Ergod. Th. & Dynam. Sys., (2023), **43**, 1666–1711
doi:10.1017/etds.2022.18

Ergodicity of Iwasawa continued fractions via markable hyperbolic geodesics

ANTON LUKYANENKO† and JOSEPH VANDEHEY‡

† *Department of Mathematics, George Mason University, 4400 University Drive, MS: 3F2, Fairfax, VA 22030, USA*
(e-mail: anton@lukyanenko.net)
‡ *Department of Mathematics, University of Texas at Tyler, Tyler, TX 75799, USA*
(e-mail: jvandehey@uttyler.edu)

(*Received* 4 *March* 2020 *and accepted in revised form* 7 *February* 2022)

Abstract. We prove the convergence and ergodicity of a wide class of real and higher-dimensional continued fraction algorithms, including folded and α-type variants of complex, quaternionic, octonionic, and Heisenberg continued fractions, which we combine under the framework of Iwasawa continued fractions. The proof is based on the interplay of continued fractions and hyperbolic geometry, the ergodicity of geodesic flow in associated modular manifolds, and a variation on the notion of geodesic coding that we refer to as geodesic marking. As a corollary of our study of markable geodesics, we obtain a generalization of Serret's tail-equivalence theorem for almost all points. The results are new even in the case of some real and complex continued fractions.

Key words: continued fractions, geodesic coding, ergodicity, complex continued fractions, Iwasawa continued fractions, Heisenberg continued fractions
2020 Mathematics Subject Classification: 37D40 (Primary); 11K50, 37A45 (Secondary)

> *'. . .attempts to find a precise relation between the cutting sequence of [a geodesic] γ and the continued-fraction expansions of endpoints of suitable lifts of γ are fraught with minor discrepancies.'*
>
> —Caroline Series [57]

1. *Introduction*

1.1. *Background.* Since the early work by Lagrange and Gauss linking regular continued fractions (CFs) to algebra and dynamical systems, an extensive and ongoing effort has focused on expanding the scope of CF theory to new algorithms. While regular CFs represent the fractional part $x - \lfloor x \rfloor$ of a real number $x \in \mathbb{R}$ as a descending

iterated fraction:

$$\cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}} \tag{1.1}$$

with positive integer digits, a menagerie of one-dimensional CF variants have been formed by modifying various aspects of this simple construction: whether by changing positive quantities to negative, altering the set of allowable digits, or selecting a different set of numbers to have expansions. (See §1.2 for an introduction to many of these variants.)

After over 200 years of study, the one-dimensional CFs are largely well understood. Most of them inherit the essential properties of regular CFs from the viewpoints of algebra, dynamics, and geometry: Lagrange's theorem, shift map ergodicity, and Diophantine interpretation, respectively. The study of one-dimensional CFs has been facilitated by a connection to hyperbolic geometry, pioneered by Artin [3] and developed by Series [57], Katok and Ugarcovici [33–35], and others. In particular, Artin observed that the Gauss map for regular CFs can be identified with a section of geodesic flow in a finite cover of the modular surface; leading to extensive developments in both CF theory and the study of geodesics on hyperbolic manifolds.

In trying to extend these properties beyond one-dimensional CFs, one is immediately confronted by the question of how to generalize one-dimensional CFs to more than one dimension. Several algorithms, such as those of Jacobi–Perron, Brun, and Selmer, act by building up the CF expansion to several different real values simultaneously [55]. Other algorithms, such as the Hurwitz complex CF algorithm [30] or the Heisenberg CF algorithm studied previously by the authors [38], treat points in these spaces as single entities with a single continued fraction expansion. (Yet another type of CF-like algorithm deriving more from geometric properties can be seen in [24, 27].) This is analogous to how complex points can be understood either via their real and complex part (that is, essentially in $\mathbb{R}^2$) or as an element in complex space (in $\mathbb{C}$). In this paper, we will generally be interested in the latter form of higher-dimensional CF expansion, as it has a more natural connection to hyperbolic geometry.

The story of these higher-dimensional CFs has been markedly different from their real CF cousins. Despite interest in these topics stretching back to the 1850s [22, 23], only a small number of algorithms are known to be well behaved. Among them is the A. Hurwitz complex CF [30], which represents a complex number z with real and imaginary parts both in $[-1/2, 1/2)$ as a descending iterated fraction:

$$\cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}, \quad a_i \in \mathbb{Z}[\mathbb{i}] \setminus \{0, \pm 1, \pm \mathbb{i}\}. \tag{1.2}$$

(See §1.2.6 for a full description.) Proofs of, for example, ergodicity for these well-behaved algorithms are extremely delicate [45]: the space of the algorithm has a serendipitous decomposition, which results in a finite range property among other features, and this

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

allows high-powered results (such as those in [32, 56]) to be applied. Should the algorithm be perturbed, even slightly (see §1.2.6 and Figure 2), the decomposition will break down and the methods will no longer apply.

As a major goal of this paper is to prove properties like ergodicity for a larger variety of higher-dimensional CFs (including perturbed variations of standard algorithms), let us discuss some of the roadblocks to using traditional techniques. First of all, one does not expect the structure of the cylinder sets (the sets of numbers whose expansions all start with the same sequence of digits) to have a simple structure, so methods like those cited above will not apply. Second, the natural extension of even some one-dimensional CF variants (see [2, §7]) as well as simple higher-dimensional CFs (see [18, 28]) is already fractal in nature, which makes it difficult to prove results about the natural extension, let alone about the simpler algorithm. Third, making a precise connection between CF digits and geodesic coding is 'fraught with minor discrepancies' and in its strong form would imply properties such as Serret's tail equivalence theorem [50] that are known to fail for higher-dimensional CFs. Indeed, the geodesic coding approach has long been considered 'intrinsically two-dimensional' [1].

In this paper, we develop a softer version of geodesic coding, which we refer to as *geodesic marking*. In a typical geodesic coding (what we describe here is an arithmetic coding, in the terminology of Katok and Ugarcovici [33], where codings formed by cutting sequences are related), we look at a given geodesic from two perspectives: first, we have a bi-infinite sequence formed by the continued fraction digits of both the forward and backward endpoints of the geodesic, and second, we have a bi-infinite sequence of intersections of our geodesic with a particular cross-section. A shift in one sequence should correspond to a shift in the other. In particular, returning to the cross-section after flowing along the geodesic should move the CF expansion forward one digit. Our geodesic *marking* still has the two bi-infinite sequences, but now returning to the cross-section can move the CF expansion forward several digits at a time. Thus the first-return map to the cross-section now corresponds to a jump transformation for the continued fraction. This jump transformation, in practice, skips over strings of small digits. (What counts as a small digit could be made effective, but we do not do so here.) It should be noted that small digits appear to cause some of the roadblocks mentioned above: cylinder sets associated with small digits tend to be irregular, while those of large digits are far better behaved, for instance. So, in essence, geodesic markings skip over the troublesome parts of CF algorithms.

Geodesic marking provides a robust connection between these higher-dimensional CFs and hyperbolic geometry which is preserved even under perturbation of the algorithm. The following theorem illustrates some significant cases to which our work applies.

THEOREM 1.1. *Folded complex CFs, folded Hurwitz quaternionic CFs, folded octonionic CFs, and folded Heisenberg CFs, as well as their α-type variants, are convergent and ergodic.*

In particular, this illustrates how our work applies to several different spaces (complex numbers, quaternionic numbers, octonionic numbers, and the Heisenberg group) and many

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

systems within those spaces (folded and α-type variants are discussed in more detail in §1.2, see also Figure 2). Our results also apply to several one-dimensional CF algorithms, such as folded real CFs and some of Nakada's α-CFs (see §§1.2.3 and 1.2.4).

While convergence follows standard arguments, the ergodicity statement is a substantial breakthrough for higher-dimensional CFs, where it was only previously known for specific complex CF variants, such as the A. Hurwitz and J. Hurwitz CFs. Our approach furthermore provides a flexible, unifying method for understanding both one-dimensional and higher-dimensional CFs.

Theorem 1.1 follows from a more general result concerning CFs on boundaries of rank-one symmetric spaces of non-compact type, which we refer to as *Iwasawa inversion spaces*. CFs were first extended to this setting by the authors in [**38**], where a CF theory on the non-commutative Heisenberg group was proposed. In [**9**], Chousionis–Tyson–Urbanski, studying conformal iterated function systems, defined Iwasawa continued fractions on the closely related *Iwasawa groups* (see §1.4). Here, we extend the definition of Iwasawa CFs to an Iwasawa CF *algorithm* associating a digit sequence to each point in an Iwasawa inversion space and leverage the connection to hyperbolic geometry to prove the following theorem.

THEOREM 1.2. *Every discrete and proper Iwasawa CF is convergent. Moreover, if it is complete, then it is ergodic.*

We will postpone the full definitions of these terms until §2, but will provide some insight into them now. Discreteness simply says that the modular group $\mathcal{M}$ associated with our CF algorithm acts discretely on the corresponding hyperbolic space. It is necessary to ensure that we have a finite-volume hyperbolic manifold (generalizing the modular surface) in which to look at geodesic flow. Properness says that the only points under consideration for our CF algorithm have norm bounded away from 1. This guarantees that CF expansions converge quickly, among other properties. Properness also helps us avoid indifferent fixed points in our dynamical system, which have been noted before to cause infinite invariant measures [**13**]. Completeness says that the set of digits for our CF algorithm is maximal in an appropriate sense, the upshot of which is that the sequence of digits in the CF expansion of a point is functionally the only expansion the hyperbolic geometry can see.

In the case where a system is not complete, we can still obtain a partial result.

THEOREM 1.3. *Let $T : K \to K$ be the shift map for a discrete and proper Iwasawa CF with $n \geq 1$ central symmetries. Then T has at most n ergodic components.*

A full description of central symmetries will appear in §3.7. For the moment, we can consider centrally symmetric systems as ones where the system is incomplete due to the appearance of hidden symmetries, such as $x \mapsto -x$, as in §1.2.4.

The use of geodesic marking, as opposed to classical geodesic coding, is critical to Theorem 1.2. As noted above, one typical corollary of geodesic coding is Serret's tail-equivalence theorem for every point. That is, two points lie in the same orbit of the modular group $\mathcal{M}$ if and only if the tails of their CF digit sequences agree. However,

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

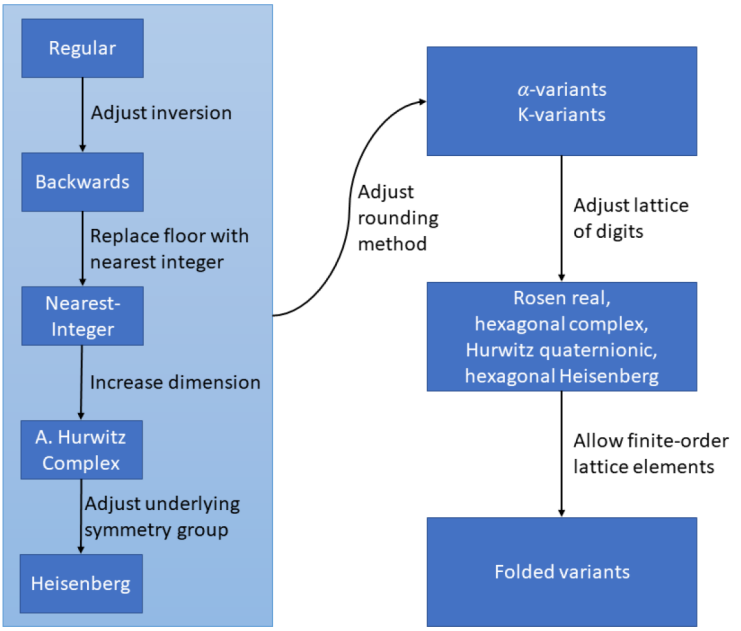


FIGURE 1. Different Iwasawa CF algorithms can be thought of as variations on the regular CF algorithm: the equations for the shift map, digit extraction algorithm, and recombination algorithm remain the same, while the underlying data are adjusted.

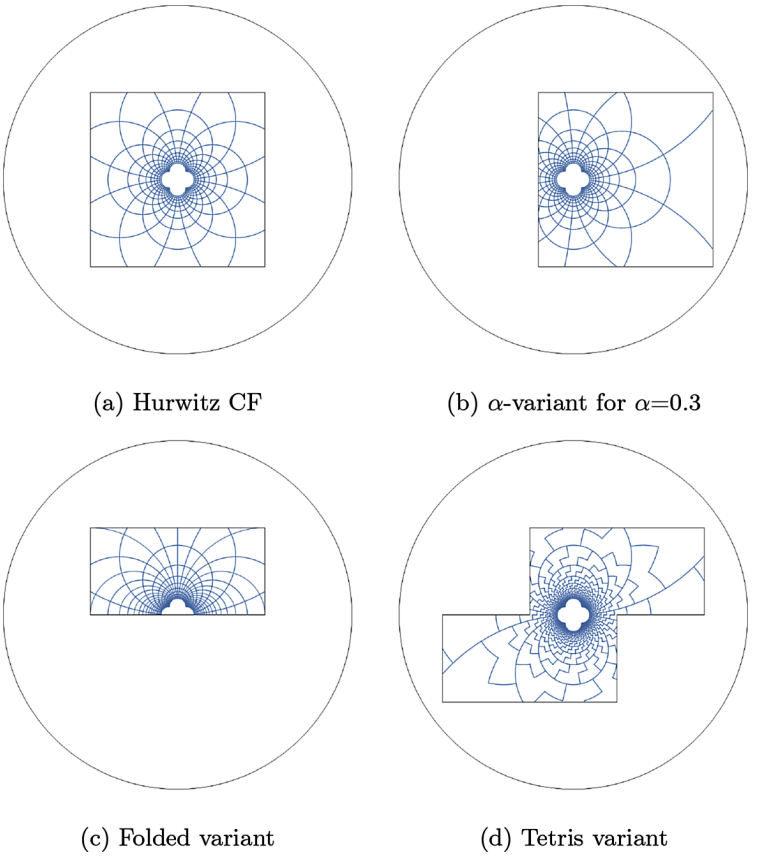
for the A. Hurwitz complex CFs, Lakein [36] provides an explicit counterexample to tail-equivalence. (Lakein's counterexample makes use of an element not belonging to $\mathcal{M}$, but this can be remedied by multiplying his choice of A by $\mathtt{i}$.) Thus, one would not expect for geodesic coding to be available in this case. Geodesic marking, however, avoids certain points which exhibit pathological behavior, those with all small CF digits. This allows for an ergodicity result and leads to the following almost everywhere (a.e.) tail equivalence result for Iwasawa CFs (proven as Theorem 6.17), which is novel for all higher-dimensional algorithms including *folded* A. Hurwitz complex CFs:

THEOREM 1.4. *Almost surely, two points in a complete, discrete, and proper Iwasawa CF are tail-equivalent if and only if they are $\mathcal{M}$-translates of one another.*

The question of tail-equivalence is being actively researched even for one-dimensional CFs, see [5, 50]. The importance of small digits versus large digits to tail-equivalence has been noted before in [48].

1.2. *Key examples of CF algorithms.* We now describe a number of well-known variants of continued fractions, primarily in the one-dimensional case, which are of interest to us. We will discuss the algorithms in an increasing order of complexity (see Figure 1 for a diagram), pointing out the variations that motivate the definition of Iwasawa CFs: namely, the choice of underlying space, inversion, digit sequence, and fundamental domain for the corresponding lattice; as well as the definitions of properness, completeness, and

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press



(a) Hurwitz CF (b) α-variant for α=0.3

(c) Folded variant (d) Tetris variant

FIGURE 2. Four variants of the Hurwitz complex CF algorithm. The fundamental domain K in each case is displayed inside the unit circle (fixed by the inversion ι_c), and is decomposed into rank-1 cylinder sets. The lattice $\mathcal{Z} = \mathbb{Z}^2$ is extended by the rotation $(x, y) \mapsto (-x, -y)$ in the folded variant.

discreteness. A more thorough discussion of the class of Iwasawa CFs is provided in §3, along with a more complete list of known Iwasawa CF algorithms in Table 1.

1.2.1. *Regular CFs.* The *regular continued fraction* representation of a number $x \in [0, 1)$ represents it as a limit

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}},$$

where $a_i \in \mathbb{N}$. (In the introduction, we ignore the behavior of points with finite CF expansion for simplicity.) The digits a_i are extracted from x by repeated applications of the Gauss map $T(x) = 1/x - \lfloor 1/x \rfloor$:

$$a_i = \left\lfloor \frac{1}{T^{i-1}x} \right\rfloor.$$

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

The Gauss map is famously ergodic with an invariant measure given by the density $(1/\log 2)(1/(1+x))$. (See [14] for a fuller treatment.)

In the framework of Iwasawa CFs, regular CFs are described using the following data:

(1) the underlying space X is $\mathbb{R}$;
(2) the inversion used is $\iota(x) = 1/x$;
(3) the allowed digits are elements of the lattice $\mathcal{Z} = \mathbb{Z}$;
(4) the set of 'fractional points' is $K = [0, 1)$, which tiles $\mathbb{R}$ under integer translations.

As we show below, many standard and novel algorithms can be described by adjusting the above data and leaving the formulas above essentially unchanged.

1.2.2. *Backwards CFs.* The *backwards CF* (sometimes called Rényi CF) reverses the domain of the Gauss map to produce the Rényi map

$$T_R(x) = T_G(1-x) = \frac{1}{1-x} - \left\lfloor \frac{1}{1-x} \right\rfloor.$$

The CF digits of $x \in [0, 1)$ are then extracted in an analogous manner to the one used for standard CFs, via

$$a_i = \left\lfloor \frac{1}{1 - T^{i-1}x} \right\rfloor, \tag{1.3}$$

and recombined as

$$x = 1 - \cfrac{1}{a_1 + 1 - \cfrac{1}{a_2 + \cdots}}.$$

The shift map T_R is ergodic, but due to the presence of indifferent fixed points, the corresponding invariant measure is infinite [1].

With a small adjustment, backwards CFs fit into the framework of Iwasawa CFs, as follows.

The mapping $x \mapsto (1-x)$ conjugates backwards CFs to an equivalent system known as the D-backwards CF with $D = [0, 1)$, see Masarotto [41]. The resulting shift map is then given by $T_D(x) = -1/x - \lfloor -1/x \rfloor$. Adjusting Masorotto's notation by using negative integer digits $a_i < -1$, we take

$$a_i = \left\lfloor \frac{-1}{T^{i-1}x} \right\rfloor \tag{1.4}$$

and recombine the digits as

$$x = \cfrac{-1}{a_1 + \cfrac{-1}{a_2 + \cdots}}.$$

All three CF algorithms discussed so far are real algorithms looking at points in $[0, 1]$ which use integers for their digits. The only difference between them is the choice of inversion: $x \mapsto 1/x$ for regular CFs, $x \mapsto 1/(1-x)$ for backwards CFs, and $x \mapsto -1/x$ for D-backwards CFs.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

In the Iwasawa CF formalism, we will assume that inversions send 0 to ∞ and preserve the unit circle. While the backwards CF algorithm *a priori* does not fit this requirement, the conjugate D-backwards system is an Iwasawa CF.

We will make use of both of the allowed inversions $\iota_+(x) = 1/x$ and $\iota_-(x) = -1/x$ throughout the paper.

Interestingly, backwards continued fractions are the more natural system within the framework of Iwasawa CFs. The inversion ι_- is an orientation-preserving linear-fractional mapping, and is an element of the modular group $PSL(2, \mathbb{Z})$, while ι_+ is orientation-reversing, which forces us to consider the larger group $PGL(2, \mathbb{Z})$. This leads us to the question of *completeness* of the digit set, see §1.2.4.

1.2.3. *Nearest-integer and α-type CFs.* The next set of CF algorithms adjusts the set of 'fractional' points and the corresponding rounding method, while also allowing variation in the choice of inversion.

A nearest-integer CF replaces the unit interval $[0, 1)$ with the interval $[-1/2, 1/2)$, and the floor function $\lfloor\cdot\rfloor$ with the nearest-integer mapping $[\cdot]$. There are three standard systems known as nearest-integer CFs, of which two fit directly into the Iwasawa CF framework, and the third is semi-conjugate to an Iwasawa CF. The first two systems are constructed by choosing the inversion function ι to be either $\iota_+(x) = 1/x$ or $\iota_-(x) = -1/x$. The corresponding shift map is given by $T(x) = \iota(x) - [\iota(x)]$, and one has the digits $a_i = [\iota(T^{i-1}x)]$, which are still integers. The third system is based on the shift map $T(x) = |1/x| - [|1/x|]$ and a more complicated system of digits, which we will discuss more extensively in §1.2.4. Due to non-injectivity of the mapping $x \mapsto [|1/x|]$, this third variant does not fit the Iwasawa CF framework.

The α-type CFs, with $\alpha \in [0, 1]$, form a family of CF algorithms that interpolate between regular and nearest-integer CFs by operating with the interval $[-\alpha, 1 - \alpha)$ and the corresponding rounding function $[x]_\alpha = [x + \alpha]$. The forward shift is given by $T(x) = \iota(x) - [\iota(x)]_\alpha$, where ι is chosen from ι_+, $\iota = \iota_-$, or $x \mapsto |1/x|$. As above, the first two choices fit the Iwasawa CF framework, while the third variant does not. All three families of systems are known to be ergodic for all $\alpha \in [0, 1]$.

- Ergodicity of the ι_+ variant for $\alpha \notin \{0, 1\}$ follows from our results and for $\alpha \in [0, 1]$ was simultaneously shown by [49].
- Ergodicity of the ι_- variant for $\alpha \notin \{0, 1\}$ is new in this paper (cf. [2]). The cases $\alpha = 0$ and $\alpha = 1$ are also ergodic, since $\alpha = 0$ gives the backwards CF and $\alpha = 1$ gives a system that is conjugate to the regular CF.
- Ergodicity of the $x \mapsto |1/x|$ variant was recently proven in [49].

