathbb{C}(z).$$

Secondly φ has a natural interpretation as meromorphic function on the extended complex plane $\mathbb{C}_\infty = \mathbb{C} \cup \{\infty\}$ in the sense of complex analysis:

$$\varphi_{\text{fun}} : \begin{cases} \mathbb{C}_\infty & \to & \mathbb{C}_\infty \\ z & \mapsto & \frac{az+b}{cz+d}. \end{cases}$$

In a more formal context, this correspondence can also be seen the following light: The group of Möbius transformations acts on the set $\mathbb{C}_\infty$ in the sense of Defintion 1.28 by $\varphi_{\text{alg}} z := \varphi_{\text{fun}}(z)$. Now, the homomorphism of Theorem 1.31 is in fact an isomorphism which assigns each Möbius transformation φ_{alg} a permutation of the set $\mathbb{C}_\infty$ which is exactly the meromorphic function φ_{fun}. Last but not least we have seen in Theorem 1.35 that we can also regard φ as equivalence class of matrices:

$$\varphi_{\text{lin}} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}_\sim \in \mathrm{PGL}_2(\mathbb{C}).$$

Whenever there is no danger of confusion, we will from now on switch between these different views on Möbius transformations seamlessly and exploit concepts of algebra, function theory and linear algebra alternately.

Lemma 1.37. *The group of Möbius transformations is generated by the following basic types of transformations:*

- *Translations:* $z \mapsto z + \alpha$ $\alpha \in \mathbb{C}$
- *Dilations:* $z \mapsto \rho z$ $\rho > 0$
- *Rotations:* $z \mapsto e^{i\theta} z$ $\theta \in (-\pi, \pi]$
- *Inversion:* $z \mapsto \frac{1}{z}$

Proof. Let $\varphi(z) = \frac{az+b}{cz+d}$ be an arbitrary Möbius transformation. In the case when $c = 0$, we may further assume w.l.o.g. that $d = 1$ such that the transformation simply writes $\varphi(z) = az + b$. Obviously this is dilation and rotation by the factor a followed by translation by b.

Let's consider the more interesting case when $c \neq 0$. Without restriction we may assume that $c = 1$, such that

$$\varphi(z) = \frac{az+b}{z+d} = a + \frac{b-ad}{z+d}.$$

Also in this case it is easy to see that φ is composed of translation by d, inversion, dilation and rotation by the factor $b - ad$ and a final translation by a. □

Now that we have defined the group of Möbius transformations, it is worth to get a better geometric intuition about how these maps act on the complex

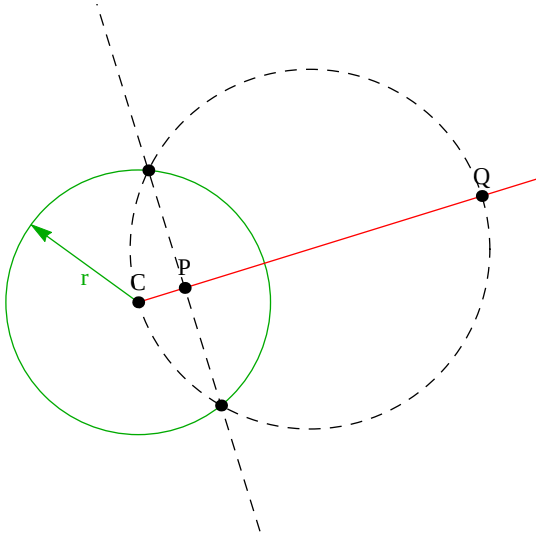

Figure 1.1: Circle inversion with respect to the green reference circle. For a point P within the reference circle, the inverse Q is constructed by first drawing a ray from C through P (red). The line perpendicular to this ray through the point P intersects the reference circle in two points. These two points together with C determine a circle (dashed) which intersects the ray in the point Q. The distances CP and CQ satisfy the relation $CP \cdot CQ = r^2$.

plane. Lemma 1.37 gives a first insight, as translations, dilations and rotations are quite easy to understand. Also the map $z \mapsto \frac{1}{z}$ has a geometric interpretation, namely as circle inversion followed by a reflection.

In 2-dimensional geometry, *circle inversion* with respect to a reference circle with center C and a radius r takes each point P on the plane to a point Q lying on the ray from C through P. Its distance from C is determined by $CP \cdot CQ = r^2$. The image of C is defined to be the point at infinity (and vice versa). Roughly speaking, the inversion turns the circle "inside out", i.e. points inside the reference circle are bijectively mapped to points outside while rays from the circle center are invariant under the circle inversion – see also Figure 1.1. A short introduction to circle inversion can be found in Mumford [9], p. 54ff. For a more comprehensive treatment see Schwerdtfeger [15].

Coming back to the concrete map $z \mapsto \frac{1}{z}$, it can now be interpreted the following way: Circle inversion with regard to the unit circle maps each $z \in \mathbb{C}$ to $\frac{z}{|z|^2} = \frac{1}{\bar{z}}$. Then reflection across the real axis (i.e. complex conjugation) takes $\frac{1}{\bar{z}}$ to $\frac{1}{z}$.

Summing up, all the basic types of Möbius transformations mentioned in Lemma 1.37 have a very direct geometric interpretation. Still, arbitrary

Möbius transformations (especially those involving at least one inversion) are hard to describe in a similar geometric and intuitive way. Fortunately there is another characterization of Möbius transformations which is both, elegant and visually accessible.

1.2.1 Stereographic projection

This section is about the great work of Douglas Arnold and Jonathan Rogness, "Möbius transformations revealed" [1], in which the authors give a characterization of Möbius transformations in terms of stereographic projections and rigid motions of spheres in 3D-space.[2]

In order to introduce stereographic projection, we first have to embed $\mathbb{C}$ into $\mathbb{R}^3$. We do so by using the map

$$\iota : \left\{ \begin{array}{ccc} \mathbb{C} & \to & \mathbb{R}^3 \\ z & \mapsto & (\mathrm{Re}\,(z), \mathrm{Im}\,(z), 0) \end{array} \right. , \tag{1.18}$$

which means that we identify the complex plane with the plane $x_3 = 0$ in $\mathbb{R}^3$. Additionally we equip $\mathbb{R}^3$ with the standard Euclidean norm:

$$\|x\|_2 := \sqrt{x_1^2 + x_2^2 + x_3^2}, \quad x \in \mathbb{R}^3. \tag{1.19}$$

Definition 1.38 (Admissible sphere)**.** A *sphere* with center $c \in \mathbb{R}^3$ and radius $r > 0$ is the set $S := \{x \in \mathbb{R}^3 \mid \|x - c\|_2 = r\}$. Its *north-pole* is the unique point $n \in S$ with maximal x_3-coordinate. A sphere whose north pole lies in the upper half-space $H := \{x \in \mathbb{R}^3 \mid x_3 > 0\}$ is called an *admissible sphere*.

Definition 1.39 (Stereographic projection)**.** Let S be an admissible sphere and $n \in S$ be its north pole. If $x \in S \setminus \{n\}$, then denote by $L_{x,n}$ the unique line joining x with n and let y be the intersection point of $L_{x,n}$ with the plane $\iota(\mathbb{C})$. Now, *stereographic projection* with regard to S is the map $P_S : S \to \mathbb{C}_\infty$ defined as $P_S(x) := \iota^{-1}(y)$ for $x \neq n$ and $P_S(n) := \infty$.

Remark 1.40. Stereographic projection with regard to any admissible sphere $S \subseteq \mathbb{R}^3$ maps S bijectively to $\mathbb{C}_\infty$. Of course the most natural choice for S in this context is the unit sphere $S_1 := \{x \in \mathbb{R}^3 \mid \|x\|_2 = 1\}$, having the advantage that stereographic projection follows the simple formula

$$P_{S_1} : \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \mapsto \frac{x_1 + i x_2}{1 - x_3}. \tag{1.20}$$

[2] See also http://www.ima.umn.edu/∼arnold/moebius

Also reverse stereographic projection can be done easily using the unit sphere:

$$P_{S_1}^{-1} : z \mapsto \frac{1}{|z|^2 + 1} \begin{pmatrix} 2\,\mathrm{Re}\,(z) \\ 2\,\mathrm{Im}\,(z) \\ |z|^2 - 1 \end{pmatrix}. \tag{1.21}$$

By pointwise identifying $\mathbb{C}_\infty$ with S_1 using the above two mappings, we obtain the *Riemann sphere* model of the extended complex plane. One of its benefits is that certain functions $\mathbb{C}_\infty \to \mathbb{C}_\infty$ can be interpreted nicely as simple rotations of the Riemann sphere.

Example 1.41. Consider the map $f : z \mapsto \frac{1}{z}$ and the map $V : S_1 \to S_1$, which rotates the Riemann sphere around the x_1 axis by 180°:

$$V : x \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}. \tag{1.22}$$

The first row of Figure 1.2 shows the upper half-plane U on the left and the unit disk D on the right together with their images under reverse stereographic projection. We see that both, U and D, correspond to a certain "halfsphere" of the Riemann sphere. If we rotate these halfspheres around the x_1 axis (leaving the points $\{\pm 1\}$ fixed) and continuously perform stereographic projection, we see that after a half turn (in the last row of Figure 1.2) we end up with the lower half-plane $f(U)$ and the set of points z with $|z| > 1$, $f(D)$.

It is worth noting that this correspondence between rotation of the Riemann sphere by 180° and the map $f : z \mapsto \frac{1}{z}$ does not just hold for the specially chosen sets U and D, but indeed pointwise for every $z \in \mathbb{C}_\infty$. In other words, we have

$$f = P_{S_1} \circ V \circ P_{S_1}^{-1}.$$

We show this by a simple calculation, using the fact that $\frac{1}{z} = \frac{\overline{z}}{|z|^2}$ as well as (1.21) and noting that for the case $z = \infty$ limits have to be introduced appropriately:

$$\begin{aligned}(P_{S_1}^{-1} \circ f)(z) = \frac{1}{\frac{1}{|z|^2} + 1} \begin{pmatrix} 2\,\mathrm{Re}\left(\frac{1}{z}\right) \\ 2\,\mathrm{Im}\left(\frac{1}{z}\right) \\ \frac{1}{|z|^2} - 1 \end{pmatrix} &= \frac{|z|^2}{1 + |z|^2} \begin{pmatrix} 2\frac{1}{|z|^2}\,\mathrm{Re}\,(\overline{z}) \\ 2\frac{1}{|z|^2}\,\mathrm{Im}\,(\overline{z}) \\ \frac{1-|z|^2}{|z|^2} \end{pmatrix} \\ &= \frac{1}{|z|^2 + 1} \begin{pmatrix} 2\,\mathrm{Re}\,(z) \\ -2\,\mathrm{Im}\,(z) \\ -|z|^2 + 1 \end{pmatrix} = (V \circ P_{S_1}^{-1})(z).\end{aligned}$$

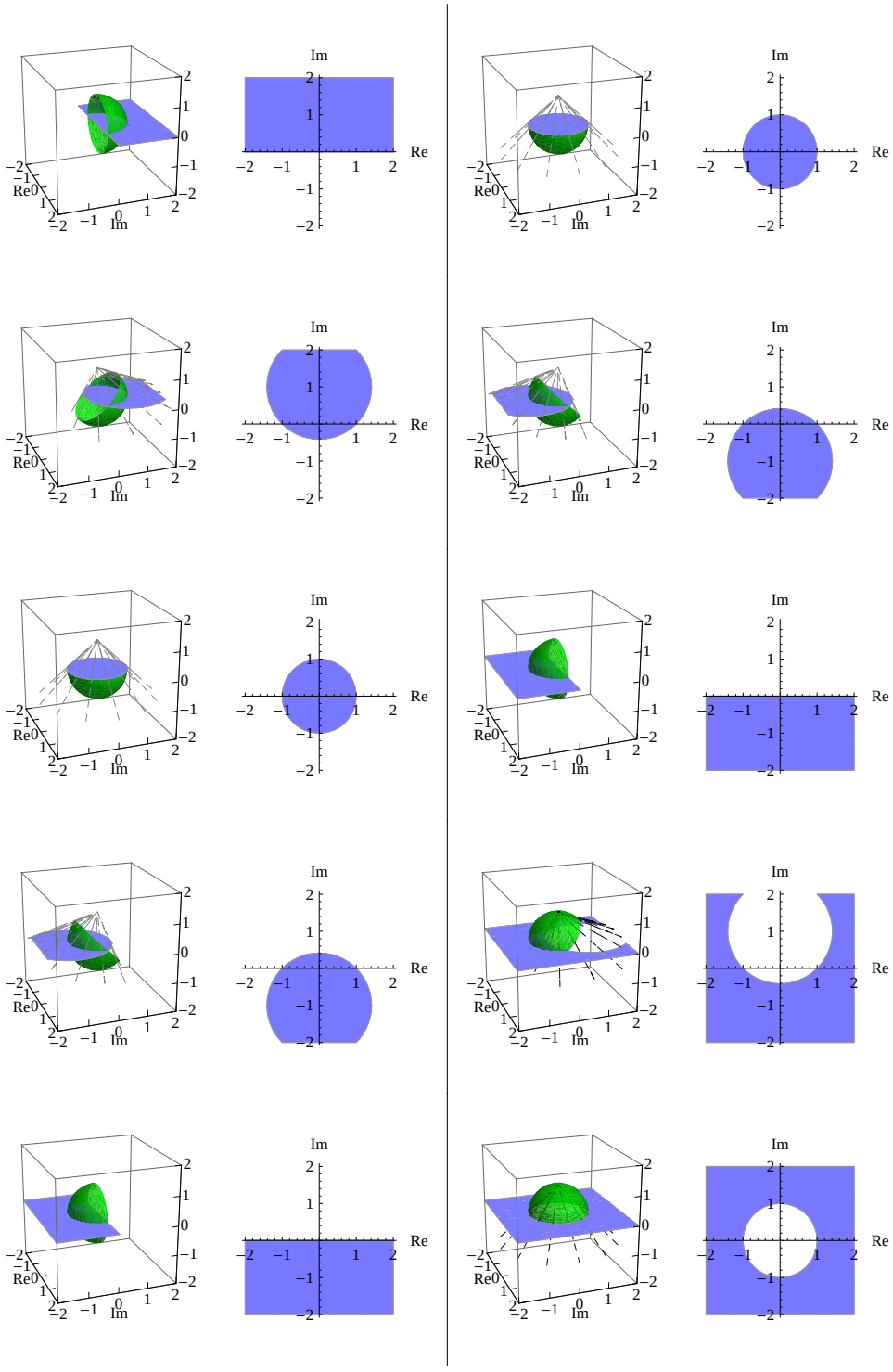

Figure 1.2: Inversion $z \mapsto \frac{1}{z}$ can be interpreted as rotation of the Riemann sphere by 180° around the x_1 axis. It maps the upper to the lower half-plane (left) and turns the unit disk inside out (right).

Remark 1.42. Other examples for transformations corresponding to a half-turn of the Riemann sphere are $z \mapsto -z$ (rotation around the x_3 axis) and $T : z \mapsto -\frac{1}{z}$ (rotation around the x_2 axis; as product of half turns around the x_1 and x_3 axes). T is a function which maps the upper half-plane to itself and which will be important in the study of the modular group in Section 2.1.

Example 1.43 (Modified Cayley transform)**.** In Figure 1.3, the action of yet another interesting transformation can be seen, which maps the upper half-plane to the unit disk. It is given by

$$\Phi(z) = \frac{iz+1}{z+i} \tag{1.23}$$

and can be either considered as a quarter turn of the Riemann sphere around the x_1 axis or as "half of an inversion", since indeed $\Phi^2(z) = \Phi(\Phi(z)) = \frac{1}{z}$ (compare with left column of Figure 1.2). In contrast to its better known brother, the *Cayley transform* given by

$$\Psi(z) = \frac{z-i}{z+i} = -i\Phi(z), \tag{1.24}$$

Φ leaves the points $\{\pm 1\}$ fixed, which often beneficial for visualization purposes. We will call Φ the *modified Cayley transform*.

Remark 1.44. Also the Cayley transform Ψ can be seen as rotation of the Riemann sphere by 120° around the axis which is spanned by the vector $(1, -1, 1) \in \mathbb{R}^3$. A figure illustrating this fact can be found for example in Mumford [9], p. 88.

We have now seen exemplarily that quite a few interesting maps are induced by rotations of the Riemann sphere. If we are willing to allow not only rotations but also translations of the Riemann sphere, we indeed obtain a new characterization of Möbius transformations, as stated in the next theorem.

Definition 1.45. A *rigid motion* of $\mathbb{R}^3$ is an affine transformation of $\mathbb{R}^3$ which is obtained purely by composition of rotations and translations.

Theorem 1.46 (Möbius transformations revealed)**.** *A function $\varphi : \mathbb{C}_\infty \to \mathbb{C}_\infty$ is a Möbius transformation, if and only if it can be obtained by reverse stereographic projection of $\mathbb{C}_\infty$ to an admissible sphere $S \subseteq \mathbb{R}^3$, followed by a rigid motion T of $\mathbb{R}^3$ which maps S to another admissible sphere TS, followed by stereographic projection from TS back to $\mathbb{C}_\infty$:*

$$\varphi = P_{TS} \circ T \circ P_S^{-1}. \tag{1.25}$$

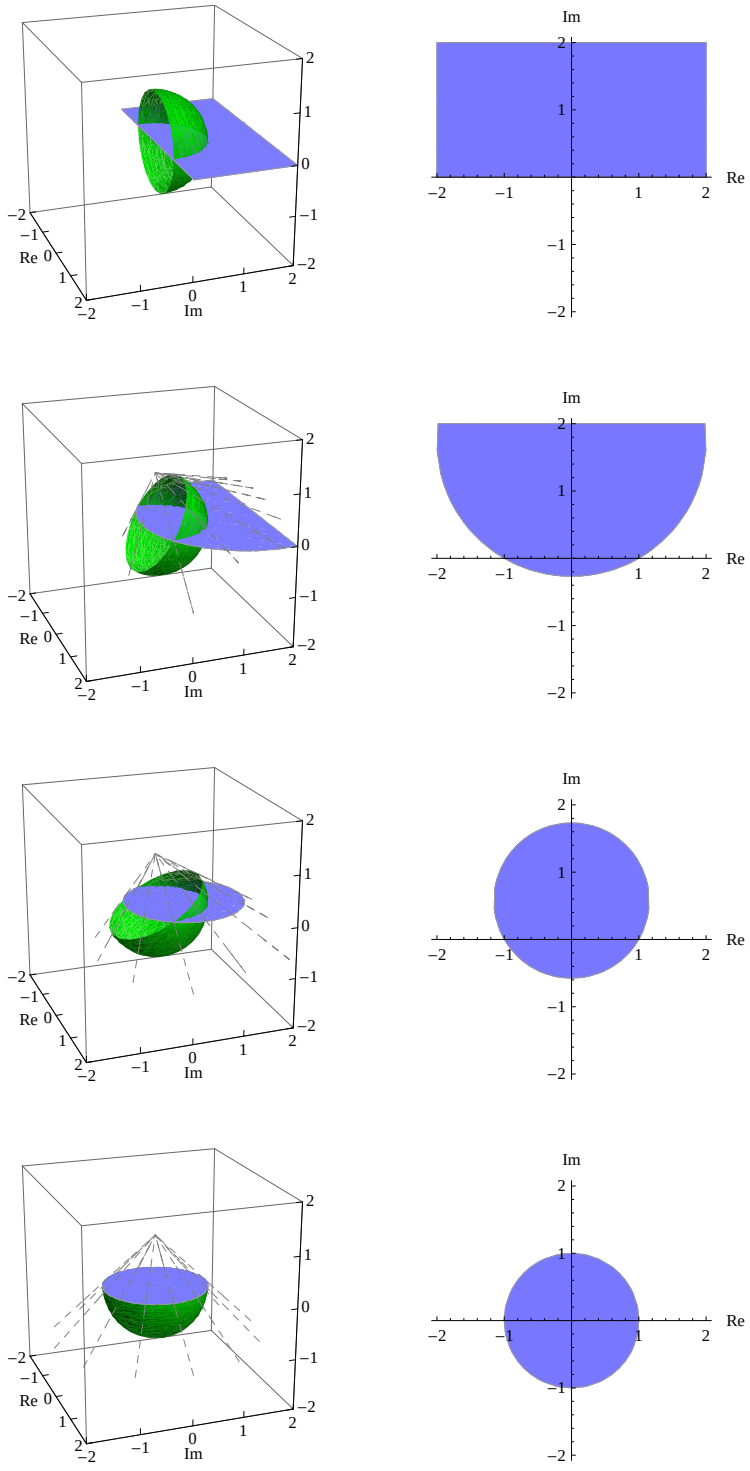

Figure 1.3: The modified Cayley transform $z \mapsto \frac{iz+1}{z+i}$ maps the upper half-plane to the unit disk, leaving the points $\{\pm 1\}$ fixed. It can be considered as a quarter turn of the Riemann sphere around the x_1 axis.

Sketch of proof. The fact that all maps of form (1.25) are indeed Möbius transformations can be seen either by direct calculation or by the observation that φ corresponds to the map $P_S^{-1} \circ \varphi \circ P_S = P_S^{-1} \circ P_{TS} \circ T$ from S to itself. If we identify S with the Riemann sphere, we can regard φ as a holomorphic automorphism of the Riemann sphere which is therefore a Möbius transformation.

It remains to show that every given transformation can indeed be realized in the form (1.25). For this purpose, we first consider the four basic types of Möbius transformations:

Translation: The map $z \mapsto z + \alpha,\ \alpha \in \mathbb{C}$ can be realized by choosing an arbitrary admissible sphere S and setting $T = T_\alpha : x \mapsto x + \iota(\alpha)$, which simply translates S in a direction parallel to the $x_3 = 0$ plane by $\iota(\alpha)$.

Dilation: The map $z \mapsto \rho z,\ \rho > 0$ can be obtained by choosing an arbitrary admissible sphere with north pole n and setting $T = D_\rho : x \mapsto x + (0, 0, (\rho - 1)n_3)$, which moves S up- ($\rho > 1$) or downwards ($\rho < 1$) in x_3 direction.

Rotation: The map $z \mapsto e^{i\theta} z,\ \theta \in (-\pi, \pi]$ can be realized by choosing an arbitrary admissible sphere and setting

$$T = R_\theta : x \mapsto \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

which rotates S around the x_3 axis by an angle of θ.

Inversion: We have seen in Example 1.41 that the map $z \mapsto \frac{1}{z}$ can be realized by choosing S as the unit sphere S_1 and setting $T = V$ – as defined in (1.22) – which is a rotation of S_1 around the x_1 axis by an angle of 180°.

Now let $\varphi(z) = \frac{az+b}{cz+d}$ be an arbitrary Möbius transformation. As in the proof of Lemma 1.37, if $c = 0$ we may assume without restriction that $d = 1$ and therefore $\varphi(z) = az + b$. Clearly this map can be realized in the form (1.25) by starting with an arbitrary admissible sphere S and taking T as the composition of rotation by $\arg(a)$, dilation by $|a|$ (in either order), followed by translation by b:

$$T := T_b \circ D_{|a|} \circ R_{\arg(a)}.$$

If $c \neq 0$, we can scale the coefficients a, b, c, d such that $c = 1$ and write φ in the form

$$\varphi(z) = a + \frac{b - ad}{z + d}.$$

Now we choose S to be the sphere with unit radius centered at the point $\iota(-d)$ and we compose T out of the following rigid motions: First, translation by d transforms S to the unit sphere S_1. Thus we can indeed apply V as the second transformation in order to perform an inversion. Finally we apply rotation and dilation by the factor $b - ad$ followed by a translation by a:

$$T := T_a \circ D_{|b-ad|} \circ R_{\arg(b-ad)} \circ V \circ T_d. \qquad \square$$

1.2.2 Generalized circles

From the geometric point of view Möbius transformations have the beautiful property that they preserve generalized circles. *Generalized circles* are either circles (in the usual sense) or lines on the complex plane $\mathbb{C}$. They can also be thought of circles on the Riemann sphere (i.e. the extended complex plane $\mathbb{C}_\infty$ projected to the unit sphere S_1, see Remark 1.40), where lines on the complex plane stand in a one-to-one correspondence to circles through the point ∞ on the Riemann sphere. In order to give an exact definition, we follow an idea taken from Schwerdtfeger [15] and make the following considerations:

A circle with center $m \in \mathbb{C}$ and radius $r > 0$ can be described as the set of points $z \in \mathbb{C}$ for which

$$|z - m| = r.$$

This is obviously equivalent to

$$|z - m|^2 = (z - m)\overline{(z - m)} = r^2$$

and

$$z\overline{z} - m\overline{z} - \overline{m}z + m\overline{m} - r^2 = 0. \tag{1.26}$$

The generalization comes into play if we multiply this last equation by a constant $A \in \mathbb{R}$

$$Az\overline{z} - Am\overline{z} - A\overline{m}z + Am\overline{m} - Ar^2 = 0$$

and introduce constants B, C and D appropriately such that we can write it in the form

$$Az\overline{z} + B\overline{z} + Cz + D = 0. \tag{1.27}$$

Note that D is real and $B = \overline{C}$ are complex conjugates. From an equation of form (1.27) we can read off the center and radius of the corresponding circle by

$$m = -\frac{B}{A}, \tag{1.27a}$$

$$r = \sqrt{m\overline{m} - \frac{D}{A}} = \sqrt{\frac{BC - AD}{A^2}}. \tag{1.27b}$$

Clearly we can only do so, if $A \neq 0$ and $BC - AD > 0$.

In the case when $A = 0$, equation (1.27) can be written as

$$\operatorname{Re}\left(\frac{C}{|C|}z\right) = -\frac{D}{2\,|C|},$$

which defines a line on the complex plane. We see this by considering the simpler equation $\operatorname{Re}(z) = -\frac{D}{2|C|}$ first (we omit the factor $\frac{C}{|C|}$), which obviously defines a line parallel to the imaginary axis through the real point $-\frac{D}{2|C|}$. Then we observe that the multiplication with $\frac{C}{|C|}$ just rotates this line around the origin by an angle which is given by $-\arg(C) = \arg(B)$.

Note that equation (1.27) can also be written in matrix form:

$$\begin{pmatrix}\overline{z} & 1\end{pmatrix} \cdot \begin{pmatrix}A & B\\ C & D\end{pmatrix} \cdot \begin{pmatrix}z\\ 1\end{pmatrix} = 0.$$

If we substitute $z = u/v$, with $u, v \in \mathbb{C}$, $v \neq 0$ and scale by $\overline{v} \cdot v = |v|^2 > 0$, we obtain the equivalent equation

$$\begin{pmatrix}\overline{u} & \overline{v}\end{pmatrix} \cdot \begin{pmatrix}A & B\\ C & D\end{pmatrix} \cdot \begin{pmatrix}u\\ v\end{pmatrix} = 0. \tag{1.28}$$

By introducing the convention to identify $\infty \in \mathbb{C}_\infty$ with the formal quotient $u/0$, for arbitrary $u \in \mathbb{C} \setminus \{0\}$, equation (1.28) makes sense for all $z = u/v \in \mathbb{C}_\infty$.

Finally we emphasize that the matrix in equation (1.28) has a negative determinant, because of the condition $BC - AD > 0$ from above. Moreover it is a Hermitian matrix – a notion which we will shortly recall:

Definition 1.47 (Hermitian matrix)**.** Let $n > 0$ and $M \in \mathbb{C}^{n\times n}$. The matrix

$$M^{\mathrm{H}} := \overline{M}^{\mathrm{T}}$$

obtained by complex conjugation and transposition of M is called *Hermitian transpose* or *conjugate transpose* of M. If M has the property $M^{\mathrm{H}} = M$, it is called a *Hermitian matrix*.

Having now the right vocabulary and properties at hand, we can give an exact definition for generalized circles.

Definition 1.48 (Generalized circle)**.** Let $M \in \mathbb{C}^{2\times 2}$ be a Hermitian matrix with $\det(M) < 0$. A *generalized circle*, for short *g-circle*, is the set of solutions $u/v \in \mathbb{C}_\infty$ with $u, v \in \mathbb{C}$, not both zero, to

$$\begin{pmatrix}\overline{u} & \overline{v}\end{pmatrix} \cdot M \cdot \begin{pmatrix}u\\ v\end{pmatrix} = 0. \tag{1.29}$$

Since there should be no danger of confusion, we will from now on use the same name for a generalized circle and its corresponding Hermitian matrix (which is uniquely determined up to a nonzero real scalar factor).

Remark 1.49. Definition 1.48 does not depend on the choice of u and v. If u/v is a solution to (1.29) and $u'/v' = u/v$, i.e. $u' = \lambda u$ and $v' = \lambda v$ for some $\lambda \in \mathbb{C} \setminus \{0\}$, then also

$$\begin{pmatrix} \overline{u'} & \overline{v'} \end{pmatrix} \cdot M \cdot \begin{pmatrix} u' \\ v' \end{pmatrix} = |\lambda|^2 \begin{pmatrix} \overline{u} & \overline{v} \end{pmatrix} \cdot M \cdot \begin{pmatrix} u \\ v \end{pmatrix} = 0.$$

Moreover we see that the point $\infty = 1/0$ lies on the g-circle $M = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ exactly when its left upper matrix entry A is zero. This is consistent with stereographic projection: The matrix entry A is zero if and only if M corresponds to a line on the complex plane. The image of this line under reverse stereographic projection is a circle on the Riemann sphere going through its north-pole, which we have identified with the point ∞.

Going back to the our starting point, the equation $|z - m| = r$, we can replace the equality sign '=' with '$<$' or '$\leq$' and repeat the above considerations without any additional changes. This naturally leads to the notions of open and closed generalized disks.

Definition 1.50 (Generalized disk)**.** Let $M \in \mathbb{C}^{2\times 2}$ be a Hermitian matrix with $\det(M) < 0$. An *open generalized disk* is the set of solutions $u/v \in \mathbb{C}_\infty$, with $u, v \in \mathbb{C}$ – not both zero, to

$$\begin{pmatrix} \overline{u} & \overline{v} \end{pmatrix} \cdot M \cdot \begin{pmatrix} u \\ v \end{pmatrix} < 0 \tag{1.30}$$

and a *closed generalized disk* is the set of solutions $u/v \in \mathbb{C}_\infty$ to

$$\begin{pmatrix} \overline{u} & \overline{v} \end{pmatrix} \cdot M \cdot \begin{pmatrix} u \\ v \end{pmatrix} \leq 0. \tag{1.31}$$

In both cases we will use the term (open/closed) *g-disk* for shortness.

