Topics in Conformal Geometry and Dynamics

Edward Thomas Crane

Trinity College.

This dissertation is submitted for the degree
of Doctor of Philosophy

University of Cambridge

December 2003

For Jacqueline

Acknowledgements

I would like to thank my supervisor, Keith Carne, for all the careful attention and encouragement he gave me, listening patiently while I explained all sorts of unpromising ideas and pointing me towards more fruitful ones. His confidence in me was a great help.

Matthew Fayers put a great deal of care into proofreading an early draft of this thesis, eliminating hundreds of grammatical and typographical errors and helping me to improve the style and presentation. Ben Green asked me a question that led me to think about the question answered by Chapter 4. I have had many useful mathematical conversations with Dinesh Markose and Alan Beardon, among others. This thesis was typeset using Latex, with the AMS Latex packages and Paul Taylor's diagrams package.

I would like to thank the Engineering and Physical Sciences Research Council for the research studentship which funded this PhD research.

Trinity College has been a wonderful home and has given me a first rate mathematical education over the course of seven years. I also wish to to thank Trinity for financial assistance in the form of a research scholarship.

Most importantly, I would like to thank Jacqui for her love and support while I have been doing my PhD. I could not have done it without her!

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. This thesis is not substantially the same as any that has been submitted, or is concurrently being submitted, for any other degree, diploma, or other qualification.

Edward Crane

Contents

List of notation

$\mathbb{N}$ .... the natural numbers, which begin at 1.

$\mathbb{C}$ .... the complex numbers.

$\widehat{\mathbb{C}} = \mathbb{P}^1(\mathbb{C})$ .... the Riemann sphere.

$\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$ ..... the open unit disc in $\mathbb{C}$.

$\overline{\mathbb{D}} = \{z \in \mathbb{C} : |z| \leq 1\}$ .... the closed unit disc in $\mathbb{C}$.

$d_{\text{hyp}}(z, w)$ .... the distance between z and w in the natural complete hyperbolic metric on a hyperbolic Riemann surface.

$\mathcal{B}$ .... the Borel σ-algebra.

$\mathcal{B}(X)$ .... the space of real Borel-measurable functions on X.

$\mathcal{B}(X, Y)$ .... the set of Borel measurable functions from X to Y.

$C(X, Y)$ .... the space of continuous maps from X to Y, with the compact-open topology.

$C(X)$ .... the Banach space of bounded continuous real-valued functions on a topological space X, with the supremum norm.

$C_0(X)$ .... the closed subspace of functions in $C(X)$ that vanish at infinity.

$\mathcal{M}(X)$ .... the space of signed tight (regular) Borel measures on the topological space X.

$\mathcal{M}^+(X)$ .... the subspace of $\mathcal{M}(X)$ consisting of positive measures.

$\mathcal{P}(X)$ .... the space of tight Borel probability measures on X.

$\Gamma(n)$ .... The subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ represented by matrices congruent to the identity modulo n.

$\Gamma_0(n)$ .... The subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ whose elements are represented by matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where $c \equiv 0 \pmod{n}$.

$\Gamma_0(m,n)$ .... The subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ whose elements are represented by matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where $c \equiv 0 \pmod{m}$ and $b \equiv 0 \pmod{n}$.

Introduction

The chapters of this thesis concern themselves with various aspects of a single theme: conformal mappings between Riemann surfaces have strongly controlled geometry and this has interesting dynamical consequences when we compose sequences of conformal mappings.

Chapter 1 covers various technical prerequisites. It is intended to make the thesis reasonably self-contained. The reader is likely to be familiar with some of this material.

In Chapter 2 we study a particular class of *iterated function systems.* An iterated function system is a kind of random dynamical system. In an ordinary discrete-time dynamical system, we have a state space X and a map $f : X \to X$ which we apply repeatedly (iterate). We are typically interested in the long-term statistical behaviour of an orbit

$$x_0, \quad x_1 = f(x_0), \quad x_2 = f(x_1), \quad x_3 = f(x_2), \ \dots \ .$$

An iterated function system differs in that the map F_n that must be applied to step from x_{n-1} to x_n is not fixed, but is a random variable; the maps at different times are chosen independently, all from the same probability distribution. Now the orbit is random; it forms a Markov chain. We recommend two readable and up-to-date introductions to the subject [27, 69].

We were motivated by [3], which studied iterated function systems in which the maps F_n are drawn from a finite set of analytic self-maps of the open unit disc in the complex plane. In that paper a stability result for the corresponding Markov Chain was proved using the Schwarz–Pick Lemma, a contractivity property of such maps. We improve the sufficient condition for stability in this situation. We then define the class of *non-uniformly contracting* iterated function systems by abstracting the contractivity property; as its author remarked, the method of [3] applies to give a stability theorem for this class. We give an alternative proof, formalising a coupling argument by the use of probability metrics. We then prove the continuous dependence of the stationary distribution on the maps and probabilities. After proving a weak law of large numbers for this class, we conclude by giving a necessary and sufficient condition for stability in terms of the *reverse iterates*. These are the maps obtained by composing the random maps F_n in reverse order.

In chapter 3, we study another type of random dynamical system that belongs to complex dynamics. This is the iteration of multivalued algebraic functions, or *holomorphic correspondences*. This is a relatively new subject, but already a number of different aspects have been explored in the literature. The fact that the setting is a surface with a moduli space should lead to an interaction between dynamical systems and arithmetic algebraic geometry. We study some ergodic properties of the iteration of the critically finite correspondences defined by Bullett [19]. These correspondences are actually arithmetic algebraic objects, also called modular correspondences. A stability result for critically finite correspondences that we hoped was new has recently appeared elsewhere [25]; we present our proof anyway, as is somewhat different. We then use some ideas from Chapter 2 to prove a stability result for a larger class of correspondences which are analogous to critically finite rational maps. As far as we know this result is new. Finally we study some concrete examples of correspondences. These are correspondences on certain elliptic curves with complex multiplication; they are rigid in that the topology of the correspondence determines the analytic isomorphism class of the elliptic curve. Each of these correspondences gives rise to a natural family of random dynamical systems parameterised by the underlying elliptic curve. We also use these correspondences to construct examples for the second stability result mentioned above.

Chapter 4 continues the first part of our theme, namely the restrictions on the metric geometry of a map imposed by conformality. We study how polynomial mappings in one complex variable distort Euclidean areas of sets. Using classical potential theory we prove that if p is monic and K is a Lebesgue measurable subset of $\mathbb{C}$, then

$$\left(\frac{\mathrm{Area}\,(p^{-1}(K))}{\pi}\right)^{\deg p} \leq \frac{\mathrm{Area}(K)}{\pi}\,.$$

In chapter 5 we study another question concerning the Euclidean geometry of polynomials in one complex variable. This is Smale's Mean Value Conjecture, a well-known conjecture which sets bounds on where a polynomial p maps any point $z \in \mathbb{C}$ in terms of the locations of the critical points ζ_i of p and their images under p. It states that if $p'(z) \neq 0$ then

$$\min_i \left|\frac{p(z) - p(\zeta_i)}{(z - \zeta_i)\,p'(z)}\right| \leq 1\,.$$

The result is known with 1 replaced by 4 on the right-hand side, or better constants in terms of the degree of p. Our result on the Conjecture itself is a reduction: we show that it suffices to give a proof for pairs (p, z) for which the values in the minimum on the left-hand side are all equal. Unfortunately we have not yet been able to make any use of this reduction to improve the constant. We then proceed to formulate a Mean Value Conjecture for rational maps, and prove it with weaker constant. Finally we study a special case of these conjectures, that in which all the critical points are fixed.

Chapter 1

Preliminaries

In this chapter we collect many of the definitions and technical results that we will need. None of the material presented in this chapter is new, with the possible exception of Lemma 1.8 and Lemma 1.10, which we are not aware of in the literature.

1.1 Analysis

1.1.1 Proper metric spaces

A metric space (X, d) is *proper* if every closed ball of any positive radius is compact, which happens if and only if for every $x \in X$ the distance function $d(x, \cdot)$ is a proper map. For example, a Riemannian manifold is proper if and only if it is complete, by the Hopf–Rinow Theorem. Proper metric spaces are the appropriate setting for most of the new results in Chapter 2. Every proper metric space is complete and locally compact and has a countable compact exhaustion so is separable. These properties allow us to use machinery from functional analysis, and ensure various desirable properties of probability metrics. Several of our proofs will make explicit use of the compactness of closed balls. Throughout Chapter 2 (X, d) stands for a proper and non-empty metric space. In contrast Y stands for a topological space, subject to conditions that vary from paragraph to paragraph.

1.1.2 Tight measures and the weak topology

Let Y be a topological space and μ be a finite Borel measure μ on Y. μ is *regular* if for every Borel set $A \subset X$,

$$\mu(A) = \sup\{\mu(K) : K \text{ compact}, K \subset A\}.$$

μ is *tight* if this equation holds when $A = X$. If μ is regular then for every Borel $A \subset X$ we have

$$\mu(A) = \inf\{\mu(V) : V \text{ open}, V \supset A\}.$$

When Y is metrizable, μ is tight if and only if μ is regular [29, Theorem 7.1.3]. A *law* on Y is a Borel probability measure. $\mathcal{P}(Y)$ denotes the set of regular laws on Y.

A *Polish space* is a topological space which can be metrized to be separable and complete.

Theorem 1.1 (Ulam). *[29, Theorem 7.1.4]*
Let Y be a Polish space. Then every finite Borel measure on Y is tight.

A topological space is called *universally measurable* if every law on it is tight. As a consequence of Ulam's theorem, every proper metric space is universally measurable. In fact a topological space is universally measurable if it is Borel-measurably isomorphic to a Borel subset of a Polish space [29, 57].

Let Y be a locally compact non-empty Hausdorff space. Let $C(Y)$ denote the Banach space of bounded continuous functions on Y, with the supremum norm. $C_0(Y)$ is the closed subspace of $C(Y)$ consisting of functions f that *vanish at infinity*, i. e. for any $\epsilon > 0$ there exists a compact set K outside which $|f| < \epsilon$. Let $\mathcal{M}(Y)$ denote the space of regular finite signed Borel measures on Y, a Banach space with the total variation norm.

Theorem 1.2 (Riesz Representation Theorem).
For any locally compact non-empty Hausdorff space Y, $\mathcal{M}(Y)$ is the dual of $C_0(Y)$ with respect to the integration pairing.

This statement may be obtained by considering complex conjugates in [63, Theorem 6.19]. We will use the *weak* topology on $\mathcal{M}(Y)$ that arises from this representation. Convergence in norm implies but is not implied by weak convergence. The weak topology on $\mathcal{M}(Y)$ is Hausdorff and only depends on the topology of Y, not on the metric.

A finite signed Borel measure integrates all bounded continuous functions, but if Y is not compact, there are bounded linear functionals on $C(Y)$ which are not represented by integration against a measure (take a Banach limit of point evaluations along a sequence that tends to infinity).

Lemma 1.3.
Let Y be a locally compact non-empty Hausdorff space and μ_n a tight law for each $n \geq 0$. Then $\mu_n \to \mu_0$ weakly against $C_0(Y)$ if and only if $\mu_n \to \mu_0$ weakly against $C(Y)$.

The proof of Lemma 1.3 uses local compactness via Urysohn's Lemma. If Y is not compact then a sequence of probability measures may converge weakly to a positive measure of mass less than 1, perhaps even 0; in that case we do not have weak convergence against $C(Y)$.

A subset $S \subset \mathcal{P}(Y)$ is *uniformly tight* if for each $\epsilon > 0$ there exists a compact set $J \subset Y$ such that $\nu(J) > 1 - \epsilon$ for every $\nu \in S$.

Lemma 1.4 (Le Cam). *[29, Theorem 11.5.3]*
Let (Y, d) be a locally compact metric space, and $mu_n \in \mathcal{P}(Y)$ for $n \geq 0$. If $\mu_n \to \mu_0$, then the set $\{\mu, \mu_1, \mu_2, \dots\}$ is uniformly tight.

Lemma 1.5 (Prohorov). *[29, Theorem 11.5.4]*
Let (Y, d) be a Polish space and $S \subset \mathcal{P}(Y)$. Then S is relatively compact if and only if it is uniformly tight.

1.2 Kantorovich metrics

Let (Y, d) be a separable metric space. Define the subspace $\mathcal{P}_d \subset \mathcal{P}(X)$ by

$$\mathcal{P}_d = \left\{ \mu \in \mathcal{P}(X) : \int d\left(y_0, y\right) \, d\mu(y) \; < \; \infty \right\} .$$

$\mathcal{P}_d$ is independent of the choice of $y_0 \in Y$. For any $\nu_1, \nu_2 \in \mathcal{P}_d$, let $\mathcal{P}^{\nu_1,\nu_2}$ stand for the space of laws on $Y \times Y$ with marginals ν_1 and ν_2. The *Kantorovich distance* $K_d\left(\nu_1, \nu_2\right)$ is defined by

$$K_d\left(\nu_1, \nu_2\right) \; = \; \inf \int d(x, y) \, dm(x, y) \;\; : m \in \mathcal{P}^{\nu_1,\nu_2} \qquad (1.1)$$

We write $\mathrm{Lip}(Y)$ for the space of Lipschitz functions $f : Y \to \mathbb{R}$, with seminorm $\|f\|_L$, the best Lipschitz constant of f. The *Wasserstein distance* is defined by

$$\gamma\left(\nu_1,\nu_2\right)=\sup\left\{\int f(y)d\left(\nu_1-\nu_2\right)(y)\ :\ \|f\|_L\leq 1\right\}.$$

Theorem 1.6 (Properties of K_d).

- $K_d\left(\nu_1,\nu_2\right)=\gamma\left(\nu_1,\nu_2\right)$.

- *The infimum in equation 1.1 is attained.*

- K_d *is a metric on* $\mathcal{P}_d(Y)$.

- *The following are equivalent:*

  1. $K_d\left(\mu_n,\mu\right)\to 0$ *as* $n\to\infty$*;*

  2. $\mu_n\to\mu$ *weakly* and *for some (and hence any) fixed* $x_0\in X$*, we have*
  $$\int d\left(x,x_0\right)d\mu_n(x)\to\int d\left(x,x_0\right)d\mu(x)\,;$$

  3. $\mu_n\to\mu$ *weakly* and *for some (and hence any) fixed* $x_0\in X$*, we have*
  $$\lim_{R\to\infty}\sup_n\int d\left(x,x_0\right)\,\mathbf{1}_{\left(d\left(x,x_0\right)>R\right)}\,d\mu_n(x)\ =\ 0\,.$$

- *A subset* $A\subset\mathcal{P}_d$ *is relatively compact with respect to the metric* K_d *if and only if it is relatively compact in the weak topology and*
$$\lim_{R\to\infty}\sup_{m\in A}\int d\left(x,x_0\right)\,\mathbf{1}_{\left(d\left(x,x_0\right)>R\right)}\,dm(x)\ =\ 0\,.$$

- *If* (Y,d) *is complete, then* $\left(\mathcal{P}_d,K_d\right)$ *is also complete.*

The first three facts are proved in [29, §11.8]. The remaining results are special cases of [61, Theorems 6.2.1, 6.3.1, 6.3.2 & 6.3.3]. Note that if (X, d) has finite diameter then the integral conditions in the convergence criteria are redundant, so K_d metrizes the weak topology on $\mathcal{P}(X)$.

Lemma 1.7.
Suppose that (X, c) is a proper metric space. Then every ball of finite radius in the metric space $(\mathcal{P}_c, K_c)$ is uniformly tight.

Proof. Consider the ball $B(\mu, R)$. Given any $\epsilon > 0$, there exists a compact set $K \subset X$ with $\mu(K) > 1 - \epsilon$. K is contained in some c-ball $B(x_0, r) \subset X$. The ball $K' = B(x_0, r + R/\epsilon)$ is compact because (X, c) is proper. If a probability measure ν has $\nu(X \setminus K') > 2\epsilon$ then $K_c(\mu, \nu) > R$. □

Suppose that we wish to study a particular probability measure that happens not to lie in $\mathcal{P}_c$. Then we change the metric on X to

$$c'(x, y) = \varphi(c(x, y)) .$$

When $\varphi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is continuous, strictly increasing and concave with $\varphi(0) = 0$, this yields a new metric space (X, c'), with the same topology as (X, c). Indeed, the balls in the c'-metric of radius less than $\sup \varphi$ are precisely the balls of finite radius in the c-metric. If c is unbounded but φ is bounded then $\mathcal{P}_{c'} = \mathcal{P}(X)$. However, in that case (X, c') is not proper, so Lemma 1.7 cannot be applied. On the other hand if (X, c) is proper and $\varphi(t) \to \infty$ as $t \to \infty$ then (X, c') is also proper.

Lemma 1.8.

Given a separable metric space (X, c) and finitely many probability measures $\mu_1, \ldots \mu_k$ on X, there exists a continuous, strictly increasing and concave function $\varphi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ such that $\varphi(0) = 0$, $\varphi(t) \to \infty$ as $t \to \infty$ and such that all the μ_i are in $\mathcal{P}_{c'}$, where $c'(x, y) = \varphi(c(x, y))$.

Proof. Define

$$g_i(t) = \mu_i\left(\{x \in X : c(x, x_0) \geq t\}\right).$$

These are decreasing functions such that $g_i(t) \to 0$ as $t \to \infty$. We will construct φ as a piecewise linear continuous function the vertices of whose graph are at (t_n, n) for $n \in \mathbb{N}$. We will define inductively the values t_n at which $\varphi(t_n) = n$. Note that φ is concave if and only if $(t_n - t_{n-1})$ is an increasing sequence. Set $t_0 = 0$, $t_1 = 1$ and then for $n \geq 2$ set

$$t_n = \inf\left\{t \,:\, g_i(t) \leq (n+1)^{-3} \text{ for } i = 1, \ldots, k \text{ and } t_n - t_{n-1} \geq t_{n-1} - t_{n-2}\right\}.$$

Then φ is certainly a continuous, strictly increasing, concave function with $\varphi(0) = 0$ and $\varphi(t) \to \infty$ as $t \to \infty$. Moreover, for $n \geq 2$ we have

$$\int \varphi\left(c\left(x, x_0\right)\right) \mathbf{1}\left[n \leq \varphi\left(c\left(x, x_0\right)\right) \leq n+1\right] d\mu_i(x) \leq (n+1) \cdot g_i\left(t_n\right) \leq (n+1)^{-2},$$

so

$$\int \varphi\left(c\left(x, x_0\right)\right) d\mu_i(x) < \infty,$$

i. e. $\mu_i \in \mathcal{P}_{c'}$ where $c' = \varphi(c)$. □

1.3 Prohorov metrics

The weak topology on $\mathcal{P}(X)$ is usually metrized by the Prohorov metric. Although we will not use it, we discuss it here for two reasons. We will need it in appendix B to explain the context of our results in the literature on IFSs, and the comparison with the Kantorovich metric serves to point out the advantages of the Kantorovich metric for our applications.

Let (X, c) be a separable metric space. We denote the ϵ-neighbourhood of a subset A by A^ϵ. The *Prohorov distance* π_c between two elements $\mu, \nu \in \mathcal{P}(X)$ is defined to be infimum of all $\epsilon > 0$ such that for all Borel $A \subset X$

$$\mu(A) \leq \nu\left(A^\epsilon\right) + \epsilon \quad \text{and} \quad \nu(A) \leq \mu\left(A^\epsilon\right) + \epsilon.$$

[16] gives more information about the Prohorov metric. The important properties are that π_c is a metric on $\mathcal{P}(X)$ and that it metrizes the weak topology. Moreover, if (X, c) is separable and complete then so is $(\mathcal{P}(X), \pi_c)$. The following result gives an alternative way to define π_c, close in spirit to our first definition of the Kantorovich metric K_c.

Lemma 1.9 (Strassen–Dudley Theorem).
Suppose $\pi_c(\mu, \nu) < \alpha$. Then there exists a measure $m \in \mathcal{P}^{\mu,\nu}$ such that $m(\{(x, y) : c(x, y) > \alpha\}) < \alpha$.

Note that the measure m acts as a *certificate* of the fact that $\pi_c(\mu, \nu) < \alpha$. For a proof, see [16, Theorem 6.9].

The following inequalities relate the Prohorov and Kantorovich metrics associated to a separable metric space (X, c).

Lemma 1.10.

1. $\pi_c\left(\mu_1, \mu_2\right) \leq \sqrt{K_c\left(\mu_1, \mu_2\right)}$.

2. $K_c\left(\mu_1, \mu_2\right) \leq \pi_c\left(\mu_1, \mu_2\right)\left(1 + \operatorname{diam}_c(X)\right)$.

Proof.

1. A minimizing measure on $X \times X$ for K_c is a certificate as in the Strassen–Dudley Theorem.
2. The certificate m provided by the Strassen–Dudley Theorem has Kantorovich integral at most equal to the right-hand side.

□

The analogue of Lemma 1.7 for Prohorov metrics fails.

1.4 Ergodic theory

A measure-preserving dynamical system (m. p. d. s.) is defined as a quadruple $(W, \mathcal{B}, \nu, T)$, where W is a set, $\mathcal{B}$ is a σ-algebra of subsets of W, ν is a positive measure on $\mathcal{B}$, $\mathcal{B}$ is complete with respect to ν (i.e. contains all subsets of ν-null sets), and $T : W \to W$ is a function such that $T^{-1}(\mathcal{B}) \subset \mathcal{B}$ and $T_*\nu = \nu$. We allow T to be defined only ν-a.e. We will often make use of the following result [58, p. 34].

Theorem 1.11 (Poincaré's Recurrence Theorem).
Let $(W, \mathcal{B}, \nu, T)$ be a m. p. d. s. and suppose that $\nu(A) > 0$. Then for ν-a.e. $x \in A$, the sequence $(T^n(x))_{n\in\mathbb{N}}$ returns to A infinitely often.

Let $(W_1, \mathcal{B}_1, \nu_1, T_1)$ and $(W_2, \mathcal{B}_2, \nu_2, T_2)$ be m. p. d. s. Then $\alpha : W_2 \to W_1$ is a *homomorphism* or *factor map* if

$$\begin{aligned} \alpha^{-1}(\mathcal{B}_1) &\subset \mathcal{B}_2\,, \\ \alpha_*(\nu_2) &= \nu_1\,, \\ \alpha \circ T_2 &= T_1 \circ \alpha, \qquad \nu_2\text{-a.e.} \end{aligned}$$

We say that T_2 is an *extension* of T_1 and that T_1 is a *factor* of T_2.

If there also exists a homomorphism $\beta : W_1 \to W_2$ such that $\alpha \circ \beta = \mathrm{id}$, ν_1-a.e. (or equivalently $\beta \circ \alpha = \mathrm{id}$, ν_2-a.e.) then we say α is an *isomorphism*. We say that a homomorphism or a measurable function is *essentially unique* if any two possible values of it agree a.e. with respect to the relevant measure.