Generalizing further, one can replace the unit interval with any measurable set K that tiles $\mathbb{R}$ under integer translations and write $T(x) = \iota(x) - [\iota(x)]_K$, where $[x]_K$ denotes the unique integer satisfying $x - [x]_K \in K$. Such systems fall under the framework of Iwasawa CFs. Our results imply that the CF is convergent and the shift map is ergodic as long as K is proper: that is, the closure of K is contained in the open unit ball $(-1, 1)$. Note that regular and backward CFs are not proper, but are nonetheless convergent and ergodic.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

1.2.4. *Folded CFs.* We now discuss in more detail the nearest-integer system based on the shift map $T(x) = |1/x| - [|1/x|]$. Because the mapping $x \mapsto |1/x|$ is 2-to-1, the standard approach is to keep track both of the integer digit and the choice made when taking the absolute value

$$b_i = [|1/x|], \quad c_i = \operatorname{sign}(x),$$

so that one reconstructs

$$x = \cfrac{c_1}{b_1 + \cfrac{c_2}{b_2 + \cdots}}.$$

To maintain similarity to previous algorithms, we combine the integer digit b_i and the sign c_i into a single datum, namely the linear mapping $a_i(x) = c_i(x + b_i)$. This allows us to rewrite the fraction in the format

$$x = \cfrac{1}{a_1\left(\cfrac{1}{a_2(\cdots)}\right)} = \lim_{n\to\infty} \iota_+ a_1 \iota_+ a_2 \ldots \iota_+ a_n(0),$$

where each a_i is now a function and we take the convention of suppressing parentheses and composition signs.

We thus transition from thinking of digits as elements of $\mathbb{Z}$ to thinking of them as automorphisms of $\mathbb{R}$. In the Iwasawa CF framework, we will assume that these automorphisms are isometries of the underlying space, which is indeed the case here.

Therefore, for the algorithm under discussion, we are now interested in digits in the expanded lattice $\mathcal{Z}$ generated by integer translations and negation, that is, $\mathcal{Z} = \langle x \mapsto x + 1, x \mapsto -x\rangle$.

Inconveniently, moving the set of 'fractional' points $K = [-1/2, 1/2)$ around by the group $\mathcal{Z}$ causes overlaps, and we therefore exclude this CF variant from the class of Iwasawa CFs.

Adjusting to the interval under consideration to $K = [0, 1/2]$ provides a non-overlapping tiling of $\mathbb{R}$ (that is, K is a fundamental domain for the action of $\mathcal{Z}$), giving the *folded CF* (see Marmi–Moussa–Yoccoz [40]) that now does fit in the Iwasawa CF framework.

The folded CF algorithm is defined by the following data:

(1) the underlying space X is $\mathbb{R}$;
(2) the inversion used is $\iota_+(x) = 1/x$;
(3) the group $\mathcal{Z}$ of allowed digits is generated by $x \mapsto x + 1$ and $x \mapsto -x$;
(4) the set of 'fractional points' is $K = [0, 1/2)$, which tiles $\mathbb{R}$ under the action of $\mathcal{Z}$.

Given these data, we obtain a rounding function $x \mapsto [x] \in \mathcal{Z}$ that now provides the unique linear mapping $[x] \in \mathcal{Z}$ combining an integer translation and possibly a negation such that $[x]^{-1}(x) \in [0, 1/2)$. For example, we have that $[5.1](x) = 5 + x$ and $[5.1]^{-1}(x) = x - 5$, while $[5.7](x) = -(x - 6)$ and $[5.7]^{-1}(x) = -(x - 6)$.

For a point $x \in [0, 1/2)$, we can then write the forward shift map as $T(x) = [1/x]^{-1}(1/x)$ and extract the digits as $a_i = [1/T^{i-1}(x)]$. The point x is reconstructed

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

from the digits by writing

$$x = \lim_{n\to\infty} \frac{1}{a_1\left(\dfrac{1}{a_2(\ldots a_n(0))}\right)},$$

or more compactly as $x = \lim_{n\to\infty} a_1\iota a_2\iota \ldots a_n(0)$.

The absolute value mapping from $(-1/2, 1/2)$ to $(0, 1/2)$ then provides a semiconjugacy between the $|\cdot|$-based nearest-integer fractions and the folded CFs. Ergodicity passes down (but not up!) through semiconjugacy, so folded CFs are ergodic. See Marmi–Moussa–Yoccoz [40] for the corresponding invariant measure.

While Marmi–Moussa–Yoccoz do describe folded variants of all α-CFs, it is only the nearest-integer variant $\alpha = 1/2$ that fits within the Iwasawa CF framework, since the other systems continue to operate with fractional sets K that are not fundamental domains for any relevant lattice.

As it turns out, the folded CFs also arise naturally from the regular CF construction, where we have $\mathcal{Z} = \mathbb{Z}$ and $\iota = \iota_+$. Since the shift map combines both elements of $\mathcal{Z}$ and the mapping ι, analysis of the shift map revolves around understanding the group $\mathcal{M} = \langle \mathbb{Z}, \iota_+ \rangle$. The group $\mathcal{M}$ includes the negation mapping $x \mapsto -x$ since

$$\frac{1}{1 + \dfrac{1}{-1 + \dfrac{1}{1+x}}} = -x, \tag{1.5}$$

so that the subgroup $\mathcal{Z}' \subset \mathcal{M}$ of linear transformations (that is, the stabilizer of ∞) is the group $\mathcal{Z}' = \langle x \mapsto x + 1, x \mapsto -x \rangle$. We thus have that the group of allowed digits $\mathcal{Z} = \mathbb{Z}$ is smaller than the natural group $\mathcal{Z}'$ of linear transformations, giving what we call an *incomplete* system. Expanding the set of digits to $\langle \mathbb{Z}, x \mapsto -x \rangle$ while also contracting the fundamental domain to $[0, 1/2)$ provides a completion of the system, again giving us the folded fractions.

1.2.5. *Rosen CFs.* We finish the discussion of one-dimensional CFs with the Rosen CFs, whose definition is motivated by connections to hyperbolic geometry of triangle groups.

To define Rosen CFs, one takes the group $\mathcal{Z}$ of allowed digits to be $(2\cos(\pi/q))\mathbb{Z}$, and the set of 'fractional points' to be $K = [-\cos(\pi/q), \cos(\pi/q))$.

Together with the inversion ι_-, the lattice $\mathcal{Z}$ generates a Hecke group, which acts discretely on the hyperbolic plane (with the case $q = 2$ reducing to the modular group $PSL(2, \mathbb{Z})$).

From here, the choice of $\iota = \iota_-$ (as used by [43]) would provide an Iwasawa CF algorithm; and our results imply that the corresponding shift map is ergodic. We emphasize that the discreteness of $\mathcal{M} = \langle \mathcal{Z}, \iota_- \rangle$ within the isometry group of hyperbolic space plays a key role in our proof, and that other choices of multiplier in front of $\mathcal{Z}$ would yield badly behaved systems.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Lastly, we note that Rosen's original CF algorithm instead is based on the mapping $x \mapsto |1/x|$, and is shown to be ergodic (in fact, weak Bernoulli) in [6]. This algorithm is not encompassed by the Iwasawa CF framework.

1.2.6. *Complex CFs.* We now briefly touch on higher-dimensional CFs, in the planar case. For more higher-dimensional CFs, including quaternionic and Heisenberg CFs, see the discussion in §3.

Our primary example is the A. Hurwitz complex CF, first defined in [31]. It is described by the following Iwasawa CF data:

(1) the underlying space X is $\mathbb{C}$;
(2) the inversion used is $\iota(z) = 1/z$;
(3) the group $\mathcal{Z}$ of allowed digits is the group of Gaussian integers, $\mathbb{Z}[\mathbb{i}]$;
(4) the specified fundamental domain K of $\mathcal{Z}$ is the unit square centered at the origin.

Thus, the shift map is given by $T(z) = 1/z - [1/z]$, where $[\cdot]$ finds the nearest Gaussian integer; the digits are extracted via $a_i = [1/T^{i-1}z]$, and reconstructed as

$$z = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}}$$

It is common to write the system in real coordinates, with corresponding data:

(1) the underlying space X is $\mathbb{R}^2$;
(2) the inversion used is $\iota(x, y) = (x, -y)/(x^2 + y^2)$, (we will denote such conjugate-reflections by ι_c);
(3) the group $\mathcal{Z}$ of allowed digits is the group $\mathbb{Z}^2$;
(4) the specified fundamental domain K of $\mathcal{Z}$ is the unit cube centered at the origin, i.e. $[-1/2, 1/2) \times [-1/2, 1/2)$.

Both the real and complex descriptions of the Hurwitz CF are quite natural: the mappings ι and $\mathbb{Z}^2$ both lift to isometries of *real* hyperbolic 3-space, while the corresponding modular group $\mathcal{M} = \langle \mathcal{Z}, \iota \rangle$ is shown to be discrete by embedding into $PSL(2, \mathbb{Z}[\mathbb{i}])$, see Proposition 3.15.

Ergodicity of the Hurtwitz CF was shown by Nakada in [45] (cf, [26]).

As in the case of ι_+ real CFs, the system is not complete, since the stabilizer of ∞ in $\mathcal{M}$ contains the unexpected mapping $z \mapsto -z$:

$$\cfrac{1}{1 + \cfrac{1}{-1 + \cfrac{1}{1+z}}} = -z.$$

As in the case of real folded fractions, we can create a folded variant by extending $\mathcal{Z}$ to include negation and reducing K correspondingly. For example, one could take $K = [-1/2, 1/2) \times [0, 1/2)$. In general, we will call a CF algorithm a *folded variant* if $\mathcal{Z}$ is expanded to the stabilizer of ∞ in $\mathcal{M}$, and K is similarly reduced.

One can likewise create α-type variants by shifting the location of the fundamental domain, that is, replacing K with $K + \alpha$; or create more exotic variants by choosing

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

a different fundamental domain entirely, e.g. by choosing a tetromino to create the Tetris CFs. However, it is not the case that we can arbitrarily shift folded variants. For example, the set $[-1/2, 1/2) \times [-1/4, 1/4)$ is not a fundamental domain for the group $\langle \mathbb{Z}[\mathbb{i}], (x, y) \mapsto (-x, -y) \rangle$.

See Figure 2 for illustrations of these algorithms and some of their cylinder sets. The finite range condition appears to fail in these cases. We recover ergodicity for folded variants. For centrally symmetric systems like the Tetris variant, we are able to bound the number of ergodic components by 2. The α-variant with $\alpha = 0.3$ shown in the figure is not complete and not centrally symmetric with respect to $z \mapsto -z$, and thus none of the results of this paper apply to it.

1.3. *Theorem 1.2 in a special case.* We now outline our proof of ergodicity in the case of nearest-integer CFs with inversion ι_-, where some simplifications are possible (cf. Remark 1.5). Ergodicity is certainly not new in the nearest-integer case, and connections to geodesic flow have been used since at least the work of Adler–Flatto [**1**]. (From a historical perspective, using ergodicity of geodesic flow to prove the ergodicity of a CF alogrithm is backwards. The ergodicity of regular CFs was shown first, and the ergodicity of geodesic flow in the modular surface was proven using this [**3**, **25**].) For a more thorough treatment of these techniques in the regular CF case, we recommend [**19**, §9.6].

We start by viewing $\mathbb{R}$ as the real axis in $\mathbb{C}$, and interpret the upper half-plane as the hyperbolic plane $\mathbb{H}^2_{\mathbb{R}}$. Both the integer shifts $\mathcal{Z}$ and the inversion ι on $\mathbb{R}$ extend to the half-plane, where they now act by isometries. The modular group $\mathcal{M}$ generated by $\mathcal{Z}$ and ι acts on $\mathbb{H}^2_{\mathbb{R}}$ discretely, and gives rise to a tiling of the space by translates of the tile $\mathcal{T}$ bounded by the vertical lines $x = \pm 1/2$ and the unit circle $\mathbb{S}$. Notably, each of the lines $x = \pm 1/2$ are equal to $M\mathbb{S}$ for an appropriate $M \in \mathcal{M}$. We will study hyperbolic geodesics γ, which takes the form of either a vertical line or a semi-circle that intersects $\mathbb{R}$ at right angles.

We will derive ergodicity for the CF shift map from the ergodicity of the geodesic flow on the modular surface $\mathcal{M}\backslash\mathbb{H}^2_{\mathbb{R}}$, which we can think of as the tile $\mathcal{T}$ with 'opposite sides' identified. That is, the sides $x = \pm 1/2$ are identified by the translation $z \mapsto z + 1$, and the two halves of the circular arc at the bottom are identified via $z \mapsto -1/z$. By Mautner's Theorem 6.3, geodesic flow in $\mathbb{H}^2_{\mathbb{R}}$ is ergodic. In particular, a generic geodesic γ is dense in $\mathcal{M}\backslash\mathbb{H}^2_{\mathbb{R}}$, see Figure 3.

It appears to be intuitively clear that, for a geodesic $\gamma \subset \mathbb{H}$, the continued fraction expansion of the forward endpoint $\gamma_+ \in \mathbb{R}$ can be immediately read off from the sequence of tiles that γ traverses in $\mathbb{H}^2$, or, equivalently, from the sequence of elements of $\mathcal{M}$ that are used to normalize it back to the starting tile. Indeed, it appears that the inversion corresponds to γ crossing $\mathbb{S}$ and the digits count the number of vertical lines crossed before returning to $\mathbb{S}$ after an inversion. Our goal will be to formalize this relationship in sufficient detail to prove the ergodicity of the shift map T from the ergodicity of the geodesic flow on $\mathcal{M}\backslash\mathbb{H}^2_{\mathbb{R}}$, doing so without relying on two-dimensional geometry, which has been central to previous approaches.

Let γ be a vertical geodesic as in Figure 3. Let $a_1 = [-1/\gamma_+]$ be the first nearest-integer CF digit of γ_+ and $M_1^{-1}(z) = -1/z - a_1$ the corresponding element of $PSL(2, \mathbb{Z})$

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

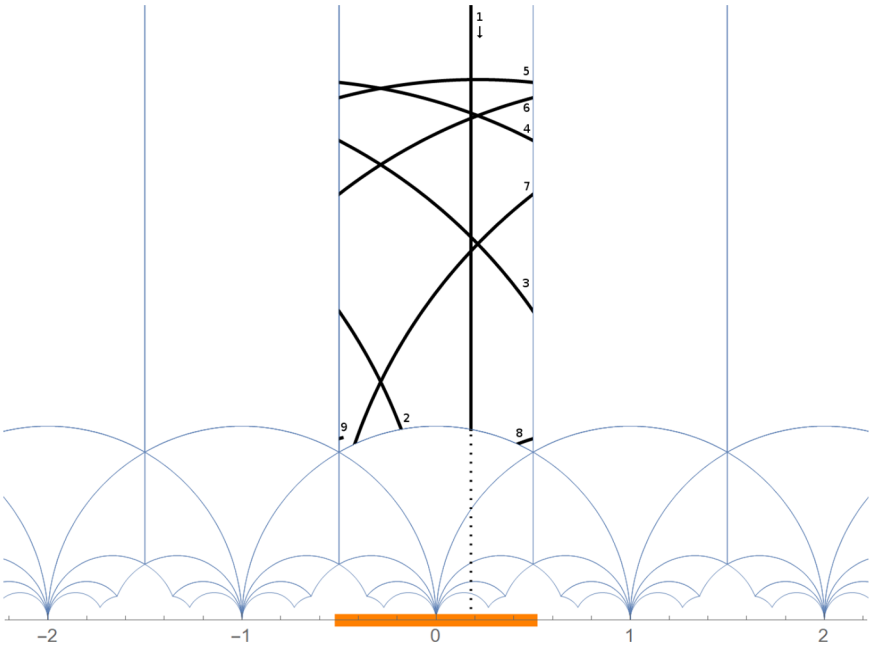


FIGURE 3. Wall-crossings of the vertical geodesic $x = 0.1795$ can be used to renormalize it to always stay within the fundamental domain for $\mathcal{M}$.

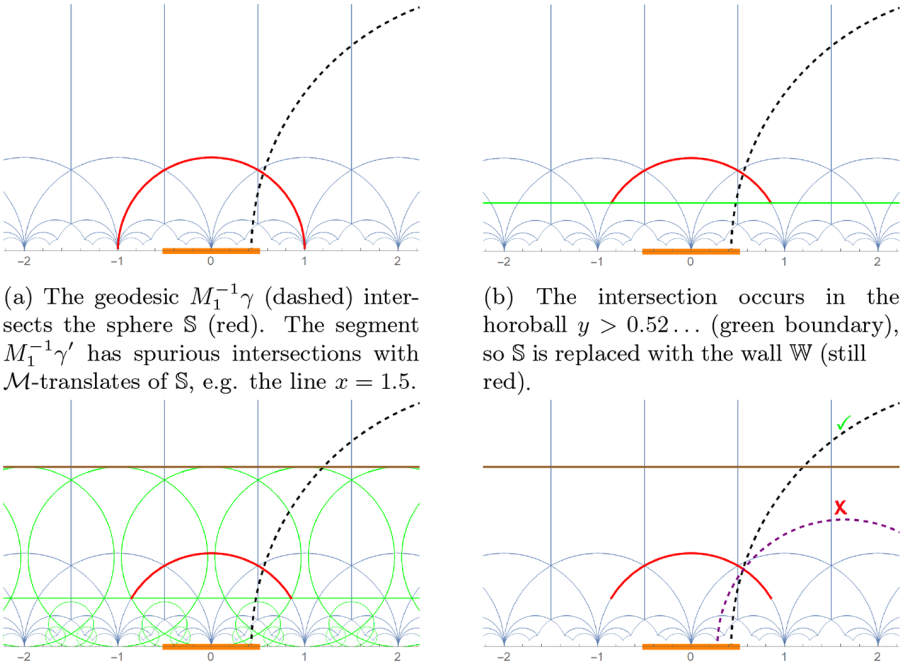
enacting the nearest-integer CF shift $T(\gamma_+)$. Applying M_1^{-1} to all of γ, we obtain Figure 4a. We denote the natural elements of $PSL(2,\mathbb{Z})$ enacting T^i by M_i^{-1}.

Consider now the subsegment γ' of γ strictly between the intersection with $\mathbb{S}$ and before the intersection with $M_1\mathbb{S}$. The intersection of γ with $M_1\mathbb{S}$ can be used to recover the first CF digit of γ_+, since we have that $M_1^{-1}(z) = -1/z + a_1$. However, intersections of γ' with other $\mathcal{M}$-translates of $\mathbb{S}$ do not correspond to digits of γ_+. We wish to find a subset of $\mathbb{S}$ for γ to intersect with that does *not* detect these 'spurious' intersections of γ' seen in Figure 4a, but does continue to detect (most of) the crossings of γ with $M_i\mathbb{S}$. In this way, intersections of γ with $\mathcal{M}\mathbb{S}$ correspond strongly to iterations of the shift map T^i on γ_+, and so the behavior of geodesic flow will strongly correlate to the behavior of the shift map T. We will do this in the unit tangent bundle of $\mathbb{H}^2_{\mathbb{R}}$, by restricting the allowed unit vectors over $\mathbb{S}$. The process is summarized in Figure 4.

We first quickly prove that $M_1^{-1}\gamma$ in fact crosses $\mathbb{S}$. Indeed, we have that $M_1^{-1}\gamma_+$ is inside $\mathbb{S}$, while $|a_i| \geq 2$, so that $M_1^{-1}(\gamma_-) = M_1^{-1}(\infty) = -a_i$ is outside of $\mathbb{S}$. While this bound appears to deteriorate to $|M_i^{-1}\gamma_-| \geq 1$ with additional iterations, by looking at the permissible digits, one shows that $|M_i^{-1}\gamma_-|$ is bounded below by the golden ratio ϕ. See Remark 1.5 for the more general approach to this step.

In Figure 4(a), we see that γ' has (at least) two intersections that we do *not* want to code: the intersection with the sphere centered at the point $(1, 0)$ and the intersection with the vertical line $x = 1.5$. The first of these is avoided simply by restricting to the vectors in $T^1\mathbb{S}$ that point towards $K = [-1/2, 1/2)$ (that is, whose corresponding geodesics terminate in K). By completeness, any (non-identity) $\mathcal{Z}$-translate of these vectors must land outside of

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press



(a) The geodesic $M_1^{-1}\gamma$ (dashed) intersects the sphere $\mathbb{S}$ (red). The segment $M_1^{-1}\gamma'$ has spurious intersections with $\mathcal{M}$-translates of $\mathbb{S}$, e.g. the line $x = 1.5$.

(b) The intersection occurs in the horoball $y > 0.52\ldots$ (green boundary), so $\mathbb{S}$ is replaced with the wall $\mathbb{W}$ (still red).

(c) Translates of the horoball by $\mathcal{M}$ to other cusps have horoheight bounded above by $y = 1.923\ldots$ (brown line).

(d) Any geodesics that have small horoheight are excluded from the unit tangent bundle of $\mathbb{W}$ to produce $\mathcal{C}_{\mathbb{W}}$.

FIGURE 4. Constructing the section $C_{\mathbb{W}}$ of geodesic flow.

K. In particular, the corresponding vectors on the sphere centered at $(1, 0)$ point to the interval $[1/2, 3/2)$, whereas we know that $M_1^{-1}\gamma_+ \in K$. Thus this intersection is avoided.