Remark 1.51. In contrast to generalized circles, the defining matrix $M = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ of a generalized disk is unique up to a *positive* real scalar factor. Switching from M to $-M$ turns the g-disk inside out while its border (the g-circle M) is left intact. Dependent on the sign of the left upper matrix entry A, we can distinguish three types of g-disks:

Case $A > 0$**:** The g-disk corresponds to a disk in the usual sense within $\mathbb{C}$. Its center is given by (1.27a) and its radius by (1.27b).

Case $A = 0$: The g-disk corresponds to a half-plane of $\mathbb{C}$, which is obtained from the left half-plane, i.e. $\mathrm{Re}\,(z) < 0$ (resp. ≤ 0), by translation by $\frac{-D}{2|C|}$ followed by a rotation around the origin by $\arg(B)$. The point ∞ is member of the generalized disk, if and only if it is closed.

Case $A < 0$: The open g-disk M corresponds to the set complement (within $\mathbb{C}_\infty$) of the closed g-disk $-M$ (discussed in the first case, $A > 0$). Accordingly, the closed g-disk M is the complement of the open g-disk $-M$. Open and closed g-disks with $A < 0$ contain the point ∞.

It is now easy to show that g-circles and g-disks are preserved under Möbius transformations.

Theorem 1.52. *Let φ be a Möbius transformation and $M \in \mathbb{C}^{2\times 2}$ be a Hermitian matrix with $\det(M) < 0$. The image of the g-circle (resp. open/closed g-disk) M under the Möbius transformation φ corresponding to the matrix $P \in \mathrm{GL}_2(\mathbb{C})$ is the g-circle (resp. open/closed g-disk)*

$$(P^{-1})^H \cdot M \cdot P^{-1}. \tag{1.32}$$

Proof. Let us write $P = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, such that the corresponding Möbius transformation φ has the form

$$\varphi\left(\frac{u}{v}\right) = \frac{au + bv}{cu + dv}.$$

Define $u' := au + bv$ and $v' := cu + dv$. We need to show that

$$\begin{pmatrix} \overline{u} & \overline{v} \end{pmatrix} \cdot M \cdot \begin{pmatrix} u \\ v \end{pmatrix} \begin{cases} = 0 \\ < 0 \\ \leq 0 \end{cases}$$

if and only if

$$\begin{pmatrix} \overline{u'} & \overline{v'} \end{pmatrix} \cdot (P^{-1})^{\mathrm{H}} \cdot M \cdot P^{-1} \cdot \begin{pmatrix} u' \\ v' \end{pmatrix} \begin{cases} = 0 \\ < 0 \\ \leq 0. \end{cases}$$

But this follows immediately from

$$P \cdot \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} au + bv \\ cu + dv \end{pmatrix} = \begin{pmatrix} u' \\ v' \end{pmatrix}. \qquad \square$$

Chapter 2

The modular group

In this chapter and also later, we adopt the notation of Schoeneberg [14].

Definition 2.1 (Modular transformation)**.** A Möbius transformation A of the form

$$A(z) = \frac{az+b}{cz+d}, \quad a,b,c,d \in \mathbb{Z}, \quad ad-bc = 1$$

is called *(inhomogeneous) modular transformation.*

Theorem 2.2 (Modular group)**.** *The set of modular transformations forms a subgroup of the group of Möbius transformations and can be identified with the projective special linear group* $\mathrm{PSL}_2(\mathbb{Z})$*. This group is called the* modular group*.*

Proof. The proof is very similar to that of Theorem 1.35. The only thing which has to be changed is the homomorphism π defined in (1.17). Its domain now is $\mathrm{SL}_2(\mathbb{Z})$, the group of 2-by-2 matrices over $\mathbb{Z}$ with determinant 1, rather than $\mathrm{GL}_2(\mathbb{C})$ (or $\mathrm{SL}_2(\mathbb{C})$). It follows by the first isomorphism theorem (Theorem 1.9), that the modular group is isomorphic to $\mathrm{SL}_2(\mathbb{Z})/\ker(\pi) \cong \mathrm{PSL}_2(\mathbb{Z})$. The fact that the modular group is a subgroup of the group of Möbius transformations is now also evident, since $\mathrm{PSL}_2(\mathbb{Z}) \leq \mathrm{PSL}_2(\mathbb{C}) \cong \mathrm{PGL}_2(\mathbb{C})$ (see Example 1.27). □

Remark 2.3. The elements of $\mathrm{SL}_2(\mathbb{Z})$ are often called *homogeneous modular transformations*, whereas the transformations of $\mathrm{PSL}_2(\mathbb{Z})$ are called *inhomogeneous transformations*. Strictly seen, an inhomogeneous transformation has to be denoted as $[M]_\sim$, which is the equivalence class in $\mathrm{PSL}_2(\mathbb{Z})$ of a matrix $M \in \mathrm{SL}_2(\mathbb{Z})$. It is clear, that $[M]_\sim$ is nothing but the set $\{\pm M\}$ and again for easier notation, we will from now on simply write either M or $-M$ instead of $[M]_\sim$. Additionally we will denote equivalence of matrices by $\sim$, i.e. if $\lambda = \pm 1$, then $M \sim \lambda M$.

2.1 Generators and relations

In group theory it is an important question, which presentations (in view of Definition 1.17) can be given for a fixed group G. This section is devoted to the investigation of this question in the case of the modular group.

Before we start, we introduce the following notation for rounding real numbers to integers:

Definition 2.4 (Rounding functions)**.** For $x \in \mathbb{R}$, the *floor function* $\lfloor x \rfloor$, the *ceiling function* $\lceil x \rceil$ and the *nearest integer function* $\operatorname{nint}(x)$ are defined as

$$\lfloor x \rfloor := \max\{k \in \mathbb{Z} \mid k \leq x\}, \tag{2.1}$$

$$\lceil x \rceil := \min\{k \in \mathbb{Z} \mid k \geq x\}, \tag{2.2}$$

$$\operatorname{nint}(x) := \operatorname{sgn}(x) \left\lceil |x| - \frac{1}{2} \right\rceil. \tag{2.3}$$

Remark 2.5. It is quite common to define the nearest integer function such that half-integers are always rounded to even numbers. For our purposes this makes no essential difference and we prefer the above definition – rounding half-integers toward zero – for its simpler closed form.

Additionally we will need three modular transformations very frequently from now on:

Definition 2.6. We denote by U, T and R the following modular transformations:

$$U : z \mapsto z + 1$$
$$T : z \mapsto -\frac{1}{z}$$
$$R = TU : z \mapsto -\frac{1}{z+1}$$

Remark 2.7. Unfortunately in literature there is no consensus on the notation of these transformations. We use the notation of Schoeneberg [14] here, but other notations are frequent. For example in Klein/Fricke [7], the symbol S is used instead of U and in other literature as well as on Wikipedia, additionally the roles of S and T are swapped.

Theorem 2.8. *The modular group is generated by the elements* $U : z \mapsto z + 1$ *and* $T : z \mapsto -\frac{1}{z}$.

Proof. Let $A : z \mapsto \frac{az+b}{cz+d}$ be an arbitrary modular transformation. Our goal is to show that A can be written as product of the transformations U and T. For

this purpose it is more convenient to view these transformations as elements of $\mathrm{PSL}_2(\mathbb{Z})$, namely

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad U = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad T = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Let's first consider the two special cases, when a or c are zero. If $a = 0$, it follows from $ad - bc = 1$, that $-b = c = \pm 1$. Therefore we have (equivalence of matrices is again denoted by $\sim$)

$$A \sim cA = \begin{pmatrix} 0 & -1 \\ 1 & cd \end{pmatrix} = TU^{cd}.$$

Similarly, $c = 0$ gives $a = d = \pm 1$ and

$$A \sim aA = \begin{pmatrix} 1 & ab \\ 0 & 1 \end{pmatrix} = U^{ab}.$$

In the more general case, when a and c are both nonzero, $ad - bc = 1$ implies that a and c are coprime and the Euclidean algorithm therefore yields a finite sequence of equations

$$\begin{aligned} a &= q_0 \cdot c + r_1 \\ c &= q_1 \cdot r_1 + r_2 \\ r_1 &= q_2 \cdot r_2 + r_3 \\ &\vdots \\ r_{n-1} &= q_n \cdot r_n + r_{n+1} \\ &= q_n \cdot (\pm 1) + 0. \end{aligned}$$

Note that the quotients q_j and the remainders r_j depend on the choice of the rounding method for integer division (this will be discussed in more detail in Remark 2.11).

We can use the above sequence of equations to reduce the Matrix A by successively multiplying powers of U and T from the left. Just note that multiplication with U^k adds k times the second row to the first row, whereas T swaps the rows and changes the sign of one arbitrary row.[1] If we concentrate only on the first column of A and apply the first few transformations

$$\begin{pmatrix} a \\ c \end{pmatrix} \overset{U^{-q_0}}{\longmapsto} \begin{pmatrix} r_1 \\ c \end{pmatrix} \overset{T}{\mapsto} \begin{pmatrix} c \\ -r_1 \end{pmatrix} \overset{U^{q_1}}{\longmapsto} \begin{pmatrix} r_2 \\ -r_1 \end{pmatrix} \overset{T}{\mapsto} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \overset{U^{-q_2}}{\longmapsto} \begin{pmatrix} r_3 \\ r_2 \end{pmatrix} \overset{T}{\mapsto} \begin{pmatrix} r_2 \\ -r_3 \end{pmatrix} \mapsto \dots,$$

[1] This freedom of choice is again due to the fact that the matrices M and $-M$ represent the same element in $\mathrm{PSL}_2(\mathbb{Z})$.

we soon recognize the general mapping rule, which is

$$\begin{pmatrix} r_{j-1} \\ r_j \end{pmatrix} \overset{TU^{-q_j}}{\longmapsto} \begin{pmatrix} r_j \\ -r_{j+1} \end{pmatrix} \quad \text{for even } j \text{ and}$$
$$\begin{pmatrix} r_{j-1} \\ -r_j \end{pmatrix} \overset{TU^{q_j}}{\longmapsto} \begin{pmatrix} r_j \\ r_{j+1} \end{pmatrix} \quad \text{for odd } j.$$

When we set $r_{-1} := a$ and $r_0 := c$, this rule is applicable for $0 \leq j \leq n$. Obviously the described procedure ends with

$$\cdots \overset{T}{\mapsto} \begin{pmatrix} r_n \\ \pm r_{n+1} \end{pmatrix} = \begin{pmatrix} \pm 1 \\ 0 \end{pmatrix}.$$

Because we know the first column of the resulting matrix and its determinant, which is 1, we can conclude that for some $k \in \mathbb{Z}$ it must have the form

$$\pm \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix} \sim U^k.$$

By setting $e_n := (-1)^n q_n$ we therefore have

$$TU^{-e_n} TU^{-e_{n-1}} \cdots TU^{-e_1} TU^{-e_0} A = U^k,$$

or equivalently – noting that $T^{-1} = T$,

$$A = U^{e_0} TU^{e_1} \cdots TU^{e_{n-1}} TU^{e_n} TU^k,$$

which gives the desired representation of A in terms of U and T in the case when a and c are both nonzero. □

Corollary 2.9. *The special linear group* $\mathrm{SL}_2(\mathbb{Z})$ *is generated by the matrices* $U = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ *and* $T = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. *The projective special linear group* $\mathrm{PSL}_2(\mathbb{Z})$ *is generated by the equivalence classes* ${}_{\pm}U$ *and* ${}_{\pm}T$.[2]

Proof. The second statement is obviously a simple reformulation of Theorem 2.8. As a consequence, for every matrix $M \in \mathrm{SL}_2(\mathbb{Z})$ there exists a product of matrices T and U which evaluates to either M or $-M$. In the case when the product is $-M$, multiplication with the additional factor $T^2 = -\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ gives a product for M. This also proves the first statement. □

It is worth formulating the algorithm used in the proof of Theorem 2.8 explicitly in the following corollary.

[2]Here we use the notation ${}_{\pm}M := M_{\sim} = \{\pm M\} \in \mathrm{PSL}_2(\mathbb{Z})$ for clearer distinction between the matrix $M \in \mathrm{SL}_2(\mathbb{Z})$ and its corresponding equivalence class in $\mathrm{PSL}_2(\mathbb{Z})$.

Corollary 2.10 (The T-U algorithm)**.** *An arbitrary modular transformation $A : z \mapsto \frac{az+b}{cz+d}$ can be represented as product of the transformations $U : z \mapsto z+1$ and $T : z \mapsto -\frac{1}{z}$, by performing the following steps:*

1. *Apply the Euclidean algorithm to a and c with the first division being $a = q_0 \cdot c + r_1$ (q_0 may be ≤ 0) and let n be the number of the last division (start counting from 0). Call the arising quotients $q_0, q_1, \ldots, q_n$.*

   Note that if $a = 0$, the Euclidean algorithm will terminate after the first iteration, yielding the one-element quotient sequence $q_0 = 0$ and $n = 0$. If $c = 0$, it will terminate immediately, giving an empty sequence of quotients and $n = -1$.

2. *For $j \in \{0, 1, \ldots, n\}$ set $e_j := (-1)^j q_j$.*

3. *Calculate the matrix product $TU^{-e_n}TU^{-e_{n-1}} \cdots TU^{-e_1}TU^{-e_0}A$ and multiply by ± 1 in order to obtain a representation with positive diagonal elements. Read off the right-upper entry and call it k.*

The transformation A can now be written as

$$A = U^{e_0}TU^{e_1} \cdots TU^{e_{n-1}}TU^{e_n}TU^k. \tag{2.4}$$

Remark 2.11. The product representation (2.4) is not unique. In fact there is quite some freedom of choice for the quotients $q_0, q_1, \ldots, q_n$ in the above T-U algorithm. With the convention $r_{-1} := a$ and $r_0 := c$, one usually sets $q_j := \left\lfloor \frac{r_{j-1}}{r_j} \right\rfloor$ for $j \geq 0$. The remainders, which are determined by

$$r_{j+1} = r_{j-1} - q_j \cdot r_j \quad \text{for } j \geq 0, \tag{2.5}$$

are then all nonnegative and form a strictly monotonic decreasing sequence,

$$r_1 > r_2 > \ldots > r_n > r_{n+1} = 0.$$

All quotients, with the possible exception of q_0, are positive. In contrast to that, if we choose for each j an arbitrary rounding direction (up- or downwards) and set q_j to either $\left\lfloor \frac{r_{j-1}}{r_j} \right\rfloor$ or $\left\lceil \frac{r_{j-1}}{r_j} \right\rceil$, then in general also negative remainders (and consequently negative quotients) will occur. Still, the absolute values of the remainders form a strictly monotonic decreasing sequence,

$$|r_1| > |r_2| > \ldots > |r_n| > |r_{n+1}| = 0.$$

Therefore the Euclidean algorithm terminates and yields a correct result.[3] Depending on the choice of the rounding directions we will in general obtain different product representations of A. Not only this, we can even go one step further and violate the constraint $|r_j| > |r_{j+1}|$ for a finite number of indices j by choosing a completely random $q_j \in \mathbb{Z}$. As long as the remainders are calculated through (2.5), all product representations obtained in this way are correct.

Example 2.12. Consider the modular transformation $A \in \mathrm{PSL}_2(\mathbb{Z})$ given by

$$A = \begin{pmatrix} 13 & 5 \\ -8 & -3 \end{pmatrix}.$$

Applying algorithm 2.10 and setting $q_j := \left\lfloor \frac{r_{j-1}}{r_j} \right\rfloor$ for all j yields the product representation

$$A = U^{-2}TU^{-2}TU^{1}TU^{-2}T.$$

In contrast to that, setting $q_j := \left\lceil \frac{r_{j-1}}{r_j} \right\rceil$ in each step leads to

$$A = U^{-1}TU^{1}TU^{-1}TU^{1}TU^{-2}T.$$

Last but not least, always rounding to the nearest integer, that is setting $q_j := \mathrm{nint}\left(\frac{r_{j-1}}{r_j}\right)$, gives the shortest representation:

$$A = U^{-2}TU^{-3}TU^{-3}T.$$

Note that this is true in general: Rounding to the nearest integer always leads to a product representation with minimal number of factors T and U^k. According to Ore [10], already Leopold Kronecker has shown that among all variants of the Euclidean algorithm, the one which uses rounding to the nearest integer requires a minimum number of steps. We will come back to this fact once more in Remark 3.7.

We have now seen that T and U generate the modular group and different product representations for a given modular transformation A can be found using the above T-U algorithm. Of course the questions arises, which relations (in sense of Definition 1.17) lie behind the ambiguity of these product representations. For example, it is easy to see that $T^2 = 1$ and $(TU)^3 = 1$ are relations which are satisfied by T and U. It is the goal of the following

[3] The determined greatest common divisor is possibly negative, which is still admissible, because if d is a greatest common divisor of $n, m \in \mathbb{Z}$, i.e. $\forall d' \in \mathbb{Z} : (d' \mid n, m \Rightarrow d' \mid d)$, then also $-d$ is.

paragraphs to show that these two relations are in fact the only ones in the sense that all other relations are derived from these. We will do this by proving the following theorem:

Theorem 2.13 (Unique T-R representation)**.** *Let $T : z \mapsto -\frac{1}{z}$ and $R = TU : z \mapsto -\frac{1}{z+1}$. Every modular transformation $A \in \mathrm{PSL}_2(\mathbb{Z})$ can be written uniquely in the form*

$$A = R^{k_1} T R^{k_2} T \cdots R^{k_{n-1}} T R^{k_n} \tag{2.6}$$

with $n \in \mathbb{N}$, $k_2, \ldots, k_{n-1} \in \{\pm 1\}$ and $k_1, k_n \in \{0, \pm 1\}$.

From the uniqueness of the product representation (2.6) it follows that we have in fact a *presentation* of the modular group in the sense of Definition 1.17:

Corollary 2.14 (T-R presentation)**.** *The modular group is generated by the elements $T : z \mapsto -\frac{1}{z}$ and $R : z \mapsto -\frac{1}{z+1}$ and can be presented as*

$$\mathrm{PSL}_2(\mathbb{Z}) \cong \left\langle T, R \mid T^2 = R^3 = 1 \right\rangle. \tag{2.7}$$

Therefore $\mathrm{PSL}_2(\mathbb{Z})$ is isomorphic to the free product of a cyclic group of order 2 and a cyclic group of order 3.

Proof. It is easy to see that the relations $T^2 = R^3 = 1$ are indeed satisfied:

$$\begin{aligned} T^2 &= \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ R^3 &= \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}^3 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

Moreover the elements of $\langle T, R \mid T^2 = R^3 = 1 \rangle$ are precisely the group words of the form (2.6), as we have seen in Examples 1.19 and 1.22. □

Before we turn to the proof of Theorem 2.13, we first make one helpful definition and study its consequences.

Definition 2.15. For a modular transformation $A : z \mapsto \frac{az+b}{cz+d}$, we define the predicates t, r and s as well as a "grading" n:

$$t(A): \quad ac \geq 0 \quad \wedge \quad bd \geq 0 \tag{2.8}$$

$$r(A): \quad a^2 + ac \leq 0 \quad \wedge \quad b^2 + bd \leq 0 \tag{2.9}$$

$$s(A): \quad c^2 + ac \leq 0 \quad \wedge \quad d^2 + bd \leq 0 \tag{2.10}$$

$$n(A) := a^2 + b^2 + c^2 + d^2. \tag{2.11}$$

Note that t, r, s and n are well-defined, since they do not change their value if we switch from the matrix $A \in \mathrm{PSL}_2(\mathbb{Z})$ to the equivalent matrix $-A$, e.g. $t(A) \Leftrightarrow t(-A)$ and $n(A) = n(-A)$. Moreover the predicates t, r, and s partition the elements of the modular group into three classes:

Lemma 2.16. *Let $A : z \mapsto \frac{az+b}{cz+d}$ be an arbitrary modular transformation. Then one and only one of the predicates $t(A)$, $r(A)$ and $s(A)$ is satisfied.*

Proof. We start by considering the two easiest cases first: If A is the identity transformation or $A = T$, we have $t(A)$, $\neg s(A)$ and $\neg r(A)$.

For all other cases, we note that at least three of the coefficients a, b, c, d are nonzero. Therefore, if one the predicate $t(A)$ is satisfied, then at least one of the two inequalities involved is *strictly* fulfilled, i.e. $ac > 0$ or $bc > 0$. Having this said, it is easy to see that $t(A) \Rightarrow \neg r(A) \wedge \neg s(A)$. Thus it remains to show

$$\neg t(A) \Rightarrow r(A) \,\dot\vee\, s(A),$$

for all transformations with at least three nonzero coefficients (here, $\dot\vee$ denotes logical exclusive or). If $t(A)$ is false, we have $ac < 0$ or $bd < 0$. Since both cases are completely symmetric, we may assume without restriction that $ac < 0$. Note that $ac < 0$ and $ad - bc = 1$ implies $bd \leq 0$ because otherwise if $bd > 0$, both nonzero terms ad and bc would have different signs and their difference could not be 1. We conclude the proof by distinguishing three cases:

Case $a^2 < c^2$: From $ac < 0$ it follows that $a^2 + ac < 0$ (i) and $c^2 + ac > 0$ (ii). Additionally, from $ad - bc = 1$ we can conclude that $b^2 \leq d^2$, because otherwise $|ad|$ would differ from $|bc|$ by more than 1. Therefore we also have $b^2 + bd \leq 0$ (iii). Taking these pieces together, we have $(ii) \Rightarrow \neg s(A)$ and $(i) \wedge (iii) \Rightarrow r(A)$.

Case $a^2 > c^2$: This case is complementary to the first one: Because of $ac < 0$ we have $a^2 + ac > 0$ (i) and $c^2 + ac < 0$ (ii). The equation $ad - bc = 1$ here implies $b^2 \geq d^2$ and $d^2 + bd \leq 0$ (iii). Thus we have $(i) \Rightarrow \neg r(A)$ and $(ii) \wedge (iii) \Rightarrow s(A)$.

Case $a^2 = c^2$: Note that this case is only possible with $a = -c = \pm 1$ (as a and c are coprime). Hence we have $a^2 + ac = c^2 + ac = 0$. However, $ad - bc = 1$ implies $b^2 \neq d^2$ and therefore we have $b^2 + bd \leq 0 \,\dot\vee\, d^2 + bd \leq 0$. So, also in this case $r(A) \,\dot\vee\, s(A)$ holds. □

We have not yet seen the real benefit and meaning of the predicates t, r and s. The following lemma will imply that they do nothing but indicate the leftmost symbol in the unique T-R product representation (2.6) of A.

To be precise, if we denote this leftmost symbol by $\alpha(A)$, we will see that $t(A) \Leftrightarrow \alpha(A) = T$, $r(A) \Leftrightarrow \alpha(A) = R$ and $s(A) \Leftrightarrow \alpha(A) = R^{-1}$. Moreover we will show that the grading $n(A)$ grows monotonically with the number of symbols R and R^{-1} in the product representation of A.

Lemma 2.17. *The predicates t, r, s and the grading n satisfy the following relations:*

(i) $t(A) \Leftrightarrow r(RA) \Leftrightarrow s(R^{-1}A)$

(ii) $t(A) \wedge t(TA) \Leftrightarrow A \in \{1, T\}$

(iii) $n(A) = n(TA)$

(iv) $t(A) \Rightarrow n(A) < n(RA) \ \wedge \ n(A) < n(R^{-1}A)$

Proof. For a better overview, we first write out the matrices corresponding to TA, RA and $R^{-1}A$. Since $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, $T = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, $R = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ and $R^{-1} = \left(\begin{smallmatrix} 1 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ these are:

$$TA = \begin{pmatrix} -c & -d \\ a & b \end{pmatrix}, \quad RA = \begin{pmatrix} -c & -d \\ a+c & b+d \end{pmatrix}, \quad R^{-1}A = \begin{pmatrix} a+c & b+d \\ -a & -b \end{pmatrix}.$$

ad (i): This is shown easily via two simple calculations:

$$\begin{aligned} r(RA) &\Leftrightarrow c^2 - c(a+c) \le 0 \wedge d^2 - d(b+d) \le 0 \\ &\Leftrightarrow ac \ge 0 \wedge bd \ge 0 && \Leftrightarrow t(A) \\ s(R^{-1}A) &\Leftrightarrow a^2 - a(a+c) \le 0 \wedge b^2 - b(b+d) \le 0 \\ &\Leftrightarrow ac \ge 0 \wedge bd \ge 0 && \Leftrightarrow t(A) \end{aligned}$$

ad (ii): It is immediate to see that $t(A) \wedge t(TA)$ is equivalent to $ac = bd = 0$. Clearly, 1 and T are the only two transformations satisfying this condition.

ad (iii): $n(A) = n(TA)$ is trivial.

ad (iv): Note that $ad - bc = 1$ implies that at least one of the numbers a and b (resp. c and d) is nonzero. Moreover, $t(A) \Rightarrow ac \ge 0 \wedge bd \ge 0$, and thus we have

$$\begin{aligned} n(RA) - n(A) &= c^2 + 2ac + d^2 + 2bd > 0 \quad \text{and} \\ n(R^{-1}A) - n(A) &= a^2 + 2ac + b^2 + 2bd > 0. \end{aligned}$$

□

We can now formulate an algorithm which yields a product representation of any arbitrary modular transformation in terms of the generators R and T.

Theorem 2.18 (The T-R algorithm)**.** *For a modular transformation $A : z \mapsto \frac{az+b}{cz+d}$, a product representation of the form (2.6) can be found by performing the following steps:*

1. *Start with $k := 0$ and set $A_0 := A$.*
2. *If $A_k = 1$ go to step 5.*
3. *Define M_k as follows:*

$$t(A_k) \Rightarrow M_k := T, \quad r(A_k) \Rightarrow M_k := R, \quad s(A_k) \Rightarrow M_k := R^{-1}.$$

4. *Set $A_{k+1} := M_k^{-1} A_k$, increment k by one and continue with step 2.*
5. *If $k = 0$, then $A = 1$, which is the empty product. Otherwise, the desired product representation is $A = M_0 M_1 \cdots M_{k-1}$.*

Proof. Note that by Lemma 2.16 the rule for the definition of the transformations M_k from step 3 is unambiguous and the described algorithm therefore yields a unique sequence of equations

$$\begin{aligned} A_1 &= M_0^{-1} A_0 \\ A_2 &= M_1^{-1} A_1 \\ A_3 &= M_2^{-1} A_2 \\ &\vdots \end{aligned}$$

The relations (i) and (ii) of Lemma 2.17 guarantee that for every pair of subsequent transformations M_k, M_{k+1}, one of them is T and the other is either R or R^{-1}. Additionally, the relations (iii) and (iv) imply $n(A_k) > n(A_{k+2})$. Since $n(1) = n(T) = 2$ is a lower bound for $n(A_k)$, the described procedure must terminate after a finite number m of iterations and the product $A = M_0 M_1 \cdots M_{m-1}$ is indeed of the desired form (2.6). □

Now have all tools in hand for the proof of Theorem 2.13. Note that two alternative proofs can be found in Schoeneberg [14], §4 and in Klein/Fricke [7], p. 452ff.[4]

[4] Volume 1, part 2, chapter 9, §1.

Proof of Theorem 2.13. Let $A : z \mapsto \frac{az+b}{cz+d}$ be an arbitrary modular transformation. The existence of a product representation of the form (2.6) is ensured by the above T-R algorithm.

In order to prove also its uniqueness, it is sufficient to show that the identity map has a unique product representation (namely the empty product). From the relations (iii) and (iv) of Lemma 2.17 we see that any product P of the form (2.6) containing at least one factor R or R^{-1} has a grading $n(P) > n(1) = 2$ and therefore $P \neq 1$. The only products which are free of factors R and R^{-1} are T and the empty product. Since $T \neq 1$, the identity map can indeed only be represented by the empty product. □

Example 2.19. We apply the algorithm of Theorem 2.18 to the modular transformation from Example 2.12,

$$A = \begin{pmatrix} 13 & 5 \\ -8 & -3 \end{pmatrix}.$$

After substitution of $R^{-1} = R^2$, we end up with the R-T product representation

$$A = R^2TR^1TR^2TR^1TR^2TR^2.$$

Note that alternatively we could just as well have started with one of the T-U product representations of Example 2.12 and substitute $U = TR$ and $U^{-1} = R^2T$. Cancellation of terms $T^2 = 1$ and $R^3 = 1$ then would, by Theorem 2.13, necessarily lead to the same T-R product representation.

Remark 2.20. We conclude this section with the final remark that the T-R algorithm successively reduces a given matrix $A \in \mathrm{PSL}_2(\mathbb{Z})$ by multiplication from the left with T, R or R^{-1}. Of course, by using a dual approach also multiplication from the right can be used. All we have to do in order to adapt the algorithm appropriately is to substitute the predicates t, r, s by predicates t', r', s' and to change the definition in step 4 from $A_{k+1} := M_k^{-1}A_k$ to $A_{k+1} := A_kM_k^{-1}$. But how do the predicates t', r' and s' have to be defined?

For this consideration we denote the leftmost symbol in the R-T product representation (2.6) of A by $\alpha(A)$ and the rightmost symbol by $\omega(A)$. In the case of the empty product, we define $\alpha(1) := \omega(1) := T$. We have already seen that $t(A) \Leftrightarrow \alpha(A) = T$, $r(A) \Leftrightarrow \alpha(A) = R$ and $s(A) \Leftrightarrow \alpha(A) = R^{-1}$. In correspondence to that, we see that we have to define

$$\begin{array}{lllll} t'(A): & \omega(A) = T & \Leftrightarrow & \alpha(A^{-1}) = T & \Leftrightarrow \quad t(A^{-1}), \\ r'(A): & \omega(A) = R & \Leftrightarrow & \alpha(A^{-1}) = R^{-1} & \Leftrightarrow \quad s(A^{-1}), \\ s'(A): & \omega(A) = R^{-1} & \Leftrightarrow & \alpha(A^{-1}) = R & \Leftrightarrow \quad r(A^{-1}). \end{array}$$

Written out explicitly, this gives for a matrix $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{PSL}_2(\mathbb{Z})$:

$$t'(A): \quad ab \leq 0 \quad \wedge \quad cd \leq 0 \tag{2.12}$$
$$r'(A): \quad a^2 - ab \leq 0 \quad \wedge \quad c^2 - cd \leq 0 \tag{2.13}$$
$$s'(A): \quad b^2 - ab \leq 0 \quad \wedge \quad d^2 - cd \leq 0 \tag{2.14}$$

Note that also the Lemmas 2.16 and 2.17 remain valid, if the predicates t, r, s are substitued by t', r', s' and the order of matrix multiplication is reversed (i.e. RA, TA, ... have to be replaced by AR, AT, ...).