Any m. p. d. s. $(W, \mathcal{B}, \nu, T)$ has an extension $(\hat{W}, \hat{\mathcal{B}}, \hat{\nu}, \hat{T})$ such that $\hat{T}$ is invertible and such that any homomorphism ψ from an invertible m. p. d. s. to $(W, \mathcal{B}, \nu, T)$ is essentially uniquely expressible as $\psi = \phi \circ \chi$. This is called the *natural extension* of $(W, \mathcal{B}, \nu, T)$. Being a universal object it is unique (up to an essentially unique isomorphism).[1] If T is ergodic with respect to ν then $\hat{T}$ is ergodic with respect to $\hat{\nu}$.

The standard construction of the natural extension is to take $\hat{W}$ to be the space of infinite backward orbits of T, i. e. sequences $(w_n)_{n=0}^{\infty}$ such that $w_n = Tw_{n+1}$ for all n, with the σ-algebra inherited from the product σ-algebra on $W^{\mathbb{N}}$. The map $\hat{T}$ is given by T acting co-ordinatewise, and $\hat{T}^{-1}$ is the left-shift. There is a unique measure $\hat{\nu}$ such that for each co-ordinate projection map $\pi_m : \hat{W} \to W$, we have $(\pi_m)_* \hat{\nu} = \nu$. Indeed, a compactness argument on trees of pre-images shows that this condition specifies $\hat{\nu}(A_m)$ consistently for sets of the form $A_m = \{(w_n)_{n=0}^{\infty} \in \hat{W} : w_m \in A\}$; the Carathéodory Extension Theorem then gives existence and uniqueness.

[1] However, a universal object does depend on the category with respect to which it is universal! Some authors do not insist that $\mathcal{B}$ be complete with respect to ν; their natural extension may have a smaller σ-algebra than ours.

1.5 Geometric function theory

We denote by $\mathbb{C}$ the complex numbers, by $\widehat{\mathbb{C}}$ the Riemann sphere, and by $\mathbb{D}$ the open unit disc in $\mathbb{C}$.

Theorem 1.12 (Uniformisation of Riemann surfaces).
Let R be any Riemann surface. Then R is the quotient of precisely one of the Riemann surfaces $\widehat{\mathbb{C}}$, $\mathbb{C}$, and $\mathbb{D}$ by a group of automorphisms that acts freely and properly discontinuously; this group is unique up to conjugation.

The universal cover of any Riemann surface has a unique complex structure that makes the covering map analytic, so the theorem amounts to proving that every simply-connected Riemann surface is isomorphic to precisely one of $\widehat{\mathbb{C}}$, $\mathbb{C}$ and $\mathbb{D}$. The only case in which the universal cover is $\widehat{\mathbb{C}}$ is when $R = \widehat{\mathbb{C}}$. The only quotients of $\mathbb{C}$ are $\mathbb{C}$ itself, the once-punctured plane, and the tori $\mathbb{C}/\Lambda$ for lattices Λ in $\mathbb{C}$. The remaining Riemann surfaces covered by $\mathbb{D}$ are called *hyperbolic*. In particular, any domain in $\mathbb{C}$ which omits at least two points of $\mathbb{C}$ is hyperbolic, as is any Riemann surface whose fundamental group is non-abelian. The automorphisms of the disc $\mathbb{D}$ are also isometries of the Poincaré metric on $\mathbb{D}$, which is the unique complete conformal Riemannian metric on $\mathbb{D}$ of constant curvature -1. (A Riemannian metric on a Riemann surface is conformal when with respect to some (hence any) local co-ordinate z it has the form $ds^2 = \rho(z)|dz|^2$, where ρ is some smooth positive function.) It follows that any hyperbolic Riemann surface R has a unique complete conformal metric with curvature -1, which we call the hyperbolic metric on R. We denote the corresponding distance function d_R or d_{hyp} when there will be no confusion over which Riemann surface is being referred to. Let $f : R \to S$ be an analytic map between hyperbolic Riemann surfaces. The derivative $Df(z)$ is a scalar multiple of an isometry of the tangent spaces at z and $f(z)$; this scalar is written $f^{\#}(z)$ and called the hyperbolic derivative of f at z.

Lemma 1.13 (Schwarz–Pick Lemma).
Let $f : R \to S$ be an analytic map between two hyperbolic Riemann surfaces. Then f does not increase distances, i. e. for all $z, w \in R$,

$$d_S(f(z), f(w)) \leq d_R(z, w)\,.$$

If there are two distinct points z and w for which equality holds then f is a local isometry. Moreover, for each $z \in R$, $f^{\#}(z) \leq 1$, with equality at any single point if and only if f is a local isometry.

The following generalisation is apparently due to Nehari [55].

Lemma 1.14 (Branched Schwarz Lemma).
Let $f, b : \mathbb{D} \to \mathbb{D}$ be analytic, with $f(0) = 0 = b(0)$, and suppose that b is a finite Blaschke product. If the valency of f is at least the valency of b at each point of $\mathbb{D}$, then $|f'(0)| \leq |b'(0)|$. If $|b'(0)| \neq 0$ then equality holds if and only if $f \cong \lambda b$, where λ is a constant of modulus 1.

Chapter 2

Iterated function systems

2.1 Iterated function systems

2.1.1 Definitions

Let Y be a topological space with Borel σ-algebra $\mathcal{B}$. Let $C(Y, Y)$ denote the space of continuous maps from Y to Y, and $\mathcal{B}(Y, Y)$ the set of Borel-measurable maps from Y to Y.

An *Iterated Function System* (IFS) $\mathcal{F}$ consists of the following data: a topological space Y, a probability space $(\Omega, \mathcal{M}, \mathbb{P})$, and a function $f : \Omega \to \mathcal{B}(Y, Y)$. We will abuse notation by using f also to stand for the corresponding map $\Omega \times Y \to Y$. We insist that the latter map be measurable (i. e. for every $A \in \mathcal{B}$, $f^{-1}(A) \in \mathcal{M} \times \mathcal{B}$). We will think of f as a random map of Y into itself. We call the IFS $\mathcal{F}$ *continuous* when $f : \Omega \to C(Y, Y)$. In the special case in which Ω is finite, we call $\mathcal{F}$ an *N-map IFS*. Then we denote the possible values of f by $f_1, f_2, \ldots, f_N$, taken with probabilities $p_1, p_2, \ldots, p_N$

respectively.

We interpret $\mathcal{F}$ dynamically by constructing a sequence of independent random maps $(F_i)_{i=1}^{\infty}$, each map distributed as f is. We think of the map F_i being applied to Y at time i. The (random) orbit of a (possibly random) point $x_0 \in Y$ is the random sequence (x_n), where

$$x_n = F_n \circ F_{n-1} \circ \cdots \circ F_1 (x_0) \ .$$

This may also be thought of as the orbit of a discrete-time Markov chain $(x_i)_{i=0}^{\infty}$, where the probability of a transition into a Borel set $A \subset Y$ is defined by

$$\mathbb{P} (x_{i+1} \in A | x_i = y) = P(y, A) = \mathbb{P} (f(y) \in A) \ .$$

$P(y, A)$ is a measurable function of y [50, Lemma 2.1]. To fix notation, the probability space on which the sequence (F_n) is defined is the unilateral shift space

$$\Omega = \{1, 2, \dots N\}^{\mathbb{N}}$$

with σ-algebra $\mathcal{M}$ obtained by completing the product σ-algebra with respect to the Bernoulli measure $B_{\mathbf{p}}$. Then

$$F_n = f_{\omega_n} \ .$$

In order to interpret $\mathcal{F}$ in terms of dynamical systems, we define the left-shift map $\sigma : \Omega \to \Omega$ given by

$$\sigma : (\omega_1, \omega_2, \omega_3, \dots) \mapsto (\omega_2, \omega_2, \omega_3, \dots) \ .$$

and the *skew-product* map $\tau : Y \times \Omega \to Y \times \Omega$ given by

$$\tau (x, (\omega_1, \omega_2, \omega_3, \dots)) = (f_{\omega_1}(x), (\omega_2, \omega_3, \omega_4, \dots)) \ .$$

2.1.2 Linear operators associated to an IFS

Let $\mathcal{B}(Y)$ denote the space of Borel-measurable real-valued functions on the topological space Y. The composition of Borel-measurable functions is Borel-measurable, so we can associate to any Borel-measurable map $h : Y \to Y$ a pull-back operator $h^* : \mathcal{B}(Y) \to \mathcal{B}(Y)$, defined by $h^*(g) = g \circ h$. If h is continuous then h^* also acts on $C(Y)$. The push-forward of a Borel measure μ on Y by h is defined by $h_*(\mu)(E) = \mu\left(h^{-1}(E)\right)$ for each Borel set $E \subset Y$. Since $h^{-1} : \mathcal{B} \to \mathcal{B}$ is an algebra homomorphism, $h_*(\mu)$ is another Borel measure. If h is continuous then h_* takes tight measures to tight measures.

The *Perron–Frobenius* or *transfer operator* $\mathcal{F}_*$ of an IFS $\mathcal{F}$ on Y acts on $g \in \mathcal{B}(Y)$ by

$$\left(\mathcal{F}^* g\right)(y) := \mathbb{E}\left(f^*(g)(y)\right) = \int g \circ f(y)\, d\mathbb{P}(f)\,,$$

where f is the random map in the definition of $\mathcal{F}$ and $\mathbb{E}$ is the expectation with respect to $\mathbb{P}$. Since $g \circ f$ is Borel-measurable, Fubini's Theorem tells us that $\mathcal{F}^*(g) \in \mathcal{B}(Y)$. The *Markov operator* $\mathcal{F}_*$ acts on the space of all Borel measures on Y by

$$\left(\mathcal{F}_* \mu\right)(A) := \mathbb{E}\left(f_*(\mu)(A)\right) = \mathbb{E}\left(\mu\left(f^{-1}(A)\right)\right) = \int \mathbb{P}(f(x) \in A)\, d\mu(x)\,,$$

where A is any Borel subset of Y. We say that $\mathcal{F}$ has the *weak Feller property* if $\mathcal{F}^*$ acts on $C(Y)$. We will call $\mathcal{F}$ *good* if it has the weak Feller property and $\mathcal{F}_*$ acts on $\mathcal{P}(Y)$. The weak Feller property does not imply that $\mathcal{F}_*$ sends tight measures to tight measures, although if Y is universally measurable then this is automatic. Every continuous N-map IFS is good. If $\mathcal{F}$ is a good IFS then $\mathcal{F}_*$ is the adjoint of $\mathcal{F}^*$ (with respect to the duality described by theorem 1.2), i. e. for every $g \in C(Y)$,

$$\int g(y)\ d\left(\mathcal{F}_*\mu\right)(y) \;=\; \int \left(\mathcal{F}^* g\right)(y)\ d\mu(y)\,.$$

Furthermore, if $\mathcal{F}$ is good then $\mathcal{F}^*$ is a bounded operator of norm 1 on $C(Y)$. In chapter 3 we will use IFSs that are good but not continuous.

A finite Borel measure μ on Y is *invariant* for $\mathcal{F}$ when $\mathcal{F}_*\mu = \mu$. Note that μ is invariant for $\mathcal{F}$ if and only if μ is a stationary distribution for the associated Markov chain, which happens if and only if $\mu \times B_{\mathbf{p}}$ is an invariant measure for τ. When μ is invariant for $\mathcal{F}$, the Cartesian projection map $\pi : Y \times \Omega \to \Omega$ makes the m. p. d. s. $(Y \times \Omega, \mathcal{B} \times \mathcal{M}, \mu \times B_{\mathbf{p}}, \tau)$ into an extension of the m. p. d. s. $(\Omega, \mathcal{M}, B_{\mathbf{p}}, \sigma)$.

The Krylov–Bogolubov Theorem states that any continuous self-map of a non-empty compact space has an invariant law. The following standard result is the analogue for IFSs.

Lemma 2.1 (Krylov–Bogolubov for IFSs).
Let Z be a non-empty compact metric space and let $R : C(Z) \to C(Z)$ be a bounded linear operator such that $R(1) = 1$. Then the adjoint R^* acts on $\mathcal{P}(Z)$ and has a fixed point there. In particular, any good IFS on Z has an invariant law.

Proof. R^* is continuous with respect to the weak topology on $\mathcal{P}(Z)$. The Banach–Alaoglu Theorem says that the unit ball of $\mathcal{M}(Z)$ is weakly compact; $\mathcal{P}(Z)$ is a closed subset of this ball. The Schauder–Tychonov Fixed Point Theorem [64, Theorem 5.28] says that any continuous self-map of a non-empty compact convex subset of a locally convex topological vector space has a fixed point. □

It is sometimes of interest to *construct* an invariant law. Choose any $\mu_0 \in \mathcal{P}(Z)$ and for each $n \in \mathbb{N}$ define

$$\mu_n = \frac{1}{n+1} \sum_{k=0}^{n} R^{*k} \mu_0 \,.$$

Since $\mathcal{P}(Z)$ is weakly compact, there is a subsequence μ_{n_i} that converges weakly to $\mu \in \mathcal{P}(Z)$, say.

$$R^* \mu_n - \mu_n = \frac{1}{n+1} \left(R^{*n+1} \mu_0 - \mu_0 \right) \,.$$

The norm of this is at most $2/(n+1)$ so it converges weakly to 0. Thus $R^* \mu_n \to \mu$ weakly as $n \to \infty$. Since R^* is continuous, $R^* \mu_n \to R^* \mu$ as $n \to \infty$, so $R^* \mu = \mu$.

When X is not compact, $\mathcal{P}(X)$ is not weakly closed, let alone compact, and an IFS $\mathcal{F}$ on X may fail to have invariant law.

$T : \mathcal{P}(X) \to \mathcal{P}(X)$ is *asymptotically stable* if there is a T-invariant law μ such that for every $\nu \in \mathcal{P}(X)$, we have $T^n \nu \to \mu$ weakly as $n \to \infty$. An asymptotically stable operator evidently has a unique invariant law. We call an IFS $\mathcal{F}$ asymptotically stable when $\mathcal{F}_*$ is asymptotically stable, which happens if and only if for every $g \in C_0(Y)$, the sequence $\left(\mathcal{F}^*\right)^n g$ converges pointwise to a constant function.

2.2 Motivation and outline of results

2.2.1 Improving the Contraction Mapping Theorem

Let (X, d) be a complete and non-empty metric space. Suppose that $f : X \to X$ is a contraction, i. e. for some $K < 1$ and all $x, y \in X$, $d(f(x), f(y)) \leq Kd(x, y)$. The Contraction Mapping Theorem says that f has a unique fixed point x_0 which attracts all orbits, i. e. for all $y \in X$, $f^n(y) \to x_0$ as $n \to \infty$. The conclusion obviously fails if we allow $K = 1$. For a more delicate result, define $f : X \to X$ to be *strictly distance-decreasing* if for all distinct $x, y \in X$, $d(f(x), f(y)) < d(x, y)$. Such maps are sometimes known as *contractive*, but we will avoid that term as is it often used in looser senses. For strictly distance-decreasing maps we have the following analogue of the Contraction Mapping Theorem. It appears in [12]; here we give a proof to serve as a model for the proofs of later results.

Proposition 2.2.
Suppose (X, d) is a proper metric space and $f : X \to X$ is strictly distance-decreasing. Then either f has a unique fixed point which attracts all points of X, or the orbit of each point tends to infinity (and this convergence is locally uniform).

Proof. Firstly note that f is continuous. Suppose that the orbit of some point $x \in X$ does not tend to infinity, i. e. $f^{\circ n}(x)$ is in some compact set K for infinitely many n. Then the orbit has a convergent subsequence $f^{\circ n_k}(x) \to x_0$ as $k \to \infty$. Because f is continuous, $f^{\circ(1+n_k)}(x) \to f(x_0)$ as $n \to \infty$. However, the distances $d\left(f^{\circ n}(x), f^{\circ(n+1)}(x)\right)$ form a strictly decreasing sequence whose limit is $d(x_0, f(x_0))$, by continuity of f and of the distance function.

If $d\left(x_0, f\left(x_0\right)\right) = 0$ then x_0 is a fixed point. In that case, for any other point $y \in X$, $d\left(f^{\circ n}(y), x_0\right)$ is a strictly decreasing sequence. Suppose for a contradiction that it converges to $\epsilon > 0$. By the properness condition, the orbit of y is precompact so we may choose a convergent subsequence $f^{\circ m_k}(y) \to y_0$. Then $d\left(y_0, x_0\right) = \epsilon$, so $d\left(f\left(y_0\right), x_0\right) < \epsilon$, so by continuity of f, we have $d\left(f^{1+m_k}(y), x_0\right) < \epsilon$ for k sufficiently large, a contradiction. Hence the fixed point attracts all orbits.

Finally we must rule out the possibility that $d\left(x_0, f\left(x_0\right)\right) = \epsilon > 0$. If this were the case, then

$$d\left(f\left(x_0\right), f\left(f\left(x_0\right)\right)\right) < \epsilon$$

and by continuity of f we would have for sufficiently large k

$$d\left(f^{\circ(1+n_k)}(x), f^{\circ(2+n_k)}(x)\right) < \epsilon\,,$$

a contradiction.

The alternative is that for each compact set K, the orbit of x only visits K finitely many times, i. e. $f^{\circ n}(x) \to \infty$. In that case, for any compact set L, $f^{\circ n}(L) \cap L = \emptyset$ for all but finitely many n. Indeed, let R be the diameter of L and let $N_R(L)$ be the closed R-neighbourhood of L, which is compact by the properness condition. Fixing some $x \in L$, the above intersection is certainly empty when $f^{\circ n}(x) \notin N_R(L)$, because f does not increase distances. □

Example 2.1. Let (X, d) be the set of non-negative integers with the metric

$$d(m,n) = \frac{1}{2} + 2^{-\max(m,n)} \quad \text{for } m \neq n.$$

Let $f : 0 \mapsto 0$ and $f : m \mapsto m+1$ for $m > 0$. Then f is strictly distance-decreasing. The topology is discrete, so X is locally compact. The point 0 is fixed, but the orbit of 1 is not precompact. Here X is complete, separable and locally compact, with a countable compact exhaustion, but the conclusion of Proposition 2.2 fails. Properness is genuinely required.

2.2.2 The Wolff–Denjoy Theorem

Let $\mathbb{D}$ be the open unit disc in the complex plane, equipped with the usual hyperbolic metric, which is proper. Suppose that $f : \mathbb{D} \to \mathbb{D}$ is analytic but is not a conformal automorphism of $\mathbb{D}$. The Schwarz–Pick Lemma tells us that f is strictly distance-decreasing, so proposition 2.2 applies. In fact, a stronger conclusion can be obtained, using not the one-point compactification of $\mathbb{D}$, but a bigger compactification, $\overline{\mathbb{D}}$.

Theorem 2.3 (Wolff–Denjoy).
Suppose that $f : \mathbb{D} \to \mathbb{D}$ is analytic. Unless f is an automorphism with a fixed point, then there is an $\alpha \in \overline{\mathbb{D}}$ such that for all $z \in \mathbb{D}$, $f^n(z) \to \alpha$ as $n \to \infty$.

Note that the theorem does not assume that f extends continuously to $\overline{\mathbb{D}}$. A simple proof of this theorem given by Beardon [12] is reproduced in [23, §IV.3].

2.2.3 Ambroladze's theorem

Let $\mathcal{F}$ be an N-map IFS on $\mathbb{D}$ whose maps are analytic. Suppose that all the maps are automorphisms of $\mathbb{D}$. If they have a common fixed point then there are infinitely many invariant measures and each orbit stays on a circle. If they have no common fixed point then it is known that there is no invariant measure on $\mathbb{D}$. However, the automorphisms extend to Möbius maps of $\widehat{\mathbb{C}}$ and the extended IFS has at least one invariant measure on the unit circle, although it need not be asymptotically stable. Henceforth we assume that with positive probability the random map $f : \mathbb{D} \to \mathbb{D}$ is not an automorphism.

Theorem 2.4 (Ambroladze). *[3]*
Let $\mathcal{F}$ be an N-map IFS of analytic endomorphisms of $\mathbb{D}$, and suppose that at least one map, say f_1, is not an automorphism. Then there is a measure μ such that for any Borel law ν, the iterates $\mathcal{F}_*^n \nu$ converge weakly to μ. Either $\mu = 0$ and $\mathcal{F}$ has no invariant law, or μ is the unique invariant law for $\mathcal{F}$. A sufficient condition for an invariant law to exist is that $f_1(\mathbb{D})$ is relatively compact in $\mathbb{D}$.

In particular an IFS of this type is automatically asymptotically stable if it has an invariant law.

2.2.4 Outline of Results

Theorem 2.4 is a generalisation to IFSs of proposition 2.2 for the special case of analytic self-maps of $\mathbb{D}$. In §2.3 we make precise the suggestion in [3] of generalising Theorem 2.4 to more general metric spaces and maps. We introduce the class of *non-uniformly contracting* IFSs and prove Theorem 2.6, which is a generalisation to these IFSs both of Theorem 2.4 and of Proposition 2.2. Ambroladze's proof applies almost verbatim to prove Theorem 2.6; here we give an alternative proof using a Kantorovich metric, in preparation for using a similar method in the iteration of correspondences. In §2.3.2 we study some consequences for IFSs of asymptotic stability, giving a weak law of large numbers for typical orbits of $\mathcal{F}$, and for all orbits in the non-uniformly contracting case.

An alternative way to generalise theorem 2.4 would be to produce a Wolff–Denjoy Theorem for analytic IFSs on $\mathbb{D}$. Suppose that $\mathcal{F}_*^n \mu \to 0$ as $n \to \infty$, weakly against $C_0(\mathbb{D})$. Is there necessarily a law $\tilde{\mu}$ supported on the unit circle $\partial\mathbb{D}$ such that $\mathcal{F}_*^n \nu \to \tilde{\mu}$ weakly against continuous functions on $\overline{\mathbb{D}}$? We make some minor progress on this question in §2.4.1.

In §2.4.2 we improve the sufficient condition for stability in Theorem 2.4. Theorem 2.16 gives a geometric sufficient condition which will be useful in Chapter 3 because it interacts well with covering maps.

Define the *reverse iterates* of the IFS $\mathcal{F}$ to be the random maps

$$G_n = F_1 \circ F_2 \circ \cdots \circ F_n \, .$$

One of the two proofs in [3] of the sufficient condition for existence of an invariant law in Theorem 2.4 is based on the following result.

Theorem 2.5 (Letac's Principle). *[47].*
Suppose that the sequence of reverse iterates $(G_n)_{n=1}^{\infty}$ almost surely converges pointwise to a (random) constant map. Then the distribution of that constant limit is the unique invariant measure *for $\mathcal{F}$.*

We will use Letac's Principle in proving Theorem 2.16. In §2.5 we ask which asymptotically stable IFSs can be proven to be asymptotically stable using Letac's Principle. We show in Corollary 2.25 that if an N-map analytic IFS on $\mathbb{D}$ has an invariant law then the sequence of reverse iterates almost surely converge pointwise to a (random) constant limit.