The second spurious intersection in our example requires more work, and we rule it out by making two observations about $M_1^{-1}\gamma'$ that are predicated on the use of horoheight and horoballs (shown in green in Figure 4, cf. Ford circles). We may measure horoheight either from ∞, in which case horoheight is simply the y coordinate and a horoball is a set of the form $\{(x, y) : y > y_0\}$, or from a (rational) point on the x-axis, in which case horoheight can be thought of as the depth into the corresponding cusp, and horoballs appear as Euclidean disks tangent to the x-axis. For our first observation, the fact that $|M_1^{-1}\gamma_+| \le 1/2$ and $|M_1^{-1}\gamma_-| \ge \phi$ implies that the intersection of $M_1^{-1}\gamma'$ with $\mathbb{S}$ occurs away from the x-axis, in the smaller 'wall' region (see Figure 4b):

$$\mathbb{W} = \left\{z \in \mathbb{S} \ : \ \mathrm{Im}(z) > \sqrt{\tfrac{3}{2}(5\sqrt{5} - 11)} \approx 0.52\right\},$$

and that the intersection with $M_1^{-1}\mathbb{S}$ is likewise bounded away from the x-axis. That is, $M_1^{-1}\gamma'$ is contained in a horoball $\mathcal{B} = \{y > \epsilon\}$ for some $\epsilon > 0$. For our second observation, consider a mapping $M \in \mathcal{M}$ that sends the line $x = 1.5$ to $\mathbb{S}$. Normalizing $M_1^{-1}\gamma'$ further by M, we see (Figure 4c) that $MM_1^{-1}\gamma' \subset M\mathcal{B}$ is now contained in one of the horoballs based at a finite rational point. In particular, this provides (see Corollary 5.6)

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

an upper bound $c = (\frac{3}{2}(5\sqrt{5} - 11))^{-1/2} \approx 1.923$ on how far $MM_1^{-1}\gamma'$ travels away from the x-axis. We can reject the intersection of $MM_1^{-1}\gamma'$ with $\mathbb{S}$ (and thus the intersection of $M_1^{-1}\gamma'$ with the line $x = 1.5$) by restricting to the vectors in $T^1\mathbb{W}$ that are returning from a cusp excursion towards ∞ of depth at least c. We will denote by $C_{\mathbb{W}}$ the set of such vectors that also point towards K. This is our desired refinement of $\mathbb{S}$ (Figure 4d).

Our choice of $C_{\mathbb{W}}$ by construction avoids all spurious intersections, but may also inadvertently ignore some intersections corresponding to small digits or even entire geodesics. We say that a geodesic γ is *markable* if it intersects $\mathcal{M}$-translates of $C_{\mathbb{W}}$ infinitely often in both the past and future, and it is easy to see that almost every geodesic is markable, see Corollary 6.6. The markable geodesic Theorem 5.1 records the desired link between the CF digits of a markable geodesic γ and its intersections with $\mathcal{M}$-translates of $C_{\mathbb{W}}$: intersections occur only with walls of the form $M_i C_{\mathbb{W}}$, the intersections occur in the desired order, and no other intersections occur.

With the section $C_{\mathbb{W}}$ in hand, we return to the question of ergodicity. We begin by working through a number of closely related functions acting on different spaces, pulling ergodicity from one function to the next. The ergodicity of geodesic flow on the modular surface $\mathcal{M}\backslash\mathbb{H}^2_{\mathbb{R}}$ implies the ergodicity of the first return map to the projection of $C_{\mathbb{W}}$ onto the modular surface. We can then lift this first return map back to $\mathbb{H}^2_{\mathbb{R}}$ to obtain an isomorphic and thus equally ergodic map $\psi : C_{\mathbb{W}} \to C_{\mathbb{W}}$ (Proposition 6.9). We then conjugate this system with the projection π from the unit tangent bundle of $\mathbb{H}^2_{\mathbb{R}}$ to $\hat{\mathbb{R}} \times \hat{\mathbb{R}}$ that takes any geodesic γ to its forward and backward endpoints (γ_+, γ_-), obtaining an ergodic mapping $\Psi = \pi \circ \psi \circ \pi^{-1}$ on $\pi(C_{\mathbb{W}})$. This map acts by taking a point (γ_+, γ_-) to $(M_i^{-1}\gamma_+, M_i^{-1}\gamma_-)$ for some i. Thus, in the first coordinate, Ψ acts by T^i, where i may depend on the value of γ_+, that is, this is a jump transformation associated with T.

Although we could conclude that this jump transformation is ergodic, the ergodicity of a jump transformation does not imply the ergodicity of the original map in general. So to recover the ergodicity of T, we step back to $\hat{\mathbb{R}} \times \hat{\mathbb{R}}$. Namely, we consider a natural-extension-like function $\hat{T}$ on $K \times \hat{\mathbb{R}}$ such that the action of $\hat{T}$ on the first coordinate is simply T. We show that the action of $\hat{T}$ on $\overline{K} = \bigcup_{i=0}^{\infty} \hat{T}^i \pi(C_{\mathbb{W}})$ is well behaved (Lemma 6.10) and that, in fact, Ψ is simply the map induced by restricting $\hat{T}$ to $\pi(C_{\mathbb{W}})$ (Lemma 6.11). Induced maps have far greater structure than jump transformations and so we are able to conclude the ergodicity of $\hat{T}$ on $\overline{K}$ from the ergodicity of Ψ (Lemma 6.12) and from there conclude the ergodicity of T by restricting to the first coordinate.

Remark 1.5. There are two sources of complexity in the full proof of Theorem 1.2. The first is that we would like to work in sufficient generality to include Heisenberg continued fractions. This requires working with hyperbolic spaces defined over complex numbers and quaternions, and obtaining some new results about inversions for the corresponding horospherical coordinates with boundary, see Theorem 2.11. The second source of complexity is the fact that, even for simple CF algorithms, the point $M_i^{-1}\gamma_-$ need not remain outside of the sphere $\mathbb{S}$, and the properness assumption is necessary to guarantee that some intersections do occur. For example, the α-CF algorithm with $\alpha > 2/3$ and inversion ι_+ would send the geodesic with endpoints $\gamma_+ = \frac{2}{3}$ and $\gamma_- = \infty$ to the geodesic with endpoints $M_1\gamma_+ = 1/2$ and $M_1\gamma_- = -1$, which does not intersect $\mathbb{S}$. However, $M_i\gamma$

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

cannot remain under $\mathbb{S}$ forever: applying the identity $|1/x - 1/y||x||y| = |x - y|$, we see that additional iterations of the shift map must pull the endpoints of $M_i\gamma$ apart and push $M_i\gamma_-$ out of the unit circle. The need to wait several iterations before a collision is detected then complicates the construction of the wall region $\mathbb{W}$ and the section $C_{\mathbb{W}}$.

1.4. *Further remarks.* Iwasawa CFs are the most general setting for our methods, which rely heavily on the fact that Iwasawa inversion spaces are boundaries of rank-one symmetric spaces of non-compact type. Indeed, Iwasawa inversion spaces are precisely the spaces with this property, with the exclusion, due to the break down of vector-space-based techniques, of the exceptional $\mathbb{X}^1_{\mathbb{O}}$ that can be defined over the non-associative octonions. Our notion of *Iwasawa inversion space* differs slightly from the notion of *Iwasawa groups* of [**9**], which excludes $\mathbb{X}^n_{\mathbb{R}}$ and allows $\mathbb{X}^1_{\mathbb{O}}$.

We remark further that boundaries of rank-one symmetric spaces of non-compact type are arguably the most general setting for geometric CFs and Diophantine theory: they are characterized [**12**, **37**] as homogeneous geodesic locally compact spaces admitting both a dilation (a notion of fraction) and a well-behaved inversion. (The Cygan metric we work with is not itself geodesic, but gives rise to a geodesic path metric.)

The present work suggests the following further directions of study.

Question 1. Under what conditions is the invariant measure for the CF shift map finite or (piecewise) analytic?

Question 2. Is the CF shift map mixing?

Question 3. Does Theorem 1.2 hold for incomplete Iwasawa CFs or for improper Iwasawa CFs with weak contact with the unit sphere (such as J. Hurwitz CFs)?

Question 4. Can one characterize periodic Iwasawa CF expansions, analogously to the quadratic surd characterization of periodic regular CFs in $\mathbb{R}$ (cf. [**63**])?

Question 5. Can one describe the set of exceptions to tail-equivalence in Theorem 1.4 (cf. [**36**])?

Question 6. What Iwasawa CF algorithms are not represented in Table 1?

1.5. *Outline of the paper.* Following this introduction, in §2 we provide the general theory and definitions for Iwasawa inversion spaces. In §3, we define Iwasawa CFs, give further examples (including Table 1), and study conditions that guarantee discreteness, properness, and completeness. In §4, we quickly prove the convergence of Iwasawa CFs. In §5, we will build up the theory surrounding markable geodesics, culminating in the markable geodesic theorem. In §6, we use the markable geodesic theorem to prove the ergodicity of the CF shift map for an Iwasawa CF expansion and, in applications of this result, prove Theorems 1.3 and 1.4.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

2. *General theory*

We now outline the structure of Iwasawa inversion spaces $\mathbb{X} = \mathbb{X}_k^n$, the associated upper half-spaces $\mathbb{H}_k^{n+1}$, and the continued fraction algorithms that can be built on $\mathbb{X}$ using this structure. We encourage the reader to skip this section on the first reading, following the intuition of the Euclidean space $\mathbb{X} = \mathbb{X}_{\mathbb{R}}^n = \mathbb{R}^n$ and hyperbolic half-space $\mathbb{H} = \mathbb{H}_{\mathbb{R}}^{n+1}$ lying above it.

2.1. *Iwasawa inversion spaces.* Abstractly, an Iwasawa inversion space $\mathbb{X}$ is an Iwasawa N-group associated by the Iwasawa (KAN) decomposition with a non-exceptional rank one semi-simple Lie group G and the parabolic boundary at infinity of the rank-one symmetric space G/K. We now recall the explicit construction and Euclidean-like structure of these spaces. Most of the contents of this section can be found in [9, 21, 51].

Fix an associative division algebra k over the reals—the real, complex, or quaternionic numbers (when working over quaternions, we will use the convention $p/q := pq^{-1}$)—and an integer $n \geq 1$. (It appears that one could also consider the exceptional case of octonions, but we will not do so here.) Recall that k has a real part $\mathrm{Re}(k)$ isomorphic to $\mathbb{R}$ and a complementary imaginary part $\mathrm{Im}(k)$ satisfying $\dim_{\mathbb{R}}(\mathrm{Im}(k)) = \dim_{\mathbb{R}}(k) - 1$. We denote the standard norm of an element of k or k^n by $\|\cdot\|$, and refer to $\|\cdot\|$-preserving k-linear automorphisms of k^n as *unitary* transformations.

Remark 2.1. For $k = \mathbb{R}$, one has $\mathrm{Im}(k) = \{0\}$. Note that $\mathrm{Im}(k)$ remains a subset of k; in particular, we do not identify $\mathrm{Im}(k)$ with $\mathbb{R}$ when $k = \mathbb{C}$. We furthermore exclude non-holomorphic transformations such as $z \mapsto \overline{z}$ from the unitary group, purely for notational convenience.

Definition 2.2. (Iwasawa inversion space) The *Iwasawa inversion space* $\mathbb{X} = \mathbb{X}_k^n$ is the set $k^n \times \mathrm{Im}(k)$ with coordinates (z, t) and group law

$$(z, t) * (z', t') = (z + z', t + t' + 2\mathrm{Im}\langle z, z'\rangle),$$

where the inner product of the vectors z, z' is given by $\langle z, z'\rangle = \sum_i \overline{z_i} z_i'$.

Over the reals, $\mathbb{X}_{\mathbb{R}}^n$ reduces to $\mathbb{R}^n$ with $*$ acting by the usual vector addition. For $k \neq \mathbb{R}$, $\mathbb{X}_k^n$ is a step-2 nilpotent group (one uses $*$ to emphasize the non-commutativity), with identity $(0, 0)$, and the inverse of a group element (z, t) given by $(-z, -t)$.

One gives $\mathbb{X}$ a gauge $|\cdot|$ and Cygan metric d (also known in different contexts as the Korányi metric or gauge metric) by defining

$$|(z, t)| := \|\|z\|^2 + t\|^{1/2}, \quad d((z, t), (z', t')) := |(-z, -t) * (z, t)|.$$

The Cygan metric is largely analogous to the Euclidean metric, insofar as its automorphisms include analogs of translations (left multiplication by an element of $\mathbb{X}$ is an isometric isomorphism); dilations (for each $r > 0$, the mapping $\delta_r(z, t) = (rz, r^2 t)$ is a group isomorphism that rescales the metric by factor r); and rotations (unitary automorphisms of k^n extend to isometric group isomorphisms of $\mathbb{X}$).

However, the metric is fractal for $k \neq \mathbb{R}$: it is not a path metric (cf. the closely associated Carnot–Carathéodory path metric) and gives $\mathbb{X}$ Hausdorff dimension $n \dim_{\mathbb{R}}(k) +$

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

$2(\dim_{\mathbb{R}}(k) - 1)$ which is not equal to its topological dimension $(n+1)\dim_{\mathbb{R}}(k) - 1$. The latter is due to the fact that large metric balls are stretched by δ_r along the t direction, while small ones are flattened out along the z direction.

The *Korányi* inversion $\iota_- : \mathbb{X} \setminus \{0\} \to \mathbb{X} \setminus \{0\}$ is defined by

$$\iota_-(z,t) = \left(\frac{-z}{\|z\|^2 + t}, \frac{-t}{\|z\|^4 + \|t\|^2} \right).$$

The Korányi inversion is a natural generalization of the mapping $x \mapsto -1/x$, and in particular satisfies the following pair of identities for $h, h' \in \mathbb{X} \setminus \{0\}$, [12]:

$$|\iota_- h| = \frac{1}{|h|}, \quad d(\iota_- h, \iota_- h') = \frac{d(h,h')}{|h||h'|}. \tag{2.1}$$

In particular, ι_- sends each sphere $S(0,r)$ to the sphere $S(0,1/r)$, and preserves the unit sphere. We prove the identities in a broader context in Theorem 2.11.

More generally, $\mathbb{X}$ admits inversions of the form

$$\iota(z,t) = \left(\frac{-A(z)}{\|z\|^2 + t}, \frac{-(\det A)t}{\|z\|^4 + \|t\|^2} \right),$$

where A is a unitary transformation of k^n. We show in Lemma 2.10 that all inversions satisfy generalizations of (2.1).

2.2. *Upper half-space.* Fix an Iwasawa inversion space $\mathbb{X} = \mathbb{X}_k^n$. We extend the structure and Cygan metric of $\mathbb{X}$ to k^{n+1} as follows, motivated by Parker [51].

Definition 2.3. Extend the Heisenberg group law to $k^n \times k = k^{n+1}$ as

$$(z,w) * (z',w') = (z+z', w+w'+2\mathrm{Im}\langle z, z'\rangle),$$

and the gauge and metric as

$$|(z,w)| = \|\|z\|^2 + \|\mathrm{Re}(w)\| + \mathrm{Im}(w)\|^{1/2}, \quad d((z,t),(z',t')) := |(-z,-t) * (z,t)|.$$

Remark 2.4. In the case $k = \mathbb{R}$, the Heisenberg group law on k^{n+1} reduces to $(z,w) * (z',w') = (z+z', w+w')$, and the gauge reduces to the Euclidean-like $|(z,w)| = (\|z\|^2 + \|w\|)^{1/2}$. One could adjust Definition 2.3, by taking a square root along the $\mathrm{Re}(w)$ direction, so that it agrees with the Euclidean metric in the real case. We will not do so.

Definition 2.5. The *upper half-space* $\mathbb{H}_k^{n+1} \subset k^{n+1}$ is the set

$$\mathbb{H}_k^{n+1} = \{(z,w) \in k^n \times k \ : \ \mathrm{Re}(w) > 0\},$$

satisfying $\partial\mathbb{H} = \mathbb{X}$.

One gives $\mathbb{H}$ two natural metrics: the restriction of the Cygan metric d on k^{n+1} (this was introduced by Parker in [51] for $\mathbb{H}^2_{\mathbb{C}}$ and generalized by Cao–Parker to $\mathbb{H}^2_{\mathbb{O}}$ in [8]); and the negatively curved hyperbolic metric $d_{\mathbb{H}}$, defined via an embedding into $\mathbb{P}(k^{n+2})$. Unless otherwise noted, $\mathbb{H}$ will always be equipped with the metric $d_{\mathbb{H}}$.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Definition 2.6. (Projective embedding) Let $\phi : k^{n+1} \to k^{n+2}$ be given by $\phi(z, w) = (1, \sqrt{2}z, w + \|z\|^2)$, and set $\Phi = \mathbb{P} \circ \phi : k^{n+1} \to \mathbb{P}(k^{n+2})$.

Consider the Hermitian form $\langle \cdot, \cdot \rangle_J$ of signature $(n + 1, 1)$ defined on k^{n+2} by

$$J = \begin{bmatrix} 0 & 0_n & -1 \\ 0_n & \mathrm{id}_n & 0_n \\ -1 & 0_n & 0 \end{bmatrix},$$

and let $\mathcal{S} = \{(1 : a : b) \ : \ \|a\| < 2\mathrm{Re}(b)\} \subset \mathbb{P}(k^{n+2})$ be the Siegel region. One can show that Φ induces a bijection between $\mathbb{H}$ and $\mathcal{S}$, and furthermore $\mathcal{S}$ is the projectivization of the negative cone of J. This induces an action of the projective unitary group $G = \mathbb{P}U(J)$ on $\mathbb{H}$, cf. §4.

Definition 2.7. (Hyperbolic metric) The hyperbolic metric $d_{\mathbb{H}}$ on $\mathbb{H}$ is the unique G-invariant Riemannian metric on $\mathbb{H}$ with sectional curvature pinched in the range $[-1, -1/4]$ if $k \neq \mathbb{R}$ or equal to -1 if $k = \mathbb{R}$.

For $\mathbb{H} = \mathbb{H}^2_{\mathbb{R}}$, $d_{\mathbb{H}}$ agrees with the familiar metric $\frac{1}{y}ds$ if one takes $x = z$ and $y = w^2$. One has $\Phi(\mathbb{H}) = \{(1 : a : b) \ : \ 2b > a^2\} \subset \mathbb{RP}^2$, and a projective change of coordinates recovers the Klein disk model of $\mathbb{H}^2_{\mathbb{R}}$ with its $SO(2, 1)$-invariant metric.

In general, the Siegel region is projectively equivalent to a unit ball in projective space $\mathbb{P}(k^{n+2})$. The mapping $\Phi|_{\mathbb{X}} : \partial\mathbb{H} \to \partial\Phi(\mathbb{H})$ omits a single point, which we identify with the point ∞ in the one-point compactification of k^{n+1} (and its subsets $\mathbb{X}$ and $\overline{\mathbb{H}}$).

2.3. *Inversion theorem.* Returning to the Cygan metric, we record two connections to the projective embedding.

LEMMA 2.8. (Parker [51]) *Suppose $p, q \in \overline{\mathbb{H}}$, with either p or q in $\mathbb{X} = \partial\mathbb{H}$. Then the Cygan metric satisfies $d(p, q) = \|\langle \phi(p), \phi(q) \rangle_J\|^{1/2}$.*

LEMMA 2.9. *Let $h \in \overline{\mathbb{H}}$ and denote $\phi(h) = (1, a, b)$. Then $|h| = \|b\|^{1/2}$.*

Proof. This is immediate from Definitions 2.3 and 2.6. □

With the above machinery, we can provide a simple description of the Korányi inversion, extended to $\overline{\mathbb{H}}$, and prove the inversion identities (2.1).

LEMMA 2.10. *The Korányi inversion $\iota_- : \overline{\mathbb{H}} \setminus \{0\} \to \overline{\mathbb{H}} \setminus \{0\}$, given by the mapping*

$$(z, w) \mapsto \left(\frac{-z}{\|z\|^2 + w}, \frac{\overline{w}}{\|\|z\|^2 + w\|^2} \right),$$

is induced by the matrix $J \in G$. That is, setting $\phi(z, w) = (1, a, b)$, one has $\phi(\iota_-(z, w)) = (1, -a/b, 1/b) = \phi(z, w)/-b$, and in $\mathbb{P}(k^{n+2})$, one has $\Phi(\iota_-(z, w)) = J\Phi(z, w)$.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Proof. We have $\phi(z, w) = (1, \sqrt{2}z, |z|^2 + w)$, so that $J\phi(z, w) = (-(|z|^2 + w), \sqrt{2}z, -1)$. Up to a factor of $-(\|z\|^2 + w)$, this is equivalent to

$$\left(1, \sqrt{2}\frac{-z}{\|z\|^2 + w}, \frac{1}{\|z\|^2 + w}\right) = \left(1, \sqrt{2}\frac{-z}{\|z\|^2 + w}, \left\|\frac{-z}{\|z\|^2 + w}\right\|^2 + \frac{\overline{w}}{\|\|z\|^2 + w\|^2}\right),$$

which in turn is equal to $\phi(\iota_-(z, w))$ as desired. □

THEOREM 2.11. (Inversion theorem) *Let $h \in (\mathbb{H} \cup \mathbb{X}) \setminus \{0\}$ and $h' \in \mathbb{X} \setminus \{0\}$. The following identities hold for the Korányi inversion ι_-, Cygan metric d, and gauge $|\cdot|$:*

$$|\iota_- h| = \frac{1}{|h|} \quad \text{and} \quad d(\iota_- h, \iota_- h') = \frac{d(h, h')}{|h||h'|}. \tag{2.2}$$

Proof. Write $\phi(h) = (1, a, b)$ and $\phi(h') = (1, a', b')$. By Lemma 2.10, $\phi(\iota_-(h)) = (1, -a/b, 1/b)$, and the first identity thus follows from Lemma 2.9.

Since $h' \in \mathbb{X}$, Lemma 2.8 gives $d(h, h') = \|\langle \phi(h), \phi(h')\rangle_J\|^{1/2}$ and $d(\iota_- h, \iota_- h') = \|\langle \iota_- \phi(h), \iota_- \phi(h')\rangle_J\|^{1/2}$. Using Lemmas 2.10 and 2.9,

$$d(\iota_- h, \iota_- h') = \left\|\left\langle \frac{\phi(h)}{-b}, \frac{\phi(h')}{-b'}\right\rangle_J\right\|^{1/2} = \frac{d(h, h')}{\|b\|^{1/2}\|b'\|^{1/2}} = \frac{d(h, h')}{|h||h'|},$$

which provides the second identity. □

Remark 2.12. Surprisingly, Lemma 2.8 and the second identity of Theorem 2.11 fail when both h and h' lie in $\mathbb{H}$.