2.2 Fundamental sets and regions

We have seen in Remark 1.36 that considering Möbius transformations as meromorphic functions $\mathbb{C}_\infty \to \mathbb{C}_\infty$ very naturally induces a group action of $\mathrm{PGL}_2(\mathbb{C})$ on $\mathbb{C}_\infty$. Clearly this group action is also given for any subgroup of $\mathrm{PGL}_2(\mathbb{C})$ and in particular for the modular group: For $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{PSL}_2(\mathbb{Z})$ the group action is given by

$$Az := \frac{az+b}{cz+d}. \tag{2.15}$$

As usual we call two points $z, w \in \mathbb{C}_\infty$ *equivalent*, in symbols $z \sim w$, if there is a transformation $A \in \mathrm{PSL}_2(\mathbb{Z})$ such that $Az = w$. We are now interested in subsets of $\mathbb{C}_\infty$ containing exactly one point from each equivalence class of the relation $\sim$:

Definition 2.21 (Fundamental set)**.** Let G be a group acting on the set $\mathcal{S}$. Denote equivalence of points in $\mathcal{S}$ under G by $\sim$, as in (1.14). A subset $\mathcal{F}^* \subseteq \mathcal{S}$ is called a *fundamental set* with respect to the group action of G on $\mathcal{S}$, if $\mathcal{F}^*$ contains exactly one point from each orbit $Gx = [x]_\sim$, i.e. the map $\mathcal{F}^* \to \mathcal{S}/\sim$, $x \mapsto [x]_\sim$ is bijective.

Remark 2.22. Fundamental sets always exist, but they are not unique in any way. If $\mathcal{F}^*$ is a fundamental set, then for every subset $X \subseteq \mathcal{F}^*$ and for every $g \in G$, also the set $(\mathcal{F}^* \setminus X) \cup gX$ is fundamental.

Remark 2.23. If $\mathcal{F}^*$ is a fundamental set for the group action of G on $\mathcal{S}$, then by definition for every $x \in \mathcal{S}$ there is a $g \in G$ such that $gx \in \mathcal{F}^*$. Clearly for $g, h \in G$ and $x \in \mathcal{S}$ we have

$$gx = hx \quad \Leftrightarrow \quad h^{-1}gx = x \quad \Leftrightarrow \quad h^{-1}g \in G_x.$$

For this reason the element g with $gx \in \mathcal{F}^*$ is uniquely determined exactly when the stabilizer G_x is trivial. Obviously $gx \in \mathcal{F}^*$ is equivalent to $x \in g^{-1}\mathcal{F}^*$,

which allows us to reformulate the above fact the following way: The images of the fundamental set $\mathcal{F}^*$ under the group action of G cover the whole set $\mathcal{S}$, that is

$$\mathcal{S} = \bigcup_{g \in G} g\mathcal{F}^*.$$

A point $x \in \mathcal{S}$ is covered once, i.e. there is a unique $g \in G$ with $x \in g\mathcal{F}^*$, if and only if G_x is trivial.

If – like in the case of $\mathbb{C}_\infty$ – the set $\mathcal{S}$ is equipped with a topology, it is often advantageous to use a concept slightly different to fundamental sets:

Definition 2.24 (Fundamental region)**.** Let G be a group acting on a set $\mathcal{S}$ which is (a subset of) a topological space. A nonempty open subset $\mathcal{F} \subseteq \mathcal{S}$ is called *fundamental region* with respect to the group action of G on $\mathcal{S}$, if it contains no distinct points equivalent under G and if the neighborhood of every boundary point of $\mathcal{F}$ contains a point of $\mathcal{S} \setminus \mathcal{F}$ which is equivalent to a point within $\mathcal{F}$.

Remark 2.25. The relation between fundamental sets and fundamental regions is the following: If $\mathcal{F} \subseteq \mathcal{S}$ is a fundamental region and if the images under G of its topological closure $\mathrm{cl}(\mathcal{F})$ cover the whole set $\mathcal{S}$, i.e.

$$\mathcal{S} = \bigcup_{g \in G} g\,\mathrm{cl}(\mathcal{F}),$$

then a fundamental set $\mathcal{F}^*$ can always be obtained from $\mathcal{F}$ by adjoining certain boundary points of $\mathcal{F}$ and we have $\mathcal{F} \subseteq \mathcal{F}^* \subseteq \mathrm{cl}(\mathcal{F})$.

However, a fundamental region may not always exist: Consider the group of translations $z \mapsto z + \alpha$, $\alpha \in \mathbb{R}$, acting on the complex plane $\mathbb{C}$. No matter how we arrange a fundamental set $\mathcal{F}^*$, take for example the imaginary axis $\mathrm{Re}\,(z) = 0$, the interior of $\mathcal{F}^*$ will always be empty and thus cannot be a fundamental region. In fact, a necessary and sufficient condition for the existence of a fundamental region is that the group action is *discontinuous* on $\mathcal{S}$, i.e. there exists and *ordinary* point $x \in \mathcal{S}$, meaning that there is no sequence of the form $(g_n y)_{n \geq 0}$ with distinct $g_n \in G$ and fixed $y \in \mathcal{S}$ which converges to x. Otherwise, if such an ordinary point does not exist, then every open subset of $\mathcal{S}$ necessarily contains distinct equivalent points. For further details see Lehner [8], Chapter IV, 1B.

Note that in Schoeneberg [14], the term "fundamental region" is used for any set X containing a fundamental set $\mathcal{F}^*$ plus some or all of its (remaining) boundary points. We prefer the above definition for being more close to the commonly used meaning of "region", i.e. a topologically connected and open set. However, we emphasize that according to our definition a fundamental region does not need to be topologically connected.

The goal of the remainder of this section is to identify fundamental regions and fundamental sets for the action of the modular group on $\mathbb{C}_\infty$. For this purpose it is instructive to first consider homogeneous modular transformations and their natural action on $\mathbb{C}^2$.

2.2.1 The action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{C}^2$

The homogeneous modular group $\mathrm{SL}_2(\mathbb{Z})$ naturally acts on the vector space $\mathbb{C}^2$ by matrix-vector multiplication. Written out explicitly, for $A \in \mathrm{SL}_2(\mathbb{Z})$ and $x = \binom{u}{v} \in \mathbb{C}^2$, this group action is given by

$$Ax = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} au + bv \\ cu + dv \end{pmatrix}.$$

We equip $\mathbb{C}^2$ with the standard Euclidean norm and its induced topology:

$$\left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|_2 := \sqrt{|u|^2 + |v|^2}$$

Let us denote the orbit of $x \in \mathbb{C}^2$ under $\mathrm{SL}_2(\mathbb{Z})$ by

$$O_x := \mathrm{SL}_2(\mathbb{Z})x = \{Ax \mid A \in \mathrm{SL}_2(\mathbb{Z})\} \subseteq \mathbb{C}^2. \tag{2.16}$$

We now wish to find a fundamental set with respect to the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{C}^2$. For this purpose, we need to choose exactly one vector from each of the different orbits O_x. In order to restrict the number of candidate vectors to an easily manageable number, we could try to first look just at vectors with minimal norm in O_x. The problem with this idea is that in general the orbit O_x might contain vectors of arbitrary small norm – in other words $\min \|O_x\|_2$ might in general not necessarily exist. However, in many cases it does:

Lemma 2.26 (Existence of $\min \|O_x\|_2$)**.** *Let $x = \binom{u}{v} \in \mathbb{C}^2$ be a vector with u, v being linear independent over $\mathbb{R}$. For every $r > 0$, there are only finitely many points $y \in O_x$ with $\|y\|_2 \leq r$. In particular, $m := \min \|O_x\|_2$ exists and $m > 0$.*

Proof. Since the complex numbers u and v are linear independent over $\mathbb{R}$, they span a non-degenerate parallelogram $P_x := \{tu + sv \mid t, s \in [0, 1)\}$ on the complex plane. Translation of P_x by integer multiples of u and v covers every point in $\mathbb{C}$ exactly once. The set

$$L_x := \{au + bv \in \mathbb{C} \mid a, b \in \mathbb{Z}\} \tag{2.17}$$

consists precisely of the vertices of all of these translated parallelograms. For $r > 0$, denote by D_r a disk of raduis r in $\mathbb{C}$ and by B_r a ball of radius r in $\mathbb{C}^2$ (both centered about the origin):

$$D_r := \{z \in \mathbb{C} \mid |z| \leq r\}, \tag{2.18}$$
$$B_r := \{y \in \mathbb{C}^2 \mid \|y\|_2 \leq r\}. \tag{2.19}$$

Obviously $L_x \cap D_r$ is finite for every $r > 0$. Now let $r > 0$ be sufficiently large such that the set $O_x \cap B_r$ is not empty (e.g. $r = \|x\|_2$) and observe

$$O_x \cap B_r \subseteq (L_x \cap D_r)^2.$$

We see that also the set $O_x \cap B_r$ is finite and $m = \min \|O_x\|_2 = \min \|O_x \cap B_r\|_2$ therefore exists. Since u and v are linear independent over $\mathbb{R}$ (and in particular over $\mathbb{Z}$), $0 \notin L_x$ and consequently $0 \notin O_x$, which is why $m > 0$. □

We now turn to the question how we can effectively determine an element of O_x with minimal norm. The task is the following: Given a vector $x = \binom{u}{v} \in \mathbb{C}^2$ with u, v linear independent over $\mathbb{R}$, find a matrix $B \in \mathrm{SL}_2(\mathbb{Z})$ such that $\|Bx\|_2$ is minimal.

In Corollary 2.9 we have seen that $\mathrm{SL}_2(\mathbb{Z})$ is generated by the matrices $T = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $U = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. The idea is now to successively multiply x with appropriate powers of T and U to obtain vectors of smaller and smaller norm. We do this by first finding an integer $k_0 \in \mathbb{Z}$, such that $\left\|U^{-k_0}x\right\|_2$ is minimal. Then we multiply with T and repeat the process for finding $k_1 \in \mathbb{Z}$ minimizing $\left\|U^{k_1}TU^{k_0}x\right\|_2$ and so on. The procedure ends when $k_n = 0$ for some $n > 0$. Note that the integers k_j can be determined easily:

Lemma 2.27. *Let $x = \binom{u}{v} \in \mathbb{C}^2$ with $v \neq 0$. The statements*

(i) $k \in \mathbb{Z}$ *minimizes* $\left\|U^{-k}x\right\|_2 = \left\|\begin{pmatrix} u - kv \\ v \end{pmatrix}\right\|_2$,

(ii) $k \in \mathbb{Z}$ *minimizes* $|u - kv|$,

(iii) $k \in \mathbb{Z}$ *minimizes* $\left|\frac{u}{v} - k\right|$,

(iv) $\left|\mathrm{Re}\left(\frac{u}{v}\right) - k\right| \leq \frac{1}{2}$,

(v) $k = \mathrm{nint}\left(\mathrm{Re}\left(\frac{u}{v}\right)\right)$,

satisfy the relations $(i) \Leftrightarrow (ii) \Leftrightarrow (iii) \Leftrightarrow (iv)$ and $(v) \Rightarrow (iv)$.

Proof. Trivial. □

Let us now suppose that the described procedure comes to an end, i.e. $k_n = 0$ for some $n > 0$. Set $B := TU^{k_{n-1}} \cdots TU^{k_0}$ and $y = \binom{y_1}{y_2} := Bx$. It follows from $k_n = 0$ and from the choice of k_{n-1} that we have

$$\|y\|_2 \leq \left\|U^k y\right\|_2 \quad \text{and} \quad \|y\|_2 \leq \left\|U^k T^{-1} y\right\|_2 \quad \text{for all } k \in \mathbb{Z}. \tag{2.20}$$

Using Lemma 2.27 – (i) $\Leftrightarrow$ (iv) – we see that (2.20) is equivalent to

$$\left|\operatorname{Re}\left(\frac{y_1}{y_2}\right)\right| \leq \frac{1}{2} \quad \text{and} \quad \left|\operatorname{Re}\left(\frac{y_2}{y_1}\right)\right| \leq \frac{1}{2},$$

which can easily be rewritten to

$$|y_1\overline{y_2} + \overline{y_1}y_2| \leq \min\{|y_1|^2, |y_2|^2\}. \tag{2.21}$$

The question arises, whether y is just a "local minimum" in the sense (2.20) or if (2.21) already implies the global minimality of y, i.e. $\|y\|_2 = \min \|O_x\|_2$. The following theorem will give us insight on this.

Theorem 2.28. *Let $A \in \mathrm{SL}_2(\mathbb{Z})$ be an arbitrary homogeneous modular transformation and let the* grading $n(A)$ *be defined as in (2.11). Let $x = \binom{u}{v} \in \mathbb{C}^2$ with $uv \neq 0$. Then the following statements hold:*

(i) If $|u\overline{v} + \overline{u}v| \leq \min\{|u|^2, |v|^2\}$, then $\|x\|_2 \leq \|Ax\|_2$.

(ii) If $|u\overline{v} + \overline{u}v| \leq \min\{|u|^2, |v|^2\}$ and $n(A) > 3$, then $\|x\|_2 < \|Ax\|_2$.

(iii) If $|u\overline{v} + \overline{u}v| < \min\{|u|^2, |v|^2\}$ and $n(A) > 2$, then $\|x\|_2 < \|Ax\|_2$.

Proof. Let us denote $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. We need to show $\|x\|_2 \leq \|Ax\|_2$, that is

$$\begin{aligned} |u|^2 + |v|^2 \quad \leq \quad & |au + bv|^2 + |cu + dv|^2 = \\ & (au + bv)(a\overline{u} + b\overline{v}) + (cu + dv)(c\overline{u} + d\overline{v}) = \\ & (a^2 + c^2)\,|u|^2 + (b^2 + d^2)\,|v|^2 + (ab + cd)(u\overline{v} + \overline{u}v), \end{aligned}$$

which is equivalent to

$$(a^2 + c^2 - 1)\,|u|^2 + (b^2 + d^2 - 1)\,|v|^2 \geq -(ab + cd)(u\overline{v} + \overline{u}v).$$

Now we find an upper bound of the right hand side by taking its absolute value and using $|u\overline{v} + \overline{u}v| \leq \min\{|u|^2, |v|^2\} =: m$. The same time, m also helps us with a lower bound of the left hand side:

$$(a^2 + b^2 + c^2 + d^2 - 2) \cdot m \geq |ab + cd| \cdot m.$$

Since m is nonzero ($uv \neq 0$), it can be canceled. Moreover, $ad - bc = 1$ implies that the terms ad and bc can never have opposite signs. In other words we always have $(ad)(bc) \geq 0$. Obviously also $(ab)(cd) \geq 0$, i.e. also the terms ad and bc have non-opposite signs, which is why $|ab + cd| = |ab| + |bd|$. For this reason we can transform the last inequality to

$$\underbrace{(a^2 - |ab| + b^2)}_{\geq (|a|-|b|)^2 =: \ell} + \underbrace{(c^2 - |cd| + d^2)}_{\geq (|c|-|d|)^2 =: r} \geq 2. \tag{2.22}$$

This obviously holds for the case $n(A) = 2$. For the case $n(A) > 2$, because of $ad - bc = 1$, we see:

(a) $|a| = |b|$ implies $|c| \neq |d|$ (and vice versa). Therefore at least one of the lower bounds ℓ and r is nonzero.

(b) $|ab| = 0$ implies $|bc| \neq 0$ (and vice versa). Therefore at least one of the lower bounds ℓ and r is in fact a *strict* lower bound.

These two observations prove (2.22) and consequently assertion (i). If additionally $n(A) > 3$, we distinguish two cases:

Case $0 \in \{a, b, c, d\}$**:** Assume without restriction that $0 \in \{a, b\}$, (the case $0 \in \{c, d\}$ is completely symmetric). It follows $\{|a|, |b|\} = \{0, 1\}$ and $\{|c|, |d|\} = \{1, N\}$ with $N > 1$, since $n(A) > 3$. Therefore, in addition to observation (b), both lower bounds ℓ and r are positive.

Case $0 \notin \{a, b, c, d\}$**:** In addition to observation (a), $|ab|$ and $|cd|$ are both nonzero. Therefore ℓ and r are both strict lower bounds.

In both cases (2.22) is strictly fulfilled, which proves (ii). We note that in the case $n(A) > 3$, (iii) already follows from (ii). It therefore remains to consider just the special case $n(A) = 3$ for proving the last statement. By the same calculation as above, the condition $\|x\|_2 < \|Ax\|_2$ is equivalent to

$$(a^2 + c^2 - 1)\,|u|^2 + (b^2 + d^2 - 1)\,|v|^2 > -(ab + cd)(u\overline{v} + \overline{u}v). \tag{2.23}$$

For $n(A) = 3$ we have $\{(a^2 + c^2 - 1), (b^2 + d^2 - 1)\} = \{0, 1\}$, which means that the left hand side simplifies to either $|u|^2$ or $|v|^2$, whereas on the right hand side we have $|u\overline{v} + \overline{u}v|$ as upper bound, because of $(ab + cd) = \pm 1$. Since by assumption $|u\overline{v} + \overline{u}v| < \min\{|u|^2, |v|^2\}$, inequality (2.23) is thus satisfied. □

We can now summarize the algorithm and prove its correctness. Note that we can slightly relax the requirement on the coordinates of $x = \binom{u}{v}$ being linear independent over $\mathbb{R}$. We will see that with a minor adaption the algorithm works as well in the case when u and v are linear dependent over $\mathbb{Q}$.

Theorem 2.29. *Let $x = \binom{u}{v} \in \mathbb{C}^2$ with u, v being either linear independent over $\mathbb{R}$ or linear dependent over $\mathbb{Q}$. A matrix $B \in \mathrm{SL}_2(\mathbb{Z})$ minimizing $\|Bx\|_2$ can be found by performing the following steps:*

1. *Set $(r_{-1}, r_0) := (u, v)$ and $j := 0$.*
2. *If $r_j = 0$, then goto step 5.*
3. *Determine $q_j := \operatorname{nint}\left(\operatorname{Re}\left(\frac{r_{j-1}}{r_j}\right)\right)$.*
4. *If $j > 0$ and $q_j = 0$ go to step 5. Otherwise, set $r_{j+1} := r_{j-1} - q_j r_j$, increment j by one and continue with step 2.*
5. *Set $n := j - 1$ and for $i \in \{0, 1, \ldots, n\}$, set $e_i := (-1)^i q_i$. The desired matrix is*
$$B = U^{-e_n} T U^{-e_{n-1}} \cdots T U^{-e_0}. \tag{2.24}$$
Note that in the case $n < 0$ this product is empty and B is the identity matrix.

Proof. The algorithm gives rise to the following sequence of equations:

$$\begin{aligned} u = r_{-1} &= q_0 \cdot r_0 + r_1 \\ v = \ r_0 &= q_1 \cdot r_1 + r_2 \\ r_1 &= q_2 \cdot r_2 + r_3 \\ r_2 &= q_3 \cdot r_3 + r_4 \\ &\vdots \end{aligned}$$

Moreover this sequence of equations corresponds to the sequence of vectors

$$\binom{u}{v} \overset{TU^{-q_0}}{\longmapsto} \binom{v}{-r_1} \overset{TU^{q_1}}{\longmapsto} \binom{-r_1}{-r_2} \overset{TU^{-q_2}}{\longmapsto} \binom{-r_2}{r_3} \overset{TU^{q_3}}{\longmapsto} \binom{r_3}{r_4} \mapsto \ldots$$

With $s_j := (-1)^{\lceil \frac{j}{2} \rceil} r_j$, we can write these vectors as $x_j := \begin{pmatrix} s_{j-1} \\ s_j \end{pmatrix}$ for $j \geq 0$, in particular we have $x_0 = x$. Now, as in the theorem, let $e_j := (-1)^j q_j$ for $j \geq 0$. In this notation, we can write in general $TU^{-e_j} x_j = x_{j+1}$ for all $j \geq 0$.

In the case when u and v are linear dependent over $\mathbb{Q}$, the vector x can be written as $x = \lambda \binom{p}{q}$ with $\lambda \in \mathbb{C}$ and $p, q \in \mathbb{Z}$. We observe the striking similarity between the Euclidean and the above algorithm: If we apply the Euclidean algorithm to p and q while rounding all quotients to the nearest integer (compare Remark 2.11), we obtain a sequence of quotients $\tilde{q}_0, \tilde{q}_1, \ldots, \tilde{q}_n$ and a sequence of remainders $\tilde{r}_0, \tilde{r}_1, \ldots, \tilde{r}_n, \tilde{r}_{n+1}$ with $\tilde{r}_{n+1} = 0$. It is immediate

to see that $q_j = \tilde{q}_j$ and $r_j = \lambda \tilde{r}_j$ for all $j \geq 0$. In particular, our algorithm terminates and we end up with the vector

$$y := x_n = TU^{-e_n} TU^{-e_{n-1}} \cdots TU^{-e_0} = \lambda \begin{pmatrix} \pm \gcd(p,q) \\ 0 \end{pmatrix}$$

having minimal norm in O_x.

In the case when u and v are linear independent over $\mathbb{R}$, it follows from the choice of q_j that for each pair of subsequent vectors x_j, x_{j+1} we have $\|x_j\|_2 \geq \|x_{j+1}\|_2$. Using $r_{j+1} = r_{j-1} - q_j r_j$ we see that $\|x_j\|_2 = \|x_{j+1}\|_2$ is equivalent to

$$\left\| \begin{pmatrix} \pm r_{j-1} \\ \pm r_j \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} \pm r_j \\ \pm r_{j+1} \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} \pm r_j \\ \pm (r_{j-1} - q_j r_j) \end{pmatrix} \right\|_2 .$$

Obviously this is the case if and only if $|r_{j-1}| = |r_{j-1} - q_j r_j|$. If we divide by r_j and set $z_j := \frac{r_{j-1}}{r_j}$, we obtain

$$\begin{aligned} |z_j| = |z_j - q_j| \quad &\Leftrightarrow \quad z_j \overline{z_j} = (z_j - q_j)(\overline{z_j} - q_j) \\ &\Leftrightarrow \quad q_j \left(q_j - 2 \operatorname{Re}(z_j) \right) = 0. \end{aligned}$$

One obvious solution to this is $q_j = 0$. For the other factor, we substitute $\alpha := \operatorname{Re}(z_j)$ and use $q_j = \operatorname{nint}(\alpha)$ to see that the equation $\operatorname{nint}(\alpha) = 2\alpha$ has the unique[5] solution $\alpha = 0$ which again leads to $q_j = 0$. Summing up, we therefore have for all $j \geq 0$

$$\|x_j\|_2 \geq \|x_{j+1}\|_2 \quad \text{and} \quad \|x_j\|_2 = \|x_{j+1}\|_2 \Leftrightarrow q_j = 0. \tag{2.25}$$

According to Lemma 2.26, the set $O_x \cap K_{\|x\|_2}$ is finite and thus we cannot have $\|x_j\|_2 > \|x_{j+1}\|_2$ for infinitely many indices j. In other words, q_n must be zero for some $n \in \mathbb{N}$ and we have

$$TU^{-e_n} \cdots TU^{-e_1} TU^{-e_0} x_0 = x_n.$$

Since the vector $y := x_n$ satisfies (2.20) and consequently (2.21), we conclude from Theorem 2.28 that $\|y\|_2 = \min \|O_x\|_2$.

In both cases (linear dependence of u, v over $\mathbb{Q}$ or linear independence of u, v over $\mathbb{R}$), we trivially have $\|y\|_2 = \|T^{-1} y\|_2$ and we therefore can define B as in (2.24). □

[5]Here we benefit from our definition of nint, which rounds $\pm\frac{1}{2}$ to zero.

Let us now denote by $\hat{\mathcal{S}} \subseteq \mathbb{C}^2$ the set of all vectors whose coordinates u and v are linear independent over $\mathbb{R}$ or linear dependent over $\mathbb{Q}$. It is clear that $\hat{\mathcal{S}}$ is invariant under the action of $\mathrm{SL}_2(\mathbb{Z})$, i.e. $\mathrm{SL}_2(\mathbb{Z})\hat{\mathcal{S}} = \hat{\mathcal{S}}$. We can therefore as well consider the group action of $\mathrm{SL}_2(\mathbb{Z})$ on $\hat{\mathcal{S}}$. We have seen in Theorem 2.28 that within the region

$$\hat{\mathcal{R}} := \left\{ \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{C}^2 \;\middle|\; |u\overline{v} + \overline{u}v| < \min\{|u|^2, |v|^2\} \right\} \subseteq \hat{\mathcal{S}} \tag{2.26}$$

equivalence of points can only be established by transformations $A \in \mathrm{SL}_2(\mathbb{Z})$ with $n(A) = 2$. Clearly these transformations are exactly given by $1, T, T^2$ and T^3. On the other hand, by Theorem 2.29, every vector $x \in \hat{\mathcal{S}}$ is equivalent to a point in the topological closure of $\hat{\mathcal{R}}$. Hence we can obtain a fundamental region for the group action of $\mathrm{SL}_2(\mathbb{Z})$ on $\hat{\mathcal{S}}$ by choosing for each $x \in \hat{\mathcal{R}}$ exactly one of the equivalent vectors

$$x = \begin{pmatrix} u \\ v \end{pmatrix}, \quad Tx = \begin{pmatrix} -v \\ u \end{pmatrix}, \quad T^2x = \begin{pmatrix} -u \\ -v \end{pmatrix}, \quad T^3x = \begin{pmatrix} v \\ -u \end{pmatrix}.$$

This choice can be done quite arbitrarily – for example we can arrange the fundamental region such that the coordinates u and v of each of its points lie in a certain fixed half-plane of $\mathbb{C}$, e.g. $\mathrm{Re}\,(u) > 0$ and $\mathrm{Re}\,(v) > 0$.

Corollary 2.30. *The set*

$$\hat{\mathcal{F}} := \hat{\mathcal{R}} \cap \left\{ \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{C}^2 \;\middle|\; \mathrm{Re}\,(u) > 0,\ \mathrm{Re}\,(v) > 0 \right\}, \tag{2.27}$$

where $\hat{\mathcal{R}}$ is defined as in (2.26), is a fundamental region for the action of the homogeneous modular group $\mathrm{SL}_2(\mathbb{Z})$ on the set $\hat{\mathcal{S}}$.

2.2.2 The action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathbb{C}_\infty$

In Theorem 2.29 we have seen an algorithm which naturally gives rise to a set $\hat{\mathcal{R}} \subseteq \mathbb{C}^2$ which can easily be restricted to a subset $\hat{\mathcal{F}} \subseteq \hat{\mathcal{R}}$ being a fundamental region for the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\hat{\mathcal{S}}$. We can exploit this fact in search of a fundamental region for the action on of the (inhomogeneous) modular group $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathbb{C}_\infty$. For this purpose we project $\mathbb{C}^2$ onto $\mathbb{C}_\infty$ using the map $\pi : \mathbb{C}^2 \to \mathbb{C}_\infty$,

$$\pi : \begin{pmatrix} u \\ v \end{pmatrix} \mapsto \frac{u}{v}. \tag{2.28}$$

Let us first consider image of $\hat{\mathcal{S}}$ under π. If $\binom{u}{v} \in \hat{\mathcal{S}}$, then by definition u and v are linear independent over $\mathbb{R}$ or linear dependent over $\mathbb{Q}$. In the first case

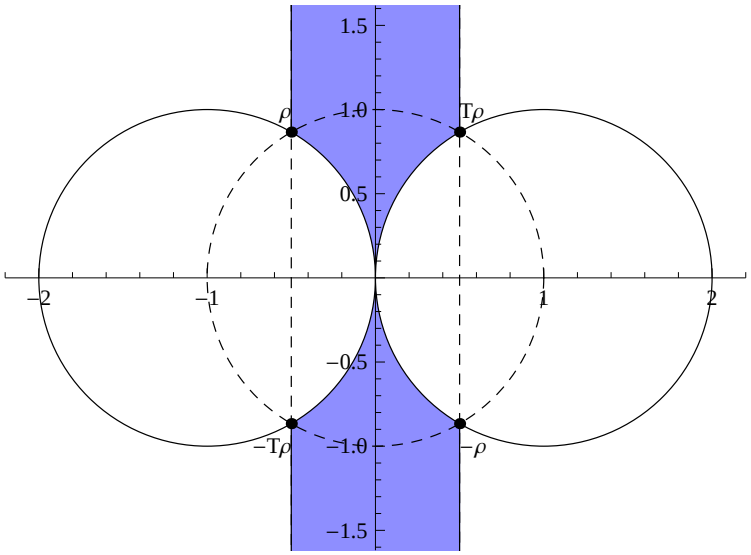

Figure 2.1: The region $\mathcal{R} \subseteq \mathbb{C}_\infty$ of numbers $z = u/v \in \mathbb{C}_\infty$ with $|u\overline{v} + \overline{u}v| < \min\{|u|^2, |v|^2\}$. It is obtained by taking the strip $\left\{z \in \mathbb{C} \;\middle|\; |\mathrm{Re}\,(z)| < \frac{1}{2}\right\}$ and cutting out two closed disks of unit radius centered about the real points ± 1. The arising vertices are labeled. As usual, T is the transformation $z \mapsto -\frac{1}{z}$ and $\rho = \exp(2\pi i/3)$ is a third root of unity.

we have $\frac{u}{v} \in \mathbb{C} \setminus \mathbb{R}$ and in the second case $\frac{u}{v} \in \mathbb{Q} \cup \{\infty\}$. This means $\frac{u}{v}$ may be everything but irrational. Denoting the set of irrational numbers by $\mathbb{I} := \mathbb{R} \setminus \mathbb{Q}$, we thus have

$$\pi(\hat{\mathcal{S}}) = \mathbb{C}_\infty \setminus \mathbb{I}.$$

Projection of the set $\hat{\mathcal{R}} \subseteq \mathbb{C}^2$ leads to the region $\mathcal{R} \subseteq \mathbb{C}_\infty$ (see also Figure 2.1),

$$\mathcal{R} := \pi\left(\hat{\mathcal{R}}\right) = \left\{\frac{u}{v} \in \mathbb{C}_\infty \;\middle|\; |u\overline{v} + \overline{u}v| < \min\{|u|^2, |v|^2\}\right\}. \tag{2.29}$$

It follows that $\mathcal{R}$ contains a fundamental region for the action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathbb{C}_\infty \setminus \mathbb{I}$. As in the homogeneous case, we see from Theorem 2.28 that equivalence of points within $\mathcal{R}$ can be established only by powers of the transformation $T \in \mathrm{PSL}_2(\mathbb{Z})$. Since $T^2 = 1$, in order to obtain a fundamental region we now need to choose for each $z \in \mathcal{R}$ just exactly one of the equivalent points z and Tz. This can for example be done such that $|z| > 1$.