2.3 Non-uniformly contracting IFSs

Definition. An N-map IFS $\mathcal{F}$ is *non-uniformly contracting* if

- Each $f_i \in \mathrm{Lip}(1)$, i. e. each f_i is non-expanding:

$$d\left(f_i(x), f_i(y)\right) \leq d(x,y) \qquad \text{for all } x, y \in X.$$

- At least one map, f_1, say, is strictly distance-decreasing, i. e.

$$d\left(f_i(x), f_i(y)\right) < d(x,y) \qquad \text{for all } x \neq y \in X.$$

  (When X is not compact, f_1 need not be Lipschitz with any constant less than 1.)

This terminology is new; we are not aware of any conflicting meanings in the literature.

Here is the generalisation of Theorem 2.4 to the setting of a non-uniformly contracting IFS.

Theorem 2.6.
Let X be a proper metric space. Let $\mathcal{F} = (f_1, \ldots, f_m : X \to X; p_1, \ldots p_m)$ be a non-uniformly contracting IFS.

1. *If $\mathcal{F}$ has an invariant law μ, then $\mathcal{F}$ is asymptotically stable. In particular, μ is the unique invariant law.*

2. *If $\mathcal{F}$ has no invariant law then for any $\nu \in \mathcal{P}(X)$, $\mathcal{F}_*^n \nu \to 0$ weakly as $n \to \infty$.*

Proof. The proof in [3] applies almost word-for-word to this generalisation, writing X in place of $\mathbb{D}$. We have to replace the $\lim_{n\to\infty}|Z_n| = 1$ with $Z_n \to \infty$ (meaning that for any compact set $K \subset X$, $Z_n \notin K$ for sufficiently large n). To generalise the proof of [3, Lemma 1], one needs the following statement.

> Suppose $(x_n) \to \infty$ as $n \to \infty$ and $d(x_n, y_n)$ is bounded, then $y_n \to \infty$ as $n \to \infty$.

This is true if and only if X is proper. The properness condition also allows us to use the original proof of [3, Lemma 3], where it is necessary that for any compact ball, the closed ball with the same centre but five times the radius is also compact.

□

Something along these lines was suggested in [3] but not made precise (no explicit condition was given on the metric space). All we claim to have added is the information that properness of X is the appropriate condition. That properness cannot be weakened much is demonstrated by example 2.1 above. The context of this result in the IFSs literature is discussed in appendix B.

2.3.1 Stability via Kantorovich metrics

We will now give an alternative proof of part 1 of Theorem 2.6, using the method of Kantorovich distances. This will serve as a model for the proofs of several asymptotic stability results.

Suppose that (X, d) is a proper metric space and $\mathcal{F}$ is a non-uniformly contracting IFS on (X, d). If $c = \varphi(d)$ is a modified metric as in §1.2, where

φ is strictly increasing, then $\mathcal{F}$ is also a non-uniformly contracting IFS on (X, c).

The following lemma is the key to our use of Kantorovich metrics.

Lemma 2.7.

Let (X, c) be a proper metric space and let $\mathcal{F}$ be a non-uniformly contracting IFS on (X, c). Then the push-forward $\mathcal{F}_*$ acts on $\mathcal{P}_c$ and is strictly distance-decreasing with respect to K_c. Also, $\mathcal{F}_*$ does not increase the Prohorov distance π_c.

Proof. Suppose $\mu \in \mathcal{P}_c$. Then

$$\begin{aligned}\int c\left(x, x_0\right) \, d\left(\mathcal{F}_*\mu\right)(x) &= \sum_{i=1}^{N} p_i \int c\left(f_i(x), x_0\right) \, d\mu(x) \\ &\leq \sum_{i=1}^{N} p_i \left(c\left(f_i\left(x_0\right), x_0\right) + \int c\left(f_i(x), f_i\left(x_0\right)\right) \, d\mu(x) \right) \\ &\leq A + \sum_{i=1}^{N} p_i \int c\left(x, x_0\right) \, ,\end{aligned}$$

where the additive constant A does not depend on the choice of μ. Thus $\mathcal{F}_*$ acts on $\mathcal{P}_c$.

Suppose that $\nu_1, \nu_2 \in \mathcal{P}_c$ are distinct. The infimum in the defining equation 1.1 is attained by a measure $m_0 \in \mathcal{P}^{\nu_1, \nu_2}$. The maps f_i act on $X \times X$ via $f_i \times f_i : (x, y) \mapsto (f_i(x), f_i(y))$, so we can define

$$\mathcal{F}_* m_0 = \sum_{i=1}^{N} p_i \, \left(f_i \times f_i\right)_* \, m_0.$$

Note that $\pi_1 \mathcal{F}_*\left(m_0\right) = \mathcal{F}_*\nu_1$ and $\pi_2 m_0' = \mathcal{F}_*\nu_2$. Since f_1 is strictly distance-decreasing, we have

$$K_c\left(\mathcal{F}_*\nu_1, \mathcal{F}_*\nu_2\right) \leq \int c(x, y) \, dm_0' < \int c(x, y) \, dm_0 = K_c(\nu_1, \nu_2).$$

Finally, we show that $\mathcal{F}_*$ does not increase π_c. It suffices to show that if $\pi_c(\nu_1, \nu_2) < \alpha$ then $\pi_c(\mathcal{F}_*\nu_1, \mathcal{F}_*\nu_2) < \alpha$. By Lemma 1.9 there is a measure $m \in \mathcal{P}^{\nu_1,\nu_2}$ such that $m(\{(x, y) : c(x, y) > \alpha\}) < \alpha$. But we also have $\mathcal{F}_*(m)(\{(x, y) : c(x, y) > \alpha\}) < m(\{(x, y) : c(x, y) > \alpha\})$ because the maps f_i do not increase distance. So $\mathcal{F}_*m$ is a certificate for the inequality $\pi_c(\mathcal{F}_*\nu_1, \mathcal{F}_*\nu_2) < \alpha$. □

We now give an example to show that $\mathcal{F}_*$ need not decrease Prohorov distances. Let δ_x and δ_y be the unit masses at distinct points $x, y \in X$ such that $c(x, y) < 1$. Then $\pi_c(\delta_x, \delta_y) = c(x, y)$. Let f_1 be a constant map with value z, where $c(x, z)$ and $c(y, z)$ both exceed $c(x, y)$, and let f_2 be the identity map. Taking positive probabilities p_1 and $p_2 = 1 - p_1 > c(x, y)$ we have a 2-map non-uniformly contracting IFS, for which $\pi_c(\mathcal{F}_*(\delta_x), \mathcal{F}_*(\delta_y)) = \pi_c(\delta_x, \delta_y)$.

Lemma 2.8.

Let $\mathcal{F}$ be a non-uniformly contracting IFS on a proper metric space (X, d). Suppose that $\mathcal{F}$ has an invariant law μ, and take any $\nu \in \mathcal{P}(X)$. Then the sequence $\mathcal{F}_*^n\nu$ is uniformly tight.

First proof. By Lemma 1.8 there is a metric $c = \varphi(d)$ such that φ is concave, strictly increasing and unbounded (so (X, c) is proper and $\mathcal{F}$ is non-uniformly contracting on (X, c)) and such that $\mu, \nu \in \mathcal{P}_c$. By Lemma 2.7 the measures $\mathcal{F}_*^n\nu$ all lie in the ball $B(\mu, K_c(\mu, \nu))$ in the K_c metric, which by Lemma 1.7 is uniformly tight. □

Second proof. Let $\epsilon > 0$. Because μ and ν are both tight, there is some compact set $K \subset X$ such that $\mu(X \setminus K) < \epsilon$ and $\nu(X \setminus K) < \epsilon$. Let

$m = \mu \times \nu \in \mathcal{P}(X \times X)$. Then $m(K \times K) > 1 - 2\epsilon$. Using the push-forward $\mathcal{F}_*$ on $\mathcal{P}(X \times X)$ defined in the proof of Lemma 2.7, we have

$$(\pi_1)_* \left(\mathcal{F}_*^n m\right) = \mathcal{F}_*^n \mu = \mu \quad \text{and} \quad (\pi_2)_* \left(\mathcal{F}_*^n m\right) = \mathcal{F}_*^n \nu\,.$$

Let $r = \operatorname{diam}(K)$ and let $N_r(K)$ be the closure of the r-neighbourhood of K, which by the properness property is also compact. If $(x, y) \in K \times K$ and $F_n \circ \cdots \circ F_1(x) \in K$ then $F_n \circ \cdots \circ F_1(y) \in N_r(K)$. It follows that $\mathcal{F}_*^n m \left(K \times N_r(K)\right)$ is at least $1 - 3\epsilon$, and hence $\mathcal{F}_*^n(\nu) \left(N_r(K)\right) > 1 - 3\epsilon$. This bound is independent of n, as required. □

We can now give our alternative proof of the first part of Theorem 2.6.

Alternative proof of Theorem 2.6(1).
Consider the modified metric $\rho = d/(1 + d)$, for which $\operatorname{diam}_\rho(X) \leq 1$, so $\mathcal{P}_\rho = \mathcal{P}(X)$. Let $\nu \in \mathcal{P}(X)$. By Lemma 2.7 the sequence $K_\rho \left(\mathcal{F}_*^n \nu, \mu\right)$ is strictly decreasing. Suppose it converges to a positive limit λ. By Lemma 2.8, we can find a weakly convergent subsequence $\mathcal{F}_*^{n_k} \nu \to \tilde{\nu}$. Because K_ρ metrizes the weak topology, we find $K_\rho(\mu, \tilde{\nu}) = \lambda$. But since $\mathcal{F}_*$ is distance-decreasing and therefore continuous, we also have $\mathcal{F}_*^{1+n_k} \nu \to \mathcal{F}_* \tilde{\nu}$. Then we have

$$\lambda = K_\rho \left(\mathcal{F}_* \tilde{\nu}, \mu\right) < K_\rho(\tilde{\nu}, \mu) = \lambda,$$

a contradiction. Since $K_\rho \left(\mathcal{F}_*^n \nu, \mu\right)$ is strictly decreasing and does not converge to any positive limit, it must converge to 0, which implies that $\mathcal{F}_*^n \nu \to \mu$ weakly. So the fixed point μ attracts all laws, as required. □

2.3.2 Ergodic properties in the stable case

A measurable dynamical system is called *uniquely ergodic* if it has only one non-zero invariant measure, up to multiplication by scalars. The invariant measure of a uniquely ergodic system is necessarily ergodic. The following lemma is a version of that statement for IFSs. It is a simple consequence of results of Ohno presented in [50, §I.1.2], but it is not explicitly stated in this way there.

Lemma 2.9.

Suppose that an N-map IFS $\mathcal{F} = (f_1, \ldots, f_N; p_1, \ldots, p_N)$ is uniquely ergodic, i. e. has a unique invariant law μ. Write $B_{\mathbf{p}}$ for the corresponding Bernoulli measure on the shift space $\{1, \ldots N\}^{\mathbb{N}}$. Then the measure $\mu \times B_{\mathbf{p}}$ is ergodic for the skew-product map τ.

Proof. Let μ be any $\mathcal{F}_*$-invariant Borel measure on X. We say that a bounded measurable function g on X is $(\mathcal{F}^*, \mu)$-invariant if $\mathcal{F}^* g = g$, μ-a.e; a set $A \subset X$ is called $(\mathcal{F}^*, \mu)$-invariant if its characteristic function χ_A is $(\mathcal{F}^*, \mu)$-invariant. We say μ is *ergodic* if every $\mathcal{F}^*$-invariant function is constant μ-a.e. It is shown in [50, §I.1.2] that μ is ergodic if and only if every $(\mathcal{F}^*, \mu)$-invariant set has μ-measure equal to 0 or 1, and that μ is ergodic if and only if $\mu \times B_{\mathbf{p}}$ is ergodic for the skew product map (in the usual sense of the term).

If A is an $(\mathcal{F}^*, \mu)$-invariant set with $0 < \mu(A) < 1$, then consider the measure μ' given by

$$\mu'(B) = \mu(A \cap B)/\mu(A)\,.$$

This defines an $\mathcal{F}_*$-invariant law on X different from μ. Hence the uniqueness of the invariant measure implies that any $(\mathcal{F}^*, \mu)$-invariant set has μ-measure either 0 or 1. This completes the proof of Lemma 2.9. □

The following corollaries tell us some circumstances in which we can plot a good picture of the invariant measure μ of an IFS $\mathcal{F}$ just by plotting sufficiently many points of a single random orbit.

Corollary 2.10.
If an N-map IFS $\mathcal{F}$ is asymptotically stable with invariant law μ then for μ-a.e. $x_0 \in X$, we have the following weak convergence of empirical measures, almost surely:

$$\frac{1}{k}\sum_{i=0}^{k-1}\delta_{x_i} \;\to \mu\,.$$

Recall that x_0 is a non-random point, whose random orbit under the IFS $\mathcal{F}$ is the sequence (x_i).

Proof. By Lemma 2.9, the measure $\mu \times B_{\mathbf{p}}$ is ergodic for the skew-product τ. For any function $g \in C_0(X)$, we have $g \circ \pi_1 \in L^1(X \times \Omega)$, so we can apply Birkhoff's Pointwise Ergodic Theorem to conclude that

$$\lim_{k\to\infty}\frac{1}{k}\sum_{i=0}^{k-1} g\left(x_i\right) = \overline{g}\left(x_0, \omega\right)$$

exists $(\mu \times B_{\mathbf{p}})$-a.s. and is a τ-invariant function, hence equal to the constant $\overline{g} := \int g\,d\mu$, for $(\mu \times B_{\mathbf{p}})$-a.e. point (x_0, ω). Now we use the fact that $C_0(X)$ is separable (this follows from the properness of X, using a countable compact exhaustion). If $g\left(x_0, \omega\right)$ exists and equals $\overline{g}$ for each element g of a countable dense subset of $C_0(X)$ (which again happens $(\mu \times B_{\mathbf{p}})$-a.e.) then it exists and equals $\overline{g}$ for every $g \in C_0(X)$, which is precisely the statement that the empirical measures converge weakly to μ. □

Corollary 2.11 (Weak law of large numbers).

Suppose that X is a proper metric space and that the N-map IFS $\mathcal{F}$ on X is non-uniformly contracting and has an invariant law. Then for every $x_0 \in X$, we have the following weak convergence of empirical measures, almost surely:

$$\frac{1}{k}\sum_{i=0}^{k-1} \delta_{x_i} \ \to \mu \,.$$

Proof. Observe that if we have two sequences (x_i) and (y_i) such that the empirical measures $\frac{1}{k}\sum_{i=0}^{k-1}\delta_{y_i}$ converge weakly to a law μ on X and such that $d\,(x_i, y_i) \to 0$ as $i \to \infty$, then the empirical measures of the sequence (x_i) must also converge weakly to μ. This follows from the fact that every element of $C_0(X)$ is uniformly continuous. By corollary 2.10 the former happens for μ-a.e. $y_0 \in X$. Let $K \subset X$ be a compact set such that $\mu(K) > 0$. Consider the set $A = K \times \{\omega \in \Omega : \, \omega_1 = 1\}$. Poincaré's Recurrence Theorem tells us that for $(\mu \times B_{\mathbf{p}})$-a.e. point $(y_0, \omega) \in A$, $\tau^n\,(y_0, \omega) \in A$ for infinitely many n. Thus we may choose a point $y_0 \in X$ for which $B_{\mathbf{p}}$-a.s. the empirical measures of (y_i) converge weakly to μ and also for infinitely many i, we have $y_i \in K$ and $F_{i+1} = f_1$. For such a y_0 and for any $x_0 \in X$, define the random sequences (y_i) and (x_i) on the same shift space $(\Omega, B_{\mathbf{p}})$. This is a simple form of coupling of the two Markov chains (x_i) and (y_i). We must show that $d\,(x_i, y_i) \to 0$ a.s. as $i \to \infty$. Since $d\,(x_i, y_i)$ is non-increasing and X is proper, x_i returns to some compact set L whenever y_i returns to K. Suppose that $d\,(x_i, y_i) \not\to 0$ as $i \to \infty$; then we could choose a sequence i_k along which $x_{i_k} \to \hat{x}$ and $y_{i_k} \to \hat{y}$ and such that $\omega_{1+i_k} = 1$. Then $d\,(x_i, y_i) \searrow d\,(\hat{x}, \hat{y})$ as $i \to \infty$, but $d\,(f_1(\hat{x}), f_1(\hat{y})) < d(\hat{x}, \hat{y})$, so for k sufficiently large, $d\,(x_{1+i_k}, y_{1+i_k}) < d(\hat{x}, \hat{y})$, which is a contradiction. □

2.4 Analytic IFSs on $\mathbb{D}$

2.4.1 Wolff–Denjoy Conjecture for analytic IFSs on $\mathbb{D}$

Suppose $\mathcal{F}$ is an N-map IFS of analytic maps of $\mathbb{D}$, and suppose that $\mathcal{F}$ has no invariant law. Theorem 2.4 then says that for any measure $\nu \in \mathcal{P}(\mathbb{D})$ we have $\mathcal{F}_*^n \mu \to 0$ weakly against $C_0(\mathbb{D})$, so the sequence $\mathcal{F}_*^n \mu$ eventually leaves every compact subset of $\mathcal{P}(\mathbb{D})$. The weak topology on $\mathcal{P}(\mathbb{D})$ coincides with the subspace topology of $\mathcal{P}(\mathbb{D}) \subset \mathcal{P}\left(\overline{\mathbb{D}}\right)$. In the case of a 1-map IFS, the Wolff–Denjoy Theorem says that in $\mathcal{P}\left(\overline{\mathbb{D}}\right)$ we have $\mathcal{F}_*^n \nu \to \delta_\alpha$ as $n \to \infty$, where δ_α is the unit point mass at $\alpha \in \overline{\mathbb{D}}$.

Conjecture 1 (Wolff–Denjoy for IFSs).
Let $\mathcal{F}$ be an N-map analytic IFS on $\mathbb{D}$. Then there exists a measure $\mu \in \mathcal{P}\left(\overline{\mathbb{D}}\right)$ such that for every $\nu \in \mathcal{P}(\mathbb{D})$, we have $\mathcal{F}_*^n \nu \to \mu$ weakly against $C\left(\overline{\mathbb{D}}\right)$.

Even in the special case where the maps extend continuously to $\overline{\mathbb{D}}$, the extension of $\mathcal{F}$ to $\overline{\mathbb{D}}$ may not be asymptotically stable. For example, there are many invariant measures on $\partial\mathbb{D}$ for the map $z \mapsto z^2$.

Lemma 2.12.
Suppose that $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ is an unbounded concave function with $\varphi(0) = 0$, and let ρ be the metric $\varphi\left(d_{\text{hyp}}\right)$ on $\mathbb{D}$, so that $(\mathcal{P}_\rho(\mathbb{D}), K_\rho)$ is a proper metric space. Let $\mu_n, \nu_n \in \mathcal{P}_\rho(\mathbb{D})$ for each $n \in \mathbb{N}$ and let $\mu \in \mathcal{P}(\partial\mathbb{D})$. If $\mu_n \to \mu$ weakly in $\mathcal{P}\left(\overline{\mathbb{D}}\right)$ as $n \to \infty$ and the sequence $K_\rho\left(\mu_n, \nu_n\right)$ is bounded, then $\nu_n \to \mu$ weakly in $\mathcal{P}\left(\overline{\mathbb{D}}\right)$.

Corollary 2.13.
Suppose that $\mu_0 \in \mathcal{P}(\mathbb{D})$, that $\mu \in \mathcal{P}(\partial\mathbb{D})$, and that there is a sequence

$n_k \to \infty$ such that $\mathcal{F}_*^{n_k}\mu_0 \to \mu$ weakly in $\mathcal{P}\left(\overline{\mathbb{D}}\right)$. Then for any *$\nu \in \mathcal{P}(\mathbb{D})$, we have $\mathcal{F}_*^{n_k}\nu \to \mu$ weakly in $\mathcal{P}\left(\overline{\mathbb{D}}\right)$. Thus it suffices to prove conjecture 1 for just one $\nu \in \mathcal{P}(\mathbb{D})$.*

Corollary 2.14.
If $\mu_0 \in \mathcal{P}(\mathbb{D})$ and $\mu \in \mathcal{P}(\partial\mathbb{D})$ then in $\mathcal{P}\left(\overline{\mathbb{D}}\right)$ we have $\mathcal{F}_*^{n_k}\mu_0 \to \mu$ as $k \to \infty$ if and only if $\mathcal{F}_*^{1+n_k}\mu_0 \to \mu$ as $k \to \infty$.

Proof of Lemma 2.12 and Corollaries 2.13 and 2.14.
Let K_E be the Kantorovich metric on $\mathcal{P}\left(\overline{\mathbb{D}}\right)$ associated to the Euclidean metric. Since $\overline{\mathbb{D}}$ has finite Euclidean diameter, K_E metrizes the weak topology on $\mathcal{P}\left(\overline{\mathbb{D}}\right)$, so we have $K_E(\mu_n, \mu) \to 0$ as $n \to \infty$. Now

$$K_E(\nu_n, \mu) \leq K_E(\mu_n, \mu) + K_E(\mu_n, \nu_n)\ ,$$

so it suffices to prove that $K_E(\mu_n, \nu_n) \to 0$. Let $0 < R < 1$, and let $\mathbb{D}_R$ be the open disc about 0 of hyperbolic radius R. We claim that for any $x \in \mathbb{D}$ and $y \in \mathbb{D} \setminus \mathbb{D}_R$, and for R large enough,

$$|x - y| \leq c(R)\rho(x, y)\,, \tag{2.1}$$

where $c(R) \to 0$ as $R \to \infty$. To show this, suppose that $\epsilon > 0$; we will find R_0 such that if $R \geq R_0$ then (2.1) holds with ϵ in place of $c(R)$. Since $|x - y| \leq 2$, this is immediate when $\rho(x, y) \geq 2/\epsilon$. Since φ is unbounded, this occurs for $d_{\text{hyp}}(x, y) \geq S$, where S is a constant depending on ϵ. On the other hand, since φ is concave, there is a constant $A > 0$ such that $\rho(x, y) \geq A d_{\text{hyp}}(x, y)$ whenever $d_{\text{hyp}}(x, y) < S$. Suppose that $d_{\text{hyp}}(0, y) \geq R$. Then $d_{\text{hyp}}(0, z) \geq R - S$ for every point z on the hyperbolic geodesic γ from x

to y. Therefore for R sufficiently large, (say $R \geq R_0$), the hyperbolic density at each point on γ is at least $1/(A\epsilon)$. Putting this together we have

$$\rho(x,y) \;=\; \varphi(d_{\text{hyp}}(x,y)) \;\geq\; A\, d_{\text{hyp}}(x,y) \;\geq\; A\frac{1}{A\epsilon}\int_\gamma |dz| \;\geq\; \frac{1}{\epsilon}|x-y|\,,$$

as required.