Compositions of diagonal elements of G (as well as certain conjugation actions) with the Korányi inversion continue to satisfy the conclusions of Theorem 2.11. We define the following.

Definition 2.13. An *inversion* is a Möbius transformation $\iota : \mathbb{X} \setminus \{0\} \to \mathbb{X} \setminus \{0\}$ satisfying the conclusions of Theorem 2.11.

It follows from the classification of isometries of $\mathbb{H}$ that every inversion factors as a composition of a rotation and the Korányi inversion.

LEMMA 2.14. *If ι is an inversion, then there exists a unitary mapping $f : k^n \to k^n$ such that $\iota = f \circ \iota_-$.*

Proof. Since ι is a Möbius transformation, it extends to an isometry of $\mathbb{H}$. The mapping $f = \iota\iota_-$ is an isometry of $\mathbb{H}$ that fixes the points 0 and ∞. It therefore maps the geodesic γ joining 0 and ∞ to itself. Since ι_- and ι fix the point $(0, 1) \in \gamma$ by the first part of (2.2), the same must be true for f. Thus, ι is represented in $U(J)$ by a matrix of the form

$$\begin{bmatrix} 0 & 0_n & -1 \\ 0_n & A & 0_n \\ -1 & 0_n & 0 \end{bmatrix}, \tag{2.3}$$

where A is a unitary matrix over k^n. □

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

In addition to the (negative) Korányi inversion ι_-, we will also be interested in the positive inversion ι_+ corresponding to the matrix $A = -I_n$ in (2.3), and the conjugation inversion ι_c corresponding to the diagonal matrix A with diagonal entries $(-1, 1, 1, \ldots, 1)$. For example, for $p = (x, y, z) \in \mathbb{R}^3$, one has $\iota_-(p) = -p/\|p\|^2$, $\iota_+(p) = p/\|p\|^2$, and $\iota_c(p) = (x, -y, -z)/\|p\|^2$. Note that under the standard identification of $\mathbb{C}$ with $\mathbb{R}^2$, the mapping $z \mapsto 1/z$ corresponds to the inversion ι_c.

2.4. *Isometries, lattices, and fundamental domains.* We thus have an Iwasawa inversion space $\mathbb{X}$ and associated hyperbolic space $\mathbb{H}$, with the unitary group G acting on $\mathbb{H}$ by isometries with respect to the Riemannian metric $d_{\mathbb{H}}$, and by generalized Möbius transformations on $\widehat{X} = \mathbb{X} \cup \{\infty\}$. One shows that G is in fact the holomorphic isometry group of $\mathbb{H}$, and the group of (1-quasi-)conformal mappings of $\mathbb{X} \cup \{\infty\}$. Restricting G to the set of transformations $\mathrm{Stab}_G(\infty)$ preserving infinity provides an action on $\mathbb{X}$ that can be identified with the group of similarities of $\mathbb{X}$. This allows us to think of $\mathrm{Isom}(\mathbb{X})$ as a subgroup of $\mathrm{Isom}(\mathbb{H})$.

The group G is, in fact, a rank-one simple Lie group, with an Iwasawa decomposition $G = KAN$. One can identify the subgroup N with the space $\mathbb{X}$ (with the group structure provided above), and the subgroup A with the group of dilations $\{\delta_r : r > 0\}$. The subgroup K can be identified with the stabilizer of the point $(0, 1) \in \mathbb{H}$, and includes the Korányi inversion.

We will be interested in lattices and fundamental domains in $\mathrm{Isom}(X)$ and $\mathrm{Isom}(\mathbb{H})$, equipped with the respective Haar measures.

Definition 2.15. Let Y be a metric space with an $\mathrm{Isom}(Y)$-invariant measure. A *lattice* is a discrete subgroup $\Gamma \subset \mathrm{Isom}(Y)$ such that the quotient $\Gamma\backslash\mathrm{Isom}(Y)$ has finite measure. The lattice is *uniform* if $\Gamma\backslash\mathrm{Isom}(Y)$ is furthermore compact, and non-uniform otherwise.

A *fundamental domain* for Γ is a measurable set $K \subset Y$ such that $\mathbb{X} = \bigcup_{a\in\Gamma} aK$ and the overlap $K \cap \bigcup_{a(\neq \mathrm{id})\in\Gamma} aK$ has measure 0.

A *rounding mapping* $[\cdot] : Y \to \Gamma$ associated with Γ and K is defined, almost everywhere, by the property that for each $a \in \Gamma$ and $x \in K$, one has $[a(x)] = a$. This property defines $[\cdot]$ uniquely away from the overlap, and $[\cdot]$ provides some choice of admissible values has been made for points in the overlap.

3. *Iwasawa continued fractions*

We can now define Iwasawa continued fractions and establish some auxiliary terminology and notation.

Definition 3.1. (Iwasawa continued fraction) The Iwasawa continued fraction algorithm is defined by the following data:

(1) an associative division algebra k over $\mathbb{R}$ and integer $n \geq 1$;
(2) the associated Iwasawa inversion space $\mathbb{X} = \mathbb{X}_k^n$;
(3) an inversion ι (see Definition 2.13);
(4) a lattice $\mathcal{Z} \subset \mathrm{Isom}(\mathbb{X})$, a fundamental domain $K \subset \mathbb{X}$ for $\mathcal{Z}$, and an associated rounding mapping $[\cdot] : \mathbb{X} \to \mathcal{Z}$ (see Definition 2.15).

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Associated with an Iwasawa CF algorithm, we have the following:

(5) the hyperbolic space $\mathbb{H} = \mathbb{H}_k^{n+1}$ satisfying $\partial\mathbb{H} = \mathbb{X}$;
(6) the holomorphic isometry group G of $\mathbb{H}$;
(7) the modular group $\mathcal{M} = \langle \iota, \mathcal{Z} \rangle \subset G$;
(8) the shift map $T : K \to K$ defined by $T(0) = 0$ if $0 \in K$ and otherwise by

$$T(x) = [\iota(x)]^{-1}(\iota(x)).$$

For a point $x \in \mathbb{X}$, we can then inductively define the continued fraction digits $a_i \in \mathcal{Z}$ and forward iterates $x_i \in K$ by taking

$$a_0 = [x], \quad x_0 = a_0^{-1}(x),$$
$$a_{i+1} = [\iota(x_i)], \quad x_{i+1} = a_{i+1}^{-1}(\iota(x_i)) = T(x_i),$$

where the sequences terminate if at some point, $x_i = 0$. The (possibly finite) sequence (a_i) of elements of $\mathcal{Z}$ is the *continued fraction sequence* of x. (Note that later in the paper, we will assign a bi-infinite string of digits to pairs of points one of which is in K, resulting in a different notion of a_0. For this reason, for points in K, we will leave a_0 undefined.)

Given a sequence (a_i) of elements of $\mathcal{Z}$ (possibly arising from the above algorithm), one defines the *convergent mappings* $M_i \in \mathcal{M}$ inductively by setting M_0 to be the identity mapping and $M_{n+1} = M_n \circ \iota^{-1} \circ a_{n+1}$. (In the following, we will often suppress the $\circ$ notation for convenience.) By construction, we see that $x_0 = M_n(x_n)$. For each i, the i^{th} *convergent* of the continued fraction is then the point $M_i(0)$. Note that $T^i x_0 = x_i = M_i^{-1}(x_0)$.

We will be interested in conditions on the continued fraction algorithm that guarantee the following properties.

Definition 3.2. The continued fraction algorithm is *convergent* if the continued fraction digits of almost every point $x \in K$ produce convergents $M_i(0)$ that indeed converge to x (clearly, every finite expansion is convergent). The algorithm is *ergodic* if the shift map T is ergodic.

We will use the following definition of ergodicity.

Definition 3.3. Let (A, μ) be a measure space and $f : A \to A$ a measurable (but not necessarily measure-preserving) transformation. Then, f is said to be ergodic with respect to μ if for every measurable $B \subset A$, $\mu(f^{-1}B \triangle B) = 0$ implies that $\mu(B) = 0$ or $\mu(A \setminus B) = 0$. If $\phi : A \to A$ is a measurable flow, then ϕ is ergodic with respect to μ if for every measurable $B \subset A$, $\mu(\phi_t(B) \triangle B) = 0$ for all $t \in \mathbb{R}$ implies that $\mu(B) = 0$ or $\mu(A \setminus B) = 0$.

Remark 3.4. Note that with this definition, ergodicity with respect to a measure μ implies ergodicity with respect to any measure that is equivalent to μ. In this paper, the relevant measure (or class of equivalent measures) will always be clear from the context, and will often be a Lebesgue or Haar measure.

We will prove the convergence of the Iwasawa CFs under the assumptions of *properness* and *discreteness*.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Definition 3.5. (Properness and discreteness) The Iwasawa continued fraction is *proper* if the closure of K is bounded away from the unit sphere: $\mathrm{rad}(K) = \sup\{|x| : x \in K\} < 1$. It is *discrete* if $\mathcal{M}$ is a discrete subgroup (and therefore, by construction, a lattice) in G.

There do exist convergent Iwasawa continued fractions that are not proper, most notably regular continued fractions on $\mathbb{R}$ and J. Hurwitz continued fractions on $\mathbb{C}$. Likewise, one can construct proper but non-discrete Iwasawa continued fractions: for example, let $\mathbb{X} = \mathbb{R}$, $\mathcal{Z} = \epsilon\mathbb{Z}$, and $K = (-\epsilon/2, \epsilon/2]$. The resulting continued fraction is generally not discrete, but will be convergent by the Śleszyński–Pringsheim theorem [**59**] for $\epsilon < 1/2$.

To prove ergodicity, we will need a further assumption of *completeness*, which rules out hidden symmetries.

Definition 3.6. (Completeness) The Iwasawa continued fraction is *complete* if one has $\mathrm{Stab}_{\mathcal{M}}(\infty) = \mathcal{Z}$.

For an incomplete continued fraction, one may pass to the *completion* by replacing $\mathcal{Z}$ with the lattice $\mathrm{Stab}_{\mathcal{M}}(\infty)$ and making a corresponding modification to the fundamental domain K and rounding function $[\cdot]$. This will result in what are often termed 'folded' variants (see §1.2.4).

Definition 3.7. The Iwasawa continued fraction is *incomplete with n central symmetries* if there exists a set $\mathcal{R} \subset \mathrm{Isom}(\mathbb{X})$ such that:

(1) every element of $\mathcal{R}$ fixes 0, that is, is a rotation around the origin;
(2) the only element of $\mathcal{Z}$ to fix 0 is the identity;
(3) $\mathrm{Stab}_{\mathcal{M}}(\infty) = \langle \mathcal{Z}, \mathcal{R} \rangle$;
(4) every element of $\mathrm{Stab}_{\mathcal{M}}(\infty)$ can be written uniquely as ra for some $r \in \mathcal{R}$, $a \in \mathcal{Z}$, and uniquely as $a'r'$ for some $a' \in \mathcal{Z}$, $r' \in \mathcal{R}$; and
(5) $\mathcal{R}$ contains n elements.

The set $\mathcal{R}$ is said to be the set of central symmetries of $\mathcal{M}$. We say that the fundamental domain K for $\mathcal{Z}$ is symmetric if for any $r \in \mathcal{R}$, rK is K up to a set of measure zero.

3.1. *Further examples.* With all of our notation now in place, we may describe many examples of Iwasawa continued fractions. In Table 1, we list several types of continued fractions, and for each of them denote the Iwasawa inversion space $\mathbb{X}$ on which it exists; the lattice $\mathcal{Z}$, which will often act by left-translation by a subset of $\mathbb{X}$; the fundamental domain K; the inversion, which in all cases will be identified by a ι signature; whether it is complete and proper (the columns C and P respectively); and some basic references.

It should be noted that all cases under consideration are discrete.

In some cases where the fundamental domain is too complicated to write succinctly, we have labeled it with the Dirichlet region. In this case, we mean the set of points that are closer to 0 than to any translate of 0 under $\mathcal{Z}$, with some choice of boundary.

Remark 3.8. Note as well that the fundamental domain K for the Shallit complex CF algorithm is a rectangle with corners at $.5 - .5\mathbb{i}$, 1, $\mathbb{i}$, and $-.5 + .5\mathbb{i}$ [**10**].

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

TABLE 1. Examples of Iwasawa continued fractions. The examples in $\mathbb{X}^2_{\mathbb{R}} = \mathbb{R}^2$ are usually presented as complex CFs. See §§1.2 and 3.1 for more information about the algorithms.

Name	$\mathbb{X}$	$\mathcal{Z}$	K	ι	C	P	References
Regular	$\mathbb{X}^1_{\mathbb{R}}$	$\mathbb{Z}$	$[0, 1)$	ι_+	N	N	[57]
Backwards	$\mathbb{X}^1_{\mathbb{R}}$	$\mathbb{Z}$	$[0, 1)$	ι_-	Y	N	See §1.2.2
Nearest Integer	$\mathbb{X}^1_{\mathbb{R}}$	$\mathbb{Z}$	$[-\frac{1}{2}, \frac{1}{2})$	ι_+	N	Y	[61]
Nearest Integer (variant)	$\mathbb{X}^1_{\mathbb{R}}$	$\mathbb{Z}$	$[-\frac{1}{2}, \frac{1}{2})$	ι_-	Y	Y	[31]
Folded Nearest Integer	$\mathbb{X}^1_{\mathbb{R}}$	$\langle \mathbb{Z}, x \mapsto -x \rangle$	$[0, \frac{1}{2}]$	ι_+	Y	Y	[40]
Nakada α, $\alpha \in (0, 1)$	$\mathbb{X}^1_{\mathbb{R}}$	$\mathbb{Z}$	$[\alpha - 1, \alpha]$	ι_+	N	Y	Cf. [2, 46]
Even	$\mathbb{X}^1_{\mathbb{R}}$	$2\mathbb{Z}$	$[-1, 1)$	ι_-	Y	N	Cf. [4, 34]
Rosen for $q \in \mathbb{N}_{\geq 3}$	$\mathbb{X}^1_{\mathbb{R}}$	$\lambda\mathbb{Z}$, $\lambda = 2\cos(\pi/q)$	$[-\lambda/2, \lambda/2)$	ι_-	Y	Y	[43], cf. [6]
α-Rosen for $q \in \mathbb{N}_{\geq 3}$	$\mathbb{X}^1_{\mathbb{R}}$	$\lambda\mathbb{Z}$, $\lambda = 2\cos(\pi/q)$	$[\lambda(\alpha - 1), \lambda\alpha)$, $\alpha \in [1/2, 1/\lambda)$	ι_-	Y	Y	New, cf. [15]
Hurwitz	$\mathbb{X}^2_{\mathbb{R}}$	$\mathbb{Z}^2$	$[-\frac{1}{2}, \frac{1}{2})^2$	ι_c	N	Y	[10, 26]
Folded Hurwitz	$\mathbb{X}^2_{\mathbb{R}}$	$\langle \mathbb{Z}^2, (x, y) \mapsto (-x, -y) \rangle$	$[-\frac{1}{2}, \frac{1}{2}) \times [-\frac{1}{2}, 0]$	ι_c	Y	Y	Cf. [52]
Hurwitz Hexagonal	$\mathbb{X}^2_{\mathbb{R}}$	$\mathbb{Z}[\rho]$, with $\rho = (1 + \sqrt{-3})/2$	Dirichlet region	ι_c	N	Y	[30]
J. Hurwitz or Tanaka	$\mathbb{X}^2_{\mathbb{R}}$	$\{(a, b) \in \mathbb{Z}^2 : a + b \text{ even}\}$	Dirichlet region	ι_c	Y	N	[10, 60]
Shallit	$\mathbb{X}^2_{\mathbb{R}}$	$\mathbb{Z}^2$	See Remark 3.8	ι_c	N	N	[10]
SKT	$\mathbb{X}^2_{\mathbb{R}}$	$\mathbb{Z}[\rho]$, with $\rho = (1 + \sqrt{-3})/2$	$[0, 1)\rho \times [0, 1)\overline{\rho}$	ι_c	N	N	[58]
Bianchi, $d = 1, 2, 3, 7, 11$	$\mathbb{X}^2_{\mathbb{R}}$	$\mathcal{O}_d$, ring of integers	Dirichlet region	ι_c	N	Y	[17, 29]
3d	$\mathbb{X}^3_{\mathbb{R}}$	$\mathbb{Z}^3$	$[-\frac{1}{2}, \frac{1}{2})^3$	ι_+	N	Y	New
Quaternionic	$\mathbb{X}^4_{\mathbb{R}}$	$\mathbb{Z}^4$	$[-\frac{1}{2}, \frac{1}{2})^4$	ι_c	N	N	[22, 23]
Hurwitz Quaternionic	$\mathbb{X}^4_{\mathbb{R}}$	Hurwitz integers	Dirichlet region	ι_c	N	Y	[44]
Octonionic	$\mathbb{X}^8_{\mathbb{R}}$	Cayley integers	Dirichlet region	ι_c	N	Y	New
Heisenberg	$\mathbb{X}^1_{\mathbb{C}}$	$\mathbb{Z}^3$	$[-\frac{1}{2}, \frac{1}{2})^3$	ι_-	N	Y	[38]
Folded Heisenberg	$\mathbb{X}^1_{\mathbb{C}}$	$\langle \mathbb{Z}^3, (z, t) \mapsto (\mathrm{i}z, t) \rangle$	$[-\frac{1}{2}, 0]^2 \times [-\frac{1}{2}, \frac{1}{2})$	ι_-	Y	Y	New
Heisenberg Hexagonal	$\mathbb{X}^1_{\mathbb{C}}$	$\mathbb{Z}[\rho] \times \sqrt{3}\mathbb{Z}$	See Example 3.13	ι_-	N	Y	New
Heisenberg Quaternionic	$\mathbb{X}^1_{\mathcal{H}}$	$(\mathbb{Z}^4 \cup (\mathbb{Z} + 1/2)^4) \times \mathbb{Z}^3$	Dirichlet region	ι_-	N	N	New

The complex continued fractions, quaternionic continued fractions, and octonionic continued fractions are embedded in higher-dimensional real spaces in the standard way, $\mathbb{C} \cong \mathbb{R}^2$, $\mathcal{H} \cong \mathbb{R}^4$, and $\mathbb{O} \cong \mathbb{R}^8$. The inversion ι_c listed in all these cases is equivalent to $z \mapsto 1/z$ on $\mathbb{C}$, $\mathcal{H}$, or $\mathbb{O}$. One reason for identifying these spaces is that the existence of maximal orders, the Gaussian and Eisenstein integers in $\mathbb{C}$, the Hurwitz integers in $\mathcal{H}$, and

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

the Cayley integers in $\mathbb{O}$, give rise to lattices on $\mathbb{R}^2$, $\mathbb{R}^4$, and $\mathbb{R}^8$ that in turn generate *proper* fundamental domains K. The Hurwitz integers in $\mathcal{H}$ are given by

$$\{a + b\mathbb{i} + c\mathbb{j} + d\mathbb{k} : a, b, c, d \in \mathbb{Z} \text{ or } a, b, c, d \in \mathbb{Z} + 1/2\}. \tag{3.1}$$

The Cayley integers in $\mathbb{O}$ are defined in Ch. 9 of [11] (where they are referred to by the less common name of octavian integers), with properness of the corresponding Dirichlet region following from lemma 6 of that chapter.

We should emphasize that Table 1 does not cover all well-studied CF algorithms. For example, odd CFs [5], CFs related to triangle groups [7], CFs related to the Jacobi–Perron algorithm or other subtraction algorithms [54], regular chains [53], and general (a, b)-continued fractions [35] do not fit into our framework. The N-continued fractions [16] and u-backwards continued fraction [20] use an ι which is not an inversion by our definition; however, our proofs could be modified to compensate. Regardless, they would still not be proper.

Remark 3.9. We are not the first to encounter problems with the incompleteness of the Hurwitz CF algorithm. Pollicott [52] studied a similar folded continued fraction, albeit using conjugation in place of negation. Nakada [47] studied the full Hurwitz CF, but took as his hyperbolic space the disjoint union of two different spaces and let negation additionally act by swapping between the two.

3.2. *Discreteness and properness.* The difficulty of pushing into higher dimensions (either by taking $k \neq \mathbb{R}$ or $n \geq 2$) is in finding an appropriate lattice $\mathcal{Z}$ and fundamental domain K such that the resulting continued fraction is both discrete and proper.

The following proposition gives a useful framework for which to prove discreteness.

PROPOSITION 3.10. *Fix an Iwasawa inversion space $\mathbb{X} = \mathbb{X}_k^n$, an inversion ι that is either ι_+, ι_-, or ι_c, and a discrete subring $R \subset k$ such that $2 \in R$. Consider the subgroup $\mathcal{Z} \subset \mathrm{Isom}(X)$ consisting of left-translations by points $(z, t) \in \mathbb{X}$ such that $z \in R^n$ and $\|z\|^2 + t \in R$. Then $\mathcal{M} = \langle \mathcal{Z}, \iota \rangle \subset \mathrm{Isom}(\mathbb{H})$ is discrete.*

Example 3.11. For example, in the case of the first Heisenberg group $\mathbb{X}_{\mathbb{C}}^1$, we might chose $R = \mathbb{Z}[\mathbb{i}]$ so $z \in \mathbb{Z}[\mathbb{i}]$ and $t \in \mathbb{i}\mathbb{Z}$.

Proof. We can embed $\mathcal{M}$ as a subgroup of $GL(n + 2, k)$ by mapping ι to a matrix J_ι of the form (2.3), and left-translation by (z, t) to the matrix $A_{(z,t)}$, where

$$A_{(z,t)} = \begin{bmatrix} 1 & 0_n & 0 \\ \sqrt{2}z & \mathrm{id}_n & 0_n \\ \|z\|^2 + t & \sqrt{2}\overline{z} & 1 \end{bmatrix}. \tag{3.2}$$

It is now easy to check that $\mathcal{Z}$ is a group.