Note that for understanding the group action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathbb{C}_\infty \setminus \mathbb{I}$ it is sufficient to look at either the upper or lower half-plane of $\mathbb{C}$, since the group action on one half-plane is symmetric to the group action on the other half-plane by $A\overline{z} = \overline{Az}$. Let us therefore denote by $\mathcal{H}$ the upper half-plane and by

$\mathcal{H}^*$ the *extended upper half-plane*:

$$\mathcal{H} := \left\{ z \in \mathbb{C} \;\middle|\; \mathrm{Im}\,(z) > 0 \right\} \tag{2.30}$$

$$\mathcal{H}^* := \mathcal{H} \cup \mathbb{Q} \cup \{\infty\}. \tag{2.31}$$

Clearly $\mathcal{H}^*$ is invariant under $\mathrm{PSL}_2(\mathbb{Z})$, i.e. $\mathrm{PSL}_2(\mathbb{Z})\mathcal{H}^* = \mathcal{H}^*$ and we can also consider $\mathrm{PSL}_2(\mathbb{Z})$ acting on $\mathcal{H}^*$.

Theorem 2.31. *Let $\mathcal{H}$ and $\mathcal{H}^*$ be defined as above. The set*

$$\tilde{\mathcal{F}} := \left\{ z \in \mathbb{C} \;\middle|\; |\mathrm{Re}\,(z)| < \frac{1}{2} \text{ and } |z| > 1 \right\} \tag{2.32}$$

is a fundamental region for the action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathbb{C}_\infty \setminus \mathbb{I}$. The part of $\tilde{\mathcal{F}}$ lying in the upper half-plane $\mathcal{H}$, i.e. the set

$$\mathcal{F} := \tilde{\mathcal{F}} \cap \mathcal{H} \tag{2.33}$$

is a fundamental region for the action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathcal{H}^*$.

Proof. The second statement that $\mathcal{F} = \tilde{\mathcal{F}} \cap \mathcal{H}$ is fundamental region for $\mathrm{PSL}_2(\mathbb{Z})$ acting on $\mathcal{H}^*$ is a simple consequence of the first statement. For proving that $\tilde{\mathcal{F}}$ is fundamental, observe that $\tilde{\mathcal{F}}$ is exactly the set

$$\tilde{\mathcal{F}} = \mathcal{R} \cap \left\{ z \in \mathbb{C} \;\middle|\; |z| > 1 \right\},$$

with $\mathcal{R}$ defined as in (2.29) – compare also Figure 2.1. Obviously $\tilde{\mathcal{F}}$ is a nonempty open subset of $\mathcal{R}$, which is why two distinct points of $\tilde{\mathcal{F}}$ can be equivalent only by the transformation T. However, since $|z| > 1$ implies $|Tz| = |-1/z| < 1$, this is impossible. Therefore $\tilde{\mathcal{F}}$ contains no equivalent distinct points.

It remains to show that every $z = \frac{u}{v} \in \mathcal{S}$ is equivalent to a point of the topological closure $\mathrm{cl}(\tilde{\mathcal{F}})$ of $\tilde{\mathcal{F}}$. For this purpose apply the algorithm of Theorem 2.29 to the vector $\binom{u}{v} \in \hat{\mathcal{S}}$ in order to obtain a transformation $B \in \mathrm{PSL}_2(\mathbb{Z})$ which maps z to a point of $\mathrm{cl}(\mathcal{R})$. It then follows that at least one of the points Bz or TBz lies in $\mathrm{cl}(\tilde{\mathcal{F}})$. □

We now wish to obtain a *fundamental set* for the action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathcal{H}^*$. For this purpose we need to consider the boundary of $\mathcal{F}$ and to investigate equivalent boundary points and their associated transformations. It turns out that we can define a fundamental set for the action of $\mathrm{PSL}_2(Z)$ on $\mathcal{H}^*$ the following way:

Theorem 2.32 (The fundamental set $\mathcal{F}^*$)**.** *Denote by $\mathcal{F}$ the fundamental region from (2.33). The boundary of $\mathcal{F}$ shall be segmented into the four "boundary arcs",*

$$\begin{aligned} a &:= \left\{ -\frac{1}{2} + yi \;\middle|\; y \geq \frac{\sqrt{3}}{2} \right\} \cup \{\infty\}, \\ b &:= \left\{ +\frac{1}{2} + yi \;\middle|\; y \geq \frac{\sqrt{3}}{2} \right\} \cup \{\infty\}, \\ c &:= \left\{ i \cdot e^{+i\varphi} \;\middle|\; 0 \leq \varphi \leq \frac{\pi}{6} \right\}, \\ d &:= \left\{ i \cdot e^{-i\varphi} \;\middle|\; 0 \leq \varphi \leq \frac{\pi}{6} \right\}. \end{aligned}$$

These boundary arcs are mapped onto each other by $Ua = b$ and $Tc = d$. The set

$$\mathcal{F}^* := \mathcal{F} \cup a \cup c \tag{2.34}$$

is a fundamental set for the action of $\mathrm{PSL}_2(\mathbb{Z})$ on the extended upper half-plane $\mathcal{H}^*$.

Proof. It follows from Theorem 2.28 that equivalence of boundary points of $\mathcal{F}$ can only be established by transformations $A \in \mathrm{PSL}_2(Z)$ with $n(A) \leq 3$. The full list of candidate transformations therefore comprises of 9 transformations: T, U, TU, UT, TUT and the respective inverse transformations (note that T is self-inverse). After looking at these transformations individually, it turns out that in fact only T and U (and U^{-1}) map boundary points to boundary points. Indeed $Ua = b$ and $Tc = d$ can readily be seen. □

Remark 2.33. In Figure 2.2, we see that the fundamental region $\mathcal{F}$ can also be described in terms of generalized disks (see Definition 1.50): With the terminology of Theorem 2.32, the boundary arcs a and b are indeed "generalized arcs" of the closed generalized disks $\mathbb{A}$ and $\mathbb{B}$. Their defining matrices are

$$\mathbb{A} : \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbb{B} : \begin{pmatrix} 0 & -1 \\ -1 & 1 \end{pmatrix}.$$

The boundary arcs c and d are part of the boundary of the closed unit disk $\mathbb{D}$, given by the matrix

$$\mathbb{D} : \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

We see that the fundamental region $\mathcal{F}$ can be characterized as set complement of the union of these three closed g-disks:

$$\mathcal{F} = \mathcal{H}^* \setminus (\mathbb{A} \cup \mathbb{B} \cup \mathbb{D}). \tag{2.35}$$

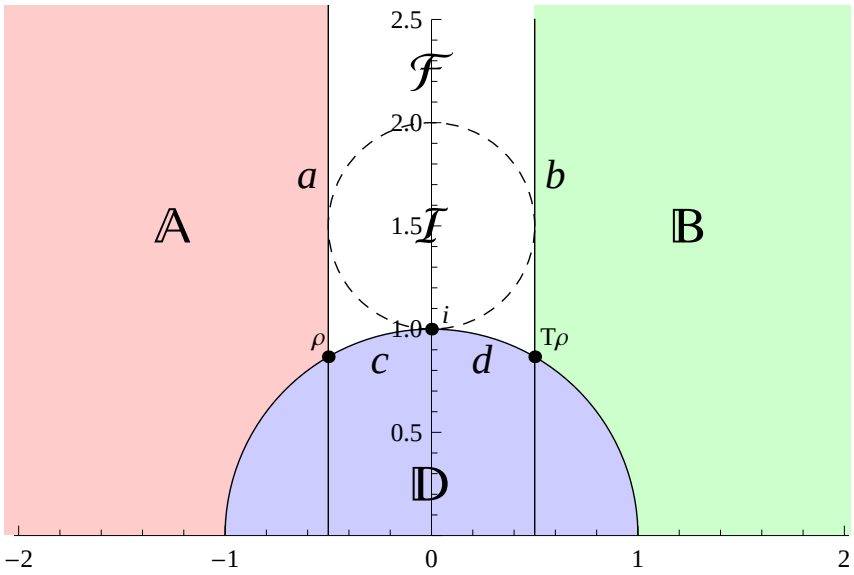

Figure 2.2: The fundamental region $\mathcal{F}$ for the action of the modular group $\mathrm{PSL}_2(\mathbb{Z})$ on the extended upper half-plane $\mathcal{H}^*$. It is bounded by "generalized arcs" a, b, c and d which correspond to the generalized disks $\mathbb{A}$, $\mathbb{B}$ and the unit disk $\mathbb{D}$. There is a unique disk $\mathcal{I} \subseteq \mathcal{F}$ which is tangent to $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{D}$.

Definition 2.34. Let the generalized disks $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{D}$ be defined as in Remark 2.33. The unique (open) g-disk $\mathcal{I} \subseteq \mathcal{F}$, which is tangent to $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{D}$ is called the *indisk* $\mathcal{I}$ of the fundamental region $\mathcal{F}$. Its defining matrix is given by

$$\mathcal{I} : \begin{pmatrix} 2 & -3i \\ 3i & 4 \end{pmatrix}. \tag{2.36}$$

2.3 The tessellation of the upper half-plane

Since $\mathcal{F}^*$ is a fundamental set for the action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathcal{H}^*$, its images under all modular transformations cover the extended upper half-plane $\mathcal{H}^*$ – compare also Remark 2.23. Thus for a point $z \in \mathcal{H}^*$ there exists a transformation $A \in \mathrm{PSL}_2(\mathbb{Z})$ such that $z \in A\mathcal{F}^*$. We can effectively determine such a transformation by adopting the algorithm of Theorem 2.29:

Theorem 2.35 (The fundamental set algorithm)**.** *Let $z \in \mathcal{H}^*$ be a point of the extended upper half-plane and let the fundamental set $\mathcal{F}^*$ be defined as in (2.34). A transformation A satisfying $z \in A\mathcal{F}^*$ can be found by performing the following steps:*

1. *Set $j := -1$ and $B_0 := 1$.*

2. *Increment j by one and set $z_j := B_j z$. If $z_j \in \mathcal{F}^*$, then goto step 6.*

3. *Set $e_j := \left\lfloor \mathrm{Re}\,(z_j) + \frac{1}{2} \right\rfloor$.*

4. *If $z_j - e_j \in \mathcal{F}^*$, set $B_{j+1} := U^{-e_j} B_j$ – else set $B_{j+1} := TU^{-e_j} B_j$.*

5. *Continue with step 2.*

6. *The desired matrix A is given by $A = {B_j}^{-1}$.*

Proof. Note that the above is essentially a reformulation of the algorithm of Theorem 2.29 with the following modifications:

(a) The algorithm is reformulated for the inhomogeneous case. The numbers z_j from above and the vectors x_j from the proof of Theorem 2.29 correspond by $z_j = \pi(x_j)$.

(b) Instead of using the nint () function, we use the above definition for determining the coefficients e_j – see step 3. This is to ensure that $z_j - e_j \in [-\frac{1}{2}, \frac{1}{2})$ – otherwise we would have problems with the termination of the algorithm for the case when some z_j lies on the boundary arc b of the fundamental domain $\mathcal{F}$ (see Figure 2.2).

(c) Theorem 2.29 yields a final vector x_n such that $w := \pi(x_n) \in \mathrm{cl}(\mathcal{R})$, where $\mathcal{R}$ is defined as in (2.29). In order to obtain a point in $\mathcal{F}^* \subseteq \mathrm{cl}(\mathcal{R})$, we need to apply to w possibly T and – if this point lies on the boundary arc b – possibly U^{-1}. We take this into account by explicitly checking whether the application of T in the last iteration of the algorithm is necessary or not (see step 4) and by the modification discussed in (b). □

Remark 2.36. The fundamental set algorithm also appears in Klein/Fricke [7], p. 212ff.,[6] in the proof that every point on the upper half-plane is equivalent to a point within $\mathcal{F}^*$. We note that the proof given there is different and more direct than ours. However, formulating and proving the algorithm for the homogeneous case first and carrying over the result to the inhomogeneous case might be considered to be more instructive.

In Figure 2.3 on the top-left, we see the images of the fundamental region $\mathcal{F}$ and its indisk $\mathcal{I}$ under the transformations of the modular group. Each of these images $A\mathcal{F}$ resp. $A\mathcal{I}$ is labeled by the T-U word representation of the corresponding transformation $A \in \mathrm{PSL}_2(\mathbb{Z})$. We call the covering of $\mathcal{H}^*$ by images of $\mathcal{F}^*$ the *modular tessellation* of the upper half-plane.

[6] Volume 1, part 2, chapter 2, §3.

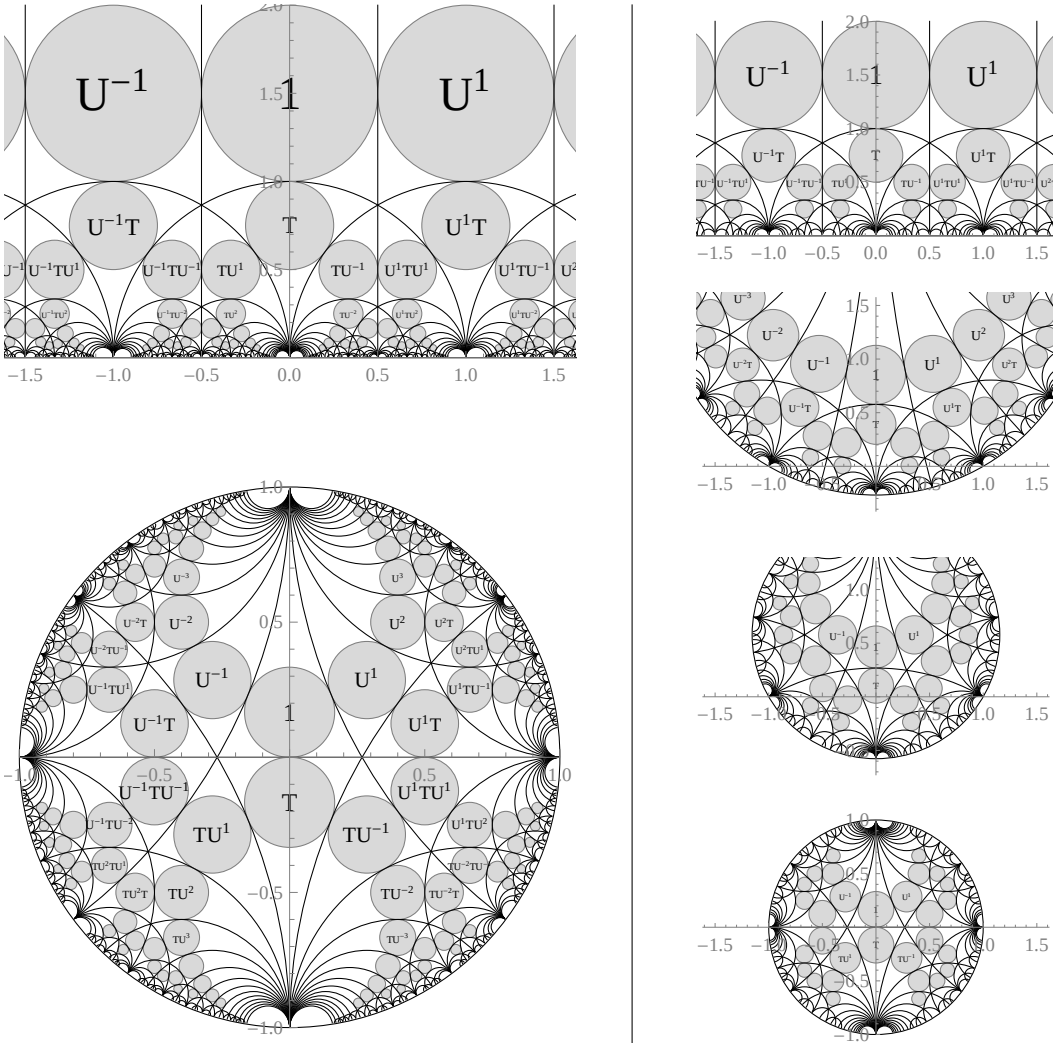

Figure 2.3: The modular tessellation. The images of the fundamental region $\mathcal{F}$ and its indisk $\mathcal{I}$ under all modular transformations (top left) are mapped to the unit circle by the modified Cayley transform Φ (bottom left). A continuous transition between these images is induced by a quarter-turn of the Riemann sphere and can be seen on the right.

For visualization purposes, the upper half-plane has the obvious disadvantage that we can never see the whole picture. We have seen in Example 1.43 that the modified Cayley transform Φ maps the upper half-plane to the unit disk. We can therefore use Φ to translate the modular tessellation to the unit disk. This means instead of looking at the regions $A\mathcal{F}$ resp. $A\mathcal{I}$ for all $A \in \mathrm{SL}_2(\mathbb{Z})$, we can alternatively depict the regions $\Phi A\mathcal{F}$ and the corresponding indisks $\Phi A\mathcal{I}$ which is done in the bottom-left picture of Figure 2.3.

For a better understanding of this transformed representation of the modular tessellation, let us identify the drawing area with $\mathbb{R}^2$ rather than with $\mathbb{C}$. Now we note that the point $(0,1)$ corresponds to the point $\infty \in \mathcal{H}^*$, the points $(\pm 1, 0)$ relate to $\pm 1 \in \mathcal{H}^*$ respectively and the point $(0,-1)$ refers to $0 \in \mathcal{H}^*$. The center $(0,0)$ represents the imaginary unit $i \in \mathcal{H}^*$. In other words, as can be seen in the right column of Figure 2.3, Φ bends the real axis to a circle[7], gluing together its ends at the point ∞ and enclosing the upper half-plane in its interior. As we have seen in Example 1.43, this continuous transition between the tessellation on the upper half-plane and its image under Φ can be explained by a quarter-turn of the Riemann sphere – compare also Figure 1.3.

Remark 2.37. In a more formal context, the bottom-left picture of Figure 2.3 can also be interpreted in two alternative ways: Firstly, we could consider a different action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathbb{C}_\infty$, which we may denote as $A * z$ for $A \in \mathrm{PSL}_2(\mathbb{Z})$ and $z \in \mathbb{C}_\infty$. For its definition we make use of the natural action of the Möbius transformation $\Phi A \Phi^{-1}$:

$$A * z := (\Phi A \Phi^{-1}) z.$$

Secondly, we could consider $\mathbb{C}_\infty$ under the natural action of a group G of Möbius transformations which is conjugate to $\mathrm{PSL}_2(\mathbb{Z})$ and whose transformations are represented by certain matrices[8] over the ring of Gaussian integers $\mathbb{Z}[i] := \{a + bi \mid a, b \in \mathbb{Z}\}$ with determinant 1:

$$G := \Phi\,\mathrm{PSL}_2(\mathbb{Z})\Phi^{-1} = \left\{ A \in \mathrm{PSL}_2(\mathbb{Z}[i]) \;\middle|\; A = \begin{pmatrix} \alpha & \beta \\ \overline{\beta} & \overline{\alpha} \end{pmatrix}_{\sim} \quad \alpha, \beta \in \mathbb{Z}[i] \right\}.$$

Clearly both, the action $*$ of $\mathrm{PSL}_2(\mathbb{Z})$ as well as the natural action of G, leave the unit disk $\Phi\mathcal{H}$ invariant. Consequently $\Phi\mathcal{F}$ is a fundamental region for both (equivalent) actions on the unit disk. In this setting, the bottom-left picture of Figure 2.3 can alternatively be interpreted as the tessellation induced by either of these two group actions.

[7] In fact, the real axis plus the point ∞ can be considered as a generalized circle.

[8] Compare also Mumford [9], p. 88, Recipe III.

2.4 Hyperbolic geometry

It is the goal of this section to introduce a general method for the construction of fundamental regions with respect to the action of $\mathrm{PSL}_2(\mathbb{Z})$ and its subgroups on $\mathcal{H}^*$ which relies on the concepts of 2-dimensional hyperbolic geometry. Brief introductions to the topic of hyperbolic geometry may also be found in Lehner [8], p. 78ff.[9] and Mumford [9], p. 377ff.

Plane (i.e. 2-dimensional) hyperbolic geometry is obtained from 2-dimensional Euclidean geometry by replacing the parallel postulate by the following axiom:

(A) *Let X be a point and L be a line on the (hyperbolic) plane, such that L does not pass through X. Then there is more than one line L' passing through X which does not meet L.*

There are various mathematical models of hyperbolic geometry. We will use a model based on the notions of generalized disks and generalized circles which we introduced in Section 1.2.2.

Definition 2.38 (Elements of hyperbolic geometry)**.** Let $\mathcal{P} \subseteq \mathbb{C}$ be a generalized disk, which serves as a model for 2-dimensional hyperbolic geometry. In this context, we will refer to $\mathcal{P}$ as *hyperbolic plane* (for short *h-plane*). The interior points of $\mathcal{P}$ are called *proper points*; the boundary points of $\mathcal{P}$ are called *improper points*. The set of all improper points (which is a generalized circle) is called the *horizon* of $\mathcal{P}$.[10] Every generalized circle C, which intersects the horizon of $\mathcal{P}$ orthogonally in two distinct (improper) points, gives rise to exactly one *hyperbolic line* (for short *h-line*) L which is given by $L = C \cap \mathcal{P}$.

In the above model of hyperbolic geometry, the angular measure is taken over from Euclidean geometry, i.e. the angle between by two h-lines which intersect each other in a proper point $z \in \mathcal{P}$ is defined as the Euclidean angle between by the (Euclidean) tangents of the h-lines at the point z. It now remains to introduce a measure for the distance between proper points in $\mathcal{P}$.

Definition 2.39 (Metric)**.** Let $\mathcal{S}$ be nonempty a set. A function $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is a *metric* on $\mathcal{S}$, if the following conditions are satisfied for all $x, y, z \in \mathcal{S}$:

(i) *Non-negativity:* $d(x, y) \geq 0$.

(ii) *Coincidence axiom:* $d(x, y) = 0$ if and only if $x = y$.

[9] Chapter II, section 12

[10] The notions of proper and improper points as well as horizon are taken over from Fenchel [3].

(iii) *Symmetry:* $d(x, y) = d(y, x)$.

(iv) *Triangle inequality:* $d(x, z) \leq d(x, y) + d(y, z)$.

Remark 2.40. From

$$0 = d(x, x) \leq d(x, y) + d(y, x) = 2d(x, y),$$

we see that non-negativity (i) is implied by the conditions (ii), (iii) and (iv).

Definition 2.41 (Isometry)**.** Let $\mathcal{S}$ be a set and d be a metric on $\mathcal{S}$. A map $\varphi : \mathcal{S} \to \mathcal{S}$ is called an *isometry*, if it leaves distances invariant, i.e. if

$$d(x, y) = d(\varphi(x), \varphi(y)) \quad \text{for all } x, y \in \mathcal{S}.$$

Remark 2.42. It is direct to see that the set of all bijective isometries forms a group under the operation of function composition. Note that the coincidence axiom (ii) implies that isometries are necessarily injective. However, in general they do not need to be surjective.

We wish to introduce a metric on the h-plane $\mathcal{P}$, such that every Möbius transformation which maps $\mathcal{P}$ onto itself is an isometry of $\mathcal{P}$. Clearly all Möbius transformations with this property form a group. We will refer to the transformations of this group as *rigid motions*[11] of the h-plane.

For the definition of such a metric we will take advantage of the so-called *cross ratio*. Note that in literature for the term "cross ratio" different notations and definitions are used. We will in this regard adhere to Carathéodory [2], whose derivation of the cross ratio's elementary properties is particularly concise and elegant.

Definition 2.43. Let $z_1, z_2, z_3, z_4 \in \mathbb{C}_\infty$ be numbers of the extended complex plane with the restriction that at most two of these numbers are equal. The *cross-ratio* $(z_1, z_2, z_3, z_4) \in \mathbb{C}_\infty$ is defined as

$$(z_1, z_2, z_3, z_4) := \frac{(z_1 - z_2)(z_3 - z_4)}{(z_1 - z_3)(z_2 - z_4)}. \tag{2.37}$$

Note that in the case of an infinite quantity $z_k = \infty$, the cross ratio shall be evaluated by formally dividing the two respective factors in the numerator and denominator of (2.37) by z_k and by substitution of expressions $\frac{1}{\infty}$ with zero.

We will need the following important properties of the cross ratio:

[11] A rigid motion is an isometry which additionally leaves angles and their orientation invariant.

Lemma 2.44 (Invariance under Möbius transformations)**.** *Let $\varphi \in \mathrm{PGL}_2(\mathbb{C})$ be a Möbius transformation and $z_1, z_2, z_3, z_4 \in \mathbb{C}_\infty$ such that their cross ratio is defined. Set $w_j := \varphi(z_j)$ for $j \in \{1,2,3,4\}$. The cross ratios of the numbers z_j and w_j are equal, i.e.*

$$(z_1, z_2, z_3, z_4) = (w_1, w_2, w_3, w_4). \tag{2.38}$$

Sketch of proof. Let us write the Möbius transformation φ as $\varphi(z) = \frac{az+b}{cz+d}$. For $i, j \in \{1,2,3,4\}$, we have

$$w_i - w_j = \frac{az_i + b}{cz_i + d} - \frac{az_j + b}{cz_j + d} = \frac{ad - bc}{(cz_i + d)(cz_j + d)} \cdot (z_i - z_j).$$

Consequently, if we define

$$A := \frac{(ad - bc)^2}{(cz_1 + d)(cz_2 + d)(cz_3 + d)(cz_4 + d)},$$

we obtain

$$(w_1 - w_2)(w_3 - w_4) = A(z_1 - z_2)(z_3 - z_4).$$

On the other hand, exploiting the symmetry of A with respect to the numbers z_j, we have

$$(w_1 - w_3)(w_2 - w_4) = A(z_1 - z_3)(z_2 - z_4).$$

In the case when all involved numbers z_j and w_j are finite (and in particular A is finite and nonzero), division of these two equations yields (2.38). The prove of the general case, when one or two of the numbers z_j (resp. w_j) are ∞, involves some inconvenient case distinctions which we will not carry out here. □

Lemma 2.45. *Let $z_1, z_2, z_3 \in \mathbb{C}_\infty$ be pairwise distinct. The function*

$$\lambda : \begin{cases} \mathbb{C}_\infty & \to & \mathbb{C}_\infty \\ z & \mapsto & (z_1, z_2, z_3, z) \end{cases} \tag{2.39}$$

is a Möbius transformation and satisfies $\lambda(z_1) = 1$, $\lambda(z_2) = \infty$, $\lambda(z_3) = 0$.

Sketch of proof. In the case when $\infty \notin \{z_1, z_2, z_3\}$, by setting $a := z_2 - z_1$, $b := z_2(z_1 - z_2)$, $c := z_3 - z_1$ and $d := z_2(z_1 - z_3)$ we can obviously write

$$\lambda(z) = \frac{(z_1 - z_2)(z_3 - z)}{(z_1 - z_3)(z_2 - z)} = \frac{az + b}{cz + d}.$$

Moreover, we have

$$ad - bc = (z_1 - z_2)(z_2 - z_3)(z_3 - z_1).$$

By assumption the numbers z_j are pairwise distinct and therefore λ is indeed a Möbius transformation. Note that for the other case, when some $z_k = \infty$, the cross ratio needs to be evaluated as specified in Definition 2.43 and the coefficients a, b, c, d need to be adapted appropriately. The statement on the images of z_1, z_2, z_3 under λ is readily verified. □

Corollary 2.46. *Denoting the extended real axis by $\mathbb{R}_\infty := \mathbb{R} \cup \{\infty\}$, the map $\lambda : \mathbb{C}_\infty \to \mathbb{C}_\infty$ from (2.39) satisfies $\lambda(z) \in \mathbb{R}_\infty$ if and only if z lies on the generalized circle which is determined by the points z_1, z_2, z_3.*

Proof. This statement is a consequence of the fact that λ is a Möbius transformation and therefore maps the g-circle determined by the points z_1, z_2, z_3 pointwise one-to-one to the extended real axis $\mathbb{R}_\infty$. □

The above properties of the cross ratio allow to define a metric on the hyperbolic plane $\mathcal{P}$ the following way:

Definition 2.47 (Poincaré metric)**.** Let $z_1, z_2 \in \mathcal{P}$ be two distinct proper points of the hyperbolic plane and let L be the unique h-line joining z_1 with z_2. Label the two (improper) endpoints of L by ∞_1 and ∞_2, such that $\infty_1, z_1, z_2, \infty_2$ are in order along the h-line L. The hyperbolic distance between z_1 and z_2 is defined as

$$d_{\text{hyp}}(z_1, z_2) := \frac{1}{2} \log \chi \quad \text{with } \chi := (z_1, \infty_2, \infty_1, z_2). \tag{2.40}$$

Additionally we define $d_{\text{hyp}}(z, z) = 0$ for all proper points $z \in \mathcal{P}$.

Theorem 2.48. *The hyperbolic distance defined in* (2.40) *is a metric on the hyperbolic plane $\mathcal{P}$ which we call the* Poincaré metric *on $\mathcal{P}$.*

Sketch of proof. The cross ratio χ in (2.40) may alternatively be written as $\chi = \lambda(z_2)$, where $\lambda(z) := (z_1, \infty_2, \infty_1, z)$ is, according to Lemma 2.45, a Möbius transformation satisfying $\lambda(z_1) = 1$ and $\lambda(\infty_2) = \infty$. Corollary 2.46 ensures that $\lambda(z)$ is real for all z lying on the h-line L. Since λ is a continuous map, it follows that λ maps the h-line segment $S \subseteq L$ which is bounded by the points z_1 and ∞_2 pointwise one-to-one to the interval $[1, \infty] \subseteq \mathbb{R}_\infty$. For this reason, the hyperbolic metric is non-negative and $d_{\text{hyp}}(z_1, z_2) \to 0$ as z_2 approaches z_1 along S. Similarly, $d_{\text{hyp}}(z_1, z_2) \to \infty$ as z_2 approaches ∞_2 along S.