Now $\mu_n \to \mu$ weakly in $\mathcal{P}\left(\overline{\mathbb{D}}\right)$ and μ is supported on $\partial\mathbb{D}$, so $\mu_n \to 0$ weakly in $\mathcal{P}(\mathbb{D})$. By the above comparison of metrics,

$$K_E\left(\mu_n,\nu_n\right) \;\leq\; 2\,\mu_n\left(\mathbb{D}_R\right) + c(R)\,K_\rho(\mu_n,\nu_n)\,,$$

which we can make arbitrarily small by choosing R and then n sufficiently large.

For corollary 2.13, use Lemma 1.8 to choose φ so that $\nu,\mu_0 \in \mathcal{P}_\rho$, then apply Lemma 2.7 to get boundedness of $K_\rho\left(\mathcal{F}_*^n\mu_0,\mathcal{F}_*^n\nu\right)$. Corollary 2.14 is proved similarly, taking $\nu = \mathcal{F}_*\mu_0$. □

To prove conjecture 1 it would be sufficient to construct a non-empty subset of $\mathcal{P}(\mathbb{D})$, forward-invariant under $\mathcal{F}_*$, with only one limit point in $\mathcal{P}(\partial\mathbb{D})$. This is how Beardon's proof of the Wolff–Denjoy Theorem works; the forward-invariant set there is a Euclidean disc tangent to the unit circle at one point, itself a limit of hyperbolic discs. Unfortunately, similar limits of Kantorovich balls in $\mathcal{P}(\mathbb{D})$ do not have unique weak limits in $\mathcal{P}(\overline{\mathbb{D}})$, so further ideas are required.

2.4.2 An improved sufficient condition for stability

Combining ideas from [3] and [13] we can obtain a better sufficient condition than the one given by Ambroladze in [3] for an N-map IFS of analytic self maps of $\mathbb{D}$ to have an invariant measure. We will show that if the image of some finite composition of the maps is a Bloch subdomain of $\mathbb{D}$ then there exists an invariant law.

Definition. Let U be a hyperbolic Riemann surface, with hyperbolic metric d. Then a subdomain $V \subset U$ is a *Bloch domain* if and only if it has finite inradius, i. e.

$$r = \sup_{x \in V} d(x, U \backslash V) < \infty.$$

Bloch subdomains are useful here because of the following result.

Theorem 2.15. *[13, Theorem 4.1]*
Suppose that U is a hyperbolic Riemann surface. If V is a Bloch subdomain of U with inradius r and $f : U \to V$ is conformal then f is Lipschitz with constant $\tanh r$ with respect to the hyperbolic metric on U.

Remark: In [13] the theorem is only stated for plane domains, but the proof applies to arbitrary hyperbolic surfaces.

Theorem 2.16.
Let U be a hyperbolic Riemann surface, and let $\mathcal{F}$ be an IFS specified by analytic maps $f_1, \ldots, f_N : U \to U$ and probabilities $p_i > 0$. Suppose that there is some finite composition of maps $f_{i_1} \circ \cdots \circ f_{i_m}$ that maps U into a Bloch subdomain of U. Then $\mathcal{F}$ is asymptotically stable.

Proof. By Theorem 2.6 we just have to show that $\mathcal{F}$ has an invariant law. We will use Letac's principle, so we have to show that the sequence of reverse iterates converges to a constant map with probability 1.

Since $f_1(U)$ is a Bloch domain, we can take $\kappa < 1$ a Lipschitz constant for f_1 with respect to the hyperbolic metric on U. Pick $x_0 \in U$. Define

$$A = \max \left\{ d\left(x_0, f_i(x_0)\right) \, : \, i = 1 \ldots N \right\} .$$

Then for any $x \in U$ we have

$$d(x, f_j(x)) \leq d(x, x_0) + d(x_0, f_j(x_0)) + d(f_j(x_0), f_j(x)) \leq 2d(x, x_0) + A.$$

Now for any $x \in U$ put $y_m = F_1 \circ \cdots \circ F_m(x)$. Let $h(t)$ be the (random) number of values $1 \leq i \leq t$ such that $F_i = f_1$. Then

$$\begin{aligned} d(y_m, y_{m+n}) &\leq \sum_{t=m}^{m+n-1} d(y_t, y_{t+1}) && (2.2) \\ &\leq \sum_{t=m}^{m+n-1} \kappa^{h(t)} d(x, f_{t+1}(x)) && (2.3) \\ &\leq (2d(x, x_0) + A) \sum_{t=m}^{m+n-1} \kappa^{h(t)} && (2.4) \end{aligned}$$

Claim. $\sum_{t=1}^{\infty} \kappa^{h(t)} < \infty$ almost surely.

It follows from the claim that y_m is almost surely a Cauchy sequence. U is a complete metric space, so we know that (y_m) converges almost surely. The inequality 2.4 above shows that the convergence is locally uniform in x; furthermore if $y'_m = F_1 \circ \cdots \circ F_m(x')$, then

$$d(y_m, y'_m) \leq \kappa^{h(t)} d(x, x').$$

It also follows from the claim that $h(t) \to \infty$ almost surely and therefore the limit function is almost surely constant.

It remains to prove the claim. The convergence of the sum is a tail event, so has probability 0 or 1; we prove it converges with positive probability. Let w_n be the number of steps between the $(n-1)$th occurence and nth occurrence of $F_i = f_1$. The w_n are non-negative and have finite expectation C, so

$$\mathbb{P}(w_n \geq \kappa^{-n/2}) \leq C.\kappa^{n/2}\,.$$

The w_n are independent, so

$$\mathbb{P}(\forall n \geq 1, w_n \leq \kappa^{-n/2}) \geq 1 - \prod_{n=1}^{\infty}(1 - C.\kappa^{n/2}) > 0,$$

because $\sum_{n=1}^{\infty} C.\kappa^{n/2}$ converges. Now

$$\sum_{t=1}^{\infty} \kappa^{h(t)} = \sum_{n=1}^{\infty} w_n \kappa^{n-1},$$

the right-hand side of which converges with positive probability, as required.

□

Note that the conditions for Theorem 2.16 may be satisfied even when no individual f_i maps into a Bloch subdomain. Bloch subdomains form a strictly larger class than relatively compact subdomains. A particular advantage is that the Bloch property is preserved under lifting to covering spaces. In Chapter 3 this will allow us to apply Theorem 2.16 to prove uniqueness of invariant measures for certain holomorphic correspondences.

2.5 Convergence of reverse iterates

2.5.1 A generalisation of Letac's Principle

Here we give a version of Letac's Principle with weaker (albeit more complicated) hypotheses than the original.

Theorem 2.17 (Generalised Letac's Principle).
Let $\mathcal{F}$ be any IFS of continuous maps. Suppose that for every sequence (n_k) in some non-empty family $\mathcal{S}$, the corresponding subsequence of reverse iterates, i. e. G_{n_k}, almost surely converges pointwise to a (random) constant. Suppose that

1. *whenever $(n_k)_{k=1}^{\infty} \in \mathcal{S}$, then also $(1+n_k)_{k=1}^{\infty} \in \mathcal{S}$, and*
2. *for any two sequences $(a_k), (b_k) \in \mathcal{S}$, there is a sequence $(c_k) \in \mathcal{S}$ such that (a_k) and (c_k) have a common infinite subsequence and (b_k) and (c_k) have a common infinite subsequence.*

Then the distribution of the limit point does not depend on the choice of (n_k) from $\mathcal{S}$ and is the unique invariant measure for $\mathcal{F}$.

The conditions here are more subtle than they appear at first sight. If G_{n_k} almost surely converges pointwise then so does G_{1+n_k}, since $F_2 \circ \cdots \circ F_{n_k+1}$ is identically distributed to G_{n_k} and appending the continuous map F_1 on the left does not affect convergence. Condition (1) is needed here because $\mathcal{S}$ need not be the family of *all* sequences such that the corresponding sequence of reverse iterates almost surely converges. (It seems unlikely that one could establish condition (2) for that family without using the conclusion of the theorem.)

Proof.

Let (a_k), (b_k), (c_k) be as in condition (2) in the statement. By hypothesis, the sequences G_{a_k}, G_{b_k} and G_{c_k} all converge pointwise almost surely to random constants a, b and c. Since (a_k) and (c_k) share a common infinite subsequence, we must have $a = c$ and likewise we must have $b = c$. Therefore the limit does not depend on the choice of sequence from $\mathcal{S}$.

Applying this result to the sequences in condition (1) in the statement, we have

$$\lim_{k\to\infty} G_{1+n_k}(x) = \lim_{k\to\infty} G_{n_k}(x) \text{ a.s.},$$

so these two limits have the same distribution.

Consider the shifted sequence of i. i. d. random maps $\widetilde{F}_i = F_{i+1}$. Note that $\left(\widetilde{F}_i\right)_{i=1}^{\infty}$ and $(F_i)_{i=1}^{\infty}$ are identically distributed. Let $\left(\widetilde{G}_i\right)$ be the sequence of reverse iterates corresponding to $\left(\widetilde{F}_i\right)$.

Because F_1 is continuous, we have

$$\lim_{k\to\infty} G_{1+n_k}(x) = F_1\left(\lim_{k\to\infty} \widetilde{G}_{n_k}(x)\right).$$

We have shown above that the distribution μ of the left-hand side equals the distribution of $\lim_{k\to\infty} G_{n_k}(x)$. By construction this is also the distribution of $\lim_{k\to\infty} \widetilde{G}_{n_k}(x)$. Since F_1 is independent of $(\widetilde{F}_i)_{i=1}^{\infty}$, we obtain $\mu = \mathcal{F}_*\mu$, as required.

For uniqueness, let μ' be any $\mathcal{F}_*$-invariant measure. Let x be a random point of X, independent of (F_i) and distributed according to μ'. Then the random variables $G_{n_k}(x)$ are all distributed according to μ'. Almost sure pointwise convergence implies weak convergence in distribution, so $\mu' = \mu$.

□

2.5.2 Converse for non-uniformly contracting IFSs

Our reason for generalising Letac's Principle in Theorem 2.17 is the following result. It says that for a *non-uniformly contracting* IFS with an invariant law, there always exists a family of sequences satisfying the conditions of Theorem 2.17. So our generalised Letac's Principle applies to every non-uniformly contracting IFS.

Theorem 2.18 (Partial Converse of Theorem 2.17).
Let (X, d) be a proper metric space. Let

$$\mathcal{F} = (f_1, \dots, f_m : X \to X \, ; \, p_1, \dots p_m)$$

be a non-uniformly contracting IFS. Suppose that $\mathcal{F}$ has an invariant law μ. Then there exists a sequence $(a_k)_{k=1}^{\infty}$ of positive integers with the following property: for any sequence $(n_k)_{k=1}^{\infty}$ of positive integers with $n_1 \geq a_1$ and $n_k - n_{k-1} \geq a_k$ for all $k \geq 2$, the corresponding subsequence of reverse iterates

$$G_{n_k} = F_1 \circ \cdots \circ F_{n_k}$$

almost surely converges locally uniformly to a (random) constant. The family of such sequences satisfies the conditions on the family $\mathcal{S}$ in Theorem 2.17, so the limit does not depend on the choice of the sequence (n_k) and the distribution of the limit is the unique $\mathcal{F}_*$-invariant measure, namely μ.

Proof.
The key to the proof is the following nesting lemma, which depends on the same hypotheses as the theorem. Let $\alpha \in X$ be an arbitrarily chosen base point, which will remain fixed for the remainder of §2.5.

Lemma 2.19.

There exists an increasing sequence of radii $(r_k)_{k=0}^{\infty}$, with $r_k \to \infty$ as $k \to \infty$, and an increasing sequence of gaps $a_k \in \mathbb{N}$ with the following property. For any sequence of times $n_k \in \mathbb{N}$ such that $n_0 = 0$ and $n_k - n_{k-1} \geq a_k$ for all $k \geq 1$, there are almost surely infinitely many (random) offsets $j \in \mathbb{N}$ such that

$$F_{n_j} = f_1 \quad \textit{and} \quad (\forall k \geq 1) \quad g_{(k+j)}\left(B\left(\alpha, r_k\right)\right) \subset B\left(\alpha, r_{k-1}\right), \tag{2.5}$$

where

$$g_k = F_{n_{k-1}+1} \circ \cdots \circ F_{n_k} .$$

We will prove Lemma 2.19 after showing how it is used to prove Theorem 2.18. The following geometric lemma is useful in both of these proofs.

Lemma 2.20. *Let K and $L_0, L_1, L_2, L_3, \ldots$ be compact subsets of a metric space (X, d), and suppose that $f : X \to X$ is strictly distance-decreasing. Suppose that for all i, $L_i \subset K$ and*

$$\operatorname{diam}\left(L_{i+1}\right) \leq \operatorname{diam}\left(f\left(L_i\right)\right) .$$

Then

$$\operatorname{diam}\left(L_i\right) \leq t_i \,,$$

where

$$t_i \to 0 \quad \textit{as } i \to \infty,$$

and t_i depends only on i, K and f.

Proof. We have

$$\operatorname{diam}\left(f(L_i)\right) \leq \varphi\left(\operatorname{diam}(L_i)\right) ,$$

where for $t \geq 0$, $\varphi(t)$ is defined by

$$\varphi(t) = \sup \{ d\left(f(x), f(y)\right) \,:\, x, y \in K \ \text{ and } \ d(x,y) \leq t \} \,. \tag{2.6}$$

It follows that for all $i \geq 1$

$$\operatorname{diam}(L_i) \leq \varphi^{\circ i}(\operatorname{diam}(K))\,.$$

Since f is continuous and the subset of $X \times X$ over which the supremum in (2.6) is taken is compact, the supremum is attained, say at (x_t, y_t). Therefore for all $t > 0$ we have $\varphi(t) < t$. We claim that for any $t_0 > 0$, the iterates $t_i = \varphi^{\circ i}(t_0) \to 0$ as $i \to \infty$. Indeed, we have $t_n \searrow r$ as $n \to \infty$, and we can choose a subsequence along which (x_t, y_t) converges to (x_r, y_r). We conclude that $d\left(x_r, y_r\right) = r$ and also $d\left(f\left(x_r\right), f\left(y_r\right)\right) = r$, a contradiction unless $r = 0$. This completes the proof of lemma 2.20. □

Now we will show how to use Lemma 2.19 to prove Theorem 2.18. The conclusion of the theorem asserts the existence of a gap sequence (a_k); this will be the sequence (a_k) given by Lemma 2.19. Let $(n_k)_{k=1}^{\infty}$ be a sequence (as in the theorem) such that $n_1 \geq a_1$ and for all $k \geq 2$, $n_k - n_{k-1} \geq a_k$. Set $n_0 = 0$ to get a sequence as in the lemma.

Let $j_1, j_2, \ldots$ be the (random) sequence of values of j for which the nesting condition (2.5) is satisfied. We have

$$\begin{aligned} &B\left(\alpha, r_0\right) \supset g_{1+j_m}\left(B\left(\alpha, r_1\right)\right) \supset g_{1+j_m} \circ g_{2+j_m}\left(B\left(\alpha, r_2\right)\right) \supset \ldots \\ &\ldots \supset g_{1+j_m} \circ \cdots \circ g_{j_{m+1}}\left(B\left(\alpha, r_{(j_{m+1}-j_m)}\right)\right) \supset g_{1+j_m} \circ \cdots \circ g_{j_{m+1}}\left(B\left(\alpha, r_0\right)\right)\,. \end{aligned} \tag{2.7}$$

Fix $a \in \mathbb{N}$ and consider the sequence of sets

$$I_m = g_a \circ \ldots g_{j_m}\left(B\left(\alpha, r_0\right)\right)\,,$$

defined for m sufficiently large, say $m \geq m_0$ (where m_0 is minimal subject to $j_{m_0} \geq a$). Since (X, d) is proper, each I_m is compact. Because of the inclusions (2.7) they are nested: whenever I_m is defined then

$$I_m \supset I_{m+1} \supset I_{m+2} \supset \dots .$$

We wish to show that almost surely their intersection is a (random) point, which we will call z_a. Since (X, d) is complete, it suffices to show that $\operatorname{diam}(I_m) \to 0$ as $m \to \infty$. This follows from Lemma 2.20. To see this, fix m for the moment and in Lemma 2.20 take $f = f_1$, $K = B(\alpha, r_0)$ and (for $0 \leq i \leq m - m_0$)

$$L_i = \left(g_{1+j_{m-i}} \circ \cdots \circ g_{j_{1+m-i}}\right) \circ \cdots \circ \left(g_{1+j_{m-1}} \circ \cdots \circ g_{j_m}\right)\left(B(\alpha, r_0)\right) . \quad (2.8)$$

When they are defined, $\operatorname{diam}(L_{i+1}) \leq \operatorname{diam}(f_1(L_i))$ because all the maps of $\mathcal{F}$ do not increase distances and the rightmost map in each 'block' in (2.8) is f_1. So

$$\operatorname{diam}(I_m) \leq \operatorname{diam}(L_{m-m_0}) \leq t_{m-m_0} ,$$

where t_{m-m_0} is the bound from Lemma 2.20 that only depends on K and f. Now we let m vary, but our choices of K and f do not vary, so the bounds t_{m-m_0} converge to 0 according to Lemma 2.20. We have shown that $\operatorname{diam}(I_m) \to 0$ as $m \to \infty$, as required.

We digress briefly here to note that if we let $a \in \mathbb{N}$ vary then we obtain as above a random sequence (z_a) of intersection points. For $a \geq 1$ they satisfy

$$z_a = g_a(z_{a+1}) .$$

We will return to the idea of a random infinite reverse orbit in the proof of Theorem 2.24.

Let $L \subset X$ be any compact set. For i sufficiently large, $L \subset B(\alpha, r_i)$ so

$$g_{1+j_m} \circ \cdots \circ g_{i+j_m}(L) \subset B(\alpha, r_0)\ .$$

Applying $g_1 \circ \cdots \circ g_{j_m}$ to both sides, we find that the compact sets $g_1 \circ \cdots \circ g_k(L)$ converge to the point z_1 as $k \to \infty$. We have shown that the sequence

$$G_{n_k} = g_1 \circ \cdots \circ g_k$$

converges locally uniformly to the constant map with (random) value z_1. This is the first conclusion of Theorem 2.18.

Next we show that the family of sequences $(n_k)_{k=1}^{\infty}$ such that $n_1 \geq a_1$ and $n_k - n_{k-1} \geq a_k$ for all $k \geq 2$ satisfies the conditions of the generalised Letac principle. Condition (1) is obviously satisfied. For condition (2), to find the sequence (c_k), just take alternately a term from the sequence (a_k) and then a term from the sequence (b_k), always ensuring that the gaps are sufficiently long. This completes the proof of Theorem 2.18. □

Proof of Lemma 2.19.

Choose any sequence of probabilities $q_1, q_2, \cdots \in (0,1)$, such that $\sum q_k < \infty$. We will define inductively an increasing sequence of natural numbers $(a_k)_{k=1}^{\infty}$ and a sequence of positive radii $(r_k)_{k=0}^{\infty}$ such that $r_k \to \infty$ as $k \to \infty$ and for $k \geq 1$ we have

$$(m \geq a_k) \implies \mathbb{P}\left[F_1 \circ \cdots \circ F_m\left(B(\alpha, r_k)\right) \subset B(\alpha, r_{k-1})\right] > 1 - q_k. \qquad (2.9)$$

Fix any positive integer b. We will show that with probability 1 the nesting event 2.5 occurs for *some* $j \geq b$. This implies that with probability 1 it happens for infinitely many j.

As in the statement of Lemma 2.19, we consider a sequence $(n_k)_{k=0}^{\infty}$ such that $n_0 = 0$ and $n_k - n_{k-1} \geq a_k$ for all $k \geq 1$. Also recall the definition of the compositions

$$g_k = F_{n_{k-1}+1} \circ \cdots \circ F_{n_k} .$$

Using the independence of the maps F_i, condition (2.9) gives

$$\mathbb{P}\left[g_k\left(B\left(\alpha, r_k\right)\right) \subset B\left(\alpha, r_{k-1}\right)\right] > 1 - q_k.$$

Since the sequence (a_k) is increasing, we have (for each positive integer j)

$$\mathbb{P}\left[g_{j+k}\left(B\left(\alpha, r_k\right)\right) \subset B\left(\alpha, r_{k-1}\right)\right] > 1 - q_k .$$

Since the maps F_i are i. i. d. , the compositions g_k are independent. Therefore for any $m \in \mathbb{N}$ we have

$$\mathbb{P}\left[(\forall k \geq m) \quad g_{j+k}\left(B\left(\alpha, r_k\right)\right) \subset B\left(\alpha, r_{k-1}\right)\right] > \prod_{k=m}^{\infty}\left(1 - q_k\right) > 0.$$

Now by independence of the maps F_n we obtain

$$\mathbb{P}\left[\exists j \geq b \text{ such that} \quad \begin{matrix} F_{n_j} = f_1 \quad \text{and for all } 1 \leq k < m \\ g_{j+k}\left(B\left(\alpha, r_k\right)\right) \subset B\left(\alpha, r_{k-1}\right) \end{matrix}\right] = 1.$$

Combining the last two equations gives

$$\mathbb{P}\left[\exists j \geq b \text{ such that} \quad \begin{matrix} F_{n_j} = f_1 \quad \text{and for all } 1 \leq k < m \\ g_{j+k}\left(B\left(\alpha, r_k\right)\right) \subset B\left(\alpha, r_{k-1}\right) \end{matrix}\right] > \prod_{k=m}^{\infty}\left(1 - q_k\right).$$

But m was arbitrary and $\prod_{k=m}^{\infty}\left(1 - q_k\right) \nearrow 1$ as $m \to \infty$, so we find

$$\mathbb{P}\left[\exists j \geq b \text{ such that} \quad \begin{matrix} F_{n_j} = f_1 \quad \text{and for all } k \geq 1 \\ g_{j+k}\left(B\left(\alpha, r_k\right)\right) \subset B\left(\alpha, r_{k-1}\right) \end{matrix}\right] = 1.$$

It remains to show that we can construct the sequences n_k and r_k as required; we choose the r_k first, and then show that the condition describing which sequences n_k we can use is of the form $n_{k+1} - n_k \geq a_k$, as in the statement of the theorem.

Let us choose the radii r_k, for $k \geq 0$. We know that $\mathcal{F}_*^n(\delta_\alpha) \to \mu$ weakly, so we also have

$$\frac{1}{n}\sum_{i=0}^{n-1} \mathcal{F}_*^i(\delta_\alpha) \to \mu \qquad \text{as } n \to \infty.$$

Therefore we can choose radii r_k and integers m_k such that for each $k \geq 1$, $m \geq m_{k-1}$ implies both

$$\mathcal{F}_*^m(\delta_\alpha)(B(\alpha, r_{k-1} - 1)) > 1 - \frac{q_k}{3} \tag{2.10}$$

and

$$\frac{1}{m}\sum_{i=0}^{m-1} \mathcal{F}_*^i(\delta_\alpha)(B(\alpha, r_{k-1} - 1)) > 1 - \frac{p_1 q_k}{6}\,. \tag{2.11}$$

At the same time we make sure that $r_i \to \infty$ as $i \to \infty$.