Unless $\sqrt{2} \in R$, the matrices $A_{(z,t)}$ will not be matrices over R itself. However, consider the discrete set S of $(n + 2) \times (n + 2)$ matrices $(a_{i,j})_{i,j=1}^{n+2}$ such that $a_{i,j} \in \sqrt{2}R$ if i or j (but not both!) is equal to 1 or $n + 2$, and otherwise $a_{i,j} \in R$. It is easy to check that S is

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

closed under multiplication. Moreover, the generators J_ι and $A_{(z,t)}$ of $\mathcal{M}$ belong to S, so that $\mathcal{M} \subset S$, so $\mathcal{M}$ must be discrete. □

For the rest of this section, we will assume that all the hypotheses of Proposition 3.10 are satisfied, so that the only remaining difficulty is proving properness.

Example 3.12. Let us consider higher-dimensional generalizations of the nearest-integers CFs. Let $k = \mathbb{R}$, $\mathbb{X} = \mathbb{X}^n_{\mathbb{R}} = \mathbb{R}^n$, for some $n \geq 1$, and $\iota = \iota_+$. The space $\mathbb{R}^n$ admits the standard lattice $\mathcal{Z} = \mathbb{Z}^n$ with fundamental domain $K = [-1/2, 1/2)^n$.

When $n = 1$, we get the usual nearest-integer CFs. When $n = 2$, we get a variant of the Hurwitz complex CFs (ι_+ acts like $z \mapsto 1/\overline{z}$). When $n = 3$, we get a three-dimensional CF which we do not believe has been studied before. However, when $n \geq 4$, the corresponding K is no longer proper.

Examples 3.11 and 3.12 fit into the framework of Proposition 3.10 very easily. However, in general, t may not belong to the ring R, but does belong to the additive subgroup R' of $\mathrm{Im}(R)$ defined by

$$R' = \{t \in \mathrm{Im}(R) : \|z\|^2 + t \in R, \text{ there exists } z \in R^n\} \subset \mathrm{Im}(R).$$

One shows that, as a set, we have $\mathcal{Z} = R^n \times R'$.

Let K_1 be the Dirichlet domain around 0 for R and let K_2 be the Dirichlet domain around 0 for R' with respect to the Euclidean metrics on k^n and $\mathrm{Im}(k)$. Then a fundamental domain for $\mathcal{Z}$ in $\mathbb{X}$ is given by $K = K_1^n \times K_2$. In particular, the radius of K is

$$\mathrm{rad}(K) = \sqrt[4]{n^2\,\mathrm{rad}(K_1)^4 + \mathrm{rad}(K_2)^2}.$$

Thus, to obtain a proper system, we require $n^2\,\mathrm{rad}(K_1)^4 + \mathrm{rad}(K_2)^2 < 1$.

Example 3.13. Suppose $k = \mathbb{C}$ and $R = \mathbb{Z}[\mathbb{i}]$. Then we have $R' = \mathbb{i}\mathbb{Z}$, $K_1 = [-1/2, 1/2)^2$, $K_2 = [-1/2, 1/2)\mathbb{i}$. In this case, $\mathrm{rad}(K_1) = 2^{-1/2}$ and $\mathrm{rad}(K_2) = 2^{-1}$. When $n = 1$, this implies that K is proper, and results in the Heisenberg continued fractions in Table 1 above. However, $\mathrm{rad}(K) < 1$ only for $n = 1$ and so this cannot be directly generalized to higher Heisenberg groups.

It is tempting to get around this by replacing R with $\mathbb{Z}[e^{2\pi\mathbb{i}/3}]$, the Eisenstein integers, as then K_1 is a hexagon with radius $3^{-1/2}$. However, this gives $R' = \sqrt{3}\mathbb{i}\mathbb{Z}$, so that $K_2 = [-\sqrt{3}/2, \sqrt{3}/2)\mathbb{i}$, and again $\mathrm{rad}(K) < 1$ only for $n = 1$.

We would, more generally, be interested in CFs on the Heisenberg group with coordinates related to the ring of integers of imaginary quadratic fields. However, if we use $R = \mathcal{O}_d$ for $d = 2, 7, 11$, then the resulting fundamental domain $K_1 \times K_2$ is not proper even when $n = 1$.

Example 3.14. Let $k = \mathcal{H}$ be the quaternions, $n = 1$, and R the Hurwitz integers (3.1), so that $R' = \mathbb{Z}[\mathbb{i}, \mathbb{j}, \mathbb{k}]$. Then $\mathrm{rad}(K_1) = 2^{-1/2}$ (see [44]) and $K_2 = [-1/2, 1/2)^3$ so $\mathrm{rad}(K_2) = \sqrt{3}/2$. In particular, if we look at $X^1_{\mathcal{H}}$, we have $\mathrm{rad}(K) = 1$, narrowly missing the properness criterion. Other nearly proper CF algorithms such as the J. Hurwitz complex CFs are known to be convergent and ergodic, so we hope to be able to extend our results to this case.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

3.3. *Completeness and incompleteness.* We now demonstrate how one can identify complete CFs, or identify symmetries of incomplete CFs.

PROPOSITION 3.15. *CF algorithms associated with $\mathbb{X}^1_{\mathbb{R}}$, $\mathcal{Z} = \mathbb{Z}$, and $\iota_+(x) = 1/x$ (e.g. regular or α-CFs) are incomplete with two central symmetries. CF algorithms associated with $\mathbb{X}^1_{\mathbb{R}}$, $\mathcal{Z} = \mathbb{Z}$, and $\iota_-(x) = -1/x$ (e.g. backwards) are complete.*

Proof. Let $\mathcal{M}_+$ and $\mathcal{M}_-$ be the modular groups associated with ι_+ and ι_-, respectively. We take advantage of the fact that one can embed $\mathcal{M}_-$ into $SL(2, \mathbb{Z})$, while $\mathcal{M}_+$ naturally embeds into the larger $GL(2, \mathbb{Z})$.

That is, we may identify elements of $\mathcal{Z}$ and the inversions $\iota_\pm$ with matrices in $GL(2, \mathbb{Z})$, acting by the usual linear fraction transformations on $\mathbb{R}$, with

$$\mathcal{Z} = \left\{ A_n = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} : n \in \mathbb{Z} \right\} \quad \iota_\pm = \begin{pmatrix} 0 & \pm 1 \\ 1 & 0 \end{pmatrix}.$$

(Note that in the standard convention, translations act by upper-triangular matrices, cf. (3.2).) To test for completeness, note that matrices in $\mathrm{Stab}_{\mathcal{M}_\pm}(\infty)$ have the form

$$\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}.$$

Since $a, d \in \mathbb{Z}$ and $|ad| = 1$, a, d must be units, so we can decompose the matrix as

$$\begin{pmatrix} a & b \\ 0 & d \end{pmatrix} = \begin{pmatrix} 1 & b(d^{-1}) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix},$$

a product of an element of $\mathcal{Z}$ and a diagonal matrix. So the only things that can potentially cause incompleteness are diagonal matrices in $\mathcal{M}$. Since the only diagonal matrices in $SL(2, \mathbb{Z})$ are $\pm I$, which act by the identity, we can conclude $\mathcal{M}_- = \mathrm{Stab}_{\mathcal{M}_-}(\infty)$.

For $GL(2, \mathbb{Z})$, the only potential additional symmetry is given by $x \mapsto -x$, corresponding to a diagonal matrix with $a = -d$. Indeed, this is contained in $\mathcal{M}_+$, represented by the word $\iota A_1 \iota A_{-1} \iota A_1$. In particular, CFs associated with ι_+ are incomplete with two central symmetries. □

A proof similar to the above also implies that the Rosen CFs are complete.

PROPOSITION 3.16. *Let k be the complex, quaternionic, or octonionic division algebra, with $\mathcal{Z}$ given by translation by Gaussian or Eisenstein integers, quaternionic or Hurwitz integers, or Cayley integers respectively. Any k-CFs associated with $\mathcal{Z}$ and an inversion of either $z \mapsto 1/z$ or $z \mapsto -1/z$ is incomplete with at least two central symmetries.*

Proof. One argues along the same lines as the proof of Proposition 3.15, embedding $\mathcal{M}$ into $GL(2, \mathcal{O}_k)$, where $\mathcal{O}_k$ is the corresponding ring of integers. If $\iota(z) = 1/z$, then $\iota A_1 \iota A_{-1} \iota A_1$ is again the central symmetry $z \mapsto -z$. If $\iota(z) = -1/z$, then the central symmetry $z \mapsto -z$ can be represented by the word $\iota A_i \iota A_{-i} \iota A_i$. In the Hurwitz complex CF case, no other central symmetries can be obtained because the matrices of $GL(2, \mathcal{O}_k)$ obtained by the embedding have determinant ± 1, and hence the only diagonal matrices have $a = \overline{d}$ or $a = -\overline{d}$. □

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

PROPOSITION 3.17. *The J. Hurwitz complex CF algorithm is complete.*

Proof. As in Proposition 3.16, we embed $\mathcal{M}$ into $GL(2, \mathbb{Z}[\mathbb{i}])$, with $\iota = \iota_+$ and $\mathcal{Z} = \{A_n : n \in (1+\mathbb{i})\mathbb{Z}[\mathbb{i}]\}$. However, by taking $\mathcal{M}$ modulo 4 and performing an exhaustive computational search, one can confirm that the central symmetry $z \mapsto -z$ never appears. □

PROPOSITION 3.18. *Standard Heisenberg continued fractions are incomplete with four central symmetries.*

Proof. Embed $\mathcal{M}$ into $GL(3, \mathbb{Z}[\mathbb{i}])$ using (3.2). Diagonal matrices then correspond to the rotations $(z, t) \mapsto (\mathbb{i}^k z, t)$. All four of these are, in fact, realized, since one has

$$\iota A_{(0,1)} \iota A_{(0,1)} \iota A_{(0,1)} = \begin{pmatrix} -\mathbb{i} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -\mathbb{i} \end{pmatrix},$$

corresponding to the rotation $(z, t) \mapsto (\mathbb{i}z, t)$. □

4. *Convergence*

Convergence in the specific case of proper and discrete Iwasawa continued fractions with $k = \mathbb{C}$, $n = 1$, and $\mathcal{Z}$ left-translations by the integer Heisenberg group was given in [38], lemma 3.19 through theorem 3.21. We now extend this to the following result.

THEOREM 4.1. *Fix a proper and discrete Iwasawa continued fraction algorithm, and let $x \in K$. If x has infinitely many CF digits, then the convergents $M_i(0)$ converge to x; otherwise, if x has exactly i CF digits, then $M_i(0) = x$.*

As the proof is nearly identical to that in [38] with some notational changes, we only highlight the general method and the new aspects of the proof. In [38], convergence is proven by extending a regular CF formula for the distance between a point and its convergents, which reads as

$$d(x, M_i(0)) = \frac{\prod_{j=0}^{i} |T^j x|}{\|q_i\|^{1/2}},$$

where q_i is the denominator of $M_i(0)$ (see below for a more precise definition). The proof of this formula extends unchanged from the Heisenberg case, and since $|T^j x| \leq \operatorname{rad} K < 1$ for proper Iwasawa CFs, convergence is immediate provided $\|q_i\|$ is bounded away from 0. It is this last point where new techniques are required. In lemma 3.20 of [38], the discreteness of the Gaussian integers was used to prove that $q_i \neq 0$, and thus, since $q_i \in \mathbb{Z}[\mathbb{i}]$, we must have that $\|q_i\| \geq 1$. However, in the general Iwasawa CF case, the rings generated by the coefficients of $\mathcal{M}$ (in a given matrix representation) need not be discrete, so a new technique is needed.

We proceed by first fixing a proper and discrete Iwasawa continued fraction algorithm. Note that we will not use properness explicitly, but it is necessary for the remainder of the proof in [38].

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Recall from §2.2 that $\mathbb{H}$ is the set $\{h = (z, w) \in k^n \times k \ : \ \mathrm{Re}(w) > 0\}$ with boundary $\partial\mathbb{H} = \mathbb{X}$. The coordinate $\mathrm{Re}(w)$ is the *horoheight (at infinity)* $\mathrm{ht}_\infty(h)$. Restricting the horoheight from below produces a *horoball at* ∞, and applying a mapping $M \in \mathcal{M}$ produces a horoball at the point $M(\infty)$. These can be defined directly using the horoheight $\mathrm{ht}_{M(\infty)}(h) := \mathrm{ht}_\infty(M^{-1}(h))$. It follows from the characterization of horoballs as limits of metric balls that horoballs are geodesically convex. We denote the horoball of height C based at a point $M(\infty)$ by $\mathcal{B}_{M(\infty)}(C) = \{h \in \mathbb{H} \ : \ \mathrm{ht}_{M(\infty)}(h) \geq C\}$.

The following generalizes the disjointness result for Ford circles.

THEOREM 4.2. *There exists $C_0 > 0$ such that for every $C \geq C_0$ and $M_1, M_2 \in \mathcal{M}$ satisfying $M_1(\infty) \neq M_2(\infty)$, the horoballs $\mathcal{B}_{M_1(\infty)}(C)$ and $\mathcal{B}_{M_2(\infty)}(C)$ are disjoint.*

Sketch of proof. The result follows from the Margulis lemma by way of the thick–thin decomposition (see, e.g. §5.10 of Thurston's notes [62]) of the quotient orbifold $\mathcal{M}\backslash\mathbb{H}$, which has a cusp corresponding to the point ∞. To see that it has this cusp, note that the translation length for elements of $\mathcal{Z} \subset \mathbb{H}$ goes to zero at large horoheight (note that one can compare actions at different horoheights by conjugating by the dilation δ_r), so that a horoball of sufficiently large horoheight must be contained in the thin part of $\mathcal{M}\backslash\mathbb{H}$. □

We can conclude, in particular, that horoballs based at points other than ∞ are quantitatively bounded with respect to horoheight from ∞.

COROLLARY 4.3. *Let $\mathcal{B} = \mathcal{B}_\infty(h_1)$ be a horoball of height h_1 based at ∞. Then for every $M \in \mathcal{M}$ satisfying $M(\infty) \neq \infty$, one has*

$$\mathrm{ht}_\infty(M(\mathcal{B})) := \sup\{\mathrm{ht}_\infty(h) \ : \ h \in M(\mathcal{B})\} \leq C_0^2/h_1 =: h_2.$$

Proof. We first show that for each $M \in \mathcal{M}$, there exists a $C_M > 0$ such that $\mathrm{ht}_\infty(M(\mathcal{B}_\infty(h))) = C_M h^{-1}$ for each $h > 0$. To verify this, we use the fact that $\mathcal{M} = \langle \mathcal{Z}, \iota \rangle$ to expand $M = \iota a_n \ldots a_1 \iota$ for $a_i \in \mathcal{Z}$, noting that initial and final translations do not affect the horoheight. However, each inversion acts, by lemmas 3.6 and 3.8 of [39], through

$$\mathrm{ht}_\infty(\iota(\mathcal{B}_\infty(h))) = 1/h, \quad \mathrm{ht}_\infty(\iota(\mathcal{B}_x(h))) = h|x|^{-2}.$$

Thus, as long as, for each i, $x_i := (a_i \iota \ldots a_1 \iota)(\infty) \neq 0$, we have

$$\mathrm{ht}_\infty(M(\mathcal{B}_\infty(h)) = h_1^{-1} \prod_{i=1}^{n} |x_i|^{-2}.$$

If at some point $x_i = 0$, then we must have $(\iota a_i \iota \ldots a_1 \iota)(\mathcal{B}_\infty(h)) = \mathcal{B}_\infty(h)$, so that digits $a_1, \ldots, a_i$ may be removed without altering the effect of M on $\mathcal{B}_\infty(h)$. With the reduction implemented, the product $C_M := \prod_{i=1}^{n} |x_i|^{-2}$ is well defined and has the desired property.

To complete the argument, note that from Theorem 4.2 we have that $h^{-1}C_M < h$ for $h = C_0$, so $C_M < C_0^2$ and $\mathrm{ht}_\infty(M(\mathcal{B})) < h_2$, as desired. □

Recall that we have an embedding $\phi : \mathbb{X} \to k^{n+2}$ given by $\phi(z, t) = (1, \sqrt{2}z, \|z\|^2 + t)$; with a corresponding embedding of $\mathcal{M}$ into $U(J) \subset GL(n + 2, k)$ acting on these

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

vectors. Isometries of $\mathbb{X}$ then embed as lower block triangular mappings of the form

$$\begin{bmatrix} a & 0_n & 0 \\ b & A & 0_n \\ c & b^\dagger & \overline{a} \end{bmatrix},$$

where $|a| = 1$ and A is a unitary transformation. The matrix associated with the inversion is given by Lemma 2.14.

Now, given a point $x \in K$ with at least m continued fraction digits (note that [38] uses the variable n instead), let q_m be the denominator of $M_m(0)$; that is, the first coordinate of the vector $M_m\phi(0)$. Thus in the matrix representation of M_m, the top-left entry is q_m and the top-right entry, in norm, is $\|q_{m-1}\|$, matching the matrix representation in lemma 3.16 of [38].

LEMMA 4.4. *Under the assumptions of Theorem 4.1, there exists $C > 0$ such that $q_m \neq 0$ implies $\|q_m\| > C$.*

Proof. By Theorem 4.2, there exists a horoball $\mathcal{B}$ based at ∞ of some horoheight C_1 such that the $\mathcal{M}$-orbit of $\mathcal{B}$ consists of disjoint horoballs. Moreover, the proof of lemma 3.9 of [39] (again, readily extended to the current setting) gives a constant s_0 such that if $q_m \neq 0$, then

$$\mathrm{ht}_\infty(M_m(\mathcal{B})) := \sup\{\mathrm{ht}_\infty(h) \ : \ h \in M_m(\mathcal{B})\} \geq s_0\|q_m\|^{-1}.$$

The disjointness requirement forces $\mathrm{ht}_\infty(M_m(\mathcal{B})) < C_1$, so $\|q_m\| > s_0/C_1 =: C$. □

From here, it remains to show that $q_m \neq 0$. This is just the content of lemma 3.20 of [38] and we can extend the argument to the general case by citing Lemma 4.4 above in place of the fact that non-zero Gaussian integers have norm at least 1.

5. *Markable geodesics*

We now study the way a geodesic γ interacts with the modular group $\mathcal{M}$ related to a proper, discrete, and complete Iwasawa continued fraction algorithm, with the goal of proving the markable geodesic Theorem 5.1 below. We will track the passage of a geodesic through $\mathcal{M}\backslash\mathbb{H}$ by detecting intersections with the unit sphere

$$\mathbb{S} = \{h \in \mathbb{H} \ : \ |h| = 1\}$$

and its images under elements of $\mathcal{M}$. We will obtain an analog of geodesic coding for certain *markable* geodesics, and then show that markability is a generic condition. Note that $\partial\mathbb{S}$ is the unit sphere in $\mathbb{X}$, and that $\iota(\mathbb{S}) = \mathbb{S}$.

THEOREM 5.1. (Markable geodesic theorem) *Fix a complete, proper, and discrete Iwasawa CF algorithm on an Iwasawa inversion space $\mathbb{X}$, with the associated hyperbolic space $\mathbb{H}$, modular group $\mathcal{M}$, and fundamental domain $K \subset \mathbb{X}$ for the lattice $\mathcal{Z} = Stab_{\mathcal{M}}(\infty)$.*

There exists a codimension-one set $C_{\mathbb{W}} \subset T^1\mathbb{H}$ and a marking that assigns to every markable geodesic satisfying $\gamma(0) \in C_{\mathbb{W}}$:

- *digits $a_i \in \mathcal{Z}$ and mappings $M_i \in \mathcal{M}$, for each $i \in \mathbb{Z}$;*
- *increasing indices $i_j \in \mathbb{Z}$ and times t_j, for each $j \in \mathbb{Z}$, with $i_0 = 0, t_0 = 0$;*

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

collectively called the marking of the geodesic γ such that the following properties exist.

(1) *(Full Coverage) The segments $[t_{j-1}, t_j]$ have length uniformly bounded below and hence cover all of $\mathbb{R}$.*
(2) *(Relation to Shift Map) For each $i \geq 1$, a_i is the i^{th} CF digit of γ_+, and M_i is the branch of T^{-i} associated with the shift map T at γ_+.*
(3) *(Cusp Detection) If, for $t \in [t_{j-1}, t_j]$, the horoheight of $\gamma(t)$ from $M\infty$ satisfies $\mathrm{ht}_{M\infty}\,\gamma(t) > h_0$, and if $M^{-1}\gamma_+ \in K$ for some $M \in \mathcal{M}$, then $M = M_{i_j}$.*
(4) *(Intersection Detection) Let $M \in \mathcal{M}$ and $t \in \mathbb{R}$. Then one has $\gamma(t) \in MC_{\mathbb{W}}$ if and only if for some j one has $t = t_j$ and $M = M_{i_j}$.*
(5) *(Shifted Gauss Equivariance) Let $k \in \mathbb{Z}$. The marking $\{a_i', M_i', i_j', t_j'\}$ associated to the markable geodesic $\gamma'(t) := M_{i_k}^{-1}\gamma(t + t_k)$ satisfies: $t_j' = t_{j+k} - t_k$, $i_j' = i_{j+k} - i_k$, $a_i' = a_{i+i_k}$, and $M_i' = M_{i_k}^{-1}M_{i+i_k}$.*

To begin with, in $\mathbb{H}^2_{\mathbb{R}}$, it is apparent from the geometry that any geodesic can only intersect $\mathbb{S}$ transversely at a single point; however, in other hyperbolic spaces, even a generic geodesic may intersect $\mathbb{S}$ at more than one point; indeed when $k \neq \mathbb{R}$, $\mathbb{H}$ does not admit *any* geodesically convex codimension-1 hypersurfaces. However, a generic geodesic intersects $\mathbb{S}$ in finitely many points, so we may speak of the *last* intersection with $\mathbb{S}$.