The symmetry of d_{hyp} can easily be verified: Note that exchanging z_1 and z_2 also swaps the roles of the improper points ∞_1 and ∞_2. We therefore need

to show that $\chi_1 := (z_1, \infty_2, \infty_1, z_2) = (z_2, \infty_1, \infty_2, z_1) =: \chi_2$ which is indeed the case:

$$\chi_1 = \frac{(z_1 - \infty_2)(\infty_1 - z_2)}{(z_1 - \infty_1)(\infty_2 - z_2)} = \frac{(z_2 - \infty_1)(\infty_2 - z_1)}{(z_2 - \infty_2)(\infty_1 - z_1)} = \chi_2.$$

By the above observations, we have $d_{\text{hyp}}(z_1, z_2) = 0$ if and only if $z_1 = z_2$. Consequently the coincidence axiom is satisfied. The proof of the fact that d_{hyp} also satisfies the triangle inequality is out of scope of this work. □

Remark 2.49. The hyperbolic distance along a h-line is additive, i.e. for three points z_1, z_2, z_3 which are in order along a h-line, we have

$$d_{\text{hyp}}(z_1, z_3) = d_{\text{hyp}}(z_1, z_2) + d_{\text{hyp}}(z_2, z_3),$$

which is a consequence of

$$\frac{(z_1 - \infty_2)(\infty_1 - z_2)}{(z_1 - \infty_1)(\infty_2 - z_2)} \cdot \frac{(z_2 - \infty_2)(\infty_1 - z_3)}{(z_2 - \infty_1)(\infty_2 - z_3)} = \frac{(z_1 - \infty_2)(\infty_1 - z_3)}{(z_1 - \infty_1)(\infty_2 - z_3)}$$

Remark 2.50 (Poincaré disk and half-plane model). If we choose the upper half-plane of $\mathbb{C}_\infty$ as model for the hyperbolic plane, i.e. $\mathcal{P} := \text{cl}(\mathcal{H})$, we obtain the *Poincaré half-plane model* of hyperbolic geometry. If instead the unit disk is chosen, $\mathcal{P} := \mathbb{D}$, then we talk of the *Poincaré disk model*.

Clearly both models are equivalent in the sense that one model can be obtained from the other by applying a Möbius map which transforms the upper half-plane to the unit disk (or vice-versa). For our purposes, the Poincaré half-plane model will be most relevant.

2.4.1 Normal polygons and fundamental regions

We can now exploit the concepts of 2-dimensional hyperbolic geometry in the search of fundamental regions for the action of $\text{PSL}_2(\mathbb{Z})$ and its subgroups on $\mathcal{H}^*$.

The idea – taken from Lehner [8] – may be shortly described as follows: Given a subgroup $G \leq \text{PSL}_2(\mathbb{Z})$, we fix a point $z_0 \in \mathcal{H}$ which has a trivial stabilizer $G_{z_0} = \{1\}$. The points z of the sought fundamental region $\mathcal{N}_{z_0}$ shall be characterized by the property $d_{\text{hyp}}(z, z_0) \leq d_{\text{hyp}}(Az, z_0)$ for all $A \in G$. It turns out that such a fundamental region $\mathcal{N}_{z_0}$ can be obtained by a simple geometric construction for which we will need the following notions:

Definition 2.51 (Hyperbolic half-plane)**.** Let D be a generalized disk, such that its boundary intersects the horizon of $\mathcal{P}$ orthogonally in two distinct (improper) points. The (nonempty) set $H := D \cap \mathcal{P}$ is called a *hyperbolic half-plane*.

Definition 2.52 (Perpendicular bisector)**.** Let $z_1, z_2 \in \mathcal{P}$ be two distinct proper points and denote by S the unique h-line segment whose endpoints are z_1 and z_2. The set of proper points w for which $d_{\text{hyp}}(w, z_1) = d_{\text{hyp}}(w, z_2)$ defines a h-line and is called the *perpendicular bisector* of S. Additionally we define $H_{z_1}(z_2)$ as the hyperbolic half-plane consisting of all points $w \in \mathcal{P}$ which are closer to z_1 than to z_2:

$$H_{z_1}(z_2) := \{w \in \mathcal{P} \mid d_{\text{hyp}}(w, z_1) < d_{\text{hyp}}(w, z_2)\}. \tag{2.41}$$

Note that the hyperbolic boundary of $H_{z_1}(z_2)$ is precisely the perpendicular bisector of the h-line segment S.

Definition 2.53 (Normal polygon)**.** Let $G \leq \mathrm{PSL}_2(\mathbb{Z})$ be a subgroup of the modular group. Let $z \in \mathcal{H}^*$ with trivial stabilizer $G_z = \{1\}$. The set

$$\mathcal{N}_z := \bigcap_{A \in G \setminus \{1\}} H_z(Az) \tag{2.42}$$

is called the *normal polygon* with respect to the group G and the point z.

Remark 2.54. The term "normal polygon" is justified by the fact that $\mathcal{N}_z$, as intersection of hyperbolic half-planes, is bounded entirely by h-line segments and therefore can be considered as a convex (possibly generalized) polygon[12] in the hyperbolic sense.

Theorem 2.55. *Let $\mathcal{N}_z$ be a normal polygon for the group $G \leq \mathrm{PSL}_2(\mathbb{Z})$ with respect to the point $z \in \mathcal{H}$. $\mathcal{N}_z$ is a fundamental region for the action of G on $\mathcal{H}^*$.*

Proof. For the proof of this fact in the more general context of *principal circle groups*, i.e. discontinuous groups of Möbius transformations fixing a given generalized circle, we refer to Lehner [8], p.146ff.[13] □

Example 2.56 (Fundamental regions for $\mathrm{PSL}_2(\mathbb{Z})$)**.** We see in Figure 2.4 that the fundamental region $\mathcal{F}$ from (2.33) can alternatively be obtained by constructing the normal polygon for a point z on the imaginary axis with $\mathrm{Im}(z) > 1$ (compare also Figure 2.2). A different fundamental region for the action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathcal{H}^*$ is displayed in Figure 2.5.

In both figures, the point z for which the normal polygon is constructed is colored red. Its equivalent points Az, $A \neq 1$ can be seen in black. For every

[12] A region bounded by infinitely many h-line segments is called a generalized polygon.
[13] Chapter IV, Section 7

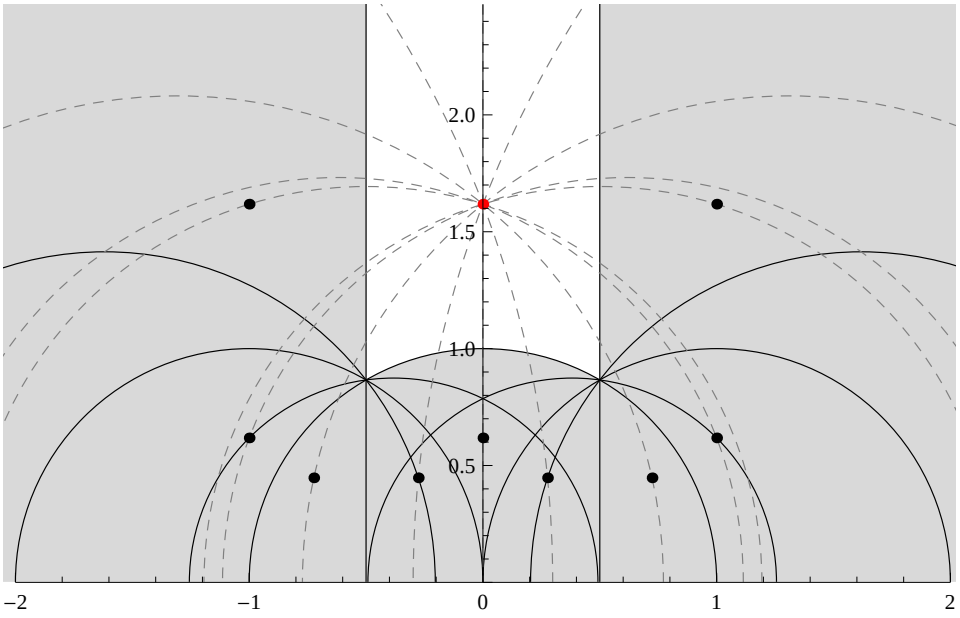

Figure 2.4: The fundamental region $\mathcal{F}$ can alternatively obtained by constructing the normal polygon with respect to a point z on the imaginary axis with $\mathrm{Im}\,(z) > 1$. Above the point $z = \varphi i$ (red) has been chosen, where $\varphi = \frac{\sqrt{5}+1}{2}$ denotes the golden ratio.

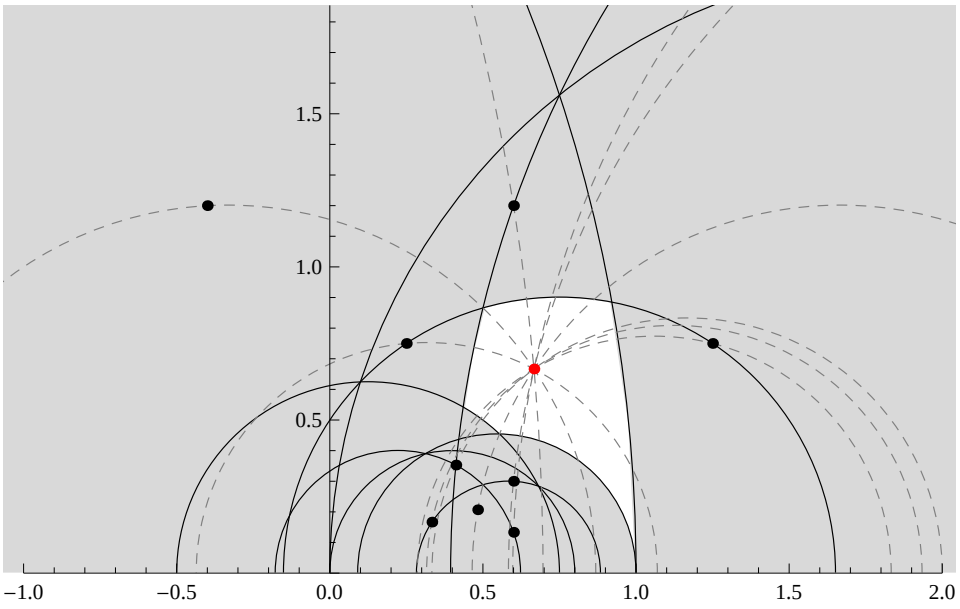

Figure 2.5: An alternative fundamental region for the action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathcal{H}^*$. It is obtained by constructing the normal polygon for the point $\frac{2}{3}(1+i)$.

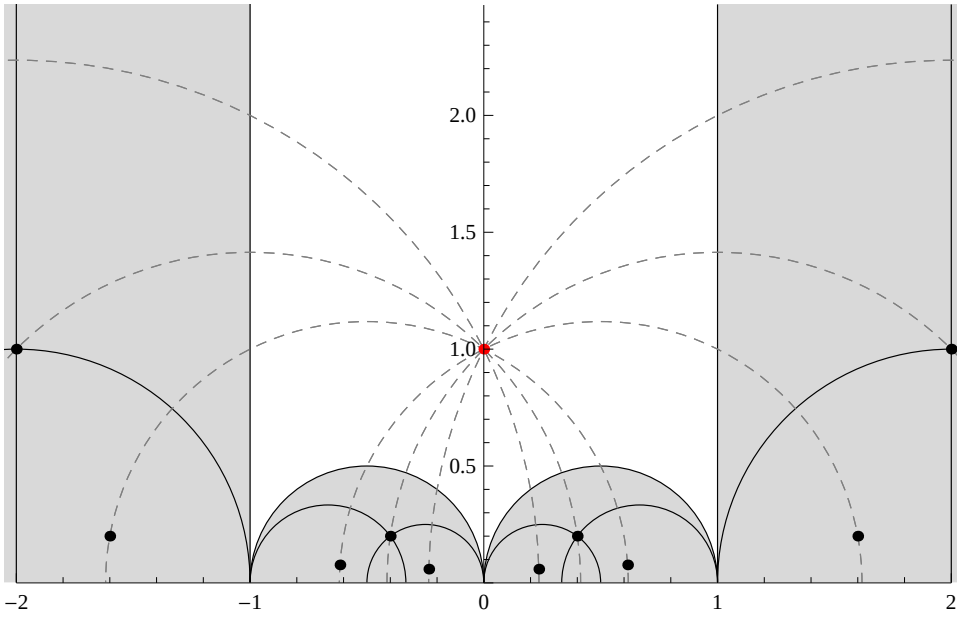

Figure 2.6: A fundamental region for the subgroup $\Gamma(2) \leq \mathrm{PSL}_2(\mathbb{Z})$. It is given by the normal polygon constructed with respect to the point i.

pair (z, Az), the corresponding perpendicular bisector (black) can be found best by following the gray dashed h-line which joins Az with z.

Note that for producing an accurate picture of the normal polygon, it was in both cases sufficient to enumerate just 9 different transformations. In case of Figure 2.4, all transformations with $n(A) \leq 3$ have been selected, where $n(A)$ denotes the grading of A as defined in (2.11).

In Figure 2.5, in order to achieve a good "locality" of the equivalent points Az around z, additionally the fundamental set algorithm (Theorem 2.35) has been utilized: The selected transformations are all of the form $A = CBC^{-1}$ with $B \in \mathrm{PSL}_2(\mathbb{Z})$, $n(B) \leq 3$ and where C denotes the transformation obtained when applying the fundamental set algorithm to z.

Example 2.57. As a demonstration of the normal polygon method in case of a proper subgroup of $\mathrm{PSL}_2(\mathbb{Z})$, we see in Figure 2.6 a normal polygon for the group $\Gamma(2) \leq \mathrm{PSL}_2(\mathbb{Z})$. For $m \in \mathbb{N}$, the *principal congruence subgroup* $\Gamma(m)$ is defined as

$$\Gamma(m) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z}) \;\middle|\; a \equiv d \equiv \pm 1,\ b \equiv c \equiv 0 \mod m \right\}.$$

For drawing Figure 2.6, all nontrivial transformations in $\Gamma(2)$ with a grading ≤ 18 (of which there are 12) have been selected. Note that for the particular

chosen point z, also "the first four" nontrivial transformations of $\Gamma(2)$ (whose grading is ≤ 6) would have been enough for obtaining the same region.

Chapter 3

Applications in visualization

3.1 The action of Möbius transformations

So far we often talked about the action of Möbius transformations on $\mathbb{C}_\infty$. However, up to now occasions have been quite rare where we literally could see them "in action". Maybe the most important exceptions to this are Figures 1.2 and 1.3, showing continuous transitions between certain sets and their images under the two transformations $z \mapsto \frac{1}{z}$ and $z \mapsto \Phi(z)$ (the modified Cayley transform), both induced by rotations of the Riemann sphere.

In this section we introduce a more direct method for visualization of such continuous transitions, a method working for arbitrary Möbius transformations and doing without stereographic projection and motion of Riemann spheres. For this purpose, we exploit the "linear algebra nature" (see Remark 1.36) of Möbius transformations. For easier notation we will not distinguish between a matrix $A \in \mathrm{GL}_2(\mathbb{C})$ and the corresponding Möbius transformation ${}_{\pm}A \in \mathrm{PGL}_2(\mathbb{C})$. In particular with A we will denote both, a matrix and its associated transformation.

For $A \in \mathrm{GL}_2(\mathbb{C})$ and $k \in \mathbb{Z}$, it follows from Theorem 1.35 that the matrix power A^k corresponds to the Möbius transformation obtained by k times composing the transformation A with itself. The idea is now to generalize the concept of matrix powers from integral to real exponents. If we do so, then for any set $S \in \mathbb{C}_\infty$ of interest, we can visualize the transition from S to its image under A simply by depicting a sequence of intermediate images $A^t S \subseteq \mathbb{C}_\infty$, with a varying parameter $t \in [0, 1]$.

In order to introduce such *generalized matrix powers*, note that for integral exponents, powers of A can be calculated by using its *Jordan normal form*.[1]

[1] Also called *Jordan canonical form*. For more details on Jordan normal forms, eigenvalues and eigenvectors see for example Hungerford [5], Chapter VII, Linear algebra.

If J is a Jordan normal form of A, then for some $P \in \mathrm{GL}_2(\mathbb{C})$ we have $A = P^{-1}JP$ and consequently

$$A^k = P^{-1}J^kP, \quad \text{for all } k \in \mathbb{Z}. \tag{3.1}$$

The matrix J has one of the two possible forms

$$(i)\ J = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \quad \text{or} \quad (ii)\ J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

and their respective matrix powers are given by

$$(i)\ J^k = \begin{pmatrix} \lambda_1^k & 0 \\ 0 & \lambda_2^k \end{pmatrix} \quad \text{or} \quad (ii)\ J^k = \begin{pmatrix} \lambda^k & k\lambda^{k-1} \\ 0 & \lambda^k \end{pmatrix}. \tag{3.2}$$

From here it is just a small step to the generalization of matrix powers to real exponents: We choose a fixed branch of the natural (complex) logarithm, for example such that the imaginary part of the logarithm ranges in the interval $(-\pi, \pi]$, i.e. $\mathrm{Im}\,(\ln z) = \arg z \in (-\pi, \pi]$ for all $z \in \mathbb{C}$. Now, for $\lambda \in \mathbb{C}$ and $k \in \mathbb{R}$, we can evaluate λ^k as $\lambda^k := \exp(k \ln \lambda)$.

Definition 3.1 (Generalized matrix power)**.** Let $A \in \mathrm{GL}_2(\mathbb{C})$ be a matrix. For $k \in \mathbb{R}$ we say B is an k-th power of A, in symbols $B = A^k$, if there is a $P \in \mathrm{GL}_2(\mathbb{C})$ such that

$$A = P^{-1}JP \quad \text{and} \quad B = P^{-1}J^kP,$$

where J is in Jordan normal form and J^k is determined by (3.2), where a fixed branch of the natural logarithm is chosen and the expressions λ^k are evaluated as $\lambda^k := \exp(k \ln \lambda)$.

Remark 3.2 (Eigenvectors and fixed points). Writing the matrix P in the form $P = (v_1 \mid v_2)$ with $v_1, v_2 \in \mathbb{C}^2$, in case (i), v_1 and v_2 are the *eigenvectors* of the matrix A corresponding to the *eigenvalues* λ_1 and λ_2 respectively. In case (ii), only v_1 is an eigenvector (v_2 is a so-called *generalized eigenvector*). It is worth noting that a vector $\binom{u}{v} \in \mathbb{C}^2$ is an eigenvector of $A \in \mathrm{GL}_2(\mathbb{C})$, if and only if $u/v \in \mathbb{C}_\infty$ is a fixed point for the Möbius transformation A:

$$A \cdot \begin{pmatrix} u \\ v \end{pmatrix} = \lambda \begin{pmatrix} u \\ v \end{pmatrix} \quad \Leftrightarrow \quad A\left(\frac{u}{v}\right) = \frac{\lambda u}{\lambda v} = \frac{u}{v}.$$

Therefore, in case (i), A is either the identity transformation or has exactly two distinct fixed points in $\mathbb{C}_\infty$. In case (ii), A has exactly one fixed point in $\mathbb{C}_\infty$.

As Definition 3.1 already suggests, generalized matrix powers are not unique but depend on the chosen branch of the natural logarithm. The following example will illustrate this:

Example 3.3. Let $A = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{C})$ be the matrix corresponding to the Möbius map $z \mapsto \frac{1}{z}$. We wish to calculate the "square root" $A^{\frac{1}{2}}$ of this transformation. The eigenvalues of A are $\lambda_1 = -1$ and $\lambda_2 = 1$; the corresponding eigenvectors are $v_1 = \left(\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right)$ and $v_2 = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$. We can therefore write A as $A = P^{-1}JP$ with $P = \left(\begin{smallmatrix} -1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ and $J = \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. Next we calculate $J^{\frac{1}{2}}$ by evaluating $\exp(\frac{1}{2}\ln\lambda_1)$ and $\exp(\frac{1}{2}\ln\lambda_2)$. Choosing a logarithm branch such that $\mathrm{Im}\,(\ln z) \in (-\pi, \pi]$ for all $z \in \mathbb{C}$, and denoting equivalence of matrices in $\mathrm{PGL}_2(\mathbb{C})$ again by $\sim$, this yields $J^{\frac{1}{2}} = \left(\begin{smallmatrix} i & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and

$$A^{\frac{1}{2}} = P^{-1}J^{\frac{1}{2}}P = \frac{1}{2}\begin{pmatrix} 1+i & 1-i \\ 1-i & 1+i \end{pmatrix} \sim \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix}.$$

We therefore see that $A^{\frac{1}{2}} = \Phi$ and the modified Cayley transform is in this sense a "square root" of the transformation $A : z \mapsto \frac{1}{z}$ – compare also Example 1.43.

In contrast to that, choosing a slightly different logarithm branch, such that $\mathrm{Im}\,(\ln z) \in [-\pi, \pi)$ for all $z \in \mathbb{C}$, we get $J^{\frac{1}{2}} = \left(\begin{smallmatrix} -i & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and therefore obtain a matrix conjugate to the above one:

$$A^{\frac{1}{2}} = \frac{1}{2}\begin{pmatrix} 1-i & 1+i \\ 1+i & 1-i \end{pmatrix} \sim \begin{pmatrix} -i & 1 \\ 1 & -i \end{pmatrix}.$$

Note that this second transformation, just like the modified Cayley transformation, can also be considered as a quarter-turn of the Riemann sphere around the x_1 axis, but it rotates in the opposite direction and maps the *lower* half-plane to the unit disk.

Example 3.4. As an application of generalized matrix powers, we can visualize the action of the modular transformations $U : z \mapsto z + 1$, $T : z \mapsto -\frac{1}{z}$ and $R : z \mapsto -\frac{1}{z+1}$ on the modular tessellation. For the parameter $t \in [0, 1]$ "intermediate actions" of these maps are given through the Möbius maps corresponding to the matrices

$$\begin{aligned}
U^t &\sim \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \\
T^t &\sim \begin{pmatrix} \cos\left(\frac{\pi t}{2}\right) & -\sin\left(\frac{\pi t}{2}\right) \\ \sin\left(\frac{\pi t}{2}\right) & \cos\left(\frac{\pi t}{2}\right) \end{pmatrix}, \\
R^t &\sim \begin{pmatrix} \sqrt{3}\cos\left(\frac{\pi t}{3}\right) - \sin\left(\frac{\pi t}{3}\right) & -2\sin\left(\frac{\pi t}{3}\right) \\ 2\sin\left(\frac{\pi t}{3}\right) & \sqrt{3}\cos\left(\frac{\pi t}{3}\right) + \sin\left(\frac{\pi t}{3}\right) \end{pmatrix}.
\end{aligned}$$

In Figures 3.1, 3.2 and 3.3 we can see the modular tessellation and its "intermediate images" under the transformations T, U and R for the parameter values $t \in \{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\}$. Again we use the modified Cayley transform to map the tessellation of the upper half-plane to the unit disk. Looking at Figure 3.2, another advantage of seeing the upper half-plane from this different angle gets apparent: Under Φ, the action of T corresponds to a simple rotation by 180° around the fixed point i. This is in close connection to Remark 1.42, where we noted that T corresponds to a half-turn of the Riemann sphere around the x_2 axis. Finally we can see in Figure 3.3 that – with a slight distortion – R corresponds to a rotation by 120° around the fixed point $\rho = \exp(2\pi i/3)$.

Note that the individual frames in Figures 3.1, 3.2 and 3.3 have been arranged such that the first column has to be read first in top-down direction and then the second column has to be read in bottom-up direction, allowing in the first row a direct comparison of the original tessellation and its image under the respective transformations.

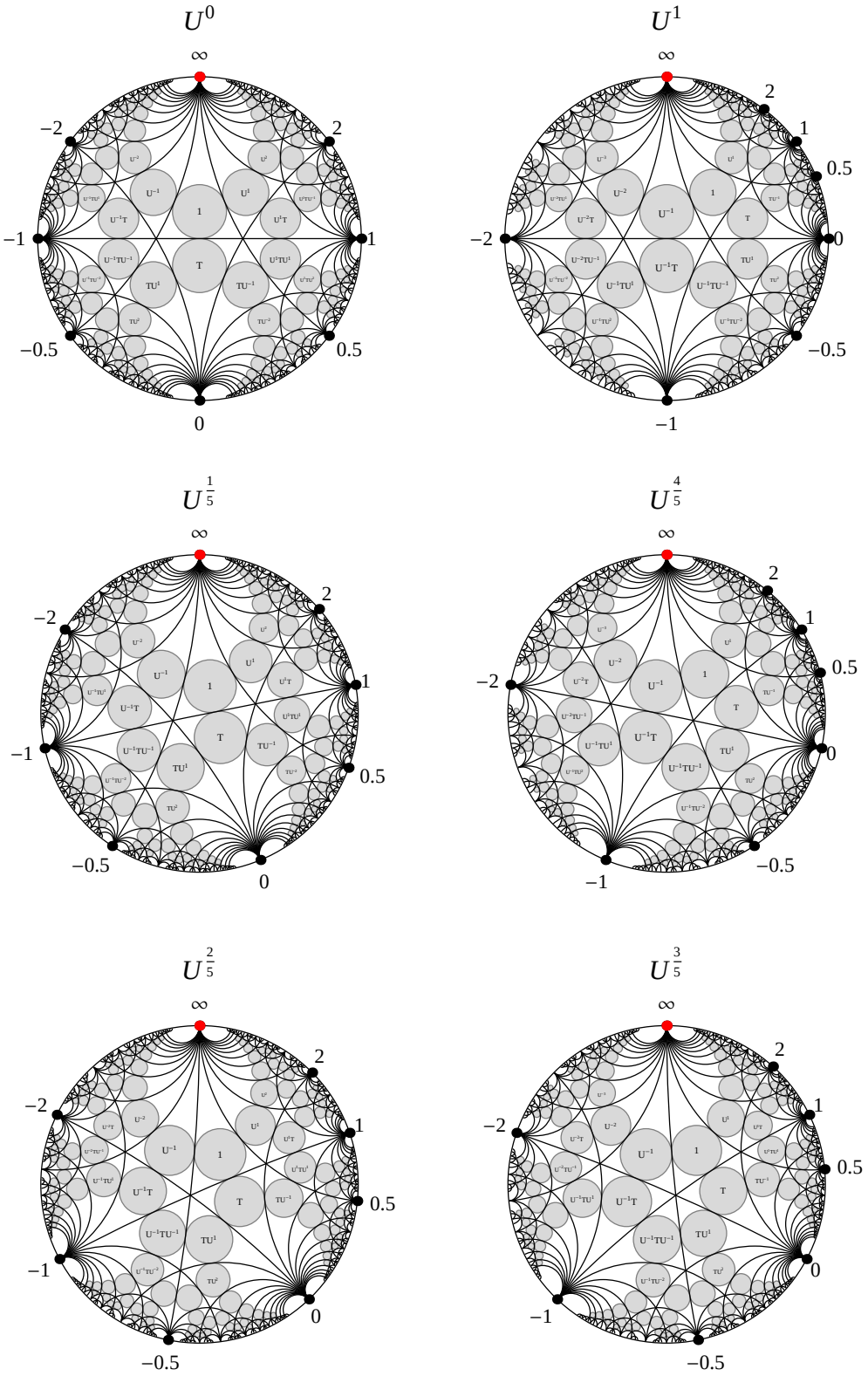

Figure 3.1: The action of $U : z \mapsto z + 1$.

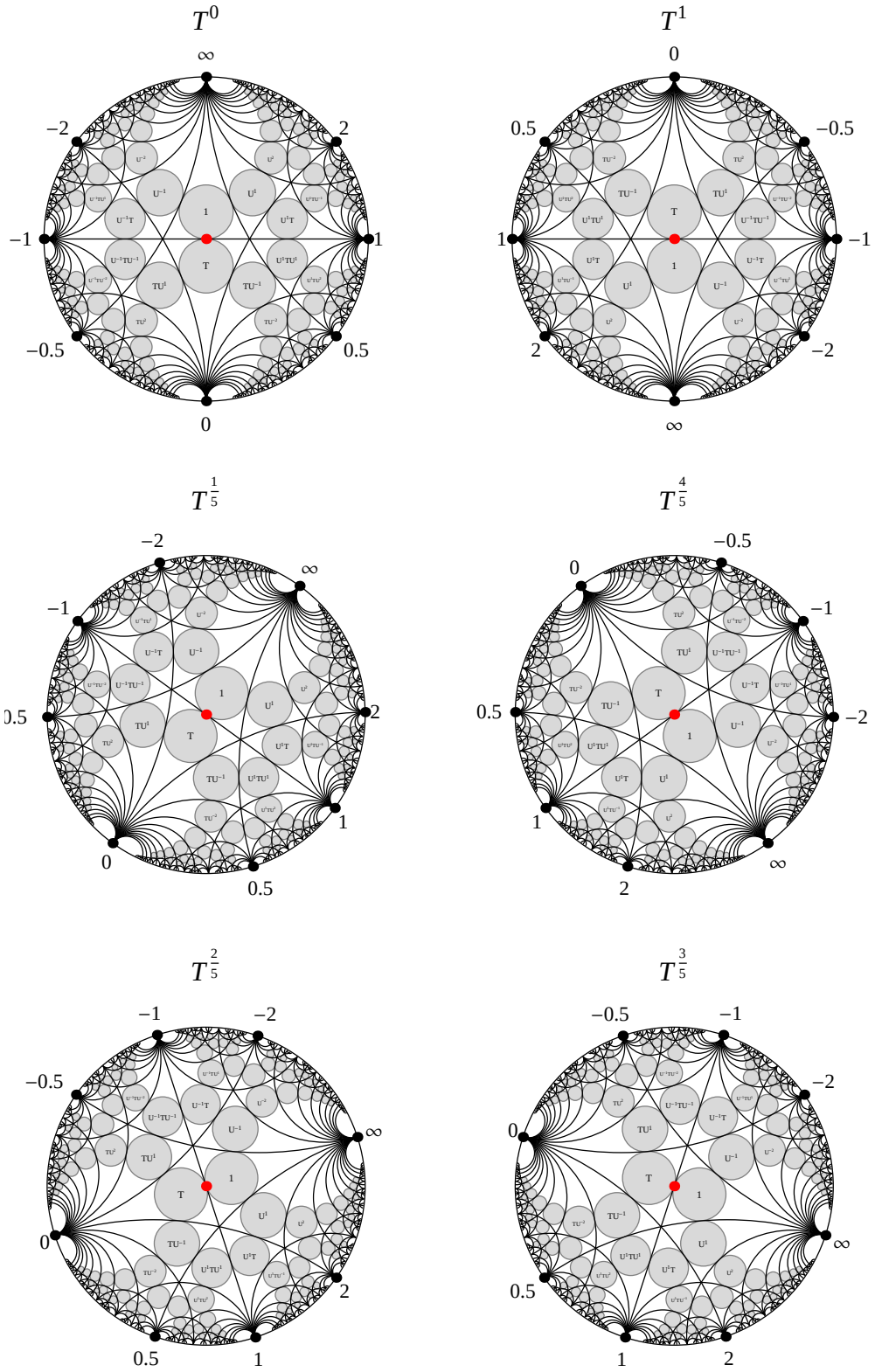

Figure 3.2: The action of $T : z \mapsto -\frac{1}{z}$.