From (2.11) it follows that

$$(\forall n \geq m_{k-1}) \qquad \mathbb{P}\left[\begin{array}{c} F_i \circ \cdots \circ F_1(\alpha) \in B(\alpha, r_{k-1} - 1) \quad \text{for} \\ \text{at least } n\left(1 - \dfrac{p_1}{2}\right) \text{ values } 0 \leq i < n \end{array}\right] > 1 - \frac{q_k}{3}. \tag{2.12}$$

Let X_n be the number of maps $F_1, \ldots, F_n$ that take the value f_1. Let $\mathcal{E}_n$ be the event that both

$$X_n \geq \frac{2np_1}{3}$$

and

$$F_i \circ \cdots \circ F_1(\alpha) \in B(\alpha, r_{k-1} - 1)$$
$$\text{for at least } n(1 - p_1/2) \text{ of the values } 0 \leq i < n. \tag{2.13}$$

When $\mathcal{E}_n$ occurs, there must be at least $\left(\frac{2np_1}{3} - \frac{np_1}{2}\right)$ values $0 \leq i < n$ such that both

$$F_i \circ \cdots \circ F_1(\alpha) \in B\left(\alpha, r_{k-1} - 1\right) \text{ and } F_{i+1} = f_1 . \tag{2.14}$$

It follows from (2.12) that

$$\mathbb{P}\left(\mathcal{E}_n\right) \geq 1 - \frac{q_k}{3} - \mathbb{P}\left(X_n \leq \frac{2np_1}{3}\right) .$$

X_n is binomially distributed with parameters (n, p_1), so for n sufficiently large,

$$\mathbb{P}\left(X_n \leq \frac{2np_1}{3}\right) < \frac{q_k}{3} .$$

Consider the compact sets

$$L_i = F_i \circ \cdots \circ F_1\left(B\left(\alpha, r_k\right)\right) .$$

When $\mathcal{E}_n$ occurs, there is a subsequence of the L_i of length at least $\frac{np_1}{6}$ satisfying the conditions of Lemma 2.20, with $f = f_1$ and $K = B\left(\alpha, r_k + r_{k-1} - 1\right)$. Supposing that n is sufficiently large, we can conclude that the event $\mathcal{E}_n$ implies

$$\operatorname{diam}\left(F_n \circ \cdots \circ F_1\left(B\left(\alpha, r_k\right)\right)\right) \ < \ 1 . \tag{2.15}$$

Putting this together, we find (for n large enough) that with probability at least $1 - q_k$, we have both (2.15) and the following, which is a consequence of (2.10):

$$F_n \circ \cdots \circ F_1(\alpha) \in B\left(\alpha, r_{k-1} - 1\right) ,$$

and therefore

$$F_n \circ \cdots \circ F_1\left(B\left(\alpha, r_k\right)\right) \subset B\left(\alpha, r_{k-1}\right) .$$

The n-tuples $(F_1, \ldots, F_n)$ and $(F_n, \ldots, F_1)$ are identically distributed. Choose a_k sufficiently large that $n \geq a_k$ is sufficiently large for all of the above conditions on n to apply, and (except when $k = 1$) also large enough that $a_k > a_{k-1}$. Then if $n \geq a_k$ we have

$$\mathbb{P}\left[F_1 \circ \cdots \circ F_m \left(B\left(\alpha, r_k\right)\right) \subset B\left(\alpha, r_{k-1}\right)\right] > 1 - q_k.$$

We have established that (2.9) is satisfied. This completes the proof of Lemma 2.19. □

2.5.3 The natural extension of the skew product

The reader may have wondered why we defined the skew product τ as an extension of the unilateral shift rather than the bilateral shift. In this section we treat the skew product $\hat{\tau}$ over the bilateral shift that arises from a stable non-uniformly contracting N-map IFS $\mathcal{F}$. Using Theorem 2.18, we will show that there is a unique $\hat{\tau}$-invariant law that makes $\hat{\tau}$ into an extension of the bilateral shift with a given Bernoulli measure, and that this extension is in fact an isomorphism; both m. p. d. s. are isomorphic to the natural extension of the original skew product τ. The invertibility of $\hat{\tau}$ as an m. p. d. s. will enable us in section §2.5.4 to apply the Birkhoff ergodic theorem to study the sequence of reverse iterates. The idea of using the natural extension of the skew product to obtain ergodic theorems about the reverse iterates has been appeared elsewhere, e.g. [30].

The bilateral shift space is

$$\hat{\Omega} = \{1, \ldots, N\}^{\mathbb{Z}},$$

with the σ-algebra $\hat{\mathcal{M}}$ which is the completion of the product σ-algebra with respect to the Bernoulli measure $B_{\mathbf{p}}$ that corresponds to the probability vector $\mathbf{p} = (p_1, \ldots, p_N)$. The left-shift map $\hat{\sigma}$ acts on a symbol sequence $\omega = (\omega_i)_{i=-\infty}^{\infty}$ thus:

$$(\hat{\sigma}\omega)_i = \omega_{i+1}\,.$$

There is a homomorphism ϕ from $\left(\hat{\Omega}, \hat{M}, \hat{B}_{\mathbf{p}}, \hat{\sigma}\right)$ to $(\Omega, \mathcal{M}, B_{\mathbf{p}}, \sigma)$, given by restriction:

$$\phi\left((\omega_i)_{i\in\mathbb{Z}}\right) = (\omega_i)_{i\in\mathbb{N}}\,.$$

Suppose that $(W, \mathcal{C}, \nu, T)$ is an invertible m. p. d. s. with a homomorphism α to $(\Omega, \mathcal{M}, B_{\mathbf{p}}, \sigma)$. Then there is an essentially unique homomorphism $\hat{\alpha} : W \to \hat{\Omega}$ such that $\alpha = \phi \circ \hat{\alpha}$. We can construct $\hat{\alpha}$ by setting

$$(\hat{\alpha}(w))_i = \left(\alpha(T^i w)\right)_0\,, \qquad \text{for } i \in \mathbb{Z}.$$

It follows that $\left(\hat{\Omega}, \hat{M}, \hat{B}_{\mathbf{p}}, \hat{\sigma}\right)$ is the natural extension of $(\Omega, \mathcal{M}, B_{\mathbf{p}}, \sigma)$.

Now suppose that (X, d) is a proper metric space with Borel σ-algebra $\mathcal{B}$. Suppose that $\mathcal{F}$ is a stable non-uniformly contracting N-map IFS on (X, d), with probability vector $\mathbf{p}$ and invariant law μ. Recall the skew-product map $\tau : X \times \Omega \to X \times \Omega$ defined in §2.1.1. We denote the Cartesian projections by

$$\pi_\Omega : X \times \Omega \to \Omega\,,$$
$$\pi_X : X \times \Omega \to X\,.$$

π_Ω is a homomorphism from $(X \times \Omega, \mathcal{B} \times \mathcal{M}, \mu \times B_{\mathbf{p}}, \tau)$ to the unilateral shift $(\Omega, \mathcal{M}, B_{\mathbf{p}}, \sigma)$.

We now define the skew product over the bilateral shift:

$$\hat{\tau} : X \times \hat{\Omega} \to X \times \hat{\Omega}\,, \quad (x, \omega) \mapsto (f_{\omega_1}(x), \hat{\sigma}\omega)\,,$$

and the Cartesian projection $\pi_{\hat{\Omega}} : X \times \hat{\Omega} \to \hat{\Omega}$.

The following diagram commutes:

$$\begin{tikzcd}
X \times \hat{\Omega} \arrow[rr, "(\mathrm{id}, \phi)"] \arrow[dr, "\hat{\tau}"] \arrow[dd, "\pi_{\hat{\Omega}}"'] & & X \times \Omega \arrow[dr, "\tau"] \arrow[dd, "\pi_{\Omega}"'] & \\
& X \times \hat{\Omega} \arrow[rr, "(\mathrm{id}, \phi)"] \arrow[dd, "\pi_{\hat{\Omega}}"'] & & X \times \Omega \arrow[dd, "\pi_{\Omega}"] \\
\hat{\Omega} \arrow[rr, "\phi"] \arrow[dr, "\hat{\sigma}"] & & \Omega \arrow[dr, "\sigma"] & \\
& \hat{\Omega} \arrow[rr, "\phi"] & & \Omega
\end{tikzcd}$$

We wish to put a probability measure on $X \times \hat{\Omega}$ so as to make the left, right, top and bottom faces of this diagram into extensions of m. p. d. s., when Ω and $\hat{\Omega}$ are given the Bernoulli measures $B_{\mathbf{p}}$ and $\hat{B}_{\mathbf{p}}$ respectively and $X \times \Omega$ is given the probability measure $\mu \times B_{\mathbf{p}}$.

Lemma 2.21.

There exists a unique Borel probability measure $\mathbb{Q}$ on $X \times \hat{\Omega}$ such that $\pi_{\hat{\Omega}}$ is a homomorphism of m. p. d. s. from $\left(X \times \hat{\Omega}, \overline{\mathcal{B} \times \hat{\mathcal{M}}}, \mathbb{Q}, \hat{\tau}\right)$ to $\left(\hat{\Omega}, \hat{M}, \hat{B}_{\mathbf{p}}, \hat{\sigma}\right)$, where $\overline{\mathcal{B} \times \hat{\mathcal{M}}}$ is the completion of the σ-algebra $\mathcal{B} \times \hat{\mathcal{M}}$ with respect to $\mathbb{Q}$. Moreover, $\pi_{\hat{\Omega}}$ is then an isomorphism between natural extensions of $(X \times \Omega, \mathcal{B} \times \mathcal{M}, \mu \times B_{\mathbf{p}}, \tau)$.

Proof. We begin by constructing a suitable measure $\mathbb{Q}$ using Theorem 2.18.

Now that we are working with a bilateral shift, it is convenient to connect the reverse iterates with the skew product. To do this, we will discard our original definition of G_n in terms of the maps $F_1, \ldots, F_n$ and instead model the reverse iterates G_n on the probability space $(\hat{\Omega}, \hat{\mathcal{M}}, \hat{B}_{\mathbf{p}})$ by extending the definition

$$F_n = f_{\omega_n} \quad \text{for each } n \in \mathbb{Z}\,,$$

and then for $n \in \mathbb{N}$ setting

$$\begin{aligned} G_n &= F_0 \circ F_{-1} \cdots \circ F_{1-n} \\ &= f_{\omega_0} \circ f_{\omega_{-1}} \circ \cdots \circ f_{\omega_{1-n}}\,. \end{aligned}$$

We have not changed the *distribution* of the sequence $(G_n)_1^\infty$, so Theorem 2.18 still applies to the sequence (G_n). Here is the connection of our new reverse iterates with the skew product:

$$\tau^n(x, \hat{\sigma}^{-n}\omega) = (G_n(\omega)(x), \omega)\,. \tag{2.16}$$

Let (n_k) be any suitable sequence for Theorem 2.18. Then we define $L(\omega)$ to be the random constant limit in Theorem 2.18, i.e.

$$L(\omega) = \lim_{k \to \infty} G_{n_k}(\alpha)\,.$$

$L(\omega)$ is a $B_{\mathbf{p}}$-a.e. defined function of ω. For all $i \in \mathbb{Z}$ and $n \in \mathbb{N}$, we have

$$G_{n+1}(\hat{\sigma}^i \omega) = f_{\omega_i} \circ G_n(\hat{\sigma}^{i-1}\omega). \tag{2.17}$$

Since the sequence $(1 + n_k)$ is also admissible for Theorem 2.18, taking the limit in (2.17) as $k \to \infty$ yields

$$L(\hat{\sigma}^i \omega) = f_{\omega_i} \circ L(\hat{\sigma}^{i-1}\omega)\,. \tag{2.18}$$

Consider the map

$$L^{\#} = (L, \mathrm{id}) : \hat{\Omega} \to X \times \hat{\omega}\,.$$

$L^{\#}$ is Borel-measurable because L is Borel-measurable, being a $B_{\mathbf{p}}$-a.e. limit of continuous functions. We define $\mathbb{Q}$ using as the push-forward by $L^{\#}$ of the Bernoulli measure:

$$\mathbb{Q} = (L^{\#})_* \, \hat{B}_{\mathbf{p}}\,.$$

To show that $\mathbb{Q}$ is a $\hat{\tau}$-invariant probability measure, it suffices to check that

$$L^{\#} \circ \hat{\sigma} = \tau \circ L^{\#}\,,$$

which is the content of (2.18). Since $\pi_{\hat{\Omega}} \circ L^{\#} = \mathrm{id}$, we have

$$\left(\pi_{\hat{\Omega}}\right)_* \mathbb{Q} = B_{\mathbf{p}}\,, \qquad \text{and also}$$

$$L^{\#} \circ \pi_{\hat{\Omega}} = \mathrm{id}\,, \quad \mathbb{Q}\text{-a.e.}\,,$$

so $\pi_{\hat{\Omega}}$ is an isomorphism as required.

The uniqueness of $\mathbb{Q}$ will follow once we show that (id, ϕ) is a homomorphism that makes $\hat{\tau}$ into the natural extension of τ. To this end, let α be a homomorphism from $(W, \mathcal{C}, \nu, T)$ to $(X \times \Omega, \mathcal{B} \times \mathcal{M}, \mu \times B_{\mathbf{p}}, \tau)$. Then $\pi_\Omega \circ \alpha$

is a homomorphism, so is essentially uniquely expressible as $\alpha = \phi \circ \beta$, where β is a homomorphism from $(W, \mathcal{C}, \nu, T)$ to $\left(\hat{\Omega}, \hat{M}, \hat{B}_{\mathbf{p}}, \hat{\sigma}\right)$. We wish to prove that

$$\alpha(w) = (L(\beta(w)), \phi(\beta(w))) , \quad \text{for } \nu\text{-a.e. } w \in W \tag{2.19}$$

for then α factors in an essentially unique way through β as

$$\alpha = (\mathrm{id}, \phi) \circ L^{\#} \circ \beta = (L, \phi) \circ \beta .$$

To prove (2.19) we use the locally uniform convergence of the reverse iterates G_{n_k} given by Theorem 2.18. Let $\epsilon > 0$. Since (X, d) is proper, we can choose a compact set $K \subset X$ such that

$$1 - \epsilon < \mu(K) = \nu(K \times \Omega) .$$

Let $(n_k)_{k=1}^{\infty}$ be a suitable sequence in Theorem 2.18. Because ν is T^{-1}-invariant, we have

$$\nu\left(\left\{w \in W : \alpha\left(T^{-n_k}(w)\right) \in K \ \text{ for infinitely many } k \in \mathbb{N}\right\}\right) > 1 - \epsilon .$$

On the other hand, $B_{\mathbf{p}}$-a.s. G_{n_k} converges uniformly on K to $L(\omega)$ as $k \to \infty$. It follows that

$$\nu\left(\{w \in W : \pi_X(\alpha(w)) = L(\beta(w))\}\right) > 1 - \epsilon ,$$

which proves (2.19) since ϵ was arbitrary.

The reader should note that the $\hat{\tau}$-invariant measure $\mathbb{Q}$ is in general not a product measure, i.e. the X and $\hat{\Omega}$ co-ordinates are not independent. □

Lemma 2.21 genuinely relies on fact that the maps are non-expanding. For a counterexample, consider the 1-map IFS given by the doubling map modulo 1, with μ being Lebesgue measure. In this case the skew product m. p. d. s. is isomorphic to the unilateral Bernoulli shift on two symbols with probabilities $\left(\frac{1}{2}, \frac{1}{2}\right)$, although the underlying shift is the Bernoulli shift on one symbol, i.e. the trivial m. p. d. s. . The natural extensions are the corresponding bilateral shifts, which are not isomorphic.

We can use L to construct the isomorphism from $\hat{\sigma}$ to the standard construction of the natural extension of τ that was explained in §1.4. To do this, we must associate to $B_{\mathbf{p}}$-almost every symbol sequence $\omega = (\omega_n)_{-\infty}^{\infty}$ a bi-infinite orbit $(x_i)(\omega)$, satisfying

$$(\forall i \in \mathbb{Z}) \qquad x_i = f_{\omega_i}(x_{i-1}) = F_i(x_{i-1}) \ . \tag{2.20}$$

For every $i \in \mathbb{Z}$, the distribution of the i-th co-ordinate x_i must be μ. These conditions are satisfied when we set

$$x_i(\omega) := L\left(\hat{\sigma}^i \omega\right) = \pi_X\left(\hat{\tau}^i(x_0, \omega)\right) \ .$$

$\hat{\tau}^{-1}$ is ergodic, since it is isomorphic to $\hat{\sigma}$. This is useful because the random backwards orbit $(x_{-n})_{n=0}^{\infty}$ is a projection of the forward orbit of $\hat{\tau}^{-1}$. In particular we can apply the Birkhoff Pointwise Ergodic Theorem to $\hat{\tau}^{-1}$ to obtain information about averages of functions along the sequence $(x_{-n})_{n=0}^{\infty}$. Since $G_n(x_{-n}) = x_0$, we can use this to obtain information about the complete sequence of reverse iterates, where previously we only had information about sparse subsequences.

Lemma 2.22.

For $B_{\mathbf{p}}$-a.e. $\omega \in \Omega$, the backward orbit $(x_{-n})_{n=0}^{\infty}(\omega)$ constructed above is uniquely determined by (2.20) together with the following property: for any compact set $K \subset X$,

$$\frac{1}{n}\sum_{j=0}^{n-1} \chi_K\left(x_{-j}\right) \to \mu(K) \text{ as } n \to \infty\,,$$

where χ_K is the characteristic function of K.

Proof. To see that the property is satisfied $B_{\mathbf{p}}$-almost surely, we apply the Birkhoff Pointwise Ergodic Theorem to the m. p. d. s. $\left(X \times \hat{\Omega}, \overline{\mathcal{B} \times \hat{\mathcal{M}}}, \mathbb{Q}, \hat{\tau}\right)$ and the L^1 function $\chi_K \circ \pi_X$, using

$$\chi_K\left(x_{-n}\right) = \chi_K \circ \pi_X \circ \hat{\tau}^{-n}\left(x_0, \omega\right) \quad \text{ and}$$

$$\int \chi_K \circ \pi_X \, d\mathbb{Q} = \textstyle\int \chi_K \, d\hat{\mu} = \mu(K)\,.$$

For the uniqueness, suppose that for some symbol sequence ω there are two orbits (x_n) and (y_n), both satisfying this criterion for every compact $K \subset X$. We may choose a compact set K such that $\mu(K) > \frac{1}{2}$. Then

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \chi_K\left(x_{-j}\right)\chi_K\left(y_{-j}\right) \;\geq\; 2\mu(K) - 1 > 0, \qquad B_{\mathbf{p}}\text{-a.s.},$$

so at infinitely many negative times we must have $d\left(x_{-j}, y_{-j}\right) \leq \operatorname{diam}(K)$. Because the maps f_i do not increase distance, we have $d\left(x_{-j}, y_{-j}\right) \leq \operatorname{diam}(K)$ for *all* $j \in \mathbb{Z}$. As was shown in the last part of the proof of Lemma 2.21, $B_{\mathbf{p}}$-a.s. there is a compact $K' \subset X$ such that $x_{-n_k} \in K'$ for infinitely many $k \in \mathbb{N}$. For these values of k, y_{-n_k} is in K'', the closed $\operatorname{diam}(K)$-neighbourhood of K', which is compact because X is proper. Meanwhile,

G_{n_k} converges uniformly on K'' to x_0, so

$$y_0 = G_{n_k}(y_{-n_k}) \to x_0 \quad \text{as } k \to \infty,$$

i.e. $y_0 = x_0$, $B_{\mathbf{p}}$-a.s.

The criterion of the lemma is unchanged if we make the index of summation run from some fixed m instead of 0; the above argument then yields $x_m = y_m$, $B_{\mathbf{p}}$-a.s. □

2.5.4 Strong converse for metrically taut spaces

Definition (Iancu and Williams). A metric space (X, d) is *taut* if for every $x, y \in X$ and every $\epsilon > 0$ there exists a sequence $x = z_0, z_1, z_2, \ldots, z_n = y$ of points such that

$$d(z_{i-1}, z_i) \leq \epsilon \quad \text{for all } i = 1, \ldots, n, \text{ and}$$
$$\sum_{i=1}^{n} d(z_{i-1}, z_i) \leq d(x, y) + \epsilon\,.$$

For example, every Riemannian manifold is a taut metric space.

Lemma 2.23. *Suppose that (X, d) is proper. Then (X, d) is taut if and only if for every $x \neq y \in X$ and $0 < t < d(x, y)$, there exists $z \in X$ with $d(x, z) = t$ and $d(z, y) = d(x, y) - t$.*

Proof. The 'if' direction is trivial. For 'only if', suppose we are given $x \neq y$ in a taut metric space (X, d) and $0 < t < d(x, y)$. Then for any $\epsilon > 0$ we obtain a chain of points z_i as in the definition of tautness, and we choose z_ϵ to be the first of these such that

$$d(x, z_\epsilon) \geq t$$

Then

$$d(x, z_\epsilon) \leq t + \epsilon$$

and

$$d(z_\epsilon) \leq d(x,y) - t + \epsilon\,.$$

Since (X, d) is proper, the ball $B(x, d(x, y))$ is compact, so we can choose a sequence $\epsilon_i \to 0$ as $i \to \infty$ such that z_{ϵ_i} converges; take z in the lemma to be the limit of this sequence. □

We will in fact only use the equivalent condition given by Lemma 2.23. Various properties of general taut metric spaces are developed in [43].

Theorem 2.24.
Suppose that (X, d) is proper and metrically taut. Let

$$\mathcal{F} = (f_1, \dots, f_N : X \to X; p_1, \dots p_N)$$

be a non-uniformly contracting IFS, on (X, d) and suppose that $\mathcal{F}$ has an invariant law μ. Then the sequence (G_k) of reverse iterates almost surely converges locally uniformly to a (random) constant.

Corollary 2.25. *Suppose that $\mathcal{F}$ is an N-map analytic IFS on $\mathbb{D}$ with an invariant law μ. Then the sequence of reverse iterates almost surely converges locally uniformly to a random constant, whose distribution is μ.*

Proof. The condition of metric tautness only comes into play in Lemma 2.27, near the end of the proof.

Firstly we rephrase the conclusion of the theorem as a large deviations statement. The sequence of maps G_n converges locally uniformly to the constant x_0 if and only if for every neighbourhood U of x_0 and every compact set $K \subset X$ we have $K \subset G_n^{-1}(U)$ for n sufficiently large.

We know that $x_{-n} \in G_n^{-1}(\{x_0\})$. Also, there is a non-decreasing sequence $(s_n)_{n=0}^{\infty}$ such that $d(x_{-n}, y) \leq s_n \implies G_n(y) \in U$. So the question is whether the sequence (s_n) grows rapidly enough that for sufficiently large n this will be satisfied for all $y \in K$. Thus we need

1. a statement about the large deviations of $d(\alpha, x_{-n})$ and
2. a statement about the growth of $s_n = d(x_{-n}, X \backslash G_n^{-1}(U))$.