LEMMA 5.2. *Let γ be a geodesic in $\mathbb{H}$ not contained in $\mathbb{S}$. Then the set of intersections $\gamma \cap \mathbb{S}$ is finite. Furthermore, if there are times t_1, t_2 such that $|\gamma(t_1)| > 1$ and $|\gamma(t_2)| < 1$, then γ does intersect $\mathbb{S}$.*

Proof. The existence of the intersection follows from the definition of $\mathbb{S}$ by $|\cdot| = 1$.

Finiteness follows by an algebraic argument. Because $\mathrm{Isom}(\mathbb{H})$ acts transitively on geodesics, we may write $\gamma = g(\gamma_2)$, where $g \in G$ and γ_2 is the geodesic joining 0 and ∞. Because g acts by projective transformations on $\mathbb{H}$, the condition $|g(\gamma_2(t))| = 1$ induces an algebraic condition on t. Thus, if the condition were to be satisfied for infinitely many t, it must be satisfied for all t, so that $\gamma \subset \mathbb{S}$, a contradiction. □

We now establish the necessary results for the proof of the markable geodesic theorem.

5.1. *Decomposing an arbitrary geodesic.* In the first stage of the proof, we will break up a geodesic γ into segments punctuated by intersections with expected images of the sphere $\mathbb{S}$, in a way that gives us control of the intermediate horoheights. For a more formal statement, see Lemma 5.7 below.

We start by restricting our attention to geodesics that intersect near the top of $\mathbb{S}$. Fix $\epsilon > 0$ such that $\epsilon + 1 < \mathrm{rad}(K)^{-1}$ (this choice comes into play in Lemma 5.4). We then have the following.

LEMMA 5.3. *Suppose γ is a geodesic ray with $|\gamma(0)| \geq 1 + \epsilon$ and $\gamma_+ \in K$. Then the horoheight of any intersection of γ with $\mathbb{S}$ satisfies $\mathrm{ht}_\infty(\gamma(t)) \geq h_2$ for some $h_2 \in (0, 1)$ depending only on ϵ.*

Proof. The existence of the intersection follows from Lemma 5.2.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

To obtain the lower bound on the horoheight of each intersection, note that γ is uniformly transverse to boundary $\mathbb{X}$ (note that we are not working in a conformal model, so γ is not necessarily perpendicular to $\mathbb{X}$), as this is true for the vertical geodesic joining 0 and ∞ and the endpoints of γ are contained in the compact set $\overline{K} \times (\overline{\mathbb{H}} \setminus B(0, 1+\epsilon))$. Thus, there is a minimal horoheight h_2 (that we may assume is in $(0, 1)$) that γ must reach as it moves away from γ_- and γ_+ before an intersection can occur. The same bound must hold for the intermediate segment by the convexity of horoballs. □

We denote the subset of $\mathbb{S}$ having horoheight at least h_2 as $\mathbb{W}$, and refer to both $\mathbb{W}$ and its images under $\mathcal{M}$ as 'walls'.

We next fix a geodesic ray γ originating in $\mathbb{W}$ and terminating in K and let $M_i \in \mathcal{M}$ be the mappings associated with the CF expansion of γ_+. We now look for intersections of γ with walls $M_i(\mathbb{W})$ by iterating the shift map on γ and identifying intersections of $M_i^{-1}(\gamma)$ with $\mathbb{W}$. This happens within finitely many iterations, with control over the intermediate digits.

LEMMA 5.4. *There is a finite collection $\mathcal{M}_0 \subset \mathcal{M}$ such that the following holds. Suppose γ is a geodesic with $\gamma(0) \in \mathbb{W}$ satisfying $\gamma_+ \in K \setminus \mathcal{M}\infty$. Then there exists a time $0 < \mathfrak{t}_1 < \infty$ and a universally bounded $\mathfrak{i}_1 > 0$ such that $M_{\mathfrak{i}_1}^{-1}(\gamma(\mathfrak{t}_1)) \in \mathbb{W}$ and $M_{\mathfrak{i}_1 - 1} \in \mathcal{M}_0$.*

At this point, for notational convenience, we will often drop parentheses when elements of $\mathcal{M}$ act on points or sets of points.

Proof. We note first that since $\gamma_+ \notin \mathcal{M}\infty$, then the continued fraction expansion of γ_+ does not terminate and so $M_i^{-1}\gamma_+$ is well defined and in K for all $i \in \mathbb{N}$.

If $|M_1\gamma(0)| \geq 1 + \epsilon$, the result is immediate by Lemma 5.3.

If not, we proceed iteratively on i, starting at $i = 1$, supposing at every stage that $|M_{i-1}^{-1}\gamma(0)| < 1 + \epsilon$ until we find the minimum positive $\mathfrak{i}_1$ for i for which

$$|M_i^{-1}\gamma(0)| \geq 1 + \epsilon. \tag{5.1}$$

Note that $M_i^{-1} = a_i^{-1}\iota M_{i-1}^{-1}$, $M_0 = \mathrm{id}$, and moreover that a_i^{-1} is an isometry of the metric d.

When $i = 1$, we have by the above observation and our definition of inversions that

$$d(M_1^{-1}\gamma_+, M_1^{-1}\gamma(0)) = d(\iota M_0^{-1}\gamma_+, \iota M_0^{-1}\gamma(0)) = \frac{d(M_0^{-1}\gamma_+, M_0^{-1}\gamma(0))}{|M_0^{-1}\gamma_+||M_0^{-1}\gamma(0)|} \tag{5.2}$$

$$\geq \frac{d(M_0^{-1}\gamma_+, M_0^{-1}\gamma(0))}{\mathrm{rad}(K)(1+\epsilon)} = \frac{d(\gamma_+, \gamma(0))}{\mathrm{rad}(K)(1+\epsilon)}. \tag{5.3}$$

In particular, since $d(\gamma_+, \gamma(0)) \geq d(K, \mathbb{W})$, this implies that

$$|M_1^{-1}\gamma(0)| \geq d(M_1^{-1}\gamma_+, M_1^{-1}\gamma(0)) - |M_1^{-1}\gamma_+| \tag{5.4}$$

$$\geq \frac{d(K, \mathbb{W})}{\mathrm{rad}(K)(1+\epsilon)} - \mathrm{rad}(K). \tag{5.5}$$

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

This lower inequality could be substantially improved if more was known about $M_0^{-1}\gamma_+$. In particular, if $|M_0^{-1}\gamma_+| \le r$ for

$$r = \frac{d(K, \mathbb{W})}{(1+\epsilon)(1+\epsilon+\mathrm{rad}(K))},$$

then we could replace the $\mathrm{rad}(K)$ in the denominator of (5.3) and (5.5) with r and obtain that $|M_1^{-1}\gamma(0)| \ge 1+\epsilon$, so that $i = 1$ itself is the minimum index for which (5.1) holds.

Now we begin the iteration. At every stage we see that

$$\begin{aligned} d(M_i^{-1}\gamma_+, M_i^{-1}\gamma(0)) &\ge \frac{d(M_{i-1}^{-1}\gamma_+, M_{i-1}^{-1}\gamma(0))}{|M_{i-1}^{-1}\gamma_+||M_{i-1}^{-1}\gamma(0)|} \\ &\ge \frac{d(M_{i-2}^{-1}\gamma_+, M_{i-2}^{-1}\gamma(0))}{|M_{i-1}^{-1}\gamma_+||M_{i-1}^{-1}\gamma(0)||M_{i-2}^{-1}\gamma_+||M_{i-2}^{-1}\gamma(0)|} \\ &\cdots \\ &\ge \frac{d(\gamma_+, \gamma(0))}{\prod_{j=0}^{i-1}|M_j^{-1}\gamma_+||M_j^{-1}\gamma(0)|}, \end{aligned}$$

and thus

$$|M_i^{-1}\gamma(0)| \ge \frac{d(K, \mathbb{W})}{(\mathrm{rad}(K)(1+\epsilon))^i} - \mathrm{rad}(K), \tag{5.6}$$

noting again that if $|M_{i-1}^{-1}\gamma_+| \le r$, then one copy of $\mathrm{rad}(K)$ in the denominator of the last inequality can be replaced with r. Thus i satisfies (5.1).

Regardless of whether $|M_{i-1}^{-1}\gamma_+| \le r$ at any stage, since $\mathrm{rad}(K)(1+\epsilon) < 1$ by the initial choice of ϵ, within a bounded number of steps independent of our choice of γ, the expression on the right of (5.6) exceeds $1+\epsilon$. Thus, there must be a uniform bound on $\mathfrak{i}_1$ such that $|M_{\mathfrak{i}_1}^{-1}\gamma(0)| > 1+\epsilon$.

Moreover, we see that if ever in our iterative process, $|M_{i-1}^{-1}\gamma_+| \le r$, then this i must be the desired value $\mathfrak{i}_1$. Thus for $i = \mathfrak{i}_1$, we must have that $|M_j^{-1}\gamma_+| > r$, $0 \le j < i-1$. However, recall that $a_{j+1} = [\iota M_j \gamma_+]$. In particular, this tells us that a_{j+1} must belong to a finite set of values for $0 \le j < i-1$, and since $M_{i-1} = \iota^{-1}a_1\iota^{-1}a_2 \ldots \iota^{-1}a_{i-1}$, there are finitely many options for what it could be. □

COROLLARY 5.5. *There exists a universal $h_1 > 0$ such that under the assumptions of the preceding lemma, we have $\mathrm{ht}_\infty(M_{\mathfrak{i}_1}^{-1}\gamma(t)) > h_1$ for all $0 \le t \le \mathfrak{t}_1$.*

Proof. We already know that $\mathrm{ht}_\infty(M_{\mathfrak{i}_1}^{-1}\gamma(\mathfrak{t}_1)) \ge h_2$ since this point is contained in $\mathbb{W}$.

Let us next consider the possible horoheights of $M_{\mathfrak{i}_1}^{-1}\gamma(0) = a_{\mathfrak{i}_1}^{-1}\iota M_{\mathfrak{i}_1-1}^{-1}\gamma(0)$. The point $\iota M_{\mathfrak{i}_1-1}^{-1}\gamma(0)$ lies in the relatively compact set $\cup\{\iota M^{-1}\mathbb{W} : M \in \mathcal{M}_0\}$, so for some h_3, we obtain $\mathrm{ht}_\infty(\iota M_{\mathfrak{i}_1-1}^{-1}\gamma(0)) > h_3$. Since translation along $\mathbb{X}$ does not affect horoheight, we likewise have $\mathrm{ht}_\infty(M_{\mathfrak{i}_1}^{-1}\gamma(0)) > h_3$.

The lemma now follows with $h_1 = \min(h_2, h_3)$ by convexity of horoballs. □

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

We are now able to characterize $M_{\mathfrak{i}_1}$ as the (essentially) unique element of $\mathcal{M}$ that can detect large horoheights along the geodesic segment between $\gamma(0)$ and $\gamma(\mathfrak{t}_1)$. Let us define an exceptional set $E \subset K$ by

$$E = K \cap \bigcup_{a \in \mathcal{Z} \setminus \{\mathrm{id}\}} aK. \tag{5.7}$$

Since K is a fundamental domain for $\mathcal{Z}$, E has measure zero.

COROLLARY 5.6. *There is an $h_0 > 1$ such that the following holds under the conditions of Lemma 5.4, and for all $0 \le t \le \mathfrak{t}_1$. If $M^{-1}\gamma_+ \in K \setminus E$, $M_{\mathfrak{i}_1}^{-1}\gamma_+ \in K \setminus E$, and $\mathrm{ht}_{M\infty}(\gamma(t)) > h_0$, then $M = M_{\mathfrak{i}_1}$.*

Proof. The geodesic segment $M_{\mathfrak{i}_1}^{-1}\gamma([0, \mathfrak{t}_1])$ is contained in the horoball $\mathcal{B} = \mathcal{B}_\infty(h_1)$, and by Corollary 4.3, there is an h_0 such that the points of $M\mathcal{B}$ have horoheight based at ∞ of at most h_0 when $M\infty \neq \infty$. In particular, this applies to the geodesic segment.

Thus, if $\mathrm{ht}_{M\infty}(\gamma(t)) > h_0$ for any $0 \le t \le \mathfrak{t}_1$, then we conclude that $M\infty = M_{\mathfrak{i}_1}\infty$ and thus that $M^{-1}M_{\mathfrak{i}_1} \in \mathrm{Stab}_{\mathcal{M}}(\infty) = \mathcal{Z}$, by completeness. Moreover, $\gamma_+ \in M(K \setminus E) \cap M_{\mathfrak{i}_1}(K \setminus E)$ so that $M(K \setminus E) \cap M_{\mathfrak{i}_1}(K \setminus E) \neq \emptyset$. Thus $(M^{-1}M_{\mathfrak{i}_1}(K \setminus E)) \cap (K \setminus E) \neq \emptyset$. By the definition of E, the only element of $\mathcal{Z}$ that takes any part of $K \setminus E$ back to itself is the identity element. Thus $M = M_{\mathfrak{i}_1}$ as desired.

We may assume without loss of generality that $h_0 > 1$. □

Iterating the above results gives us a sequence of indices $\mathfrak{i}_j$ and times $\mathfrak{t}_j$ with the following properties.

LEMMA 5.7. *Let h_0 be the constant in Corollary 5.6 and γ a geodesic ray with $\gamma(0) \in \mathbb{W}$, $\gamma(t) \notin \mathbb{W}$ for $t > 0$, and $\gamma_+ \in K \setminus \mathcal{M}(\{\infty\} \cup E)$. Then there is an increasing sequence $\mathfrak{i}_j$, $j \ge 0$, of indices starting with $\mathfrak{i}_0 = 0$ and an increasing sequence of times $\mathfrak{t}_j$, $j \ge 0$, starting with $\mathfrak{t}_0 = 0$ such that the following conclusions hold.*

(1) *For each $j \ge 0$: $\gamma(\mathfrak{t}_j) \in M_{\mathfrak{i}_j}\mathbb{W}$, while for $t > \mathfrak{t}_j$, $\gamma(t) \notin M_{\mathfrak{i}_j}\mathbb{W}$.*
(2) *For each $j \ge 1$: if $\mathfrak{t}_{j-1} \le t \le \mathfrak{t}_j$ and a matrix $M \in \mathcal{M}$ satisfies both $M^{-1}\gamma_+ \in K$ and $\mathrm{ht}_\infty M^{-1}\gamma(t) > h_0$, then $M = M_{\mathfrak{i}_j}$.*

Proof. Given γ satisfying the assumptions of the lemma, the $j = 0$ case of conclusion (1) is trivial.

Moreover, we obtain $\mathfrak{i}_1$ and $\mathfrak{t}_1$ from Lemma 5.4. There might be several choices of $\mathfrak{t}_1$ due to multiple intersections with $M_{\mathfrak{i}_1}\mathbb{W}$ (see Lemma 5.2); however, we let $\mathfrak{t}_1$ be the last of these. We then know that $M_{\mathfrak{i}_1}^{-1}\gamma(\mathfrak{t}_1) \in \mathbb{W}$, which is equivalent conclusion (1) for $j = 1$. We then obtain conclusion (2) for $j = 1$ from Corollary 5.6.

We now proceed inductively: once $\mathfrak{t}_j$ and $\mathfrak{i}_j$ are defined, we replace γ with the geodesic segment $\gamma'(t') = M_{\mathfrak{i}_j}^{-1}\gamma(t' + \mathfrak{t}_j)$ restricted to $t' \in [0, \infty)$. We then obtain $\mathfrak{t}'_1$, $\mathfrak{i}'_1$, and $M_{\mathfrak{i}'_1}$ as before, and take $\mathfrak{t}_{j+1} = \mathfrak{t}_j + \mathfrak{t}'_1$ and $\mathfrak{i}_{j+1} = \mathfrak{i}_j + \mathfrak{i}'_1$. The desired properties follow from the fact that the shift map acts as a shift on the digits of γ_+, via the identity $M_{\mathfrak{i}_{j+1}} = M_{\mathfrak{i}_j}M'_{\mathfrak{i}'_1}$.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Finally, we note that since $h_0 > 1$, if $\mathrm{ht}_\infty M^{-1}\gamma(t) > h_0$, then t cannot be any of the $\mathfrak{t}_j$ values, so there is no ambiguity in conclusion (2). □

5.2. *Decomposing a markable geodesic.* Lemma 5.7 tells us how geodesic rays leaving the wall $\mathbb{W}$ towards K return to other walls $M\mathbb{W}$, for various $M \in \mathcal{M}$. In particular, if a point on our ray has large horoheight with respect to $M\infty$, then the ray should cross the wall $M\mathbb{W}$. We now use this to define a set $C_\mathbb{W} \subset T^1\mathbb{H}$ lying over $\mathbb{W}$, where this 'if' condition becomes 'if and only if.' We will then call a geodesic *markable* if it intersects $\mathcal{M}$-translates of $C_\mathbb{W}$ infinitely often in the past and future, and show in the markable geodesic theorem (Theorem 5.1) that the behavior of a markable geodesic's cusp excursions is directly related to the continued fraction expansion of the forward endpoint. We will see in Corollary 6.6 that markable geodesics are generic.

Definition 5.8. Using the constant $h_0 > 1$ provided by Lemma 5.7, we define $C_\mathbb{W} \subset T^1\mathbb{H}$ as follows: a vector based at a point in $\mathbb{W}$ is in the set $C_\mathbb{W}$ if and only if the corresponding geodesic *line* γ satisfies:

(1) $\gamma(0) \in \mathbb{W}$, while for $t > 0$, $\gamma(t) \notin \mathbb{W}$;
(2) $\gamma_+ \in K \setminus \mathcal{M}(\{\infty\} \cup E)$, where E is the exceptional set (5.7);
(3) there exists a *spotter time* $\widehat{t} < 0$ such that $\mathrm{ht}_\infty(\gamma(\widehat{t})) > h_0$.

Critically, the third condition tells us that γ intersects some $MC_\mathbb{W}$ for $M \in \mathcal{M}$ at some time t_M if and only if there is an associated spotter time $\widehat{t}_M < t_M$ satisfying $\mathrm{ht}_\infty M^{-1}\gamma(\widehat{t}_M) > h_0$, or equivalently $\mathrm{ht}_{M\infty}\gamma(\widehat{t}_M) > h_0$.

Definition 5.9. A geodesic γ is *markable* if it intersects $\mathcal{M}$-translates of $C_\mathbb{W}$ infinitely many times in both the past and the future. Unless stated otherwise, we will also assume that $\gamma(0) \in C_\mathbb{W}$.

In the following lemma, we will show that for markable geodesics, spotter times follow a natural progression. That is, if we see a spotter time $\widehat{t}$ associated with an intersection time t, then we must move beyond t before seeing the spotter time associated with any other intersection.

LEMMA 5.10. *Let γ be a markable geodesic, and $M, M' \in \mathcal{M}$. Suppose that $\gamma(a) \in MC_\mathbb{W}$ and $\gamma(b) \in M'C_\mathbb{W}$, attested by the corresponding spotter times $\widehat{a}, \widehat{b}$. Then these must alternate order: if $a < b$, then $\hat{a} < a < \hat{b} < b$.*

Proof. We will prove an equivalent statement: if $\max(\widehat{a}, \widehat{b}) < \min(a, b)$, then $a = b$. Suppose it is false. Since γ is markable, we may assume without loss of generality that $\gamma(0) \in C_\mathbb{W}$, $0 < \hat{a} < \hat{b} < \min(a, b)$.

Let $\mathfrak{t}_j$ be the sequence in Lemma 5.7. Then for some fixed j, we have $\mathfrak{t}_{j-1} < \hat{a} \le \mathfrak{t}_j$. Conclusion (2) of the same lemma states that, since $\hat{a}$ is in the correct range and $\gamma(a) \in MC_\mathbb{W}$, we have $M = M_{i_j}$ and by the definition of $\mathfrak{t}_j$ (that is, conclusion (1) of the lemma) we have $a = \mathfrak{t}_j$. Furthermore, $\mathfrak{t}_{j-1} < \hat{b} < a = \mathfrak{t}_j$, so by the same argument $b = \mathfrak{t}_j$, as desired. □

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

We can now show that if a geodesic starts in $C_{\mathbb{W}}$, its next intersection with a translate of $C_{\mathbb{W}}$ will be captured by an iteration of the shift map.

LEMMA 5.11. *Let γ be a markable geodesic such that $\gamma(0) \in C_{\mathbb{W}}$, and suppose that the next intersection with a translate of $C_{\mathbb{W}}$ occurs at $MC_{\mathbb{W}}$. Then for some $j \geq 1$, we have $M = M_{\mathfrak{i}_j}$ and $\gamma(\mathfrak{t}_j) \in MC_{\mathbb{W}}$, where $\mathfrak{i}_j, \mathfrak{t}_j$ are defined for γ in Lemma 5.7.*

Proof. Let $t > 0$ denote the time when $\gamma(t) \in MC_{\mathbb{W}}$. We know that there must exist a spotter time $\widehat{t}$ associated with t and moreover, by Lemma 5.10, we know that $0 < \widehat{t} < t$. Let $j \geq 1$ be such that $\mathfrak{t}_{j-1} \leq \widehat{t} \leq \mathfrak{t}_j$. Then by conclusion (2) of Lemma 5.7, we have that $M = M_{\mathfrak{i}_j}$ and $\gamma(\mathfrak{t}_j) \in MC_{\mathbb{W}}$. □

We can now prove the markable geodesic theorem.

Proof of Theorem 5.1. For positive i, let a_i and M_i be the digits and mappings corresponding to the CF expansion of the forward endpoint γ_+, making property (2) immediate. We will define the remaining data iteratively.

Let $t_1 > 0$ be the first positive time when γ intersects an $\mathcal{M}$-translate of $C_{\mathbb{W}}$. Lemma 5.11 then provides an index k such that $t_1 = \mathfrak{t}_k$ and a corresponding number $\mathfrak{i}_k$ which we record as i_1 satisfying $\gamma(t_1) \in M_{i_1} C_{\mathbb{W}}$. We will now show that properties (1), (4), and (3) hold on the initial segment $[t_0, t_1]$.