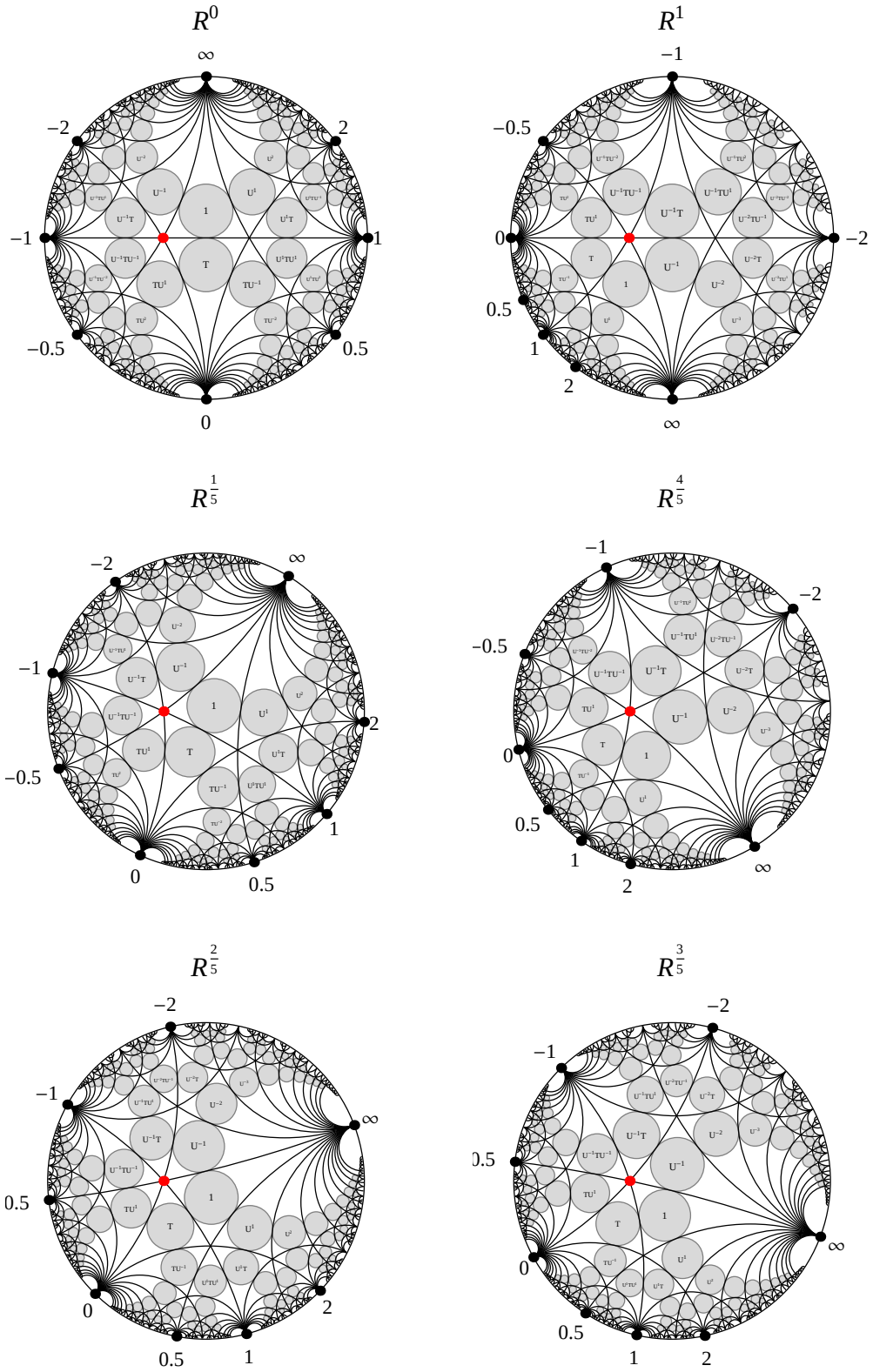

Figure 3.3: The action of $R : z \mapsto -\frac{1}{z+1}$.

3.2 Continued fractions

For an arbitrary modular transformation A, a representation as product of shifts $U^j : z \mapsto z + j$ and inversions $T : z \mapsto -\frac{1}{z}$ can be found by the T-U algorithm of Corollary 2.10. By writing out this product, for example in the case when $n = 2$, we have

$$A = U^{e_0} T U^{e_1} T U^{e_2} T U^k,$$

or more explicitly

$$A(z) = e_0 - \cfrac{1}{e_1 - \cfrac{1}{e_2 - \cfrac{1}{k+z}}}. \tag{3.3}$$

Here a close relation between modular transformations and continued fractions immediately gets apparent. In this section we will investigate this relation somewhat deeper. First we will use Pringsheim's more space-saving notation for continued fractions, namely

$$b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \dots}}} =: b_0 + \frac{a_1}{\lfloor b_1 \rfloor} + \frac{a_2}{\lfloor b_2 \rfloor} + \frac{a_2}{\lfloor b_3 \rfloor} + \dots \tag{3.4}$$

In the case when all $a_j = 1$, we adhere to the standard sequence notation for continued fractions:

$$b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \dots}} =: [b_0, b_1, b_2, \dots].$$

For determining a continued fraction representation for a real number α such that all $a_j = 1$, one usually sets $\alpha_0 := \alpha$ as well as $b_j := \lfloor \alpha_j \rfloor$ and $\alpha_{j+1} := \frac{1}{\alpha_j - b_j}$ for $j \geq 0$ until some α_j is zero (which is the case if and only if $\alpha \in \mathbb{Q}$). In this way we obtain a finite or infinite sequence of equations

$$\alpha = \alpha_0 = b_0 + \frac{1}{\alpha_1}, \quad \alpha_1 = b_1 + \frac{1}{\alpha_2}, \quad \alpha_2 = b_2 + \frac{1}{\alpha_3}, \quad \dots \tag{3.5}$$

giving rise to the continued fraction representation $\alpha = [b_0, b_1, b_2, \dots]$. The rational number $c_n := [b_0, b_1, \dots, b_n]$ obtained by truncating the continued fraction representation after the coefficient b_n, is called the n-th *convergent* of the continued fraction. If the continued fraction is infinite, i.e. $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, then we have $\lim_{n\to\infty} c_n = \alpha$.

Remark 3.5. Note that by using the method above, all the coefficients b_j with $j > 0$ are positive. A representation for α of this form is called a *regular continued fraction* – see also Perron [11], §9.

In contrast to that, if we set $b_j := \lceil \alpha_j \rceil$ for some or all of the indices j, then also negative coefficients b_j, $j > 0$, will occur and we obtain in this way a so called *semi-regular continued fraction* representation of α. This is in strong analogy to Remark 2.11 that within the Euclidean algorithm the quotients q_j can be determined by rounding r_{j-1}/r_j either up- or downward. Note that in Perron [11], §36, semi-regular continued fractions are defined such that for all $j > 0$ the coefficients b_j are positive, but allowing for $a_j \in \{\pm 1\}$. However this makes no essential difference.

If we use the nearest integer function in each step, i.e. $b_j := \operatorname{nint}(\alpha_j)$ for all j, then we have $|b_j| \geq 2$ for all $j > 0$. If additionally $\alpha \in \mathbb{Q}$, then it can be shown that the resulting continued fraction representation is one of minimal length – according to Perron [11], §39, we call finite continued fractions with this minimality property *canonical continued fractions*.

We can now reformulate Corollary 2.10 in order to construct a continued fraction representation of any given modular transformation.

Corollary 3.6. *An arbitrary modular transformation $A(z) = \frac{az+b}{cz+d}$ can be written as continued fraction*

$$A(z) = [q_0, q_1, \ldots, q_n, (-1)^{n+1}(k+z)] \tag{3.6}$$

where the integers n, $q_0, q_1, \ldots, q_n$ and k are determined by the T-U algorithm from Corollary 2.10.

Proof. By using the continued fraction representation of A given in (3.3) and by applying the definition $e_j := (-1)^j q_j$, we see

$$\begin{aligned} A(z) &= e_0 + \cfrac{-1}{e_1} + \cfrac{-1}{e_2} + \ldots + \cfrac{-1}{e_n} + \cfrac{-1}{k+z} \\ &= q_0 + \cfrac{-1}{-q_1} + \cfrac{-1}{q_2} + \ldots + \cfrac{-1}{(-1)^n q_n} + \cfrac{-1}{k+z}. \end{aligned} \tag{3.7}$$

Now for every odd $j \leq n$ we can rewrite

$$\cfrac{-1}{-q_j} + \cfrac{-1}{\ldots} \quad \text{to} \quad \cfrac{1}{q_j} + \cfrac{1}{\ldots}.$$

Thus if n is odd, every numerator -1 in (3.7) can be turned into $+1$. In the other case, when n is even, only one negative numerator at the end, $\frac{-1}{k+z}$, remains, but this can easily be rewritten to $\frac{1}{-(k+z)}$. Taking both cases together, we obtain (3.6). □

Remark 3.7. It is worth noting that determining a continued fraction representation for a rational number $p/q \in \mathbb{Q}$ with $p, q \in \mathbb{Z}$ is essentially equivalent to applying the Euclidean algorithm to the integers p and q: If we set $(r_{-1}, r_0) := (p, q)$ and substitute in (3.5) $\alpha_j = r_{j-1}/r_j$ for all $j \geq 0$, we obtain

$$\begin{aligned} \frac{r_{-1}}{r_0} &= b_0 + \frac{r_1}{r_0} \quad &\Leftrightarrow \quad r_{-1} &= b_0 \cdot r_0 + r_1 \\ \frac{r_0}{r_1} &= b_1 + \frac{r_2}{r_1} \quad &\Leftrightarrow \quad r_0 &= b_1 \cdot r_1 + r_2 \\ \frac{r_1}{r_2} &= b_1 + \frac{r_3}{r_2} \quad &\Leftrightarrow \quad r_1 &= b_2 \cdot r_2 + r_3 \\ \vdots & & \vdots & \end{aligned}$$

In other words, the coefficients b_j of the desired continued fraction representation are nothing else but the quotients of the Euclidean algorithm which we used to denote by q_j.

This observation also allows it to see the T-U algorithm of Corollary 2.10 in a different light: For a given modular transformation $A(z) = \frac{az+b}{cz+d}$, by applying the Euclidean algorithm to a and c, we effectively determine a continued fraction representation for the rational number $A(\infty) = \frac{a}{c} = [q_0, q_1, \ldots, q_n]$. If we set again $e_j := (-1)^j q_j$, it follows that correspondingly the modular transformation $P := U^{e_0} T U^{e_1} T \ldots U^{e_n} T$ maps ∞ to $\frac{a}{c}$. Since the stabilizer of ∞ is generated by the transformation U, all transformations with this property can be written as PU^k for some $k \in \mathbb{Z}$.

In particular, if we determine the quotients q_j by rounding to the nearest integer, we obtain a canonical continued fraction, i.e. a continued fraction representation of minimal length. Consequently, the corresponding T-U product representation is one with minimal number of factors T and U^k.

Corollary 3.8. *Denote the extended rational numbers by $\mathbb{Q}_\infty := \mathbb{Q} \cup \{\infty\}$ and let $r \in \mathbb{Q}_\infty$. A transformation $A \in \mathrm{PSL}_2(\mathbb{Z})$ satisfying $A(\infty) = r$ can be found by determining a continued fraction representation of r, that is $r = [b_0, b_1, \ldots, b_n]$, and setting $A := U^{e_0} T U^{e_1} T \ldots U^{e_n} T$ where $e_j := (-1)^j b_j$. In particular, the k-th convergent c_k, $k \leq n$, of this continued fraction representation can be written as $c_k = U^{e_0} T U^{e_1} T \ldots U^{e_k} T(\infty)$.*

We have now seen that there is a natural correspondence between rational numbers, continued fractions and the T-U word representations of modular transformations. In order to formalize this correspondence, let us denote by $\mathbb{Z}^\star := \bigcup_{n \geq 0} \mathbb{Z}^n$ the set of finite integer sequences (or words over the alphabet $\mathbb{Z}$). Furthermore we set $\mathbb{Q}_\infty$ and define a map $f : \mathbb{Z}^\star \to \mathbb{Q}_\infty$ by

$$f : \begin{cases} \mathbb{Z}^\star & \to & \mathbb{Q}_\infty \\ (b_0, b_1, \ldots, b_n) & \mapsto & [b_0, b_1, \ldots, b_n]. \end{cases} \tag{3.8}$$

Note that evaluation of continued fractions shall take place in $\mathbb{Q}_\infty$ with the natural conventions for treating infinite quantities, i.e. for $a \neq 0$ and $b \neq \infty$ we have

$$\frac{a}{0} = \infty, \quad \frac{b}{\infty} = 0, \quad b \pm \infty = \infty.$$

Moreover the empty continued fraction shall evaluate to ∞, i.e. $f(\epsilon) = \infty$, where $\epsilon \in \mathbb{Z}^\star$ denotes the empty sequence. We call the sequence $\beta \in \mathbb{Z}^\star$ a *continued fraction representation* for $f(\beta) \in \mathbb{Q}_\infty$.

Next we set $\Sigma := \{T, U\} \subseteq \mathrm{PSL}_2(\mathbb{Z})$ and let $\Sigma_\sim$ be the free group generated by the symbols T and U. Moreover we denote by $\langle U \rangle$ both, the subgroup of $\Sigma_\sim$ generated by the symbol U and the subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ generated by the transformation U. We now consider left cosets of $\langle U \rangle$ in $\Sigma_\sim$ and $\mathrm{PSL}_2(\mathbb{Z})$:

$$\begin{aligned} \Sigma_\sim/\langle U \rangle &:= \{\sigma\langle U \rangle \mid \sigma \in \Sigma_\sim\}, \\ \mathrm{PSL}_2(\mathbb{Z})/\langle U \rangle &:= \{A\langle U \rangle \mid A \in \mathrm{PSL}_2(\mathbb{Z})\}. \end{aligned}$$

It is important to note these are *not* factor groups, as $\langle U \rangle$ is not a normal subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ or $\Sigma_\sim$. In particular, neither $\mathrm{PSL}_2(\mathbb{Z})/\langle U \rangle$ nor $\Sigma_\sim/\langle U \rangle$ carry a "natural" group structure – we will regard them just as sets.

We now have defined all domains needed for writing $f : \mathbb{Z}^\star \to \mathbb{Q}_\infty$ as composition of three other functions, namely $f = g_3 \circ g_2 \circ g_1$, where

$$\begin{aligned} g_1 &: \quad \mathbb{Z}^\star \quad \to \quad \Sigma_\sim/\langle U \rangle, \\ g_2 &: \quad \Sigma_\sim/\langle U \rangle \quad \to \mathrm{PSL}_2(\mathbb{Z})/\langle U \rangle, \\ g_3 &: \mathrm{PSL}_2(\mathbb{Z})/\langle U \rangle \to \quad \mathbb{Q}_\infty. \end{aligned}$$

Let us first turn to the definition of g_1, which maps a continued fraction representation to a left coset of a certain T-U group word.

$$g_1 : \begin{cases} \mathbb{Z}^\star & \to & \Sigma_\sim/\langle U \rangle \\ (b_0, b_2, \ldots, b_n) & \mapsto & U^{e_0} T U^{e_1} T \ldots U^{e_n} T \langle U \rangle \end{cases} \quad \text{where } e_j := (-1)^j b_j. \tag{3.9}$$

Note that in the case of the empty sequence $\epsilon \in \mathbb{Z}^\star$ we have $g_1(\epsilon) = \langle U \rangle$.

In order to define g_2, let $\varphi : \Sigma \to \mathrm{PSL}_2(\mathbb{Z})$ be the canonical embedding, i.e. $\varphi(T) = T$ and $\varphi(U) = U$. Let $\overline{\varphi}$ be the unique extension of φ to a homomorphism $\Sigma_\sim \to \mathrm{PSL}_2(\mathbb{Z})$, according to Theorem 1.14. Note that $\overline{\varphi}$ is just the map which evaluates T-U group words to concrete elements of $\mathrm{PSL}_2(\mathbb{Z})$ in the obvious way. The function g_2 now takes left cosets in $\Sigma_\sim$ to left cosets in $\mathrm{PSL}_2(\mathbb{Z})$ by

$$g_2 : \begin{cases} \Sigma_\sim/\langle U \rangle & \to & \mathrm{PSL}_2(\mathbb{Z})/\langle U \rangle \\ \sigma\langle U \rangle & \mapsto & \overline{\varphi}(\sigma)\langle U \rangle. \end{cases} \tag{3.10}$$

Last but not least we define g_3 as the map which evaluates the transformations of a left coset $A\langle U\rangle \in \mathrm{PSL}_2(\mathbb{Z})/\langle U\rangle$ at the point ∞. Note that the result is the same for all transformations within one coset because $\langle U\rangle$ is exactly the stabilizer of ∞, i.e. $\langle U\rangle = \{B \in \mathrm{PSL}_2(\mathbb{Z}) \mid B(\infty) = \infty\}$. This allows us to define

$$g_3 : \begin{cases} \mathrm{PSL}_2(\mathbb{Z})/\langle U\rangle & \to \quad \mathbb{Q}_\infty \\ A\langle U\rangle & \mapsto \quad A(\infty). \end{cases} \tag{3.11}$$

Lemma 3.9. *The maps g_1, g_2, g_3 and f, defined as above, satisfy*

$$f = g_3 \circ g_2 \circ g_1.$$

Proof. It follows from Corollary 3.8 that the composed map $g_2 \circ g_1 : \mathbb{Z}^\star \to \mathrm{PSL}_2(\mathbb{Z})/\langle U\rangle$ takes a continued fraction representation $(b_0, b_1, \ldots, b_n)$ to a left coset $A\langle U\rangle$, such that $g_3(A\langle U\rangle) = A(\infty) = [b_0, b_1, \ldots, b_n]$. Therefore we have indeed $f = g_3 \circ g_2 \circ g_1$. □

Lemma 3.10. *The map g_1 defined in (3.9) is injective. Its image $g_1(\mathbb{Z}^\star)$ consists precisely of those cosets $\sigma\langle U\rangle \in \Sigma_\sim/\langle U\rangle$, where the T-U group word σ is such that:*

(i) The reduced form of σ never contains the symbol T^{-1}.

(ii) If the reduced form of σ is not empty, its rightmost symbol is T.

In particular there is a one-to-one correspondence between all continued fraction representations and T-U group words of this form.

Proof. The fact that g_1 is injective is obvious. It is also clear that the word

$$U^{e_0}TU^{e_1}T \ldots U^{e_n}T \tag{3.12}$$

with $n \geq 0$, $e_j \in \mathbb{Z}$ occurring in the definition of g_1 already is in reduced form and satisfies the conditions (i) and (ii). Conversely, every reduced word w satisfying (i) and (ii) has necessarily the form

$$w = U^{k_1}T^{\ell_1}U^{k_2}T^{\ell_2} \ldots U^{k_m}T^{\ell_m},$$

with $m >= 0$, $k_j \in \mathbb{Z}$ and $\ell_j \geq 1$. Because for $\ell \geq 1$, T^ℓ and $(TU^0)^{\ell-1}T$ are *identical* as words, we can for sure notate w alternatively in the form (3.12). □

Lemma 3.11. *The map g_3 defined in (3.11) is bijective. In particular there is a one-to-one correspondence between the left cosets of $\langle U\rangle$ in $\mathrm{PSL}_2(\mathbb{Z})$ and the extended rational numbers $\mathbb{Q}_\infty$.*

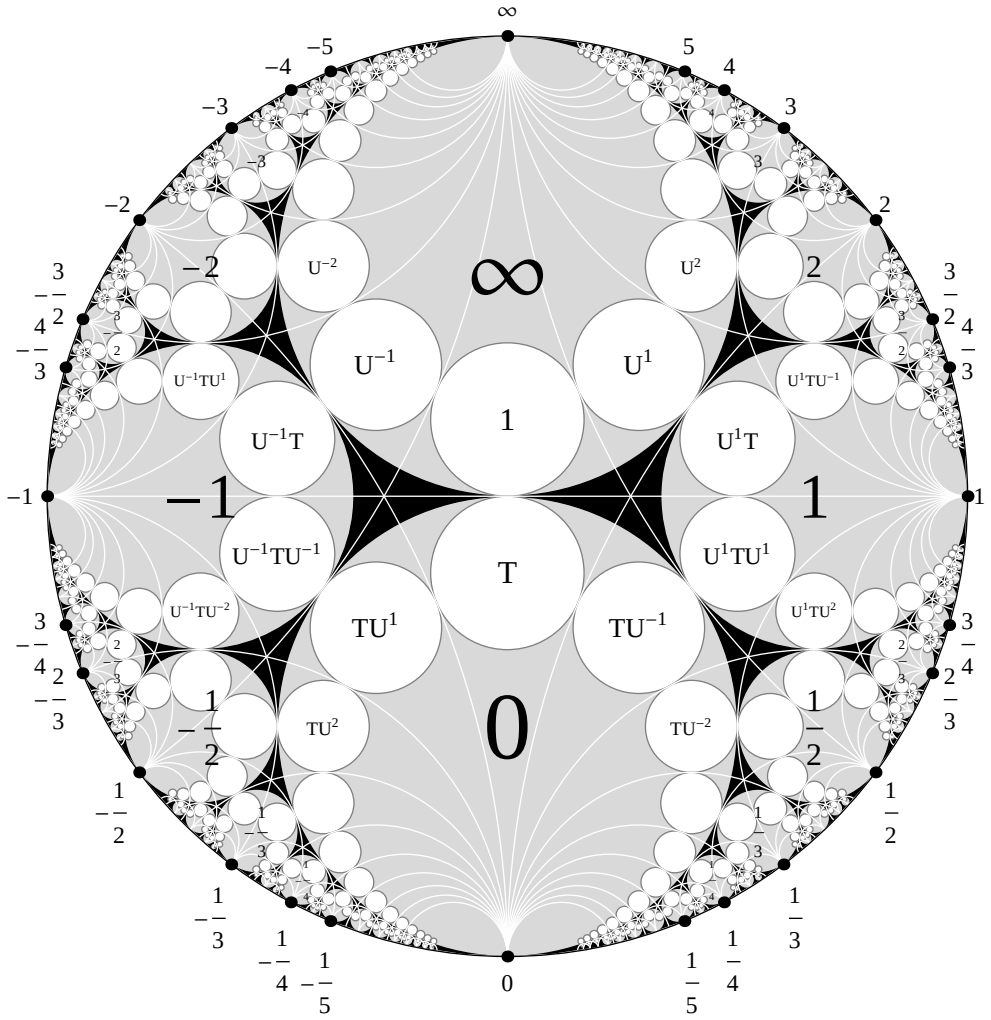

Figure 3.4: The modular tessellation under the modified Cayley transform Φ. The Ford disk $\mathcal{L}_\infty$ (light gray, labeled with "∞") encloses precisely the indisks $B\mathcal{I}$, $B \in \langle U \rangle$, in its interior. It can thus be considered as representative for the subgroup $\langle U \rangle$ in $\mathrm{PSL}_2(\mathbb{Z})$. The images of $\mathcal{L}_\infty$ under the modular group are all Ford disks $\mathcal{L}_r$ with $r \in \mathbb{Q}_\infty$. For $A \in \mathrm{PSL}_2(\mathbb{Z})$, we have $A\mathcal{L}_\infty = \mathcal{L}_r$ exactly when $A(\infty) = r$. Therefore $\mathcal{L}_r$ corresponds directly to both, the number r (the point, where $\mathcal{L}_r$ touches the extended real axis) and the left coset $A\langle U \rangle$ ($\mathcal{L}_r$ encloses precisely the indisks $B\mathcal{I}$, $B \in A\langle U \rangle$).

Proof. If $g_3(A\langle U\rangle) = g_3(B\langle U\rangle)$, then $A(\infty) = B(\infty)$. This is equivalent to $B^{-1}A \in \langle U\rangle$ or $A\langle U\rangle = B\langle U\rangle$. Therefore g_3 is injective. Since every rational number has a finite continued fraction expansion, $f = g_3 \circ g_2 \circ g_1$ is surjective. Consequently g_3 must as well be surjective. □

Looking at the modular tessellation, Figure 2.3, we see that the images of the indisk $\mathcal{I}$ of the fundamental region $\mathcal{F}$ have a natural one-to-one correspondence to modular transformations, i.e. the disk $A\mathcal{I}$ can be considered as a graphical representative for the modular transformation A. The question arises, whether there is such a visual and clear representation also for left cosets of $\langle U\rangle$ in $\mathrm{PSL}_2(\mathbb{Z})$. Indeed, we can see in Figure 3.4 – depicting the modular tessellation under the modified Cayley transform Φ – that the disks $U^k\mathcal{I}$, $k \in \mathbb{Z}$ form a "generalized ring" which asymptotically approaches the point ∞. We can enclose this ring in a generalized disk – in Figure 3.4, this enclosing disk is shown in light gray and is labeled with "∞".

Definition 3.12. The unique (open) g-disk $\mathcal{L}_\infty$ containing all the disks $U^k\mathcal{I}$, $k \in \mathbb{Z}$, in its interior and being tangent to each of them is called the *Ford disk at* ∞. It contains all points $z \in \mathbb{C}$ with $\mathrm{Im}\,(z) > 1$. In view of Definition 1.50, its defining matrix is given by

$$\mathcal{L}_\infty : \begin{pmatrix} 0 & -i \\ i & 2 \end{pmatrix}. \tag{3.13}$$

For a modular transformation $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{PSL}_2(\mathbb{Z})$, the image of $\mathcal{L}_\infty$ under A is called the *Ford disk at* $\frac{a}{c}$, $\mathcal{L}_{\frac{a}{c}} := A\mathcal{L}_\infty$.

The above definition as well as the bijective correspondence between left cosets of $\langle U\rangle$ in $\mathrm{PSL}_2(\mathbb{Z})$ and $\mathbb{Q}_\infty$ (Lemma 3.11) will get clearer in view of Figure 3.4: We can see that every Ford disk $\mathcal{L}_r$ (the light gray disks, each of them labeled with the corresponding number $r \in \mathbb{Q}_\infty$), "touches" the extended real axis $\mathbb{R}_\infty := \mathbb{R} \cup \{\infty\}$ (appearing as unit circle under the modified Cayley transform) exactly in the point r, that is

$$\mathrm{cl}(\mathcal{L}_r) \cap \mathbb{R}_\infty = \{r\} \quad \text{for all } r \in \mathbb{Q}_\infty.$$

For seeing the relation between Ford disks and left costets of $\langle U\rangle$ in $\mathrm{PSL}_2(\mathbb{Z})$, first observe that the Ford disk $\mathcal{L}_\infty$ encloses precisely all indisks $B\mathcal{I}$, $B \in \langle U\rangle$. In this sense, we can consider $\mathcal{L}_\infty$ as a graphical representative for the subgroup $\langle U\rangle$ of $\mathrm{PSL}_2(\mathbb{Z})$. Every Ford disk $\mathcal{L}_r$, $r \in \mathbb{Q}_\infty$, is the image of $\mathcal{L}_\infty$ under some transformation $A \in \mathrm{PSL}_2(\mathbb{Z})$ with $A(\infty) = r$, i.e. $A\mathcal{L}_\infty = \mathcal{L}_r$. Consequently $\mathcal{L}_r$ encloses all indisks $B\mathcal{I}$, $B \in A\langle U\rangle$ and thus can be considered as graphical representative for the left coset $A\langle U\rangle$.

Also the statement of Lemma 3.10, the one-to-one correspondence between continued fraction representations and T-U words of the form (3.12), has a nice visual interpretation, as we will illustrate in the following example.

Example 3.13. Consider the (semi-regular) continued fraction expansion of the irrational number π, obtained by using the nearest integer function for rounding the α_j, as discussed in Remark 3.5. The first few coefficients of this continued fraction are given by

$$\pi = [3, 7, 16, -294, 3, -4, 5, -15, -3, 2, 2, 2, 2, 3, -85, -3, 2, 15, 3, 14, \dots]$$

More coefficients can be found by looking up the sequence A133593 in the *On-Line Encyclopedia of Integer Sequences (OEIS)*[2]. By Corollary 3.8, the convergents of this continued fraction give rise to a sequence of Modular transformations:

$$\begin{aligned}
[3] &= \frac{3}{} &&= U^3T(\infty),\\
[3,7] &= \frac{22}{7} &&= U^3TU^{-7}T(\infty),\\
[3,7,16] &= \frac{355}{113} &&= U^3TU^{-7}TU^{16}T(\infty),\\
[3,7,16,-294] &= \frac{104348}{33215} &&= U^3TU^{-7}TU^{16}TU^{294}T(\infty),\\
&\vdots &&\vdots
\end{aligned}$$

The T-U words occurring in this sequence can now be interpreted as "path" through the set of indisks $A\mathcal{I}$, $A \in \mathrm{PSL}_2(\mathbb{Z})$ as follows:

1. Start with the Ford disk $\mathcal{L}_\infty$. Label each indisk $U^k\mathcal{I}$ (contained in $\mathcal{L}_\infty$) with the corresponding integer $k \in \mathbb{Z}$. The Ford disk $U^3T\mathcal{L}_\infty = \mathcal{L}_3$, corresponding to the convergent $c_0 = 3$, can be approached by starting at the indisk $\mathcal{I}$ (carrying the label 0), going 3 steps right to the indisk with label 3, $U^3\mathcal{I}$ (marked red in the first row of Figure 3.5) and finally applying T, that is going from there to the tangent indisk $U^3T\mathcal{I}$, lying within the Ford disk $\mathcal{L}_3$.

2. For convenience, set $A_1 := U^3T$. Within the Ford disk $\mathcal{L}_3$, label each indisk $A_1U^k\mathcal{I}$ with the *negated*[3] integer $-k$. Go from indisk $A_1\mathcal{I}$ (labeled 0) seven steps to the indisk with label 7, $A_1U^{-7}\mathcal{I}$ – see second row of

[2] https://oeis.org

[3] The negation of the sign comes from the fact that we want the indisk labels to correspond to the coefficients b_j of the continued fraction rather than to the exponents $e_j = (-1)^j b_j$ of the T-U word.