It is for the second of these that we will use the metric tautness property. We will deal first with the common building blocks for the two statements above.

The next step in the proof of Theorem 2.24 is to use the natural extension of the skew product map τ associated with $\mathcal{F}$, as in 2.5.3.

We assume that f_1 is strictly distance-decreasing. Let $K \subset X$ be compact. Now apply the Birkhoff Pointwise Ergodic Theorem in the same way as in the proof of Lemma 2.22 to the indicator function for the event

$$(x_0 \in K \text{ and } \omega_1 = 1) \ .$$

Since the co-ordinate x_0 is a function purely of the $\omega_i : i \leq 0$, and ω_1 is independent of these, we have

$$\int \chi\{x_0 \in K \text{ and } \omega_1 = 1\} \, d\hat{\mu} \; = \; p_1 \, \mu(K).$$

Hence

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \chi\{x_{-j} \in K \text{ and } \omega_{1-j} = 1\} = p_1 \mu(K) \quad B_{\mathbf{p}}\text{-a.s.}$$

So with probability 1, the random backwards orbit reasonably often lands in K when the next map is f_1.

We are now in a position to prove a suitable large deviations result for $d\,(\alpha, x_{-n})$.

Lemma 2.26.

Fix any positive distance t. Then $B_{\mathbf{p}}$-almost surely, $d\,(\alpha, x_{-n}) < nt$ for all sufficiently large n.

Proof. The proof is by contradiction. The statement "$d\,(\alpha, x_{-n}) > nt$ for infinitely many positive n" in itself does not violate the ergodic properties established above. However, if $d\,(\alpha, x_{-n}) > nt$ then $d\,(\alpha, x_{-m}) > mt/2$ for $n \leq m \leq n(1+c)$, for some $c > 0$ which depends on t but not on n. The reason is that

$$\begin{aligned} d\,(\alpha, x_{-j}) &\leq d\,(\alpha, F_{-j}(\alpha)) \quad + \quad d\left(F_{-j}(\alpha), F_{-j}\left(x_{-(1+j)}\right)\right) \\ &\leq \max_{i=1,2,\ldots,N} d\,(\alpha, f_i(\alpha)) \quad + \quad d\left(\alpha, x_{-(1+j)}\right). \end{aligned}$$

Define

$$R = \max_{i=1,2,\ldots,N} d\,(\alpha, f_i(\alpha)) < \infty\,.$$

If

$$d\,(\alpha, x_{-n}) > nt$$

then

$$d\,(\alpha, x_{-j}) > jt/2$$

whenever

$$n \leq j \leq n\left(\frac{R+t}{R+t/2}\right).$$

We may choose a large compact set $K \subset X$ such that

$$\mu(K) > \frac{R + \frac{t}{2}}{R + t} .$$

Suppose now that for infinitely many positive n we have $d\left(\alpha, x_{-n}\right) > nt$. Then for infinitely many positive n we have

$$\frac{1}{n} \sum_{j=0}^{n-1} \chi_K\left(x_{-j}\right) < \frac{R + \frac{t}{2}}{R + t} ,$$

contrary to Lemma 2.22. □

Next we use the metric tautness property to give a lower bound on the growth of the pre-images of an arbitrary neighbourhood U of x_0.

Lemma 2.27.

There is a distance $t' > 0$ such that

$$s_n = d\left(x_{-n}, X \backslash G_n^{-1}(U)\right) \geq nt'$$

for all sufficiently large n, $B_{\mathbf{p}}$-a.s.

Proof. Choose $\epsilon > 0$ so that $B\left(x_0, \epsilon\right) \subset U$. Define, for $a \leq b$,

$$\begin{aligned} r(x, a, b) &= \sup\left\{r : F_a \circ \cdots \circ F_b(B(x, r)) \subset B\left(F_a \circ \cdots \circ F_b(x), \epsilon\right)\right\} \\ &= d\left(x, X \backslash \left(F_a \circ \cdots \circ F_b\right)^{-1}\left(B(F_a \circ \cdots \circ F_b(x), \epsilon\right)\right) . \end{aligned}$$

We now show that $(r(x, a, b) - \epsilon)$ is superadditive in the sense that for $a \leq b \leq c$ we have

$$(r(x, a, c) - \epsilon) \geq (r(x, b+1, c) - \epsilon) + \left(r\left(F_{b+1} \circ \cdots \circ F_c(x), a, b\right) - \epsilon\right) . \quad (2.21)$$

To see this, let y be any point in the closed ball $B(x, S)$ where S is the right-hand side of (2.21). We have to show that $F_a \circ \cdots \circ F_c(y)$ lies in

$B\left(F_a \circ \cdots \circ F_c(x), \epsilon\right)$. The metric tautness condition tells us that there is a point z such that

$$\begin{aligned} d(x,z) &\leq r(x, b+1, c) \quad \text{and} \\ d(z,y) &\leq r\left(F_{b+1} \circ \cdots \circ F_c(x), a, b\right) - \epsilon\,. \end{aligned}$$

Then, by definition,

$$d\left(F_{b+1} \circ \cdots \circ F_c(x), F_{b+1} \circ \cdots \circ F_c(z)\right) \leq \epsilon.$$

Because the maps do not increase distance,

$$d\left(F_{b+1} \circ \cdots \circ F_c(z), F_{b+1} \circ \cdots \circ F_c(y)\right) \leq d(z,y)\,,$$

so by the triangle inequality,

$$\begin{aligned} d\left(F_{b+1} \circ \cdots \circ F_c(x), F_{b+1} \circ \cdots \circ F_c(y)\right) &\leq \epsilon + d(z,y) \\ &\leq r\left(F_{b+1} \circ \cdots \circ F_c(x), a, b\right)\ . \end{aligned}$$

Therefore we have

$$d\left(F_a \circ \cdots \circ F_c(x), F_a \circ \cdots \circ F_c(y)\right) \leq \epsilon\,,$$

which establishes (2.21).

Now we apply the superadditivity repeatedly to obtain

$$r\left(x_{-n}, 1-n, 0\right) - \epsilon \geq \sum_{j=1}^{n} r\left(x_{-j}, 1-j, 1-j\right) - \epsilon.$$

Next, we show that when $F_k = f_1$, then $r(x,k,k) - \epsilon > 0$ for every $x \in X$. Suppose not, then there is a point $x \in X$ and a sequence of points (z_n) such that $d\left(z_n, x\right) \to \epsilon$ as $n \to \infty$ and $d\left(f_1\left(z_n\right), f_1(x)\right) > \epsilon$. Because X is proper, there is a convergent subsequence with limit z, say. Then $d(z, x) = \epsilon$ but $d\left(f_1(z), f_1(x)\right) \geq \epsilon$. Since f_1 is strictly distance-decreasing, this is a contradiction.

We now apply the Birkhoff Pointwise Ergodic Theorem to the the averages of the positive function $r(\omega) = r\left(x_{-1}, 0, 0\right)$ along the forward orbits of the ergodic transformation $\hat{\sigma}^{-1}$. We find that $B_{\mathbf{p}}$-a.s.,

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} r\left(x_{-j}, j-1, j-1\right) - \epsilon \quad \geq \quad \int \mathbb{E}(r(\omega) - \epsilon)\, d\mu \quad > \quad 0.$$

(We do not know that the function r is in $L^1\left(B_{\mathbf{p}}\right)$, so we have to apply the Birkhoff Pointwise Ergodic Theorem to truncations of the function in order to get this, which is why we have a lim inf rather than a limit). Hence

$$\liminf_{n\to\infty} \frac{r\left(x_{-n}, 0, n-1\right) - \epsilon}{n} \quad > \quad 0, \qquad B_{\mathbf{p}}\text{-a.s.}$$

Therefore the conclusion of the lemma is satisfied by taking

$$t' = \int \mathbb{E}(r(\omega) - \epsilon)\, d\mu\,.$$

□

At the cost of using a less well-known result, we could shorten this proof slightly by applying Kingman's Subadditive Ergodic Theorem instead of Birkhoff's Pointwise Ergodic Theorem.

Finally we combine Lemma 2.26 and Lemma 2.27 to complete the proof of the theorem. Choose $t = 2t'$, so that for sufficiently large n we have

$$d\left(x_{-n}, X \backslash G_n^{-1}(U)\right) \geq d\left(\alpha, x_{-n}\right) + nt',$$

and $K \subset B(\alpha, nt')$ for sufficiently large n. Therefore

$$K \subset G_n^{-1}(U) \qquad \text{for sufficiently large } n,\ B_{\mathbf{p}}\text{-a.s.}$$

□

Chapter 3

Holomorphic correspondences

3.1 Introduction

In this chapter we will study the geometry and dynamics of holomorphic correspondences on compact Riemann surfaces. Polynomials and rational maps on the Riemann sphere are objects belonging to complex algebraic geometry, yet in studying their dynamics, complex analysis must be applied in ways that do not belong to algebraic geometry. This is because we are interested in describing asymptotic behaviour with respect to the usual topology rather than the Zariski topology. We will do the same with holomorphic correspondences, introducing them as objects of algebraic geometry then switching to the point of view of geometric function theory to obtain conclusions about the invariant probability measures of the associated Markov chains. In doing this we will make use of some of the results and techniques of Chapter 2.

3.1.1 Correspondences – the algebraic viewpoint

The following definitions mostly follow the exposition given in [37, §2.5], although some of the notation follows [21]. Let C and C' be (smooth, projective) algebraic curves over $\mathbb{C}$, which we may also think of as compact Riemann surfaces. The d^{th} symmetric product $(C')^{(d)}$ is a compact complex manifold [37, p. 236]. A *correspondence* $T : C \to C'$ of degree d associates to every point $p \in C$ an effective divisor $T(p)$ of degree d on C', varying holomorphically with p, in the sense that it is a holomorphic map from C to $(C')^{(d)}$. Equivalently we can specify T by its curve of correspondence

$$D = \{(p,q) : q \in T(p)\} \subset C \times C' \,.$$

Then $T(p)$ is obtained by pullback of divisors, as $T(p) = i_p^*(D)$, where i_p is the inclusion map $C' \to C \times C'$ such that $i_p(q) = (p,q)$. When D is irreducible then we call the correspondence T irreducible. Note that D need not be a smooth curve in $C \times C'$, but nevertheless there is a compact Riemann surface $\tilde{D}$ (the desingularization of D), with analytic maps $\pi_1 : \tilde{D} \to C$ and $\pi_2 : \tilde{D} \to C'$ such that

$$T(p) \;=\; \sum_{z \,\in\, \tilde{D} \,:\, \pi_1(z) = p} v_{\pi_1}(z) \,.\, \pi_2(z) \;=\; \pi_2 \circ \pi_1^*(p)\,,$$

where $v_{\pi_1}(z)$ is the valency of π_1 at z. We will sometimes think of the correspondence as a multivalued map $T = \pi_2 \circ \pi_1^{-1}$.

The *inverse* of the correspondence T is the correspondence T^{-1} given by $\pi_1 \circ \pi_2^*$. (We have reversed the roles of C and C' while keeping the same curve of correspondence.) The degree d' of T^{-1} need not be the same as that of T.

The pair (d', d) is called the *bidegree* of T. A point $z \in \tilde{D}$ is called *forward-singular* if $v_{\pi_1}(z) > 1$ and *backward-singular* if $v_{\pi_2}(z) > 1$. Accordingly a point $p \in C$ is called forward-singular if $T(p)$ is supported on fewer than d points, which happens if and only if $p = \pi_1(z)$ for some forward-singular point $z \in \tilde{D}$. Similarly $q \in C'$ is called backward-singular if and only if $T^{-1}(q)$ is supported on fewer than d' points. In [37], the point $(p, q) \in D$ is called a *coincident* point if q appears in $T(p)$ with multiplicity greater than 1. If D has a node at (p, q), then (p, q) is a coincident point but need not be forward-singular. Coincident points are important in enumerative algebraic geometry, but do not have much connection with dynamics; in contrast forward-singular and backward-singular points turn out to be very important in studying the dynamics of a correspondence.

Suppose that $R : \widehat{\mathbb{C}} \to \widehat{\mathbb{C}}$ is a rational map of degree n. Then we may consider R as a correspondence of bidegree $(n, 1)$ and R^{-1} as a correspondence of bidegree $(1, n)$. Note that a critical point of R is not a singular point, but a critical value is both a backward-singular point for the correspondence R and a forward-singular point of the correspondence R^{-1}.

Given two correspondences $T : C \to C'$ and $S : C' \to C''$, we may compose them as follows. Expressing $T(p) = \sum_{q \in C'} m(q).q$, we define $S \circ T(p) = \sum_{q \in C'} m(q)T(q)$, an effective divisor on C''. Note that unless the correspondence T has bidegree $(1, 1)$, then the composite $T^{-1} \circ T$ will not be the identity. If $d' = 1$ then $T^{-1} \circ T$ will be the correspondence $p \mapsto d.p$.

From the point of view of dynamics, we will wish to consider two correspondences $T : C \to C$ and $T' : C' \to C'$ as equivalent when there exists an isomorphism of algebraic curves $\phi : C \to C'$ such that $\phi^{-1} \circ T' \circ \phi = T$ as divisors. When this happens, we say that T and T' are conjugate.

3.1.2 The geometric function theory viewpoint

A holomorphic correspondence is a multivalued analytic map between compact Riemann surfaces which is locally defined except at a discrete set of singular points by a finite set of of analytic branches. Each connected component of the graph of the correspondence in the neighbourhood of a singular point is given analytically by a Puiseux expansion, i. e. a power series in some fractional power of a local co-ordinate.

3.1.3 Dynamics on the Jacobian

In order to iterate the correspondence T, we must have $C' = C$. To carry out iteration using only the constructions of algebraic geometry, we must extend T linearly to $T : \mathrm{Div}(C) \to \mathrm{Div}(C)$; then T^n will associate to each point of C a divisor of degree d^n. This does not constitute an interesting dynamical system because the degree of the divisors varies with time, so it is not very meaningful to compare distinct points of an orbit. A natural way to surmount this difficulty is to consider the action of T on $\mathrm{Div}_0(C)$. Consider a principal divisor (g) on C. We have

$$T((g)) = (\pi_2)_* \circ \pi_1^*((g)) = (\pi_2)_*((g \circ \pi_1)) = (h)\,,$$

where h is the meromorphic function on C such that $h(z)$ is the product of the values of $g \circ \pi_1$ over the points of $\pi_2^{-1}(z)$, repeated according to multiplicity. h is a non-zero meromorphic function. Thus T maps principal divisors to principal divisors. Hence T descends to a linear map on the Jacobian of C, which we write as $T : \mathrm{Pic}_0(C) \to \mathrm{Pic}_0(C)$.

When C has positive genus, a correspondence $T : C \to C$ is said to have *valence* $k \in \mathbb{Z}$ if the linear equivalence class of the divisor $T(p) + k.p$ does not depend on p. This happens if and only if $T : \mathrm{Pic}_0(C) \to \mathrm{Pic}_0(C)$ is multiplication by $-k$.

Not every correspondence has valency, but on a generic Riemann surface there are no correspondences without valency, since the Jacobian is a complex torus which typically has no endomorphisms other than multiplication by elements of $\mathbb{Z}$. For example, an elliptic curve (which is isomorphic to its Jacobian) has a correspondence without valency if and only if it has complex multiplication: in this case the map $T : \mathrm{Pic}_0(C) \to \mathrm{Pic}_0(C)$ is a non-integer endomorphism. Conversely any non-integer endomorphism is itself a correspondence without valency. In §3.5.2 we will construct some irreducible correspondences of bidegree $(2, 2)$ without valency.

3.1.4 Separable correspondences

A holomorphic correspondence $T : C \to C'$ of bidegree (d', d) is said to be *separable* if

$$T = \pi_3^{-1} \circ \pi_4$$

for some analytic maps

$$\pi_3 : C' \to C'' \quad \text{and} \quad \pi_4 : C \to C''\,.$$

If T is separable and the divisors $T(p)$ and $T(q)$ have a point in common, then $T(p) = T(q)$. In the dynamical case $C = C'$, we can understand everything about the dynamics of a separable correspondence by studying the associated correspondence $\pi_4 \circ \pi_3^{-1} : C'' \to C''$. This correspondence will sometimes be simpler to study because the genus of C'' is less than or equal to the genus of C.

Lemma 3.1.
If $T : C \to C$ is separable with $C'' = \widehat{\mathbb{C}}$, then T has valence 0.

Proof. In this case π_3 and π_4 are meromorphic functions and $T(p) - T(q) = \pi_3^*(\pi_4(p) - \pi_4(q))$, which is a principal divisor since $\pi_4(p) - \pi_4(q)$ is principal. □

3.1.5 Correspondences as Markov chains

Given a correspondence $T : C \to C$ of degree $d > 0$, we will consider an associated Markov chain taking values in C, defined as follows. Let X_0 be a C-valued random variable, possibly constant. Express

$$T(X_n) = \sum_{i=1}^{k} m_i \, . \, p_i \, ,$$

for distinct points $p_i \in C$, then set

$$\mathbb{P}(X_{n+1} = p_i) = m_i/d \, .$$

We also insist that the transitions at distinct times are mutually independent and independent of X_0.

In general this Markov chain cannot be expressed as an IFS of continuous (single-valued) maps. We could make finitely many branch cuts along smooth arcs. Then the orbits $(X_n)_1^\infty$ would be those of an IFS $\mathcal{F}$ of finitely many functions, each one analytic except on the chosen arcs. However the choice of arcs is rather arbitrary and because the individual functions are not continuous at the cuts we cannot apply any of the theory of continuous IFS.

We consider the curve C with the usual topology (inherited from a projective embedding), not the Zariski topology. This makes C compact. For a continuous function f on C, we define the pull-back $T^*(f)$ to be the function given by

$$T^*(f)(p) = \sum \frac{m_i}{d} f(p_i)\,,$$

where $T(p) = \sum m_i p_i$. Observe that T^* acts on $C(C)$. This is easily checked locally. In fact T^* is the Perron–Frobenius operator of the IFS $\mathcal{F}$, which is therefore good in the sense of §2.1.2. Furthermore $T^* : C(C) \to C(C)$ is a bounded linear operator. We can also describe the Markov operator of $\mathcal{F}$ purely in terms of T, as follows. Given a Borel set A, let $n(T, z, A)$ denote the number of points of $T(z)$ in A, counted according to multiplicity. Then, given a law μ on C, we define its push-forward by

$$T_*(\mu)(A) = \frac{1}{d} \int_C n(T, z, A)\, d\mu(z)\,.$$

We will call T_* the *push-forward operator* and recall that it is adjoint to T^*. We will be interested in *invariant* laws, those for which $T_*(\mu) = \mu$. Lemma 2.1 tells us that there exists at least one invariant probability measure for T.[1] Note that T_* acts continuously on $\mathcal{M}(C)$ (with the weak topology) and in particular acts continuously on $\mathcal{P}(C)$.

[1] We warn the reader that an invariant measure for T need not be invariant for T^{-1}.

3.2 Background results

3.2.1 Invariant measures for rational maps, rational semigroups and Kleinian groups

Let $T : \widehat{\mathbb{C}} \to \widehat{\mathbb{C}}$ be a rational map of degree $d \geq 2$. As a correspondence, T has bidegree $(d, 1)$. A T^{-1}-invariant measure is necessarily T-invariant (see footnote 1). In the context of T-invariant measures, the standard terminology is to say that a T-invariant measure is *balanced* when it is also T^{-1}-invariant.

T may have a finite backward-invariant set E, called the exceptional set, consisting of at most two points. Choose any point $z_0 \notin E$ and use it as the initial state of the Markov chain (z_n) associated to the inverse correspondence T^{-1}. What can be said about the statistical behaviour of the sequence $(z_n)_{n=0}^{\infty}$? This question was answered for a polynomial map by Brolin in the 1960s and then for an arbitrary rational map in the early 1980s by Lyubich [49], and at the same time by Mañé et al. [32, 52]. Write μ_n for the distribution of z_n. The laws μ_n converge weakly to a doubly-invariant law μ whose support is $\mathcal{J}$, the Julia set of T. Moreover, for every Borel set B on which T is injective, we have $\mu(T(B)) = d \cdot \mu(B)$; in other words, the Jacobian of T with respect to μ is d everywhere. μ is the unique invariant measure of maximal entropy for T, with entropy $\log d$. Also, T is an exact endomorphism with respect to the measure μ, which means that for any Borel set $A \subset \widehat{\mathbb{C}}$ with $\mu(A) > 0$, we have $\mu(T^n(A)) \to 1$ as $n \to \infty$. (Exactness is the measure-theoretic analogue of the topological transitivity of T on $\mathcal{J}$). It follows that T is strong-mixing, hence also ergodic, with respect to μ. The natural extension of the system (T, μ) is also ergodic, so if z_0 is chosen randomly according to μ then with probability one the random backward orbit $(z_n)_{n=0}^{\infty}$ approximates μ in the following sense: for any continuous function f on $\widehat{\mathbb{C}}$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(z_i) \quad = \quad \int f\, d\mu.$$

Put another way, the sequence of empirical measures

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{z_i}$$

almost surely converges weakly to μ.

Let G be a finitely generated (but not finite) Kleinian group, with a given set H of generators. Construct an iterated function system $\mathcal{F}$ using these generators and their inverses, acting on $\widehat{\mathbb{C}}$. G also acts on hyperbolic 3-space $\mathbb{H}^3$. Suppose $x \in \mathbb{H}^3$ is not fixed by any element of G, then the sequence of images of x under the reverse iterates is a random walk on the Cayley graph of (G, H) embedded in $\mathbb{H}^3$. This random walk almost surely converges to a unique point in $\partial\mathbb{H}^3 = \widehat{\mathbb{C}}$, and the distribution μ of this limit point is an invariant measure for $\mathcal{F}$ supported on the limit set of G.[2] In fact if G is non-elementary then $\mathcal{F}$ is asymptotically stable.

The topological dynamics of semigroups of rational maps have been studied by Hinkkanen and Martin [41] and others. To make sense of the notion of an invariant measure for a semigroup of rational maps, one needs to consider an iterated function system defined by rational maps $f_1, \ldots, f_N$ and probability vector $\mathbf{p} = (p_1, \ldots, p_N)$. For such an IFS, Sumi [71] proved that

[2] Of course the measure μ depends on the probability vector of $\mathcal{F}$, unlike the Patterson–Sullivan measure, which only depends on G.

among all invariant measures for the skew product, there is a unique one that maximises the relative metric entropy of the skew product over the unilateral Bernoulli shift. This maximal measure is in fact obtained by iterating the Perron-Frobenius operator for the inverse of the skew product, which involves the inverse correspondences of the f_i. The maximal relative metric entropy is

$$\sum_{i=1}^{N} p_i \log\left(\deg f_i\right) \ ;$$

the topological entropy of the skew product itself is the maximum of the entropies of these invariant measures equal to $\log \sum \deg f_i$.