Let $\hat{t}_1$ be a spotter time associated with the intersection of γ with $M_{i_1} C_{\mathbb{W}}$; that is, $\hat{t}_1 < t_1$ and $\mathrm{ht}_{M_{i_1}\infty}\, \gamma(\hat{t}_1) > h_0 > 1$. Since $\gamma(t_0) \in C_{\mathbb{W}}$ and $\gamma(t_1) \in M_{i_1} C_{\mathbb{W}}$, then by Lemma 5.10, we have that $\hat{t}_1 \in [t_0, t_1]$. Let ϵ be the distance (not depending on γ) between the horospheres $\mathrm{ht}_\infty(\cdot) = 1$ and $\mathrm{ht}_\infty(\cdot) = h_0$. Since γ is a unit speed geodesic, $t_1 - t_0 > \epsilon$, and property (1) holds for $j = 1$.

Next, the 'if' direction of property (4) is immediate for $j = 0$ and $j = 1$ from the definitions. Now suppose $t \in (t_0, t_1]$ satisfies $\gamma(t) \in MC_{\mathbb{W}}$ for some $M \in \mathcal{M}$. Then by definition of t_1, we have that $t = t_1$, and from Lemma 5.7, we have that $M = M_{i_1}$. Thus the 'only if' direction of property (4) holds for $t \in (t_0, t_1]$.

Suppose next that $t \in [t_0, t_1]$ satisfies $\mathrm{ht}_{M\infty}\, \gamma(t) > h_0$ for some $M \in \mathcal{M}$. Then by Lemma 5.7, there exists $\ell \geq 1$ and $t' > t$ such that $M = M_\ell$, and $\gamma(t') \in M_\ell \mathbb{W}$. By definition of $C_{\mathbb{W}}$ via spotter times, we obtain that $\gamma(t') \in M_\ell C_{\mathbb{W}}$. Since we assumed that t_1 is the first time that the forward ray of γ intersects $C_{\mathbb{W}}$, we have that $t_1 \leq t'$. The converse inequality is given by Lemma 5.10, since t is a spotter time associated to t', so that $t_1 = t'$ and $M = M_{i_1}$ follows from property (4). So property (3) holds for $j = 1$.

To define t_j, i_j for $j \geq 2$, we now consider a renormalized geodesic $\gamma' = M_{i_1}^{-1}\gamma$ with $\gamma'(0) = M_{i_1}^{-1}\gamma(t_1)$. We may then find t_1', i_1' for γ' as we did above and let $t_2 = t_1 + t_1'$ and $i_2 = i_1 + i_1'$. Iterating this procedure gives t_j, i_j for all $j \geq 1$. By the work above, properties (1), (3), and (4) hold on the corresponding initial segment of the renormalized geodesics and thus hold on the entire forward geodesic ray of γ. Moreover, from this definition, we see that property (5) holds for all i, j, k that are non-negative.

To define a_i, M_i for non-negative i and i_j, t_j for negative j, let t_{-1} be the smallest (in norm) negative value for which $\gamma(t_{-1})$ intersects a $\mathcal{M}$-translate $MC_{\mathbb{W}}$ of $C_{\mathbb{W}}$. Consider a renormalized geodesic $\gamma' = M^{-1}\gamma$ with $\gamma'(0) = M^{-1}\gamma(t_{-1}) \in C_{\mathbb{W}}$. Set $i_{-1} = -i_1'$,

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

$a_i = a'_{i+i_{-1}}$, and $M_i = M^{-1}M'_{i-i_{-1}}$ for $i_{-1} < i \leq 0$. Since γ' is a markable geodesic satisfying the conditions of the theorem and properties (1)–(4) hold for $\gamma'|_{[0,\infty]}$, properties (1)–(5) hold for $\gamma|_{[t_{-1},\infty)}$. Iterating this process yields the remaining definitions and properties on the backwards ray of γ (note that the full ray is covered by property (1)). □

6. *Ergodicity*

We now prove the ergodicity of the shift map by first relating the cross-section $C_{\mathbb{W}}$ studied in §5 to geodesic flow on a quotient of $\mathbb{H}$, and then to the shift map on the boundary. We start by recalling the ergodicity result for geodesic flow. This section culminates in the ergodicity part of Theorem 1.2.

Remark 6.1. All statements concerning ergodicity and measure will be made with respect to the relevant Hausdorff measure; depending on context, this can be interpreted as Haar measure, surface measure, or Lebesgue measure. Because there are no surprises along the way, we will suppress discussion of the details.

6.1. *Ergodicity of the geodesic flow.* The space $(\mathbb{H}, d_{\mathbb{H}})$ is a symmetric space with a complete Riemannian metric with pinched negative curvature. In particular, any pair of points in $\mathbb{H}$ (indeed, in $\overline{\mathbb{H}} \cup \{\infty\}$) determines a unique geodesic. Alternately, a *pointed* geodesic is determined by an element of the unit tangent bundle $T^1\mathbb{H}$, namely a point in $\mathbb{H}$ and a unit vector over it.

The geodesic flow on $T^1\mathbb{H}$ moves vectors along geodesics as follows.

Definition 6.2. (Geodesic flow) Given a vector $(h, v) \in T^1\mathbb{H}$, let $\gamma : \mathbb{R} \to \mathbb{H}$ be a unit-speed geodesic satisfying $\gamma(0) = h$ and $\gamma'(0) = v$. The time-t geodesic flow of (h, v) is then given by $\phi_t(v) := (\gamma(t), \gamma'(t)) \in T^1\mathbb{H}$.

Given a set $A \subset T^1\mathbb{H}$, one says that A is ϕ-invariant, if for each $t \in \mathbb{R}$, the symmetric difference $(\phi_t^{-1}A)\triangle A$ has measure zero. We will be interested in sets A that are furthermore invariant under a lattice $\Gamma \subset G$, that is, $\mu(\gamma(A)\triangle A) = 0$ for every $\gamma \in \Gamma$.

We can now state Mautner's ergodicity theorem (cf. Moore's extension of the result to the frame bundle [65]).

THEOREM 6.3. (Mautner's ergodicity theorem [42]) *Let Γ be a lattice in G, and $A \subset T^1\mathbb{H}$ a Γ-invariant set that is furthermore invariant under geodesic flow. Then either $\mu(A) = 0$ or $\mu(T^1\mathbb{H} \setminus A) = 0$.*

6.2. *Ergodicity of the markable cross-section.* We continue working with a fixed complete, discrete, and proper Iwasawa continued fraction algorithm. Consider the natural projection $\pi_{\mathbb{H}} : \mathbb{H} \to \mathcal{M}\backslash\mathbb{H}$.

Mautner's Theorem 6.3 immediately applies to our setting. We record this in the following lemma, which can be interpreted either in the formulation of Theorem 6.3 or, equivalently, using orbifold geodesic flow.

LEMMA 6.4. *Geodesic flow on $\mathcal{M}\backslash\mathbb{H}$ is ergodic.*

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Proof. $\mathcal{M}$ is assumed to be discrete; to show it is a lattice, we must show that there exists a finite-volume fundamental domain for $\mathcal{M}$. Let K' be the region lying over both K having horoheight at least $\epsilon > 0$, for a choice of ϵ satisfying $\mathrm{rad}(K \times [0, \epsilon])^{-2} > 1$. Given a point $h \in \mathbb{H}$, we may use $\mathcal{Z}$ to translate h so that it lies over K, and invert it if necessary to increase its horoheight multiplicatively by at least $\mathrm{rad}(K \times [0, \epsilon])^{-2}$ (see [39] for the interaction of horoheight and inversions), and translate again to place it over K. Within finitely many iterations, we obtain an image of h contained in K'. Thus, K' contains a fundamental domain for the $\mathcal{M}$ action on $\mathbb{H}$. Lastly, K' has horoheight bounded below and bounded extent along $\mathbb{X}$, so has finite hyperbolic volume. □

LEMMA 6.5. *The first-return map on $\pi_{\mathbb{H}}(C_{\mathbb{W}})$ is a.e. well defined and ergodic.*

Proof. Consider the family $\mathcal{F} \subset T^1\mathbb{H}$ of geodesic rays that pass through $C_{\mathbb{W}}$. Recalling that $C_{\mathbb{W}}$ consists of geodesics coming from large horoheight through the wall $\mathbb{W}$ and proceeding to K, it is clear $\mathcal{F}$ has positive measure. Since $\mathcal{M}$ is discrete, $\pi_{\mathbb{H}}(\mathcal{F})$ also has positive measure. Thus, by ergodicity, almost every geodesic in $\mathcal{M}\backslash\mathbb{H}$ passes through $\pi_{\mathbb{H}}(C_{\mathbb{W}})$.

Since $\pi_{\mathbb{H}}(C_{\mathbb{W}})$ is generically transverse to geodesic flow, we conclude that almost every geodesic ray in $\pi_{\mathbb{H}}(C_{\mathbb{W}})$ returns to $\pi_{\mathbb{H}}(C_{\mathbb{W}})$, and that the resulting first-return map is ergodic. □

We are now able to show that markable geodesics are generic.

COROLLARY 6.6. *Almost every geodesic γ satisfying $\gamma(0) \in C_{\mathbb{W}}$ is markable.*

Proof. By the previous lemma, the first-return mapping on $\pi_{\mathbb{H}}(C_{\mathbb{W}})$ is well defined. Thus, given a generic geodesic ray γ in $C_{\mathbb{W}}$, $\pi_{\mathbb{H}}(\gamma)$ will return to $\pi_{\mathbb{H}}(C_{\mathbb{W}})$ after some time. Lifting to $\mathbb{H}$, this implies that γ intersects $MC_{\mathbb{W}}$ for some $M \in \mathcal{M}$. Iterating the first-return map gives infinitely many intersections. Reversing the flow gives the same result for the backward orbit of γ. □

Now that we have shown that almost all geodesics are markable, we can quickly prove that $C_{\mathbb{W}}$ has no unexpected symmetries.

COROLLARY 6.7. *The restriction of $\pi_{\mathbb{H}}$ to $C_{\mathbb{W}}$ is a.e. injective.*

Proof. Suppose the statement is false, and there exists a non-identity mapping $M \in \mathcal{M}$ such that $MC_{\mathbb{W}} \cap C_{\mathbb{W}}$ has positive measure. Then by the previous corollary, there is a markable geodesic γ with $\gamma(0) \in MC_{\mathbb{W}} \cap C_{\mathbb{W}}$. However, then we have $\gamma(0) \in C_{\mathbb{W}}$ and $M\gamma(0) \in C_{\mathbb{W}}$, and it follows from the intersection detection property of Theorem 5.1 that $M = M_{i_0} = \mathrm{id}$. □

Definition 6.8. Let us define a mapping $\psi : C_{\mathbb{W}} \to C_{\mathbb{W}}$ by $\psi(\gamma)(t) = M_{i_1}^{-1}\gamma(t + t_1)$, where M_{i_1} and t_1 are given by Theorem 5.1. This is well-defined a.e.

PROPOSITION 6.9. *The mapping $\psi : C_{\mathbb{W}} \to C_{\mathbb{W}}$ is ergodic.*

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Proof. The first-return map on $\pi_{\mathbb{H}}(C_{\mathbb{W}})$ is ergodic by Lemma 6.5. Corollary 6.7 then allows us to identify $\pi_{\mathbb{H}}(C_{\mathbb{W}})$ with $C_{\mathbb{W}}$, and Theorem 5.1 tells us that ψ is indeed a lift of the first-return mapping on $\pi_{\mathbb{H}}(C_{\mathbb{W}})$. □

6.3. *Ergodicity of a natural extension and of the shift map.* At this point, we would like to project $C_{\mathbb{W}}$ onto the forward endpoint and use the ergodicity of ψ to derive the ergodicity of T. However, the transformation that ψ induces on the forward endpoint is a jump transformation associated with T and it is not the case that the ergodicity of a jump transformation implies the ergodicity of the original transformation. (See, for example, Chs. 17–19 of [**54**].) So we will instead project onto both endpoints and analyze the resulting transformation more carefully.

Throughout the rest of this section, we will assume, without directly stating it, that all statements about sets hold up to sets of zero measure and that any geodesic under consideration is markable, since this is a generic condition. We continue to work with a complete, discrete, and proper Iwasawa CF expansion.

Let $\pi : C_{\mathbb{W}} \to K \times \mathbb{X}$ be the injective map from a geodesic γ intersecting $C_{\mathbb{W}}$ to its forward and backward endpoints (γ_+, γ_-). On $\pi(C_{\mathbb{W}})$, ψ induces the isomorphic mapping $\Psi = \pi \circ \psi \circ \pi^{-1}$, which is ergodic on $\pi(C_{\mathbb{W}})$. Since, by the markable geodesic Theorem 5.1, ψ acts on a geodesic γ by the mapping M_{i_1} associated with γ_+, we conclude that $\Psi(\gamma_+, \gamma_-) = (M_{i_1}^{-1}\gamma_+, M_{i_1}^{-1}\gamma_-)$.

Let us extend the shift map T to act on $K \times \mathbb{X}$ by $\hat{T}(z, w) = (M_1^{-1}z, M_1^{-1}w)$, where $M_1 \in \mathcal{M}$ is the mapping associated with z. Since $Tz = M_1^{-1}z$, this truly is an extension. Let $\overline{K} = \bigcup_{i=0}^{\infty} \hat{T}^i \pi(C_{\mathbb{W}}) \subset K \times \mathbb{X}$.

We wish to compare how Ψ acts on $\pi(C_{\mathbb{W}})$ with how $\hat{T}$ acts on $\overline{K}$. In the following lemma, will show that the restriction $\hat{T}|_{\overline{K}}$ of $\hat{T}$ to $\overline{K}$ is well behaved.

LEMMA 6.10. *$\hat{T}|_{\overline{K}} : \overline{K} \to \overline{K}$ is surjective. Furthermore, almost every point of $\overline{K}$ returns to $\pi(C_{\mathbb{W}})$ within finitely many iterations of $\hat{T}|_{\overline{K}}$, so that we have*

$$\overline{K} = \bigcup_{i=0}^{\infty} \hat{T}|_{\overline{K}}^{-i} \pi(C_{\mathbb{W}}). \tag{6.1}$$

Proof. It is immediate from the definition of $\overline{K}$ that $\hat{T}|_{\overline{K}}\overline{K} \subset \overline{K}$. To prove the reverse containment, we wish to show that for any $(z, w) \in \overline{K}$, there exists $(z', w') \in \overline{K}$ with $\hat{T}|_{\overline{K}}(z', w') = (z, w)$.

Since $(z, w) \in \overline{K}$, there exists a smallest non-negative integer i such that $(z, w) \in \hat{T}|_{\overline{K}}^{i}\pi(C_{\mathbb{W}})$. If $i \geq 1$, then clearly there is $(z', w') \in \hat{T}|_{\overline{K}}^{i-1}\pi(C_{\mathbb{W}})$ such that $\hat{T}|_{\overline{K}}(z', w') = (z, w)$.

So suppose $i = 0$. Then $(z, w) \in \pi(C_{\mathbb{W}})$. Since Ψ is an onto map of $\pi(C_{\mathbb{W}})$ to itself, for almost every (z, w), there exists some (z'', w'') such that $\Psi(z'', w'') = (z, w)$. Thus, if we let i_1 be the index so that $\Psi(z'', w'') = (M_{i_1}^{-1}z'', M_{i_1}^{-1}w'') = \hat{T}|_{\overline{K}}^{i_1}(z'', w'')$, then we have that $(z, w) \in \hat{T}|_{\overline{K}}^{i_1}\pi(C_{\mathbb{W}})$ with $i_1 > 0$ and the argument of the previous paragraph applies.

Implicit in the last paragraph is the idea that for almost every $(z, w) \in \pi(C_{\mathbb{W}})$, $\Psi(z, w) \in \pi(C_{\mathbb{W}})$ as well, so that (z, w) returns to $\pi(C_{\mathbb{W}})$ in a finite number of

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

iterations of $\hat{T}|_{\overline{K}}$. Since every $(z, w) \in \overline{K} \setminus \pi(C_{\mathbb{W}})$ appears in some $\hat{T}|_{\overline{K}}^{i}\pi(C_{\mathbb{W}})$, say, $\hat{T}|_{\overline{K}}^{i}(z'', w'') = (z, w)$, we can also extend this to say that almost every point in $\overline{K}$ returns to $\pi(C_{\mathbb{W}})$ under a finite number of iterations.

This immediately shows that $\overline{K} \subset \bigcup_{i=0}^{\infty} \hat{T}|_{\overline{K}}^{-i}\pi(C_{\mathbb{W}})$ and the reverse inclusion is trivial. □

We restrict our attention to $\overline{K}$, setting $\hat{T} := \hat{T}|_{\overline{K}}$.

Equation (6.1) looks similar to the definition of a natural extension, so raises the following question, which we will not address.

Question 7. Is $\hat{T} : \overline{K} \to \overline{K}$ the natural extension of $T : K \to K$?

One can look at, for example, [**18**] for a discussion of the natural extension in the case of the A. Hurwitz complex CF.

Now we can state the connection between Ψ and $\hat{T}$.

LEMMA 6.11. *Ψ is the transformation induced by restricting $\hat{T}$ to $\pi(C_{\mathbb{W}})$.*

Proof. Since $\mathcal{Z}$ is countable, the set of points in K with eventually periodic continued fraction expansions is countable as well, and hence, since we are working up to measure zero, we may assume any points under consideration are not eventually periodic.

Let $(z, w) \in \pi(C_{\mathbb{W}})$ and let $i(z, w)$ be the minimal positive integer such that $\hat{T}^{i(z,w)}(z, w) \in \pi(C_{\mathbb{W}})$. The existence of $i(z, w)$ a.e. follows from Lemma 6.10. We wish to show that, where it exists, $\hat{T}^{i(z,w)}(z, w) = \Psi(z, w)$.

Let γ be the markable geodesic with endpoints (z, w), and let i_1 be the corresponding value from the marking in Theorem 5.1. Then $\Psi(z, w) = (M_{i_1}^{-1}z, M_{i_1}^{-1}w)$ and thus $\hat{T}^{i_1}(z, w) = \Psi(z, w) \in \pi(C_{\mathbb{W}})$. By the minimality of $i(z, w)$, we have that $i(z, w) \leq i_1$. We must show that $i(z, w)$ cannot be strictly less than i_1.

Suppose $i(z, w) < i_1$ and consider the mapping $M = M_{i(z,w)}$. Since $(M^{-1}z, M^{-1}w) \in \pi(C_{\mathbb{W}})$, $M^{-1}\gamma$ intersects $C_{\mathbb{W}}$. This means γ intersects $MC_{\mathbb{W}}$ and thus by the intersection detection property of Theorem 5.1, $M = M_{i_j}$ for some j. Since the two mappings are equal, we have that $T^{i(z,w)}z = M_{i(z,w)}^{-1}z = M_{i_j}^{-1}z = T^{i_j}z$. However, since we have assumed z does not have an eventually periodic expansion, this is only possible if $i(z, w) = i_j$. Additionally, since there are no positive i_j between 0 and i_1, we must have that $i(z, w) = i_1$, which completes the proof. □

While there is a close connection between the dynamical properties of a map and the dynamical properties a new map induced from the first, in general, one cannot use the ergodicity of the induced map to conclude the ergodicity of the original map; however, Lemma 6.11 when combined with (6.1) is enough to prove the following result immediately (see theorem 17.2.4 of [**54**] for full details).

LEMMA 6.12. *$\hat{T}$ is ergodic on $\overline{K}$.*

We can now project to the first coordinate to complete the proof of Theorem 1.2 (see also §1.3).

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Proof of Theorem 1.2. Let us suppose the shift map is not ergodic. Then there are complementary subsets A and B of K that are both invariant under T and have non-zero measure. We may extend these to complementary subsets A', B' of $\overline{K}$ by taking their preimages under projection to the first coordinate. Both A' and B' have positive measure since $\pi(C_{\mathbb{W}}) \subset \overline{K}$ and we claim there exists a neighborhood U of infinity in $\mathbb{X}$ such that $K \times U \subset \pi(C_{\mathbb{W}})$.

Let us now show that this set U does exist. Consider any pair (γ_+, γ_-) of endpoints of a geodesic γ, such that $\gamma_+ \in K$ and $|\gamma_-|$ is sufficiently large. In particular, if $|\gamma_-| > 1 + \epsilon$ with ϵ as in Lemma 5.3, then the conclusion of that lemma and the definition of $\mathbb{W}$ imply that the geodesic γ passes through $\mathbb{W}$. Moreover, by taking the framework of Lemma 5.3 and dilating, we see that if $|\gamma_-|$ is sufficiently large, then the geodesic must travel far into the cusp at infinity: namely, there must exist a time $\hat{t}$ such that $\mathrm{ht}_\infty(\gamma(\hat{t})) > h_0$. Thus, γ does intersect $C_{\mathbb{W}}$ and $(\gamma_+, \gamma_-) \in \pi(C_{\mathbb{W}})$ as desired. (Since we are working up to measure zero sets, we may assume that $\gamma_+ \notin \mathcal{M}(\{\infty\} \cup E)$ as well.)

Consider $\hat{T}^{-1}A'$. Any point $(z, w) \in \overline{K}$ such that $\hat{T}(z, w) \in A'$ must clearly satisfy $Tz \in A$. In other words, $z \in T^{-1}A = A$. Thus $(z, w) \in A'$, so $\hat{T}^{-1}A' \subset A'$ and likewise $\hat{T}^{-1}B' \subset B'$. Hence, A' and B' are both disjoint T-invariant subsets of $\overline{K}$ with positive measure. The ergodicity of $\hat{T} : \overline{K} \to \overline{K}$ provided by Lemma 6.12 gives the contradiction. □

Remark 6.13. We have proved ergodicity with respect to Lebesgue measure, but with the framework we have developed, we may now consider the question of absolutely continuous invariant measures as well.