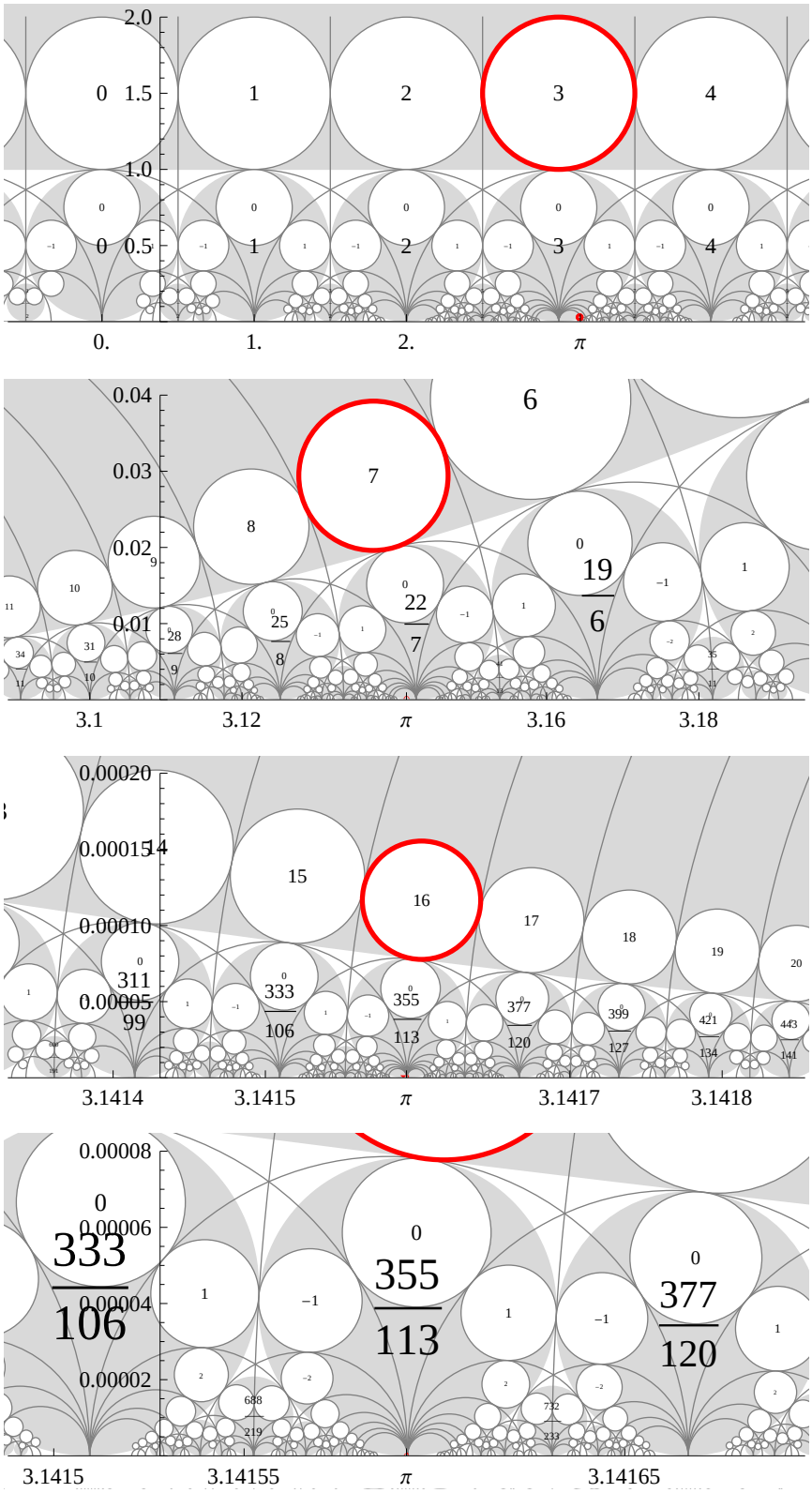

Figure 3.5: $\pi = \frac{355}{113}$ – well, almost.

Figure 3.5. Applying T again takes us to the indisk $A_1U^{-7}T\mathcal{I}$ within the Ford disk corresponding to the next convergent of the continued fraction: $\mathcal{L}_{c_1}$, $c_1 = \frac{22}{7}$.

3. Set $A_2 := U^3TU^{-7}T$. As above, label all indisks $A_2U^k\mathcal{I}$ within $\mathcal{L}_{c_1}$ with k and go from $A_2\mathcal{I}$ with label 0 sixteen steps to $A_2U^{16}\mathcal{I}$ with label 16 (Figure 3.5, 3rd row) and from there to $A_2U^{16}T\mathcal{I}$, lying within $\mathcal{L}_{c_2}$, $c_2 = \frac{355}{113}$.

   ...

Without explicating further steps of the above example, we can easily imagine how this generalizes to different continued fraction representations of arbitrary real numbers: Given a continued fraction $\alpha = [b_0, b_1, \dots]$, we start with the indisk $\mathcal{I}^{(0)} := \mathcal{I}$ contained in the Ford disk $\mathcal{L}^{(0)} := \mathcal{L}_\infty$. For every $j \geq 0$, we label the indisks contained in $\mathcal{L}^{(j)}$ with successive integers in counter-clockwise direction (if j is even) or in clockwise direction (if j is odd) in such a way that the disk $\mathcal{I}^{(j)}$ carries the label 0. Now we choose $\mathcal{I}^{(j+1)}$ as the unique indisk which is exterior to $\mathcal{L}^{(j)}$ and tangent to the disk with label b_j (within $\mathcal{L}^{(j)}$). Moreover we choose as $\mathcal{L}^{(j+1)}$ the unique Ford disk containing $\mathcal{I}^{(j+1)}$.

Considering also "intermediate" indisks (those with labels between 0 and b_j within $\mathcal{L}^{(j)}$), we describe in this way an *indisk path*, or in other words a chain of successively tangent disks, through the set of all indisks $A\mathcal{I}$, $A \in \mathrm{PSL}_2(\mathbb{Z})$. Such a path always starts at $\mathcal{I}$ and, in case of a rational number $\alpha \in \mathbb{Q}_\infty$, ends at some $U^{e_0}TU^{e_1}T \dots U^{e_n}T\mathcal{I}$, appending exactly one symbol U, U^{-1} or T to the corresponding group word when moving one step forward along the path. Obviously these paths can be considered as a visual representation of T-U words of the form (3.12) and we can draw the following conclusions:

Corollary 3.14. *Continued fraction representations are not unique in any way.*

Proof. All different indisk paths starting at $\mathcal{I}$ and ending within a given Ford disk $\mathcal{L}_r$ give rise to different continued fraction representations for the same rational number $r \in \mathbb{Q}_\infty$. Continued fraction representations are therefore not unique, unless we impose certain conditions on its coefficients b_j, as for example regularity – compare Perron [11], §9. □

Corollary 3.15. *Canonical continued fraction representations are in general not unique.*

Proof. Considering a continued fraction representation $r = [b_0, b_1, \dots, b_n]$ of a rational number $r \in \mathbb{Q}_\infty$, its length n is exactly the number of hops between tangent Ford disks in the corresponding indisk path (or the number of symbols

T in the corresponding T-U word). Let us call n the *length* of the indisk path. Note that a given Ford disk may possibly be visited more than once while walking along an indisk path. Clearly this is not the case if the path is one of minimal length from $\mathcal{L}_\infty$ to $\mathcal{L}_r$. Still, minimality with regard to path length does not imply uniqueness: Consider for example $r = \frac{2}{5}$. In Figure 3.4, we can see $\mathcal{L}_{2/5}$ as the smaller of the two Ford disks being tangent to $\mathcal{L}_{1/3}$ and $\mathcal{L}_{1/2}$. We can see that there are three paths of length 3 from $\mathcal{L}_\infty$ to $\mathcal{L}_{2/5}$, namely

$$\infty \to 0 \to \frac{1}{2} \to \frac{2}{5}, \quad \infty \to 0 \to \frac{1}{3} \to \frac{2}{5}, \quad \infty \to 1 \to \frac{1}{2} \to \frac{2}{5}.$$

Since these paths are of minimal length, they give rise to three different canonical continued fraction representations of r:

$$r = \frac{2}{5} = [0, 2, 2] = [0, 3, -2] = [1, -2, 3]. \qquad \square$$

Considering indisk paths, we can also give an alternative proof for the following fact:

Corollary 3.16. *All relations satisfied by the generators $T, U \in \mathrm{PSL}_2(\mathbb{Z})$ are derived from $T^2 = 1$ and $(TU)^3 = 1$.*

Proof. As above, in the definition of the map g_2 in (3.10), let again $\Sigma_\sim := \langle T, U \rangle$ be the free group of T-U words, and let $\overline{\varphi} : \Sigma_\sim \to \mathrm{PSL}_2(\mathbb{Z})$ be the natural evaluation map. We need to show that every relation $\sigma \in \Sigma_\sim$ (satisfying $\overline{\varphi}(\sigma) = 1$) can be written as product of words conjugate to T^2 and $(TU)^3$, that is

$$\sigma = \prod_{j=1}^{n} (\tau_j R_j {\tau_j}^{-1})^{k_j}, \tag{3.14}$$

with $k_j \in \mathbb{Z}$, $\tau_j \in \Sigma_\sim$ and R_j being either the word T^2 or $(TU)^3$. From $\overline{\varphi}(\sigma) = 1$ we see that the indisk path corresponding to σ is closed, i.e. it starts and ends at the indisk $\mathcal{I}$. Let us call such an indisk path a *loop*. Next, set $\rho := \exp(2\pi i/3)$ and denote by $[\rho]_\sim$ the orbit of ρ under $\mathrm{PSL}_2(\mathbb{Z})$. We define the number of times a loop L turns around a point $z \in [\rho]_\sim$ the *winding number* $\nu_z(L)$. Every full turn around z in positive (resp. negative) direction shall give a contribution of $+1$ (resp. -1) to $\nu_z(L) \in \mathbb{Z}$. If the winding number $\nu_z(L)$ is zero for all $z \in [\rho]_\sim$, we call L a *degenerate loop*. In this case L consists enirely of subpaths of the form

$$\mathcal{I} \to \mathcal{I}_1 \to \cdots \to \mathcal{I}_{n-1} \to \mathcal{I}_n \to \mathcal{I}_{n-1} \to \cdots \to \mathcal{I}_1 \to \mathcal{I}$$

and the corresponding T-U word can be reduced to the empty word just by repeated application of the relation $T^2 = 1$. Finally, using the *Kronecker delta* notation,

$$\delta_{z,w} := \begin{cases} 1 & \text{if } z = w, \\ 0 & \text{otherwise,} \end{cases}$$

we call a loop L a *primitive loop* for the point $z \in [\rho]_\sim$, if $\nu_w(L) = \delta_{w,z}$ for all $w \in [\rho]_\sim$. Looking at Figure 3.4 and using its indisk labeling, the relation $(TU)^3 = 1$ corresponds to the loop

$$L'_\rho : 1 \to T \to TU \to U^{-1}TU^{-1} \to U^{-1}T \to U^{-1} \to 1.$$

We see that L'_ρ goes once around the point ρ in clockwise (negative) direction. Going the loop in reverse direction yields a primitive loop for the point ρ, $L_\rho := (L'_\rho)^{-1}$, corresponding to the relation $(TU)^{-3} = 1$. Next we define such a primitive loop L_z for every $z \in [\rho]_\sim$. We do so by going from $\mathcal{I}$ to any indisk in the "neighbourhood" of z along a path P_z, then going around z (and no other point) once in positive direction and finally returning to $\mathcal{I}$ by going back the reverse path ${P_z}^{-1}$. Every such primitive loop L_z corresponds to a relation of the form $\tau(TU)^{-3}\tau^{-1} = 1$.

Let now $\sigma \in \Sigma_\sim$ be an arbitrary relation, and call S the closed indisk path corresponding to the word σ. There are only finitely many points $z \in [\rho]_\sim$ with $\nu_z(S) \neq 0$. We can therefore define a loop V as the composition of primitive loops L_z, where z runs over this finite set of points – each primitive loop L_z shall be taken as often as the winding number $\nu_z(S)$ declares:

$$V := \prod_{z \in [\rho]_\sim} L_z^{\nu_z(S)}.$$

Note that the particular order of the individual factors in this product is irrelevant for our purposes. By going through the loop S in forward and through V in backward direction, we obtain a degenerate loop SV^{-1}, i.e. $\nu_z(SV^{-1}) = 0$ for all $z \in [\rho]_\sim$. In other words, with V we have found a word $\psi \in \Sigma_\sim$ such that $\sigma\psi^{-1}$ can be reduced to 1 entirely by applying the relation $T^2 = 1$, i.e. we have $\sigma\psi^{-1}\omega^{-1} = 1$ for some product $\omega \in \Sigma_\sim$ consisting entirely of factors conjugate to the word T^2. Summing up, with $\omega\psi$ we have a product representation for the relation σ of the desired form (3.14). □

Note that the above proof is essentially a discrete variant of the proof given in Klein/Fricke [7], p. 452ff.[4] We conclude this section with some final remarks and references for further reading:

[4]Volume 1, part 2, chapter 9, §1.

Remark 3.17. The connection between Ford disks and continued fractions has been the object of many investigations in history. For example Ford [4] gives a geometric interpretation of continued fractions by identifying rational numbers p/q, where $p, q \in \mathbb{Z}$ are coprime, with circles of radius $r = (2q^2)^{-1}$ and center $C : (p/q, r)$. He also gives a reference to an interesting book of Züllig [16], containing a considerable number of concrete numerical examples and accurate figures relating continued fractions to curves along the boundary of chains of Ford disks. Note that these interpretations are essentially equivalent to the one given in Example 3.13. However, both authors do not use a concept similar to indisks, nor do they make such an explicit connection to the modular group.

Remark 3.18. Considering the continued fraction expansion of π from Example 3.13, $\pi = [3, 7, 16, -294, \dots]$, and looking at the ford disk $\mathcal{L}_{c_3}$, $c_3 = \frac{355}{113}$, in the last row of Figure 3.5, we see that (starting as usual at the indisk with label 0) we would have to go 294 (!) steps in clockwise direction to approach the Ford disk $\mathcal{L}_{c_4}$, $c_4 = \frac{104348}{33215}$. From the image one can hardly to grasp how small $\mathcal{L}_{c_4}$ must be in comparison to $\mathcal{L}_{c_3}$. In fact, the ratio of the radii[5] between these two Ford disks is approximately $1 : 86\,400$. This huge difference in size explains why the rational number $\frac{355}{133}$ is such an extraordinary good approximation for π. The absolute errors between π and the convergents c_2, c_3 and c_4 are

$$\begin{aligned} \pi - \frac{22}{7} &\approx -1.2 \cdot 10^{-3}, \\ \pi - \frac{335}{133} &\approx -2.7 \cdot 10^{-7}, \\ \pi - \frac{104348}{33215} &\approx -3.3 \cdot 10^{-10}. \end{aligned}$$

Comparing $c_2 = \frac{22}{7}$ and $c_3 = \frac{335}{113}$, we see that by increasing the numerator and denominator by a relatively small factor of about 15, we achieve more than 4000-fold increase in precision. In contrast to that, blowing up numerator and denominator further by an additional factor of roughly 300 improves precision only by a factor of approximately 800.

In Perron [11], p. 61ff., the approximations of π obtained by its *regular continued fraction expansion* are discussed in a similar fashion and some interesting historical details can be found there as well.

Geometric considerations on Ford disks can also be used to prove the following number-theoretic theorem:

[5] The radius of a Ford disk $\mathcal{L}_{p/q}$, for coprime $p, q \in \mathbb{Z}$, is $\frac{1}{2q^2}$ – see also Ford [4].

Theorem 3.19. *If $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ is an irrational number and if $k = \sqrt{5}$, then the equation*

$$\left| \frac{p}{q} - \alpha \right| < \frac{1}{k \cdot q^2} \tag{3.15}$$

is satisfied by infinitely many $p, q \in \mathbb{Z}$. If $k > \sqrt{5}$, then there are irrational numbers α, such that (3.15) is satisfied by only finitely many $p, q \in \mathbb{Z}$.

Remark 3.20. An elementary proof of Theorem 3.19, relying on continued fraction expansions of irrational numbers, can be found in Perron [11], §14. An alternative proof, based purely on geometric considerations on Ford circles, is given in Ford [4].

3.3 The modular tessellation and the exponential transformation

In the theory of modular functions, which we will touch in the next section, a map of essential importance is the transformation $z \mapsto \exp(2\pi i z)$, as it prominently occurs in Fourier expansions of modular functions – compare for example Petersson [12] and Rademacher [13]. Adopting the notation of Lehner [8], we will denote this transformation by

$$e(z) := \exp(2\pi i z). \tag{3.16}$$

Just like the Cayley transform, e maps the upper half-plane onto the unit disk $\mathbb{D}$, as we see through

$$|e(z)| = \exp(\operatorname{Re}(2\pi i z)) = \exp(-2\pi \operatorname{Im}(z)).$$

Obviously $\operatorname{Im}(z) > 0$ implies $|e(z)| < 1$. The real axis is mapped to the boundary of $\mathbb{D}$. In contrast to the Cayley transform, e is not a one-to-one map from the upper half-plane to the unit disk, as clearly $e(z) = e(z + k)$ for arbitrary $k \in \mathbb{Z}$. Instead it can be considered as a bijective map from the strip

$$S = \left\{ z \in \mathbb{C} \;\middle|\; \operatorname{Re}(z) \in [-0.5, 0.5),\ \operatorname{Im}(z) \geq 0 \right\}$$

to the punctured unit disk $\mathbb{D} \setminus \{0\}$. Note that the exponential function has an essential singularity at the point ∞, but still it is sometimes useful to define $e(\infty) := 0$. This is motivated by a continuous extension of e from the domain S to $S_\infty := S \cup \{\infty\}$ by

$$e(\infty) := \lim_{\substack{\operatorname{Im}(z) \to \infty \\ z \in S}} e(z) = 0.$$

In this way we obtain a bijective map from the set S_∞ to the closed unit disk $\mathbb{D}$. Figure 3.6 shows the modular tessellation of the upper half-plane – to be precise the part of it lying within S_∞ – mapped to the unit disk by $z \mapsto e(z)$.

Remark 3.21. In order to trace the image of the fundamental region $\mathcal{F}$ and its indisk $\mathcal{I}$ under the map e, we have to zoom in on the neighborhood of zero, which is done in Figure 3.7. The top-left frame again shows the image of the tessellation under e on the whole unit disk.

The second frame displays the images of the regions $TU\mathcal{F}$ and $TU^{-1}\mathcal{F}$ in more detail. Comparing with Figure 2.3, we see that originally these two regions do not have any boundary arc in common. However, due to the periodicity of e, the left boundary arc of $TU\mathcal{F}$ (being a segment of the line $\mathrm{Re}\,(z) = -1/2$) and the right boundary arc of $TU^{-1}\mathcal{F}$ (being a segment of the line $\mathrm{Re}\,(z) = 1/2$) have the same image under the transformation e. Therefore the images of $TU\mathcal{F}$ and $TU^{-1}\mathcal{F}$ touch each other along a certain interval on the negative real axis.

The third frame reveals some more detail on the real shape of $e(T\mathcal{F})$. Its "tip" is not round, as a short look on the first two frames might suggest, but in fact it has a kink at the point $e(\rho) = e(T\rho) \approx -0.0043$. Moreover, the image of $T\mathcal{F}$ completely surrounds the image of the fundamental region $\mathcal{F}$ as we see in the next frame:

Denoting the boundary arcs of $\mathcal{F}$ by a, b, c and d as in Figure 2.2, we can see in the fourth frame the image of the left boundary arc a and the right boundary arc b of $\mathcal{F}$ are both mapped to the same interval $[e(\rho), 0]$ on the negative real axis. The images of the unit circle arcs c and d form the drop-shaped boundary of $e(\mathcal{F})$. Note that $e(c)$ (resp. $e(d)$) is precisely the part of the boundary of $e(\mathcal{F})$ in the lower (resp. upper) half-plane.

Finally in the last row, we can see the image of the indisk $\mathcal{I}$ in detail. Note that the transformed indisk touches itself in the point $e(\pm\frac{1}{2} + \frac{3}{2}i) \approx -0.00008$. The points within $\mathcal{F}$ lying "above" $\mathcal{I}$ (including the point ∞) are mapped to just another tiny drop-shaped region surrounded entirely by $e(\mathcal{I})$.

Having studied the image of the modular tessellation under the transformation $z \mapsto e(z)$ in detail, the question arises whether this relates somehow to the image of the tessellation under the modified Cayley transform Φ which we have seen in Figure 2.3. Indeed it is possible to visualize the connection between these two images. For this purpose we take advantage of yet another Möbius transformation:

$$f(z) := \frac{i}{z+i} = -\frac{i}{2} \cdot \Phi(z) + \frac{1}{2}. \tag{3.17}$$

As we see, f can be considered as composition of the modified Cayley transform Φ, a clockwise rotation by 90°, scaling by the factor $\frac{1}{2}$ and a final translation

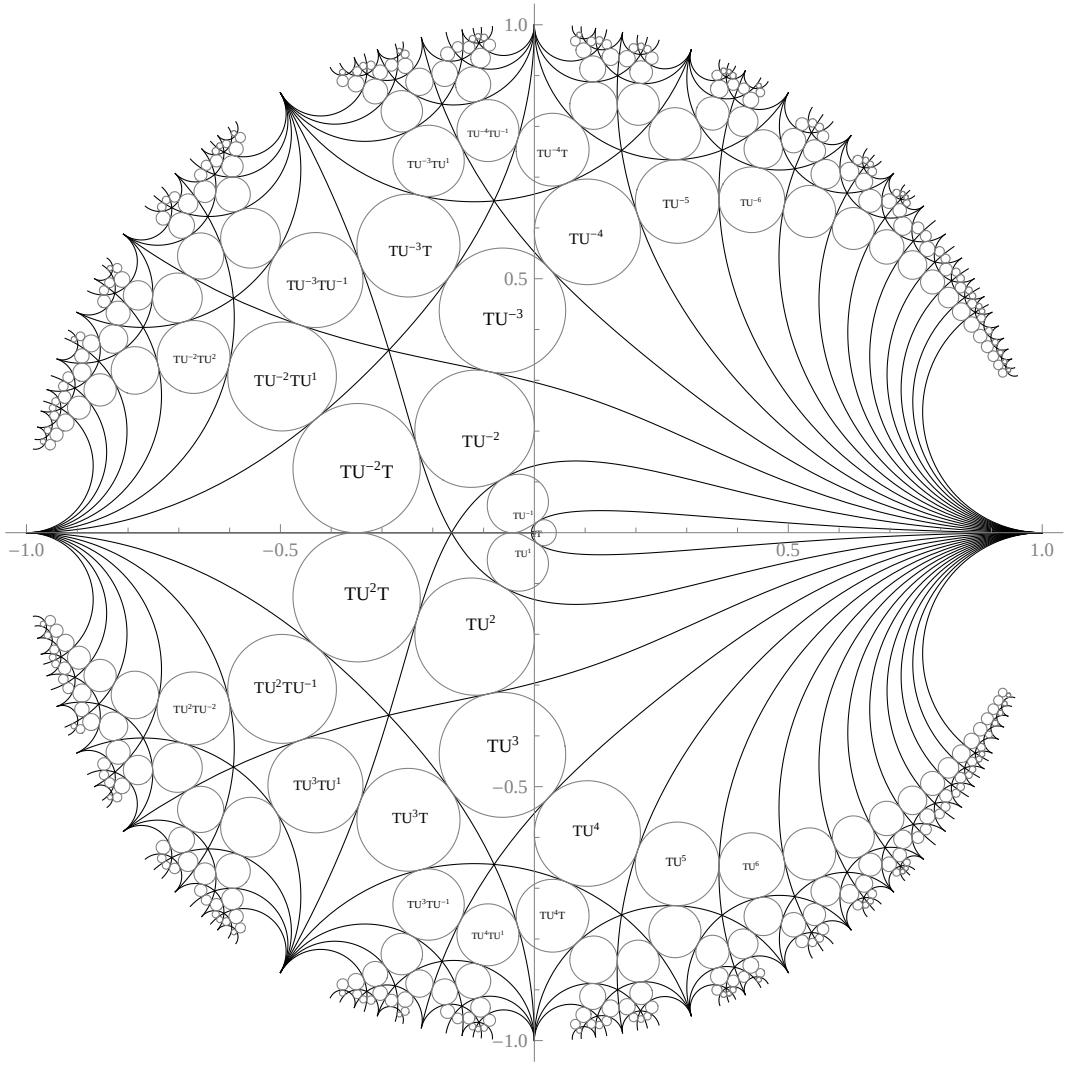

Figure 3.6: The modular tessellation under the transformation $z \mapsto e(z)$. The equation $|z| = \exp(-2\pi \operatorname{Im}(z))$ implies that points with large imaginary part are mapped very closely to 0. For this reason, the image of the fundamental region $\mathcal{F}$ cannot be seen any more in this scale. Its points, satisfying $\operatorname{Im}(z) > \frac{\sqrt{3}}{2}$, are mapped into a disk centered about the origin of a radius smaller than 0.005.

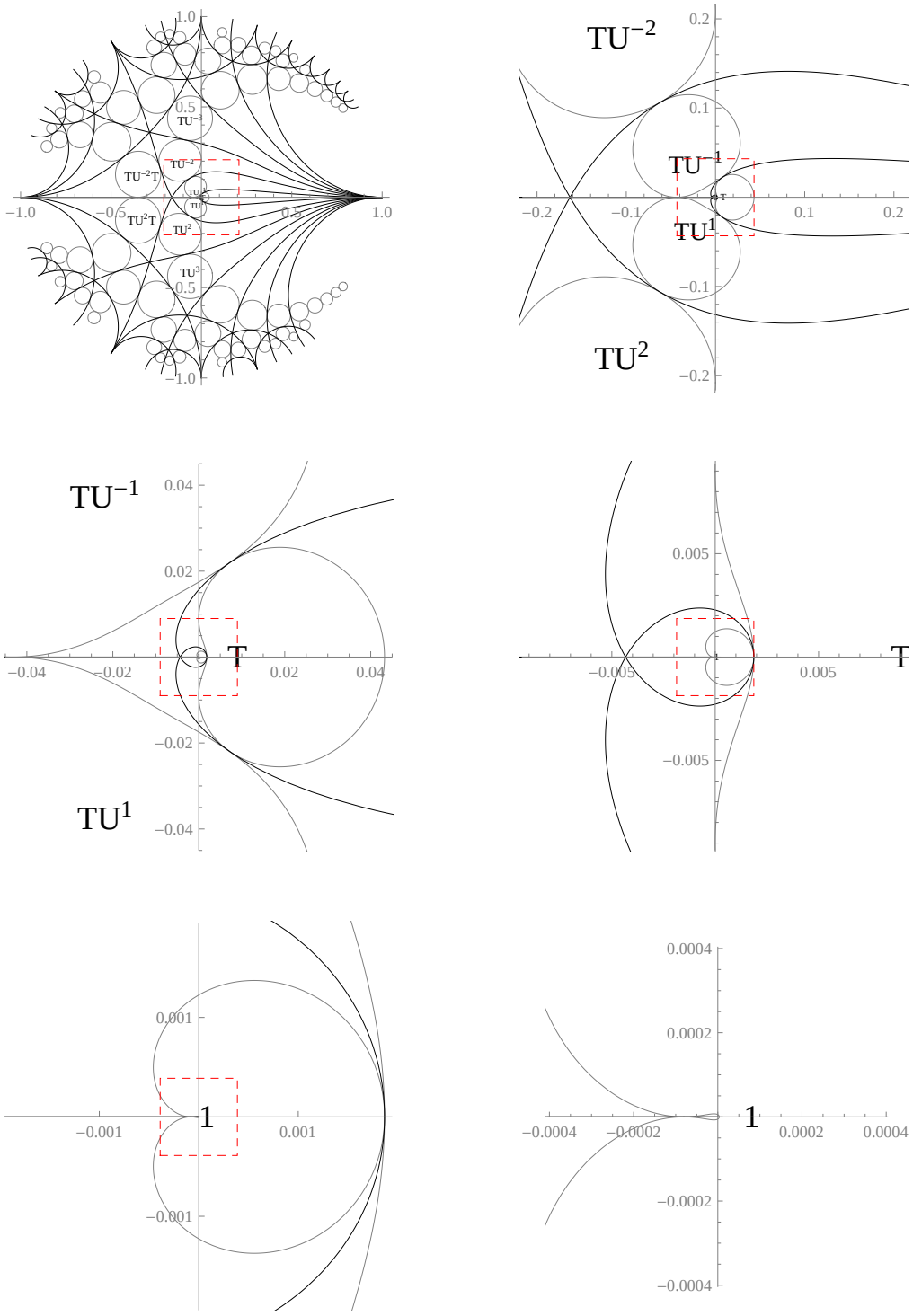

Figure 3.7: The image of the fundamental region $\mathcal{F}$ and its indisk $\mathcal{I}$ under the map $z \mapsto e(z)$ can be seen by zooming in to a close neighborhood of the point 0.

by $\frac{1}{2}$. Therefore, f maps the upper half-plane to a disk of radius $\frac{1}{2}$ centered about the real point $\frac{1}{2}$. The choice of f is not as arbitrary as it first might seem, because

$$f(0) = e(0) = 1 \quad \text{and} \quad f(\infty) = e(\infty) = 0.$$

This property allows it to establish a continuous transition between the images of the tessellation under f and e which leaves the points 0 and 1 fixed. For this purpose we may for example use the map

$$h(t, z) := f(z)^{1-t} \cdot e(tz), \tag{3.18}$$

with varying parameter $t \in [0, 1]$. Note that complex powers z^t, $z \in \mathbb{C}$, $t \in \mathbb{R}$ shall be evaluated as $\exp(t \ln z)$, by choosing a branch of the natural logarithm such that the imaginary part of the logarithm ranges in the interval $(-\pi, \pi]$. Moreover, for $z = \infty$ we define $h(t, \infty) := f(\infty) = e(\infty) = 0$.

In Figure 3.8, we can see the images the modular tessellation under the map $h(t, \cdot)$ for the parameter values $t \in \{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\}$. In the first frame, we see that the image of the tessellation under $h(0, \cdot) = f$ is indeed a rotated and scaled version of the one belonging to Φ, when omitting the part which lies outside the strip $|\mathrm{Re}\,(z)| \leq \frac{1}{2}$ (compare with Figure 2.3).

As the parameter value t is stepwise increased in the following frames, the image is slowly fanned out to the whole unit disk. Finally we end up with the image of the tessellation under $h(1, \cdot) = e$.

Remark 3.22. In a more general context, the connection between a region R of similar shape as above (i.e. a region bounded by three circular arcs having vertex angles 90°, 90° and 0°), a parabolic Möbius transformation leaving the vertex with angle 0° fixed and an exponential transformation which maps R to the unit disk, is discussed in Lehner [8], p. 68ff.