Many direct analogies can be drawn between concepts and results in the theory of iteration of rational maps and in the theory of Kleinian groups. This relationship, called Sullivan's dictionary, has substantial predictive power. Hopes have been expressed that it can be extended to include results about certain classes of holomorphic correspondences. It is not immediately obvious how to define such objects as the Julia set or limit set for a general correspondence; [21] describes how one might try to generalise the topological-dynamical descriptions of Julia set and limit set. On the other hand, since Julia sets of rational maps and limit sets of Kleinian groups both arise as the supports of maximal entropy invariant measures of correspondences, it makes sense to study invariant measures for general holomorphic correspondences. In the light of Sumi's results, the next step should be to study invariant measures and stability for the iteration of an irreducible correspondence. In this thesis we obtain stability and entropy results about invariant measures for certain special classes of irreducible correspondences, namely critically finite correspondences and forward critically finite correspondences.

3.2.2 Example: the AGM correspondence

Let $a_0 \geq b_0 > 0$. For $n \geq 0$

$$a_{n+1} = \frac{a_n + b_n}{2} \quad \text{and} \quad b_{n+1} = \sqrt{a_n b_n}.$$

Consider the canonical definite elliptic integral

$$G\left(a_0, b_0\right) = \int_0^{\pi/2} \frac{d\theta}{a_0^2 \sin^2\theta + b_0^2 \cos^2\theta}, \qquad a \geq b > 0$$

Gauss observed that

$$G\left(a_i, b_i\right) = G\left(a_{i+1}, b_{i+1}\right) \ ,$$

and that the sequences a_n and b_n converge rapidly to a common limit $M(a_0, b_0)$. This enabled him to evaluate definite elliptic integrals with unprecedented precision, since

$$G(a_0, b_0) = \frac{\pi}{2M(a_0, b_0))} \ .$$

$M(a_0, b_0)$ is called the arithmetic-geometric mean of a and b. Like Newton's method for finding a root of a polynomial, this method displays quadratic convergence: the error, once small enough, is roughly squared at each step of the iteration. If we try to extend the transformation $(a, b) \mapsto \left(\frac{a+b}{2}, \sqrt{ab}\right)$ to complex values of a and b, we immediately encounter the problem of which square root to choose. One solution is to choose both! Let us simplify the transformation by projectivizing it, to obtain

$$T : z = \frac{a_n}{b_n} \mapsto \frac{a_{n+1}}{b_{n+1}} = \frac{z+1}{2\sqrt{z}} \ .$$

This is the AGM correspondence. It is an irreducible correspondence of bidegree $(2, 2)$ on $\mathbb{P}^1(\mathbb{C}) = \widehat{\mathbb{C}}$, as we may check by exhibiting $\tilde{D} = \widehat{\mathbb{C}}$ and the maps

$$\pi_1 : z \mapsto z^2 \quad \text{and} \quad \pi_2 : z \mapsto \frac{z^2+1}{2z} \ .$$

3.2.3 Critically finite correspondences

The results of this chapter were inspired by [20] and [19]. In §3.2.3–3.2.4 we summarise the statement and proof of [19, Prop. 1] and its application to the AGM correspondence. In doing so we will set up the notation that we need for later sections.

Definition.
A holomorphic correspondence $T : C \to C$ is *critically finite* if all forward-singular and backward-singular points have finite grand orbits, i. e. there is a finite completely invariant set of points of C containing all the singular points.

Suppose that a correspondence $T : C \to C$ is critically finite, and that $A \subset C$ is the smallest completely invariant set containing all the singular points. Then $\pi_1^{-1}(A) = \pi_2^{-1}(A)$, a finite subset of $\tilde{D}$. Set $U = \tilde{D} \setminus \pi_1^{-1}(A)$. Then the restrictions $\pi_1|_U$ and $\pi_2|_U$ are unbranched covering maps of X.

For the moment we will not consider the cases where $C = \mathbb{P}^1$ and T has a complete critical orbit consting of at most two points, or C has genus 1 and T has no singular points. In all remaining cases, X and U are hyperbolic Riemann surfaces, so we may uniformise U via an unbranched covering map $\pi_U : \mathbb{H} \to U$, which is the quotient map for some Fuchsian group $G_0 < \mathrm{PSL}_2(\mathbb{R})$. Then $\pi_1 \circ \pi_U$ and $\pi_2 \circ \pi_U$ are two different unbranched covering maps of X; they are the quotient maps for Fuchsian groups G_1 and G_2 respectively, both containing G_0. Because $\mathbb{H}$ is simply-connected, we can lift $\pi_2 \circ \pi_U$ to an analytic map $M : \mathbb{H} \to \mathbb{H}$. Because $\pi_2 \circ \pi_U$ is an unbranched covering map, the lift M is a Möbius automorphism of $\mathbb{H}$, i. e. an element of

$\mathrm{PSL}_2(\mathbb{R})$. The following diagram commutes:

$$\begin{tikzcd}
\mathbb{H} \arrow[r, "M"] \arrow[d, "\pi_U"'] & \mathbb{H} \arrow[d, "\pi_U"] \\
U \arrow[r, "f"] \arrow[d, "\pi_1"'] \arrow[dr, "\pi_2"] & U \arrow[d, "\pi_1"] \\
X \arrow[r, "T"] & X
\end{tikzcd}$$

It follows that

$$G_2 = M^{-1}\, G_1\, M\,.$$

Irreducibility implies that π_1 and π_2 do not simultaneously factor through any quotient of U, so

$$G_0 = G_1 \cap G_2\,.$$

If the bidegree of T is (m, n) then G_0 has index $n = \deg \pi_1$ in G_1 and index $m = \deg \pi_2$ in G_2. When the intersection of two subgroups of $\mathrm{PSL}_2(\mathbb{R})$ or of $\mathrm{PSL}_2(\mathbb{C})$ is of finite index in each, they are said to be *commensurable*.

The hyperbolic area of U is n times that of X because π_1 is an unbranched covering map; similarly it is m times the area of X. These areas are finite, so $m = n$.

Conversely, suppose we are given a finitely generated Fuchsian group of the first kind, i.e. $G_1 < \mathrm{PSL}_2(\mathbb{R})$ such that $V = \mathbb{H}/G_1$ is a compact surface minus finitely many punctures. Also suppose that $M \in \mathrm{PSL}_2(\mathbb{R})$ is such that

$$G_0 = M^{-1} G_1 M \cap G_1$$

has finite index both in G_1 and in $G_2 = M^{-1} G_1 M$. Let the quotient map for G_1 be $q_1 : \mathbb{H} \to V$. Then

$$q_2 = q_1 \circ M : \mathbb{H} \to \mathbb{H}/G_1$$

is a quotient map for G_2. Define $U = \mathbb{H}/G_0$. Both q_1 and q_2 factor through the quotient map $\pi_U : \mathbb{H} \to U$, so there exists unique unbranched covering maps

$$\pi_i : U \to V\,, \quad i = 1, 2,$$

such that

$$\pi_i \circ \pi_U = q_i\,.$$

Now we have a correspondence

$$T : \pi_2 \circ \pi_1^{-1} : V \to V\,.$$

Now fill in the punctures of U and V to obtain compact surfaces U' and V'. Apply the Removable Singularity Theorem for punctured Riemann surfaces to extend the maps π_1 and π_2 to *branched* coverings from U' to V', thereby extending T to a correspondence $T : V' \to V'$. Such correspondences (and formal sums of them) are called *modular*[3]. The above working shows that a correspondence on a compact Riemann surface is modular if and only if it is

[3] In [26], *correspondances modulaires* are defined in a similar fashion beginning with a more general simply-connected non-compact algebraic group in place of $\mathrm{PSL}_2(\mathbb{R})$

critically finite with a finite invariant set of at least three points; this is the content of [19, Prop. 1].

One useful consequence of the above analysis is that we can easily describe the n^{th} iterate of T. For $n \geq 0$ define

$$\mathcal{S}[n] = \{g_n \circ M \circ g_{n-1} \circ M \circ \cdots \circ M \circ g_1 \circ M \circ g_0 \ : \text{all } g_i \in G_1\}\,.$$

$\mathcal{S}[n]$ is the set of lifts of branches of T^n to $\mathbb{H}$, (w. r. t. the quotient map $q_1 = \pi_1 \circ \pi_U$). The branches themselves are of the form $q_1 \circ \gamma \circ q_1^{-1}$, where $\gamma \in \mathcal{S}[n]$. We will also need a notation for the semigroup of all lifts of branches of iterates of T, so we define

$$\mathcal{S} = \bigcup_{n=0}^{\infty} \mathcal{S}[n]\,.$$

Finally we must decide when the pairs (G_1, M) and (G_1', M') give rise to conjugate correspondences. There are two choices to be made to recover such a pair from a correspondence, namely the choice of the covering group from a family of conjugate subgroups of $\mathrm{PSL}_2(\mathbb{R})$ and the choice of the lift M. Thus the pairs (G_1, M) and (G_1', M') give rise to conjugate correspondences precisely when there exist $h \in \mathrm{PSL}_2(\mathbb{R})$ and $g \in G_1$ such that $G_1' = h^{-1}\, G_1\, h$ and $M' = h^{-1}\, g\, M\, h$. Bullett remarks that composing M with an element of the normaliser of G_1 in $\mathrm{PSL}_2(\mathbb{R})$ leaves G_2 unchanged, but may alter the correspondence. So it is not enough merely to specify the groups G_1 and G_2; we really need to know the conjugating element M.

3.2.4 Lifting the AGM correspondence

The AGM correspondence $z \mapsto \frac{z+1}{2\sqrt{z}}$ is critically finite: the set $\{0, 1, -1, \infty\}$ is completely invariant. The point 1 is a critical point for both branches, the values being 1 and -1. The two branches at -1 are analytic but happen both to take the value 0 there. 0 maps to ∞ in a forward-singular fashion and ∞ maps to itself in a forward-singular fashion. The covering group of $\widehat{\mathbb{C}} \setminus \{0, 1, -1, \infty\}$ is

$$G_1 = \Gamma_0(4,2) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : \begin{array}{c} c \equiv 0 \pmod 4 \\ b \equiv 0 \pmod 2 \end{array} \right\} \Big/ \pm I\,.$$

The map M is $z \mapsto z/2$, represented by $\begin{pmatrix} 1/\sqrt{2} & 0 \\ 0 & \sqrt{2} \end{pmatrix}$, so that

$$G_0 = \Gamma(4) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \pmod 4 \right\} \Big/ \pm I\,.$$

Notice that the AGM is conjugate to its own inverse via the automorphism $z \mapsto \frac{z+1}{z-1}$ of $\widehat{\mathbb{C}}$. This automorphism lifts to $h : z \mapsto -1/z$ on $\mathbb{H}$. Note that h conjugates G_1 to $G_2 = \Gamma_0(2,4)$, conjugates G_2 to G_1 and conjugates M to M^{-1}.

3.3 Invariant measures of critically finite correspondences

Let $T : C \to C$ be a critically finite correspondence. The restriction of T to the complete critical orbit A gives a Markov chain with finitely many states, so there is at least one invariant measure supported on A; this need not also be invariant for T^{-1}. More interestingly we can restrict T to the complement $X = C \setminus A$, which has finite hyperbolic area $2\pi(2g - 2 + n)$, where g is the genus of C and $n = |A|$. Observe that the hyperbolic area measure is invariant for both T and T^{-1}, because the lift M is a hyperbolic isometry. Dividing the hyperbolic area measure by its total mass, we obtain a doubly invariant probability measure μ_h on X.

Let T be a holomorphic correspondence on a quasi-projective curve C, i.e. a compact Riemann surface punctured at finitely many points. In [26] it is shown that if T_* preserves the hyperbolic area measure associated to a punctured surface $C \backslash \{z_1, \ldots, z_n\}$ then T is in fact a modular correspondence.

3.3.1 Lattès rational maps

Before proving any theorems, let us draw an analogy between critically finite correspondences and certain special rational maps, originally constructed by Lattès. Let $\wp$ be the Weierstrass function of the elliptic curve $E = \mathbb{C}/\Lambda$, where Λ is a lattice in $\mathbb{C}$. Then $\wp$ is a degree 2 branched cover of $\widehat{\mathbb{C}}$, with four critical points at the half-lattice points. $\wp$ is an even function. Now let $z \mapsto az$ induce an endomorphism $\varphi : E \to E$ (for which we need $a\Lambda \subset \Lambda$). Suppose that $|a| > 1$ so that $\deg \phi > 1$. Then for any $x \in \widehat{\mathbb{C}}$, $\wp^{-1}(x)$ consists of two points $\pm t \in E$, and $\wp(at) = \wp(-at) = R(x)$ defines a rational map R. The Lyubich measure for R is the $\wp$-image of the normalised Euclidean area measure on E; in particular its support, the Julia set of R, is $\widehat{\mathbb{C}}$. The best way to think about the action of the Perron–Frobenius operator for R^{-1} is in terms of the Perron-Frobenius operator associated to the correspondence $z \mapsto a^{-1}z$ on E.

3.3.2 Galois correspondences

There is an easy way to produce examples of critically finite correspondences. Let $R : C \to C'$ be any branched covering map between compact Riemann surfaces and define the Galois correspondence of R by $T_R : p \mapsto R^*(R(p))$, a correspondence on C. The images of p under T_R are all the points that map to the same point as does p under R. Since R has only finitely many critical points and each grand orbit of T_R is a fibre of R (hence finite), T_R is critically finite. T_R will generally not be irreducible (it has the identity correspondence as a component), but each irreducible component T of T_R will also be a critically finite correspondence. In the terminology of [26], these are *interior* modular correspondences, while all other irreducible modular correspondences are *exterior*. A modular correspondence specified by (G, M) (as in §3.2.3) is interior if and only if the group generated by G and M is discrete. Thus an interior correspondence T has *finite dynamics*, meaning that the set of irreducible correspondences that arise as Zariski components of the graphs of iterates of T is finite. There are at most $\deg(R)$ such components; indeed at each point there are at most $\deg(R)$ branches of iterates of T.

To avoid confusion, let us mention that R need not be a Galois covering (also called a regular or a normal covering, meaning that R is the quotient map for a group of automorphisms of C). It is such a covering if and only if the components of T_R are all single-valued. So from the point of view of correspondences, the situation is most interesting when R is not a Galois covering.

3.3.3 A dynamical dichotomy

Theorem 3.2.

Let $T : C \to C$ be an irreducible critically finite correspondence with complete critical orbit A. Suppose that the Riemann surface $X = C \setminus A$ is hyperbolic. Then exactly one of the following holds.

1. *M and G_1 generate a dense subgroup of $\mathrm{PSL}_2(\mathbb{R})$. In this case the the semigroup generated by M and G_1 is also dense in $\mathrm{PSL}_2(\mathbb{R})$. In particular T has an infinite forward orbit, and if $A \neq \emptyset$ then each point of X has a forward orbit under T whose closure in C meets A.*

2. *M and G_1 generate a discrete subgroup of $\mathrm{PSL}_2(\mathbb{R})$. In this case, T is a component of a Galois correspondence, arising from a map of degree at most $42(2g - 2 + |A|)$.*

The basic dichotomy in Theorem 3.2 that $\langle G_1, M\rangle$ is either discrete or dense is not new. [19] points out that the existence of an infinite orbit of T implies that the group generated by M and G_1 is dense in $\mathrm{PSL}_2(\mathbb{R})$ and hence by a deep theorem of Margulis that G_1 is a weakly arithmetic group, i. e. some conjugate of G_1 is commensurable with $\mathrm{PSL}_2(\mathbb{Z})$. It follows that G_1 contains parabolic elements, and hence that $A \neq$, so in fact the last part of item 1 always applies. However, we did not wish to rely on Margulis' result in what is otherwise a fairly elementary theorem. The parts of Theorem 3.2 that may be new are the result that the semigroup generated by G_1 and M is dense and the result that shows that the existence of a grand orbit with more than $42(2g - 2 + |A|)$ points guarantees the existence of an infinite orbit.

Proof. Consider the semigroup $\mathcal{S}$ generated by G_1 and M, and its closure $\overline{\mathcal{S}}$. We will prove below that $\overline{\mathcal{S}}$ is in fact a subgroup of $\mathrm{PSL}_2(\mathbb{R})$, and hence is also the closure of the subgroup $\langle G_1, M\rangle$ of $\mathrm{PSL}_2(\mathbb{R})$ generated by G_1 and M. Since $\overline{\mathcal{S}}$ is closed, it is in fact a Lie subgroup of $\mathrm{PSL}_2(\mathbb{R})$ [66, Theorem VII.2.5]. Furthermore, $\overline{\mathcal{S}}$ is invariant under conjugation by any element of G_1. In particular there are two hyperbolic elements of G_1 with disjoint fixed point sets in $\widehat{\mathbb{C}}$ that both conjugate $\overline{\mathcal{S}}$ to itself. The connected component of $\overline{\mathcal{S}}$ containing the identity is invariant under conjugation by the same elements. The connected Lie subgroups of $\mathrm{PSL}_n(\mathbb{R})$ were classified in [36]; in the simple case of $\mathrm{PSL}_2(\mathbb{R})$, the only possibilities are the trivial subgroup, the one-parameter groups with either a fixed point in $\overline{\mathbb{H}}$ or a fixed geodesic in $\mathbb{H}$, and $\mathrm{PSL}_2(\mathbb{R})$ itself. Of these, only the trivial subgroup and $\mathrm{PSL}_2(\mathbb{R})$ are invariant under conjugation by two hyperbolic elements that have disjoint fixed point sets. Hence $\overline{\mathcal{S}}$ is either discrete or all of $\mathrm{PSL}_2(\mathbb{R})$.

In the discrete case, we must have $\overline{\mathcal{S}} = \langle G_1, M\rangle$. Hurwitz's Theorem tells us that the hyperbolic area of a fundamental domain for this group is at least $\pi/21$, so

$$[\langle G_1, M\rangle : G_1] \leq 42(2g-2+|A|)\,.$$

The correspondence T is a component of the Galois correspondence arising from the map

$$R : \mathbb{H}/G_1 \to \mathbb{H}/\langle G_1, M\rangle\,.$$

It remains to prove that $\overline{\mathcal{S}}$ is a subgroup of $\mathrm{PSL}_2(\mathbb{R})$. Using continuity of right-multiplication, Lemma 3.3 below implies that the inverse of any element of $\mathcal{S}$ lies in $\overline{\mathcal{S}}$. Since taking the inverse in $\mathrm{PSL}_2(\mathbb{R})$ is continuous, $\overline{\mathcal{S}}$ is also closed under inversion. □

Remark. [25] gives a proof that if a closed monoid in a simply-connected semi-simple Lie group of rank one contains a lattice then it is in fact a subgroup. This result could be used in place of Lemma 3.3 in the proof of Theorem 3.2. On the other hand the proof below of Lemma 3.3 would also provide a proof of the above-mentioned result about closed monoids that makes direct use of the finite volume of the quotient by the lattice, whereas [25] uses the hyperbolic metric, the classification of isometries and some clever conjugation arguments.

Lemma 3.3 (Recurrence for critically finite correspondences). *For any $g \in \mathcal{S}$, the identity element I of $\mathrm{PSL}_2(\mathbb{R})$ is in the closure of $\mathcal{S} \circ g$.*

Remark: This lemma has consequences for the Markov chain $(X_n)_{n=0}^{\infty}$ associated to T. In particular, for any $\epsilon > 0$, and any finite orbit segment $x_0, x_1, \ldots, x_n$ (satisfying $x_{i+1} \in T(x_i)$ for $i = 0, \ldots n-1$) we have

$$\mathbb{P}(\exists\, k > n : d(X_k, X_0) < \epsilon \,|\, X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) > 0\,.$$

Proof. Make geodesic branch cuts on $X = C \setminus A$ to leave a connected and simply-connected domain U. On U, the correspondence T is described by m branches $f_1, \ldots, f_m$, each of which is a local isometry from U into X. Because the branch cuts have total hyperbolic area 0, the branches f_i are well-defined modulo 0 as Borel-measurable maps from U to U, (where modulo 0 refers to the normalised hyperbolic area measure μ_h). The individual maps f_i need not preserve μ_h, but nevertheless μ_h is an invariant probability measure for the Borel-measurable IFS $\mathcal{F}$, whose maps are these m branches of T on U, assigned equal probabilities. Let T_1U be the unit tangent bundle of U (where U has the Riemannian metric inherited from the hyperbolic metric on X). Let $\tilde{f}_i$ be the map (f_i, Df_i) induced by f_i on T_1U, and let $\tilde{\mathcal{F}}$ be the associated measurable IFS. Write λ for the normalised Lebesgue measure on each fibre of T_1U, and $\lambda \times \mu_h$ for the product measure on T_1U. Let $B_{1/m}$ be the Bernoulli measure associated to the probability vector $(1/m, \ldots, 1/m)$. Then $B_{1/m} \times (\lambda \times \mu_h)$ is an invariant measure for the associated skew-product map

$$\tau : \{1, \ldots, m\}^{\mathbb{N}} \times T_1U \to \{1, \ldots, m\}^{\mathbb{N}} \times T_1U \,.$$

Let $\tilde{V}$ be any neighbourhood of the identity in $\mathrm{PSL}_2(\mathbb{R})$. Identify $\mathrm{PSL}_2(\mathbb{R})$ with $T_1\mathbb{H}$ by fixing some base point $(\tilde{x}_0, \tilde{v}_0) \in T_1\mathbb{H}$ and making the identification

$$\gamma \in \mathrm{PSL}_2(\mathbb{R}) \quad \leftrightarrow \quad (\gamma(\tilde{x}_0), D\gamma(\tilde{v}_0)) \,.$$

We may assume that $\tilde{x}_0$ covers a point $x_0 \in U$ which is not mapped onto any of our branch cuts under any branch of any iterate of T. The covering map $\pi_1 : T_1\mathbb{H} \to T_1C$ maps $\tilde{V}$ homeomorphically onto a neighbourhood V of a point (x_0, v_0) in T_1U.

The element $g \in \mathcal{S}$ may be represented as

$$g = \gamma_n \circ M \circ \gamma_{n-1} \circ M \circ \cdots \circ M \circ \gamma_0 \,,$$

for some (not necessarily unique) elements $\gamma_0, \ldots, \gamma_n \in G_1$. Choose one such representation. Making V smaller if necessary, this representation corresponds to applying the composition of some particular sequence of n of the maps $\tilde{f}_i$ to the neighbourhood V. That sequence defines a cylinder set $Cyl(g) \subset \{1, \ldots, m\}^{\mathbb{N}}$ with Bernoulli measure m^{-n}. Making V smaller if necessary we may assume that V is an *open* neighbourhood of (x_0, v_0). Then

V is measurable and $\lambda \times \mu_h(V) > 0$. It follows that

$$B_{1/m} \times (\lambda \times \mu_h)\,(Cyl(g) \times V) > 0\,.$$

We now apply Poincaré's Recurrence Theorem to the skew-product map τ for $\tilde{\mathcal{F}}$. It says that there exists a point in $Cyl(g) \times V$ that returns to $Cyl(g) \times V$ after some positive number of iterations of the skew-product map τ. In particular, there is an element $h \in \mathcal{S}$ such that $h \circ g \in \tilde{V}$. The lemma follows. □

3.3.4 Asymptotic stability

If $T : C \to C$ is an interior correspondence then it has infinitely many invariant probability measures. On the other hand if $T : C \to C$ is an irreducible critically finite correspondence with an infinite orbit, then Theorem 3.2 and Lemma 3.3 give us some topological information about the possible orbits of the Markov chain of T, but we can use them to prove more.