First, note that since geodesic flow preserves Haar measure on $\mathbb{H}$, there is a canonical derivation of an invariant measure for ψ on $C_{\mathbb{W}}$. This then projects to an invariant measure for Ψ on $\pi(C_{\mathbb{W}})$. Since Ψ is the transformation induced by restriction $\hat{T}$ to $\pi(C_{\mathbb{W}})$, there is again a canonical derivation of an invariant measure for $\hat{T}$ on $\overline{K}$ (see [**54**, Theorem 17.1.6]). From here, projection onto the first coordinate would give an invariant measure for T on K. All of these operations preserve the fact that they are absolutely continuous with respect to the corresponding Hausdorff measure.

Note that even though the measure on $C_{\mathbb{W}}$ and $\pi(C_{\mathbb{W}})$ is bounded, the measure on $\overline{K}$ and K may be infinite. Indeed, this occurs for the Rosen continued fractions [**20**].

6.4. *Application: ergodic components of incomplete Iwasawa CFs.* In this subsection, we will prove Theorem 1.3.

Let $\mathcal{R}$ denote the set of central symmetries of $\mathcal{M}$ (cf. Definition 3.7).

LEMMA 6.14. *Let $r \in \mathcal{R}$. Then for any $a \in \mathcal{Z}$, there exists $a' \in \mathcal{Z}$, $r' \in \mathcal{R}$ such that $a\iota r = r'a'\iota$. Moreover, if r' is the identity, then r must be as well.*

Proof. Since $a\iota r\iota^{-1} \in \mathrm{Stab}_{\mathcal{M}}(\infty)$, the decomposability assumption on $\mathcal{R}$ implies that there exist $r' \in \mathcal{R}$ and $a' \in \mathcal{Z}$ such that $a\iota r\iota^{-1}\iota = r'a'\iota$, as desired.

Let r'' denote $\iota r\iota^{-1}$. Since this fixes 0 and ∞, it must belong to $\mathcal{R}$. So if r' is the identity, then $r'a' = ar''$ implies that $a^{-1}a' = r''$. However, $\mathcal{R} \cap \mathcal{Z} = \{\mathrm{id}\}$, so r'' and hence r must be the identity. □

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

At this point, we wish to start connecting the behavior of an incomplete Iwasawa CF with n central symmetries with the behavior of its completion.

As such, let us specialize our notation. Let K be the symmetric fundamental domain for the incomplete continued fraction over $\mathcal{Z}$ and let K_c be an associated fundamental domain for the *completion* of the continued fraction over $\mathrm{Stab}_{\mathcal{M}}(\infty)$ so that $K = \bigcup_{r\in\mathcal{R}} rK_c$ up to a set of measure zero. Let T be the shift map on K that acts by ι and then an element of $\mathcal{Z}$. Let T_c be the shift map on K_c that acts by ι and then an element of $\mathrm{Stab}_{\mathcal{M}}(\infty)$.

LEMMA 6.15. *With the notation of the paragraph directly above, the map T on K is isomorphic to a skew-product $T_c \rtimes f$ on $K_c \times \mathcal{R}$ over the map T_c on K_c.*

Proof. There is an obvious isomorphism between $K_c \times \mathcal{R}$ and K given by $(z, r) \leftrightarrow rz$. The map T acts on rz by $a\iota$ for some $a \in \mathcal{Z}$. By Lemma 6.14, there exists $a' \in \mathcal{Z}$, $r' \in R$ such that $T(rz) = r'a'\iota(z)$. Let r'' be such that $r''a'\iota(z) \in K_c$, so that T can be considered as acting on the space $K_c \times R$ by

$$(z, r) \mapsto (r''a'\iota z, r'r''^{-1}).$$

Since $r''a' \in \mathrm{Stab}_{\mathcal{M}}(\infty)$, this maps (z, r) to $T_c(z)$ in the first coordinate. Let $f(z, r) = r'r''^{-1}$, so that $T = T_c \rtimes f$. To show that $T_c \rtimes f$ is truly a skew-product and finish the proof, we must show that for almost all fixed z, $f(z, \cdot)$ is an injection (and hence a bijection).

Suppose that $f(z, \cdot)$ is not an injection, so that $r_1 \neq r_2$ but $f(z, r_1) = f(z, r_2)$. This implies that $T(r_1 z) = T(r_2 z)$. Let $a_1, a_2 \in \mathcal{Z}$ be such that T acts by $a_1\iota$ on $r_1 z$ and acts by $a_2\iota$ on $r_2 z$. Then $a_1\iota r_1\iota^{-1}(\iota z) = a_2\iota r_2\iota^{-1}(\iota z)$. However, for almost all z (namely, those z not belonging to the exceptional set E (5.7)), $a_1\iota r_1\iota^{-1}$ is the unique element of $\mathrm{Stab}_{\mathcal{M}}(\infty)$ that brings ιz to K. Thus, for such z, $a_1\iota r_1\iota^{-1} = a_2\iota r_2\iota^{-1}$. Recall from the proof of the previous lemma that $\iota r_1\iota^{-1}, \iota r_2\iota^{-1} \in \mathcal{R}$. So by the uniqueness of the decomposition, we have that $\iota r_1\iota^{-1} = \iota r_2\iota^{-1}$, and hence $r_1 = r_2$. So $f(z, \cdot)$ is injective. □

Theorem 1.3 immediately follows from the next lemma.

LEMMA 6.16. *Let A be any ergodic component of K with positive measure, then the measure of A must be at least $1/|\mathcal{R}|$ (all with respect to a normalized Lebesgue measure on K).*

Proof. We may consider A as a positive measure subset of $K_c \times \mathcal{R}$ invariant under the skew-product $T_c \rtimes f$ defined in the previous lemma. Consider also the standard projection onto the first coordinate: $\pi_K : K_c \times R \to K_c$. Since T_c is the shift map associated with a discrete, proper, and *complete* Iwasawa CF expansion, it will be ergodic due to Theorem 1.2, and thus it suffices to prove that $\pi_K(A)$ is a T_c-invariant set, since it must have full measure on K_c (that is, $1/|\mathcal{R}|$).

Suppose $z \in \pi_K(A)$, so that there exists $r \in \mathcal{R}$ such that $(z, r) \in A$. Let $z' \in T_c^{-1}z$. Then, since $T_c \rtimes f$ is a skew-product, there exists (for almost all such z) $r' \in \mathcal{R}$ such that $(T_c \rtimes f)(z', r') = (z, r)$. Thus $(z', r') \in (T_c \rtimes f)^{-1}A = A$, so $z' \in \pi_K(A)$. Thus $T_c^{-1}\pi_K(A)$ is (up to measure zero), a subset of $\pi_K(A)$.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Now suppose $z \in \pi_K(A)$ and again let $r \in \mathcal{R}$ be such that $(z, r) \in A = T^{-1}A$. Thus $T(z, r) \in A$, and projecting this into the first coordinate, we see that $T_c z \in \pi_K(A)$. Thus $\pi_K(A) \subset T_c^{-1}\pi_K(A)$. This proves the two sets are equal up to measure zero, as desired. □

In certain cases, one can show that the skew-product over an ergodic transformation is itself ergodic, see [64] and related papers of the second author for some interesting examples. If we could prove such a result here, we could remove the completeness condition in the case of centrally symmetric systems.

6.5. *Application: tail equivalence.* In this section, we prove Theorem 1.4 in the following more precise formulation (note that markable geodesics are generic by Corollary 6.6).

THEOREM 6.17. (Tail equivalence of markable geodesics) *Let γ be a markable geodesic and $\gamma' = M\gamma$ with $M \in \mathcal{M}$ and $\gamma'_+ \in K$. If a_i, a'_i are the sequence of CF digits of γ_+ and γ'_+ respectively, then they have the same tail—that is, there exist some $k, k' \in \mathbb{N}$ such that $a_{k+i} = a'_{k'+i}$ for all $i \geq 1$.*

Remark 6.18. We note that the condition $\gamma'_+ \in K$ is not necessary. If it were not there, we could define $a'_0 = [\gamma'_+]$ and let the continued fraction expansion of γ'_+ start with this a'_0; however, since this a'_0 might be confused with the corresponding digit of the marking, we will not use it here.

Proof. While γ' is a markable geodesic, it may or may not pass through $C_{\mathbb{W}}$.

The result follows immediately from Theorem 5.1 if γ' *does* pass through $C_{\mathbb{W}}$: the cusp detection property gives us that for some j, $M = M_{i_j}^{-1}$. So the marking of γ' is a shift of the marking of γ. If $j \geq 0$, then $a'_i = a_{i_j+i}$ for $i \geq 1$, and if $j < 0$, then $a'_{-i_j+i} = a_i$ for $i \geq 1$.

We now assume that γ' does not pass through $C_{\mathbb{W}}$. If $|\gamma'_-| \geq 1 + \epsilon$, with ϵ as in Lemma 5.3, then we apply Lemma 5.2 to see that γ' intersects $\mathbb{W}$. Let $\gamma''(t) = \gamma'(t + t')$ be such that $\gamma''(0) \in \mathbb{W}$. However, if $|\gamma'_-| < 1 + \epsilon$, then we may apply the proof of Lemma 5.4 to γ' to find an index i_1 and corresponding time t_1 such that $M_{\mathrm{i}_1}^{-1}\gamma'(\mathrm{t}_1) \in \mathbb{W}$. (Note that the condition in the lemma that $\gamma(0) \in \mathbb{W}$ is not actually used in the proof, only that $|\gamma(0)| < 1 + \epsilon$. Moreover, since γ' is markable, we know that $\gamma'_+ \notin \mathcal{M}\infty$.) In this case, let $\gamma''(t) = M_{\mathrm{i}_1}^{-1}\gamma'(t + \mathrm{t}_1)$, so that once again $\gamma''(0) \in \mathbb{W}$.

We claim that γ'_+ and γ''_+ are tail-equivalent. This is obvious in the first case, since $\gamma'_+ = \gamma''_+$. In the second case, they are still tail-equivalent, since $\gamma''_+ = T^{\mathrm{i}_1}\gamma'_+$ and T again acts via a shift of the digits. Moreover, γ'' is still a markable geodesic, since this property is $\mathcal{M}$-invariant.

By applying the idea of the proof of Lemma 5.11, we have that γ'' intersects $M_{\mathrm{i}_j}C_{\mathbb{W}}$ at time t_j for some j. In particular, if we let $\gamma'''(t) = M_{\mathrm{i}_j}^{-1}\gamma''(t + \mathrm{t}_j)$, then by the same argument as previously, we see that γ'''_+ is tail-equivalent to γ''_+ and hence to γ'_+. In addition, γ''' now passes through $C_{\mathbb{W}}$ so our earlier argument applies and we see that γ'''_+ is tail-equivalent to γ_+, as desired. □

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press

Acknowledgements. A.L. was supported by University of Michigan NSF RTG grant 1045119. This article was written during visits by the authors to University of Texas at Tyler, George Mason University, University of Michigan, and the Ohio State University. The authors thank these institutions for their hospitality, and Simons Travel Grant and GEAR Grant NSF DMS 11-07452 for the travel funding. The authors would also like to thank Jayadev Athreya and Ralf Spatzier for their helpful comments, and the anonymous referee for the extensive suggestions that improved the exposition of this paper.

REFERENCES

[1] R. L. Adler and L. Flatto. The backward continued fraction map and geodesic flow. *Ergod. Th. & Dynam. Sys.* **4**(4) (1984), 487–492.
[2] P. Arnoux and T. A. Schmidt. Cross sections for geodesic flows and α-continued fractions. *Nonlinearity* **26**(3) (2013), 711–726.
[3] E. Artin. Ein mechanisches system mit quasiergodischen bahnen [A mechanical system with quasiergodic orbits]. *Abh. Math. Semin. Univ. Hambg.* **3**(1) (1924), 170–175.
[4] M. Bauer and A. Lopes. A billiard in the hyperbolic plane with decay of correlation of type n^{-2}. *Discrete Contin. Dyn. Syst.* **3**(1) (1997), 107–116.
[5] F. P. Boca and C. Merriman. Coding of geodesics on some modular surfaces and applications to odd and even continued fractions. *Indag. Math. (N.S.)* **29**(5) (2018), 1214–1234.
[6] R. Burton, C. Kraaikamp and T. Schmidt. Natural extensions for the Rosen fractions. *Trans. Amer. Math. Soc.* **352**(3) (2000), 1277–1298.
[7] K. Calta and T. A. Schmidt. Continued fractions for a class of triangle groups. *J. Aust. Math. Soc.* **93**(1–2) (2012), 21–42.
[8] W. Cao and J. R. Parker. Shimizu's lemma for quaternionic hyperbolic space. *Comput. Methods Funct. Theory* **18**(1) (2018), 159–191.
[9] V. Chousionis, J. Tyson and M. Urbański. Conformal graph directed Markov systems on Carnot groups. *Mem. Amer. Math. Soc.* **266**(1291) (2020), viii+155.
[10] B. Cijsouw. Complex continued fraction algorithms. *Master's Thesis*, Radboud University, 2015.
[11] J. H. Conway and D. A. Smith. *On Quaternions and Octonions: Their Geometry, Arithmetic, and Symmetry*. A K Peters, Ltd., Natick, MA, 2003.
[12] M. Cowling, A. H. Dooley, A. Korányi and F. Ricci. H-type groups and Iwasawa decompositions. *Adv. Math.* **87**(1) (1991), 1–41.
[13] K. Dajani, D. Hensley, C. Kraaikamp and V. Masarotto. Arithmetic and ergodic properties of 'flipped' continued fraction algorithms. *Acta Arith.* **153**(1) (2012), 51–79.
[14] K. Dajani and C. Kraaikamp. *Ergodic Theory of Numbers (Carus Mathematical Monographs, 29)*. American Mathematical Society, Providence, RI, 2002.
[15] K. Dajani, C. Kraaikamp and W. Steiner. Metrical theory for α-Rosen fractions. *J. Eur. Math. Soc. (JEMS)* **11**(6) (2009), 1259–1283.
[16] K. Dajani, C. Kraaikamp and N. van der Wekken. Ergodicity of N-continued fraction expansions. *J. Number Theory* **133**(9) (2013), 3183–3204.
[17] S. G. Dani. Continued fraction expansions for complex numbers—a general approach. *Acta Arith.* **171**(4) (2015), 355–369.
[18] H. Ei, S. Ito, H. Nakada and R. Natsui. On the construction of the natural extension of the Hurwitz complex continued fraction map. *Monatsh. Math.* **188**(1) (2019), 37–86.
[19] M. Einsiedler and T. Ward. *Ergodic Theory with a View Towards Number Theory (Graduate Texts in Mathematics, 259)*. Springer-Verlag, London, 2011.
[20] K. Gröchenig and A. Haas. Backward continued fractions, Hecke groups and invariant measures for transformations of the interval. *Ergod. Th. & Dynam. Sys.* **16**(6) (1996), 1241–1274.
[21] W. M. Goldman. *Complex Hyperbolic Geometry (Oxford Mathematical Monographs)*. Oxford Science Publications; The Clarendon Press; Oxford University Press, New York, 1999.
[22] W. R. Hamilton. On continued fractions in quaternions. *Philos. Mag.* **III** (1852), 371–373; **IV** (1852), 303; **V** (1853), 117–118, 236–238, 321–326.
[23] W. R. Hamilton. On the connexion of quaternions with continued fractions and quadratic equations. *Proc. R. Ir. Acad.* **5**(219–221) (1853), 299–301.

[24] G. P. Hayward. The action of the Picard group on hyperbolic 3-space and complex continued fractions. *PhD Thesis*, University of the Witwatersrand, Faculty of Science, School of Mathematics, 2014.
[25] G. A. Hedlund. A metrically transitive group defined by the modular groups. *Amer. J. Math.* **57**(3) (1935), 668–678.
[26] D. Hensley. *Continued Fractions*. World Scientific, Hackensack, NJ, 2006.
[27] S. Hersonsky and F. Paulin. Diophantine approximation for negatively curved manifolds. *Math. Z.* **241**(1) (2002), 181–226.
[28] G. Hiary and J. Vandehey. Calculations of the invariant measure for Hurwitz continued fractions. *Exp. Math.* doi: 10.1080/10586458.2019.1627255. Published online 26 June 2019.
[29] R. Hines. Badly approximable numbers over imaginary quadratic fields. *Acta Arith.* **190**(2) (2019), 101–125.
[30] A. Hurwitz. Uber die Entwicklung complexer Grössen in Kettenbrüche [On the expansion of complex quantities in continued fractions]. *Acta Math.* **11**(1–4) (1887), 187–200.
[31] A. Hurwitz. Uber eine besondere Art der Kettenbruch-Entwicklung reeller Grössen [On a special type of continued fraction expansions of real numbers]. *Acta Math.* **12**(1) (1889), 367–405.
[32] S. Ito and M. Yuri. Number theoretical transformations with finite range structure and their ergodic properties. *Tokyo J. Math.* **10** (1987), 1–32.
[33] S. Katok and I. Ugarcovici. Arithmetic coding of geodesics on the modular surface via continued fractions. *European Women in Mathematics—Marseille 2003 (CWI Tract, 135)*. Eds. K. Dajani and J. von Reis. Centrum Wiskunde & Informatica, Amsterdam, 2005, pp. 59–77.
[34] S. Katok and I. Ugarcovici. Symbolic dynamics for the modular surface and beyond. *Bull. Amer. Math. Soc. (N.S.)* **44**(1) (2007), 87–132.
[35] S. Katok and I. Ugarcovici. Applications of (a, b)-continued fraction transformations. *Ergod. Th. & Dynam. Sys.* **32**(2) (2012), 755–777.
[36] R. B. Lakein. Continued fractions and equivalent complex numbers. *Proc. Amer. Math. Soc.* **42** (1974), 641–642.
[37] E. Le Donne. A metric characterization of Carnot groups. *Proc. Amer. Math. Soc.* **143**(2) (2015), 845–849.
[38] A. Lukyanenko and J. Vandehey. Continued fractions on the Heisenberg group. *Acta Arith.* **167**(1) (2015), 19–42.
[39] A. Lukyanenko and J. Vandehey. Intrinsic diophantine approximation in Carnot groups and in the Siegel model of the Heisenberg group. *Monatsh. Math.* **192**(3) (2020), 651–676.
[40] S. Marmi, P. Moussa and J.-C. Yoccoz. The Brjuno functions and their regularity properties. *Comm. Math. Phys.* **186**(2) (1997), 265–293.
[41] V. Masarotto. Metric and arithmetic properties of a new class of continued fraction expansions. *Master's Thesis*, Universita di Padova and Leiden University, 2009.
[42] F. I. Mautner. Geodesic flows on symmetric Riemann spaces. *Ann. of Math. (2)* **65** (1957), 416–431.
[43] D. Mayer and F. Strömberg. Symbolic dynamics for the geodesic flow on Hecke surfaces. *J. Mod. Dyn.* **2**(4) (2008), 581–627.
[44] C. M. Mennen. The algebra and geometry of continued fractions with integer quaternion coefficients. *PhD Thesis*, University of the Witwatersrand, 2015.
[45] H. Nakada. On the Kuzmin's theorem for the complex continued fractions. *Keio Engrg. Rep.* **29**(9) (1976), 93–108.
[46] H. Nakada. Metrical theory for a class of continued fraction transformations and their natural extensions. *Tokyo J. Math.* **4**(2) (1981), 399–426.
[47] H. Nakada. On ergodic theory of A. Schmidt's complex continued fractions over Gaussian field. *Monatsh. Math.* **105**(2) (1988), 131–150.
[48] H. Nakada and R. Natsui. On the equivalence relations of α-continued fractions. *Indag. Math. (N.S.)* **25**(4) (2014), 800–815.
[49] H. Nakada and W. Steiner. On the ergodic theory of Tanaka–Ito type α-continued fractions. *Tokyo J. Math.* **1**(1) (2021), 1–15.
[50] G. Panti. Slow continued fractions, transducers, and the Serret theorem. *J. Number Theory* **185** (2018), 121–143.
[51] J. R. Parker. Shimizu's lemma for complex hyperbolic space. *Internat. J. Math.* **3**(2) (1992), 291–308.
[52] M. Pollicott. The Picard group, closed geodesics and zeta functions. *Trans. Amer. Math. Soc.* **344**(2) (1994), 857–872.
[53] A. L. Schmidt. Diophantine approximation of complex numbers. *Acta Math.* **134**(1) (1975), 1–85.
[54] F. Schweiger. *Ergodic Theory of Fibred Systems and Metric Number Theory*. Oxford Science Publications; The Clarendon Press; Oxford University Press, New York, 1995.
[55] F. Schweiger. *Multidimensional Continued Fractions*. Oxford University Press, New York, 2000.
[56] F. Schweiger and M. Waterman. Some remarks on Kuzmin's theorem for F-expansions. *J. Number Theory* **5**(2) (1973), 123–131.

[57] C. Series. The modular surface and continued fractions. *J. Lond. Math. Soc. (2)* **31**(1) (1985), 69–80.
[58] I. Shiokawa, R. Kaneiwa and J. Tamura. A proof of Perron's theorem on Diophantine approximation of complex numbers. *Keio Engrg. Rep.* **28**(12) (1975), 131–147.
[59] I. Śleszyński. Supplement to the Note on the Convergence of Continued Fractions. *Mathematical Collection* **14**(3) (1889), 436–438 (in Russian).
[60] S. Tanaka. A complex continued fraction transformation and its ergodic properties. *Tokyo J. Math.* **8**(1) (1985), 191–214.
[61] S. Tanaka and S. Ito. On a family of continued-fraction transformations and their ergodic properties. *Tokyo J. Math.* **4**(1) (1981), 153–175.
[62] W. Thurston. *Geometry and Topology of Three-Manifolds*. Princeton University Press, Princeton, NJ, 1980.
[63] J. Vandehey. Lagrange's theorem for continued fractions on the Heisenberg group. *Bull. Lond. Math. Soc.* **47**(5) (2015), 866–882.
[64] J. Vandehey. Non-trivial matrix actions preserve normality for continued fractions. *Compos. Math.* **153**(2) (2017), 274–293.
[65] R. J. Zimmer. *Ergodic Theory and Semisimple Groups (Monographs in Mathematics, 81)*. Birkhäuser Verlag, Basel, 1984.

https://doi.org/10.1017/etds.2022.18 Published online by Cambridge University Press