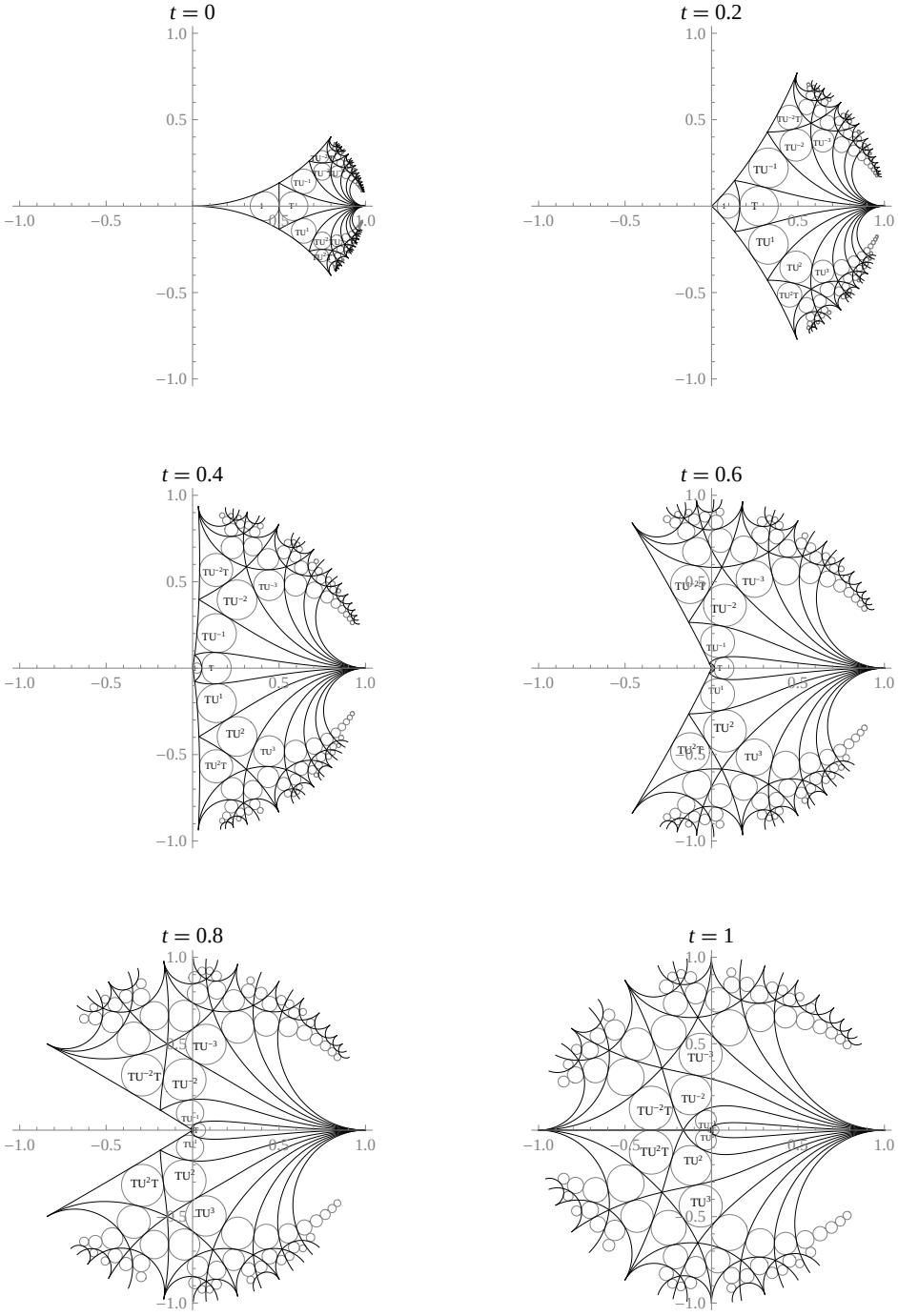

Figure 3.8: The modular tessellation on the strip $|\mathrm{Re}\,(z)| \leq \frac{1}{2}$ under the maps $z \mapsto f(z) = \frac{i}{z+i}$ (top-left) and under $z \mapsto e(z)$ (bottom-right). A continuous transition between these two images is established through the map $h(t, z) = f(z)^{1-t} \cdot e(tz)$ when the parameter $t \in [0, 1]$ is varied.

3.4 Modular functions

We devote the last section of this thesis to the visualization of modular functions. The theory of modular functions is a branch of complex analysis whose importance and beauty lies most notably in its connections to number theory. We will however not dive deeply into this theory. Instead we will content ourselves with depicting graphs of certain selected modular functions, enjoying their visual aesthetics and reading off some of its properties, like zeros and poles as well as their order. Unfortunately such illustrations of modular functions are rarely found in literature. Therefore this section should be considered as complementary material to more comprehensive treatments on the theory of modular functions given for example in Klein/Fricke [7], Lehner [8] or Schoeneberg [14].

Modular functions are meromorphic maps (i.e. maps which are holomorphic[6] except for isolated poles, or in other words, maps which can be represented as quotient of two holomorphic functions) which are defined on the upper half-plane and which are invariant under the transformations of the modular group.

Definition 3.23 (Modular function)**.** Let the upper halfplane $\mathcal{H}$ and the extended upper halfplane $\mathcal{H}^*$ be defined as in (2.30) and (2.31) respectively. A map $F : \mathcal{H}^* \to \mathbb{C}_\infty$ is called a *modular function*, if it satisfies the following conditions:

(i) F is meromorphic on the upper half-plane $\mathcal{H}$.

(ii) On $\mathcal{H}^*$, $F = F \circ A$ for all modular transformations $A \in \mathrm{PSL}_2(\mathbb{Z})$.

(iii) There is a constant $C \geq 0$ such that F has a series expansion of the form

$$F(z) = \sum_{k \geq k_0} a_k \exp(2\pi i k z), \tag{3.19}$$

with $k_0 \in \mathbb{Z}$, $a_k \in \mathbb{C}$, $a_{k_0} \neq 0$, which converges for all $z \in \mathcal{H}$ with $\mathrm{Im}\,(z) > C$. Moreover,

$$f(\infty) = \begin{cases} 0 & \text{if } k_0 > 0, \\ a_0 & \text{if } k_0 = 0, \\ \infty & \text{if } k_0 < 0. \end{cases}$$

Remark 3.24. For the proof that modular functions indeed exist we refer to Schoeneberg [14], Chapter II, §3.

[6] Holomorphic functions are also frequently called "analytic".

A modular function of essential importance is the J function, also known as the *absolute modular invariant* or *Klein's complete invariant*. One of its important properties is, that on the fundamental set $\mathcal{F}^*$ it takes on each value in $\mathbb{C}_\infty$ exactly once. In other words, J can be considered as bijective map from $\mathcal{F}^*$ to $\mathbb{C}_\infty$.

In order to discuss this one-to-one mapping between $\mathcal{F}^*$ and $\mathbb{C}_\infty$ under J, let us denote the boundary arcs of the fundamental region $\mathcal{F}$ again by a, b, c and d, as in Figure 2.2. Moreover, denote by $e := \{\lambda i \mid \lambda \geq 1\} \cup \{\infty\}$ the arc which splits $\mathcal{F}$ into two symmetric halves, i.e. two open connected components $\mathcal{F}_{\text{left}}$ and $\mathcal{F}_{\text{right}}$. For the special boundary points i, ρ and ∞ of $\mathcal{F}$ we have

$$J(i) = 1, \quad J(\rho) = 0, \quad J(\infty) = \infty.$$

Additionally, the following mappings are pointwise one-to-one:

1. The left boundary arc a and the right boundary arc b of $\mathcal{F}$ are both mapped to the set $\overline{\mathbb{R}}_{\leq 0} := \{z \in \mathbb{R} \mid z \leq 0\} \cup \{\infty\}$:
$$J(a) = J(b) = \overline{\mathbb{R}}_{\leq 0}.$$

2. The boundary arcs c and d of $\mathcal{F}$ are both mapped to the interval $[0, 1]$:
$$J(c) = J(d) = [0, 1].$$

3. The "symmetry arc" e is mapped to $\overline{\mathbb{R}}_{\geq 1} := \{z \in \mathbb{R} \mid z \geq 1\} \cup \{\infty\}$:
$$J(e) = \overline{\mathbb{R}}_{\geq 1}.$$

4. In particular, the boundaries of $\mathcal{F}_{\text{left}}$ and $\mathcal{F}_{\text{right}}$ are both mapped to $\mathbb{R} \cup \{\infty\}$.

5. Finally, the images of $\mathcal{F}_{\text{left}}$ and $\mathcal{F}_{\text{right}}$ under J are exactly the upper and lower half-plane:
$$J(\mathcal{F}_{\text{left}}) = \mathcal{H} \quad \text{and} \quad J(\mathcal{F}_{\text{right}}) = -\mathcal{H}.$$

Unfortunately, plotting the function graph of J, or more generally the function graph of any map $f : \mathbb{C}_\infty \to \mathbb{C}_\infty$ is not directly possible, as it is in fact a 4-dimensional object.[7] However there is a simple idea for getting around this problem: We assign each $z \in \mathbb{C}_\infty$ a certain color $\mathcal{C}(z)$ and obtain a picture of f by dying each point z within the domain of f in the color $\mathcal{C}(f(z))$. Our choice of the color coding is quite simple:

[7] It involves two dimensions for real and imaginary part of the function argument and two more dimensions for real and imaginary part of the function value.

(i) The tone of the color $\mathcal{C}(z)$ encodes the complex argument of z:

Red	$(\arg z = 0)$	$\to$	Orange	$\to$	Yellow	$\to$
Green	$(\arg z = \frac{\pi}{2})$	$\to$	Turquoise	$\to$		
Cyan	$(\arg z = \pi)$	$\to$	Blue	$\to$		
Violet	$(\arg z = -\frac{\pi}{2})$	$\to$	Magenta	$\to$	Red (again).	

(ii) The saturation and brightness of the color $\mathcal{C}(z)$ encodes the absolute value of z. For this purpose, we use the continuous map

$$b(r) := \begin{cases} 0 & \text{if } r = 0, \\ \frac{1}{\pi}\arctan(\ln r) + \frac{1}{2} & \text{if } r \in (0, \infty), \\ 1 & \text{if } r = \infty \end{cases}$$

to first bring $|z|$ to the interval $[0, 1]$. Note that b has the neat property $b(\frac{1}{r}) = 1 - b(r)$. Finally we define the saturation of $\mathcal{C}(z)$ as $1 - b(|z|)^2$ and its brightness as $1 - [1 - b(|z|)]^2$. This means that $\mathcal{C}(z)$ changes gradually from a perfect black (if $z = 0$) to a perfect white ($z = \infty$) as the absolute value of z grows.

Remark 3.25. Obviously the color coding may be chosen very arbitrarily. However, for our purposes there are three properties making $\mathcal{C}(z)$ a particularly good choice: First, $z \mapsto \mathcal{C}(z)$ is injective, i.e. distinct points in $\mathbb{C}_\infty$ are colored differently. Moreover, when considering colors as three-dimensional vectors of $[0, 1]^3 \subseteq \mathbb{R}^3$ with red, green and blue component, $z \mapsto \mathcal{C}(z)$ is differentiable. This guarantees smooth color transitions without visible "edges" for visualization of continuous functions. Lastly, saturation and brightness depend logarithmically on the absolute value of z, capturing well the typical growth behavior of modular functions and allowing a visual distinction of values relatively close to 0 (resp. ∞).

We now wish to plot the J function using the above idea and the color coding $\mathcal{C}(z)$. However, instead of plotting J directly on the upper half-plane, we again prefer to translate the picture from the upper half-plane to the unit disk using the modified Cayley transform, i.e. instead of visualizing $\mathcal{C} \circ J$ on the upper half-plane, Figure 3.9 shows $\mathcal{C} \circ J \circ \Phi^{-1}$ on the unit disk. In order to accentuate the symmetry of J with respect to the transformations of the modular group, additionally the image of the modular tessellation under Φ is displayed in gray lines. We can now read off the following properties from this illustration:

1. J is injective from $\mathcal{F}$ to $\mathbb{C}_\infty$: No two distinct points within the region $\Phi\mathcal{F}$ have the same color.

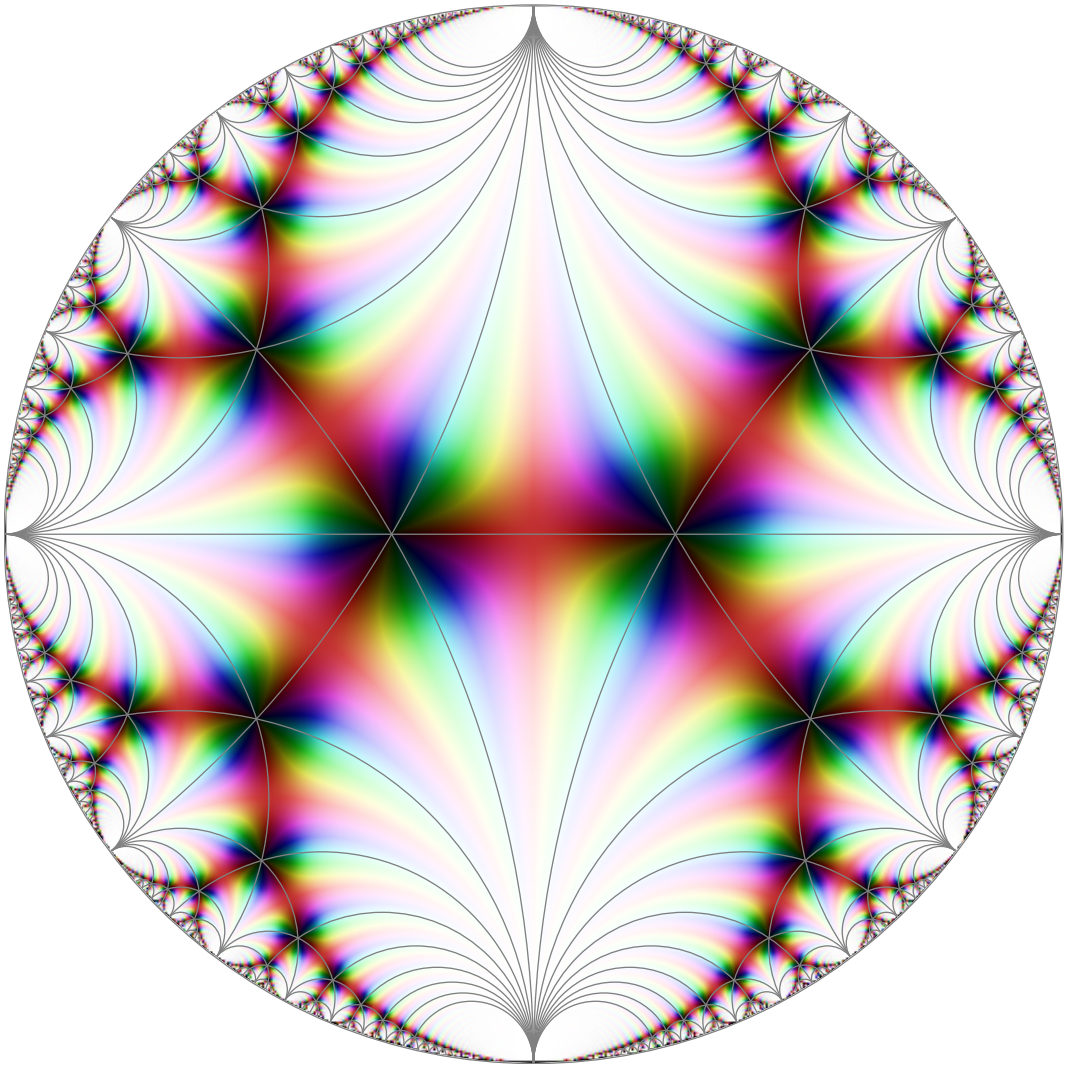

Figure 3.9: Klein's complete invariant J, defined on the upper half-plane, is visualized by dying each point z of the unit disk in the color $\mathcal{C} \circ J \circ \Phi^{-1}(z)$, where Φ^{-1} is the inverse modified Cayley transform which takes the unit disk bijectively to the upper half-plane.

2. $J(\infty) = \infty$: The point $\Phi(\infty) = i$ is colored in white.

3. $J(\mathcal{F}_{\text{left}}) = \mathcal{H}$ and $J(\mathcal{F}_{\text{right}}) = -\mathcal{H}$: The color tones in the left half of $\Phi\mathcal{F}$, that is $\Phi\mathcal{F}_{\text{left}}$, range from red, orange, yellow and green to cyan, and therefore $\arg(J(z)) \in (0, \pi)$ for $z \in \mathcal{F}_{\text{left}}$. Similarly we see $\arg(J(z)) \in (-\pi, 0)$ for $z \in \mathcal{F}_{\text{right}}$.

4. J has a zero of order 3 at every point equivalent to ρ: The points colored in black are precisely those vertices of all modular triangles[8] which lie in the interior of the unit disk. These are exactly the points equivalent to ρ. The order of a given zero (resp. pole) at a point z can be read off by counting the number of times we go through the set of all color tones while walking once around z along a sufficiently small circular path which does not enclose any other zero or pole except z. Applying this to the zero at ρ (or any of its equivalent points), this yields an order of 3, as we visit the color tone red (as well as every other color tone) exactly three times as we go once around this point.

The special importance of J lies in the fact that all modular functions can be represented as rational functions in J with complex coefficients (for the proof of this fact we refer to Schoeneberg [14], Chapter II, §3).

Theorem 3.26. *The set of modular functions is identical to the field $\mathbb{C}(J)$ of rational functions in J with coefficients in $\mathbb{C}$.*

Remark 3.27. The statement of Theorem 3.26 remains true when J is replaced by any modular function of the form $A \circ J$, where A is a Möbius transformation. In other words, we have $\mathbb{C}(J) = \mathbb{C}(A \circ J)$ for $A \in \mathrm{PGL}_2(\mathbb{C})$.

Example 3.28. In Figure 3.10, four examples for rational functions in J are given. Note that except $z \mapsto J(z)^2$, these maps are instances of modular functions of the form $A \circ J$, where A is a Möbius transformation.

[8] We consider the regions $\Phi A\mathcal{F}$, $A \in \mathrm{PSL}_2(\mathbb{Z})$ as triangles in the sense of hyperbolic geometry.

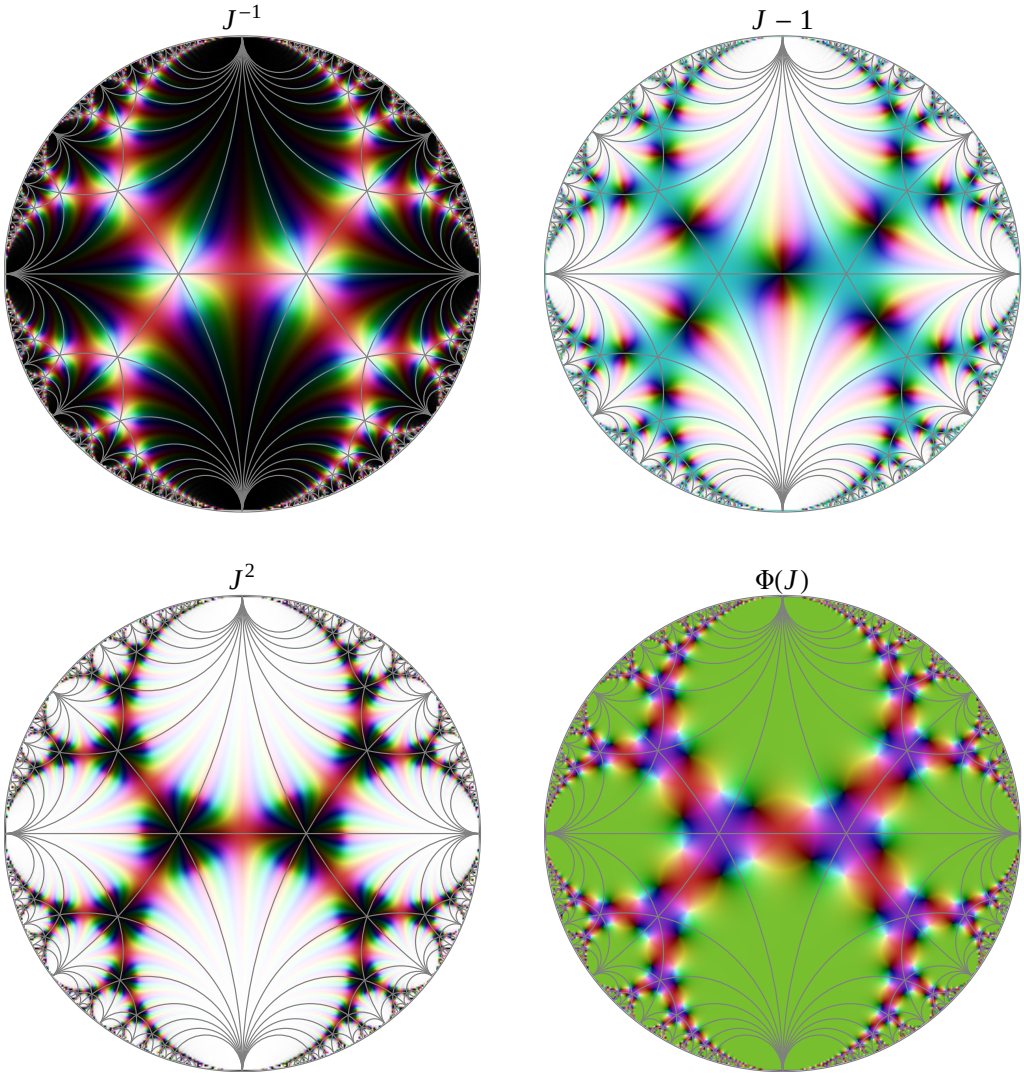

Figure 3.10: Examples for rational functions in J: The function $z \mapsto J(z)^{-1}$ (top-left) is zero at all rational points (hence the dominant black color) and has a pole of order 3 at all points equivalent to ρ. The function $z \mapsto J(z) - 1$ (top-right) has a zero of order 2 at all points equivalent to i. For $z \mapsto J(z)^2$ (bottom-left), the order of the zeros at points equivalent to ρ is six, i.e. twice the order of the zeros of the original function J at that points. Lastly, composing J with the Cayley transform, that is $z \mapsto \Phi \circ J(z)$ (bottom-right) yields a zero (resp. pole) of order 1 at every point z for which $J(z) = i$ (resp. $J(z) = -i$). Its value at rational points is $\Phi \circ J(\infty) = \Phi(\infty) = i$, explaining the dominant green coloring.

Example 3.29. As a final example, Figure 3.11 shows what happens, when J is composed with the generating function of the Fibonacci sequence. Although it is a slight digression, we will shortly give some background about the Fibonacci sequence here.

The Fibonacci sequence $0, 1, 1, 2, 3, 5, 8, \ldots$ is a so-called *C-finite sequence*, i.e. a sequence satisfying a recurrence of the form

$$\sum_{i=0}^{r} c_i a_{n+i} = 0 \quad \text{for all } n \geq 0,$$

where r is the order of the recurrence, c_i are fixed constants and a_n is the n-th element of the sequence. The Fibonacci sequence and its elements F_n satisfy the recurrence $F_{n+2} - F_{n+1} - F_n = 0$, $n \geq 0$, with initial values $(F_0, F_1) = (0, 1)$. The *generating function* of a sequence $(a_n)_{n\geq 0}$ is defined as

$$F(z) := \sum_{n\geq 0} a_n z^n.$$

For C-finite sequences, the generating function is always a rational function – see Theorem 4.3 in Kauers/Paule [6]. In particular, the generating function of the Fibonacci sequence is given by

$$F(z) := \sum_{n\geq 0} F_n z^n = \frac{z}{1 - z - z^2}.$$

More on the topic of sequences and generating functions, as well as symbolic sums and asymptotic estimates may be found in Kauers/Paule [6].

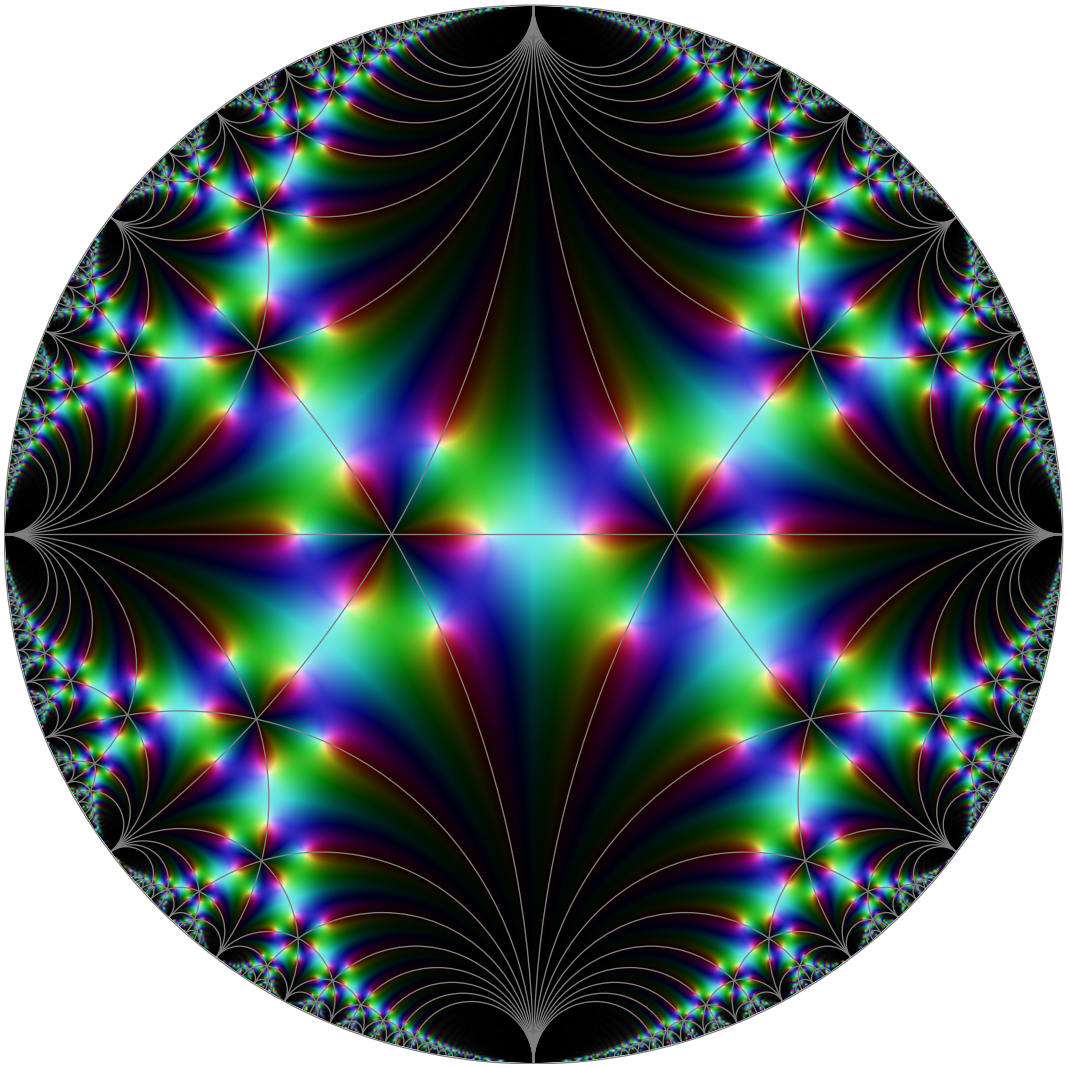

Figure 3.11: The modular function $G = F \circ J$, obtained by composition of the Klein's complete invariant J and the generating function of the Fibonacci sequence, $F(z) = \frac{z}{1-z-z^2}$, i.e. $G(z) = \frac{J(z)}{1-J(z)-J(z)^2}$, satisfies $G(\infty) = 0$, has zeros of order 3 at all points equivalent to ρ and poles of order 1 at all points $z \in \mathcal{H}$ satisfying $J(z) \in \{-\varphi, \frac{1}{\varphi}\}$, where $\varphi := \frac{1+\sqrt{5}}{2}$ denotes the golden ratio.

Index

Bibliography

[1] Douglas N Arnold and Jonathan Rogness. Möbius transformations revealed. *Notices of the AMS*, pages 1226–1231, 2008.

[2] C. Carathéodory. *Funktionentheorie*. Number Bd. 1 in Funktionentheorie. Birkhäuser, 1960.

[3] W. Fenchel. *Elementary Geometry in Hyperbolic Space*. De Gruyter studies in mathematics. Walter de Gruyter, 1989.

[4] Lester R Ford. Fractions. *The American Mathematical Monthly*, 45(9):586–601, 1938.

[5] T.W. Hungerford. *Algebra*. Graduate Texts in Mathematics. Springer, 1974.

[6] M. Kauers and P. Paule. *The Concrete Tetrahedron: Symbolic Sums, Recurrence Equations, Generating Functions, Asymptotic Estimates*. Texts & Monographs in Symbolic Computation. Springer Wien, 2011.

[7] F. Klein and R. Fricke. *Vorlesungen über die Theorie der elliptischen Modulfunktionen*. Bibliotheca Mathematica Teubneriana. Johnson Reprint Corporation, 1966.

[8] Joseph Lehner. *Discontinuous Groups and Automorphic Functions*. Mathematical Surveys and Monographs. American Mathematical Society, 1982.

[9] D. Mumford, C. Series, and D. Wright. *Indra's Pearls: The Vision of Felix Klein*. Indra's Pearls: The Vision of Felix Klein. Cambridge University Press, 2002.

[10] O. Ore. *Number Theory and Its History*. Dover Books on Mathematics. Dover Publications, 2012.

[11] O. Perron. *Die Lehre von den Kettenbrüchen*. Number Bd. 1-2 in B.G. Teubners Sammlung von Lehrbüchern auf dem Gebiete der mathematischen Wissenschaften mit Einschluss ihrer Anwendungen. B.G. Teubner, 1913.

[12] Hans Petersson. Über die Entwicklungskoeffizienten der automorphen Formen. *Acta Mathematica*, 58(1):169–215, 1932.

[13] Hans Rademacher. The Fourier series and the functional equation of the absolute modular invariant $J(\tau)$. *American Journal of Mathematics*, pages 237–248, 1939.

[14] Bruno Schoeneberg. *Elliptic modular functions*. Springer-Verlag Berlin, 1974.

[15] H. Schwerdtfeger. *Geometry of Complex Numbers*. Dover Books on Mathematics. Dover Publications, 2012.

[16] J. Züllig. *Geometrische Deutung unendlicher Kettenbrüche und ihre Approximation durch rationale Zahlen*. Orell Füssli Verlag, 1928.