Theorem 3.4.
Let $T : C \to C$ be a critically finite correspondence with complete critical orbit A, such that the Riemann surface $X = C \setminus A$ is hyperbolic. Suppose that T is not a Galois correspondence, or equivalently that T has an infinite orbit. Then the normalised hyperbolic area measure is the unique invariant probability measure for the restriction of T to X. Furthermore the restriction of T to X is asymptotically stable: for any Borel probability measure ν on X, we have $T^n_* \nu \to \mu_h$ as $n \to \infty$.

This theorem has now appeared in a preprint by Clozel and Otal [25], of which we became aware only after producing the proof below. Although we cannot claim priority for the result, we have kept it in this thesis because our proof is quite different from theirs, at least on the surface. The proof in [25] proves the locally uniform convergence of the sequence $f_n = (T^*)^n f$ to a constant function, for each bounded and uniformly continuous function f on C. A simple coupling argument is used to show that the f_n are uniformly continuous, and hence have a convergent subsequence. Then the density of $\langle G_1, M\rangle$ in $\mathrm{PSL}_2(\mathbb{R})$ and some inequalities between L^2 norms are used to show that any subsequential limit of the f_n must be constant.

As remarked after the statement of Theorem 3.2, there cannot exist a correspondence with an infinite orbit but without singular points on a compact hyperbolic Riemann surface. So A is non-empty and X has at least one cusp. Since we do not know an elementary proof of this, we give two versions of our proof of Theorem 3.4. The simpler of the two makes the assumption that A is non-empty.

Proof of Theorem 3.4. We use the method of the Kantorovich metric to obtain the asymptotic stability of T, which of course implies uniqueness of the invariant measure. The method of proof is almost identical to the argument of §2.3.1. Let $\nu \neq \mu_h$ be a Borel probability measure on X. Let d be the hyperbolic metric on X, and consider a modified metric $c = \varphi(d)$, as provided by Lemma 1.8, such that $\mu, \nu \in \mathcal{P}_c$ and (X, c) is a proper metric space. We will show that

$$K_c\left(T_*(\nu), T_*\left(\mu_h\right)\right) \leq K_c\left(\nu, \mu_h\right)\ , \tag{3.1}$$

$$\exists n \in \mathbb{N} \text{ such that } \quad K_c\left(T_*^n(\nu), T_*^n\left(\mu_h\right)\right) < K_c\left(\nu, \mu_h\right)\ . \tag{3.2}$$

This plays the rôle of Lemma 2.7. After this, the argument given in §2.3.1

works almost word-for-word, subject to replacing $\mathcal{F}_*$ by T_* and changing the hypotheses of Lemma 2.8 in an obvious way. The variable n in equation (3.2) does not interfere with the proof.

Let us prove (3.1). The infimum in the definition of $K_c\left(\nu, \mu_h\right)$ is attained by some measure m on $X \times X$, whose marginals are ν and μ_h. To each pair $(x, y) \in X \times X$ we associate a shortest directed geodesic segment $\gamma(x, y)$ from x to y in the hyperbolic metric on X. We can do this in a measurable fashion. Pushing forward by this map takes m to a measure m' on the space $\mathcal{G}X$ of directed geodesic segments in X.[4] The branches of T map geodesic segments to geodesic segments, so by weighting the branches each with probability $\frac{1}{m}$ (where the bidegree of T is (m, m)), we obtain a Markov operator P acting on the space of measures $\mathcal{M}(\mathcal{G}X)$. Project $P(m')$ back to a measure $P(m)$ on $X \times X$, whose marginals are $T_*(\nu)$ and $T_*(\mu_h)$. We have

$$\begin{aligned} K_c\left(T_*(\nu), t_*\left(\mu_h\right)\right) \leq \int c(x, y)\, d(P(m)(x, y)) & \\ = \quad \int c(x, y)\, dm(x, y) = K_c\left(\nu, \mu_h\right) \ . \end{aligned}$$

Now we prove (3.2), assuming that A is non-empty. Because $\mathcal{S}$ is dense in $\mathrm{PSL}_2(\mathbb{R})$, at least one of the images of any given minimal geodesic segment under some iterate of T must fail to be a minimal geodesic segment because it winds around a puncture and intersects itself. We produce a modifed Markov operator Q by mapping each geodesic segment to the weighted formal sum of *minimal* geodesic segments that join the endpoints of each of its images

[4] Of course $\mathcal{G}X$ is essentially the tangent bundle of X, but it is better here to think of geodesic segments than of tangent vectors.

under T. Then a countable additivity argument shows that for some $n \in \mathbb{N}$,

$$\int c(x,y)\, d\left(Q^n(m)\right)(x,y) \quad < \quad \int c(x,y)\, d\left(P^n(m)\right)(x,y)\,,$$

which implies (3.2).

Finally we prove (3.2) without assuming that A is non-empty. This proof involves a more substantial modification of the operator P. Since $\nu \neq \mu_h$, the measure m is not supported on the diagonal in $X \times X$. By a countable covering argument, there exists $(x_0, y_0) \in X \times X$ such that

$$m\left(B(x_0,\epsilon), B(y_0,\epsilon)\right) = \eta > 0\,, \quad \text{where } \epsilon = d(x_0,y_0)/6\,.$$

Recall that we defined $\mathcal{S}[n]$ to be the set of lifts of branches of T^n. Now we define

$$\mathcal{R}[n] = \{K^{-1} \circ K' : K, K' \in \mathcal{S}[n]\}\,,$$

and

$$\mathcal{R} = \cup_{n=0}^{\infty} \mathcal{R}[n]\,.$$

We may think of $\mathcal{R}[n]$ as the set of all geometrical relationships between lifts of different branches of T^n with respect to q_1.

Lemma 3.5.

Under the hypotheses of Theorem 3.4, $\mathcal{R}$ is dense in $\mathrm{PSL}_2(\mathbb{R})$.

Because R is dense in $\mathrm{PSL}_2(\mathbb{R})$, there exist $K', K \in \mathcal{S}[n]$ for some $n \in \mathbb{N}$, such that

$$d\left(K^{-1} \circ K'(x_0), y_0\right) < \epsilon \quad \text{and} \quad d\left(K^{-1} \circ K'(y_0), x_0\right) < \epsilon\,.$$

Hence

$$(K(x_0), K'(y_0)) < \epsilon \quad \text{and} \quad d\left(K(y_0), K'(x_0)\right) < \epsilon\,. \tag{3.3}$$

Let A be the set of minimal geodesics that begin in $B(x_0, \epsilon)$ and end in $B(y_0, \epsilon)$. Let $s \in A$. The images of s under the branches of T^n corresponding to K and K' are geodesic segments s_1 and s_2, both of the same length as s, which is at least 4ϵ. Let s_1' be the minimal geodesic segment from the start point of s_1 to the end point of s_2 and let s_2' be the minimal geodesic segment from the start point of s_2 to the end point of s_1. Both s_1' and s_2' have length at most 3ϵ, because of (3.3) and the triangle inequality. Replacing s_1 and s_2 by s_1' and s_2' whenever they arise in this way yields a probability measure on $\mathcal{G}X$ whose endpoint projections are the same as those of $P^n(m')$. However, the integral of the length of its segments (in the c-metric) is strictly smaller than the corresponding integral for $P^n(m')$. This suffices to prove (3.2) and completes the proof of Theorem 3.4. □

Proof of Lemma 3.5. Note that $\mathcal{R}[n]$ is closed under taking of inverses in $\mathrm{PSL}_2(\mathbb{R})$. We will show that $\overline{\mathcal{R}}$ is a subgroup of $\mathrm{PSL}_2(\mathbb{R})$. For this it suffices to prove that for any $F \in \mathcal{R}[m]$ and $G \in \mathcal{R}[n]$, $G \circ F$ is in $\overline{\mathcal{R}}$. To this end, let U be any neighbourhood of $G \circ F$ in $\mathrm{PSL}_2(\mathbb{R})$. From the definition of $\mathcal{S}[n]$ we see that for any $H \in \mathcal{S}[m]$ and $G = L^{-1} \circ L' \in \mathcal{R}[n]$, we have

$$H^{-1} \circ G \circ H = (L \circ H)^{-1} \circ (L' \circ H') \in \mathcal{R}[n+m]\,. \tag{3.4}$$

Given the neighbourhood $U \circ F^{-1}$ of G in $\mathrm{PSL}_2(\mathbb{R})$, there exists a neighbourhood V of the identity such that for all $H \in V$ we have $H^{-1} \circ G \circ H \in U \circ F^{-1}$. Indeed, the function $H \mapsto H^{-1} \circ G \circ H$ is continuous and maps the identity to G. Now suppose $F = K^{-1} \circ K'$, where $K, K' \in \mathcal{S}[m]$. We can write

$$F = (J \circ K)^{-1} \circ (J \circ K')\,,$$

where $J \in \mathcal{S}[N]$ for some large N and $J \circ K \in V$. This is possible because $\mathcal{S}$ is dense in $\mathrm{PSL}_2(\mathbb{R})$; we just have to choose $J \in V \circ K^{-1}$. Then U contains

$$(J \circ K)^{-1} \circ G \circ (J \circ K) \circ F = (L \circ J \circ K)^{-1} \circ (L' \circ J \circ K)\,,$$

which is in $\mathcal{R}[n+m+N]$. Since U was arbitrary, we see that $G \circ F \in \overline{\mathcal{R}}$.

We now know that $\overline{\mathcal{R}}$ is a closed (hence Lie) subgroup of $\mathrm{PSL}_2(\mathbb{R})$, and that it is invariant under conjugation by any element of $\mathcal{S}$, hence also by any element of $\overline{\mathcal{S}} = \mathrm{PSL}_2(\mathbb{R})$. Therefore $\overline{\mathcal{R}} = \mathrm{PSL}_2(\mathbb{R})$ as required. This completes the proof of Lemma 3.5. □

3.4 Forward critically finite correspondences

3.4.1 The inverse of a hyperbolic rational map

Consider a rational map T of degree $d \geq 2$ acting on $\widehat{\mathbb{C}}$. Let C be the set of critical points of T and let $PC = \cup_{n=1}^{\infty} T^n(C)$ be the postcritical set. Suppose for convenience that the Julia set $\mathcal{J}$ of T does not contain the point ∞ (replace T by a conjugate if necessary). When $\overline{PC} \cap \mathcal{J} = \emptyset$, we say that T is *hyperbolic*. [23, §V.2] gives basic results about hyperbolic rational maps, in particular that T is hyperbolic if and only if every critical point lies in the basin of an attracting periodic orbit of T. We will exclude from the present discussion the simple cases when T is conjugate to $z \mapsto z^d$ or to $z \mapsto z^{-d}$. Then PC contains at least three points, so $U = \widehat{\mathbb{C}} \backslash \overline{PC}$ is a hyperbolic domain. Since U contains no critical values of T, near any point of U there are d distinct analytic branches of T^{-1}, which can be analytically continued along any path in U, although they cannot be defined as single-valued maps right around any postcritical point. However, we can choose a single-valued lift of each of these branches to $\mathbb{D}$, obtaining analytic maps $f_1, \ldots, f_d : \mathbb{D} \to \mathbb{D}$ that make the following diagram commute:

$$\begin{tikzcd}
\mathbb{D} \arrow[r, "f_i"] \arrow[d, "\pi"'] & f_i(\mathbb{D}) \arrow[r, "\text{inclusion}"] \arrow[d, "\pi"'] & \mathbb{D} \arrow[d, "\pi"'] \\
U & T^{-1}(U) \arrow[l, "T"'] \arrow[r, "\text{inclusion}"] & U
\end{tikzcd}$$

Because the iterated pre-images of any point not in the exceptional set E accumulate on $\mathcal{J}$, there must be points in $T^{-1}(\overline{PC}) \setminus \overline{PC}$. Since $f_i(\mathbb{D})$ omits every lift of any such point, f_i is not an isometry, so the Schwarz–Pick Lemma tells us that f_i is strictly distance-decreasing. It follows that T is locally expanding w. r. t. the hyperbolic metric on U. In particular, on the compact and T-invariant set $\mathcal{J} \subset U$, we find that the hyperbolic derivative satisfies $T^{\#} \geq \rho > 1$. Since $\mathcal{J}$ is compact, the Euclidean metric on $\mathbb{C}$ is comparable on $\mathcal{J}$ to the hyperbolic metric on U. Hence T is expanding on $\mathcal{J}$ in the sense that there are constants $\alpha > 0$ and $A > 1$ such that $|(T^n)'(z)| > aA^n$ for all $z \in \mathcal{J}$ and all $n \in \mathbb{N}$. In fact this *only* happens when T is hyperbolic, so hyperbolic rational maps are also called *expanding* rational maps.

At this point we depart from the treatment given in [23], by considering the iterated function system $\mathcal{F}$ associated to the lifts f_i. Let $F_1, F_2, F_3, \ldots$ be a sequence of i. i. d. random maps from $\mathbb{D}$ to itself, with $\mathbb{P}(F_n = f_i) = \frac{1}{d}$.

Choose $x_0 \in \mathbb{D}$ with $z_0 = \pi(x_0)$. Now for each $n \in \mathbb{N}$, set

$$x_n = F_n \circ \cdots \circ F_1(x_0).$$

Then $z_n = \pi(x_n)$ reconstructs the Markov chain (z_n) described in 3.1.5. As remarked above, it is known that the distributions μ_n of $z_n = \pi(x_n)$ converge weakly to a probability distribution on U. We now ask whether the distributions ν_n of x_n necessarily converge to a probability distribution on $\mathbb{D}$. It is not obviously so. Indeed, $\mathbb{D}$ covers U with infinitely many sheets, so the space of measures that project via π to the limit measure μ is not even compact. However, as a corollary of Theorem 2.16, we will prove that in fact the ν_n do converge weakly to a probability measure ν on $\mathbb{D}$:

Proposition 3.6.
Let $\mathcal{F}$ be any IFS on $\mathbb{D}$ which is a lift of the restriction of T^{-1} to $\widehat{\mathbb{C}} \setminus \overline{PC}$. Then $\mathcal{F}$ is asymptotically stable.

The maps f_i all avoid all the lifts of points in $T^{-1}(\overline{PC}) \setminus \overline{PC}$. Unfortunately it is not true that the complement of these points is a Bloch domain in $\mathbb{D}$, because in U there are points of U (near the postcritical set) that in the hyperbolic metric on U are arbitrarily far from all such points. Claim: there is a subdomain $V \subset U$ such that $T^{-1}(V)$ is relatively compact in V. It follows that $\pi^{-1}(T^{-1}(V))$ is a Bloch subdomain of $\pi^{-1}(V)$ (although not relatively compact in $\pi^{-1}(V)$, and is forward-invariant for each of the maps $f_1, \ldots, f_d$. Theorem 2.16 now tells us that the restriction of $\mathcal{F}$ to $\pi^{-1}(V)$ is asymptotically stable. Of course its invariant probability measure is also an invariant probability measure for $\mathcal{F}$, so $\mathcal{F}$ itself is asymptotically stable, by Ambroladze's Theorem.

To prove the claim, we will construct the domain V by specifying its complement, a closed forward-invariant neighbourhood of the postcritical set. There are two types of critical point to consider: the periodic ones and the non-periodic ones. For any periodic cycle of T, say $z_1, z_2, \ldots, z_k$, containing at least one critical point of T, there exist small closed discs B_i about z_i such that $T(B_i)$ is contained in the interior of B_{i+1} (subscripts modulo k). Any non-periodic critical point is attracted under iteration of T to some attracting cycle [23, §V]. Fix such a cycle. Again we can find a cycle of closed discs B_i about the points of the cycle, each mapping strictly inside the next. Now for each critical point z attracted to that cycle, we have $T^m(z)$ in the interior of B_j, for some m and j. Then we can find a sequence of closed discs about the points $T(z), T^2(z), \ldots T^{m-1}(z)$, each mapping strictly inside the next and the last mapping strictly inside B_j. Do this for all the critical points. There are only finitely many critical points (at most $2d - 2$), so we have now chosen a finite collection of closed discs, whose union contains the closure of the postcritical set. The domain V is now defined to be the complement of the union of all these discs; by construction $T^{-1}(V)$ is relatively compact in V, as required.

3.4.2 Forward critical finiteness

In §3.4.1 we were able to prove asymptotic stability of a certain correspondence by lifting its branches to the universal cover of a suitable subdomain and applying iterated function system results. We will try to extend this method to apply to a larger class of correspondences $T : C \to C$. We need to find a closed proper subset of C that is forward-invariant under T and contains all the backward-singular points.

Definition.
The correspondence T is *forward critically finite* when the union of the forward orbits of the backward-singular points is finite.

We will denote by PC the postcritical set, i.e. the union of the backward-singular points and their forward orbits. Note that every postcritically finite rational map is forward critically finite, although it is only be critically finite as a correspondence if it is conjugate to $z \mapsto z^n$ or $z \mapsto z^{-n}$.

For a forward critically finite correspondence T of bidegree (m, n) on a curve C, we may delete the set PC from C. The punctured Riemann surface $X = C \setminus PC$ is typically hyperbolic, uniformised by the unit disc $\mathbb{D}$, in which case we can choose lifts of the m branches of T^{-1} to m analytic maps $\mathbb{D} \to \mathbb{D}$. If we prove that the corresponding IFS is asymptotically stable, then as a corollary the restriction of T^{-1} to $C \setminus PC$ is asymptotically stable. If we can also show that any T^{-1}-invariant probability measure must assign mass zero to PC, it will follow that T^{-1} is asymptotically stable.

3.4.3 The tangential correspondence on a quartic

Let C be a smooth curve of degree 4 in $\mathbb{P}^2(\mathbb{C})$. Every such curve is the canonical curve of its underlying Riemann surface. In particular a compact Riemann surface of genus 3 has a unique embedding of this form, up to automorphisms of $\mathbb{P}^2(\mathbb{C})$. For each point $p \in C$, the tangent line to C at p cuts out a divisor $S(p)$ of degree 4 on C, and $T(p) = S(p) - 2p$ is an effective divisor of degree 2. The correspondence T is called the *tangential* *correspondence* on C. It has bidegree $(10,2)$, which is computed using the Riemann-Hurwitz formula in [37, p. 290].

A point $p \in C$ is a *flex* of C when $p \in T(p)$. A point at which the tangent to C meets C with multiplicity 4 is called a hyperflex. At a simple flex p, one branch of T has a fixed point at p, while the other branch has a critical point at p. These are the only critical points of T. Thus the backward-singular points of T are the points q such that $T(p) = p + q$, where p is a simple flex. If p is a hyperflex of C then the graph of T has an ordinary double point at (p,p); thus p is a fixed-point for both branches of T in a neighbourhood of p. In fact it is a repelling fixed point of each branch: by putting the curve C in a suitable normal form at p, we compute that the multipliers are $-1 \pm \sqrt{-2}$.

The forward-singular points of T are the points whose tangent line is a bitangent to C. We do not count the tangent at a hyperflex as a bitangent. Note that the set of forward-singular points of T is forward-closed under T. In fact they form super-repelling cycles of length 2.

Let B be the number of bitangent lines to C, H the number of hyperflexes, and F the number of simple flexes. We always have $F + 2H = 24$ and $B + H = 28$ (see [37, p. 549]). Since $H \leq 12$, we have $B \geq 16$. Let D be the desingularised graph of T. Then $T = \pi_2 \circ \pi_1^{-1}$, where $\pi_1 : D \to C$ has degree 2 and is branched over $2B$ points, and $\pi_2 : D \to C$ has degree 10 and is branched over F points. The fact that π_1 has degree 2 and is branched shows that D is irreducible. The genus of D is $g_D = (F + 42)/2 = B + 5$.

If C is a generic smooth plane quartic, then C has no hyperflexes, 24 simple flexes and 28 bitangent lines. In this case, T has 56 forward-singular points and 24 backward-singular points and the genus of its graph is 33.

The tangential correspondence on a smooth plane curve of degree at least 4 is never separable. Indeed, if $T(p)$ and $T(q)$ have two points in common (or a common multiple point) then the tangent lines are p and q are equal, so p and q are either equal or lie on a bitangent.

To ask for the tangential correspondence on a smooth plane quartic to be forward critically finite is to ask a great deal: the forward orbit of each of the simple flexes must be finite. One way to achieve this is dicussed in the next section.

3.4.4 The Fermat quartic

The Fermat quartic is the smooth plane curve C given in homogeneous co-ordinates by $X^4 + Y^4 + Z^4 = 0$. The Fermat quartic has 12 hyperflexes, namely those points described by homogeneous co-ordinates consisting of one zero and two eighth roots of unity. C has 16 other bitangent lines, meeting C in 32 distinct points. Note that *all* the flexes of C are hyperflexes, so T has no backward-singular points. Hence T is forward critically finite.

Because π_2 is unbranched, there is no obstruction to analytic continuation of any branch of T^{-1} along any path on C. Let $\pi_D : \mathbb{D} \to D$ be an analytic universal covering map. Then $\pi_C = \pi_2 \circ \pi_D$ is a universal covering map for C with deck transformation group Γ_C isomorphic to the fundamental group of C based at $p = \pi_C(0)$. Γ_D is a subgroup of Γ_C of index 10. Γ_C and Γ_D both act freely and properly discontinuously on $\mathbb{D}$. Because it has degree 2, π_1 is a normal covering, so there is a group G_C containing Γ_D with index 2 such that the quotient map for G_C is $\pi_1 \circ \pi_D$. However, G_C does not act freely on $\mathbb{D}$; it is a so-called *Belyi uniformisation* of C.

Consider the ten distinct branches of T^{-1} in a simply-connected neighbourhood of p in C. Choose lifts of each of these branches and analytically continue them to obtain analytic maps $f_1, \dots, f_{10} : \mathbb{D} \to \mathbb{D}$. We study the IFS $\mathcal{F}$ obtained from these maps with $p_1 = \cdots = p_{10} = 1/10$.

Lemma 3.7.

The maps $f_1, \dots, f_{10}$ are all Lipschitz with some constant $\rho < 1$, with respect to the hyperbolic metric on $\mathbb{D}$.

Proof. Each map f_i has critical points at each point of 32 orbits of Γ_D in the unit disc (These orbits depend on i.) Since D is compact, it has finite diameter in the hyperbolic metric. Thus each point $z \in \mathbb{D}$ is within distance $\operatorname{diam}(D)$ of a critical point of f_i. The present lemma now follows from the branched Schwarz lemma (Lemma 1.14). □

The monodromy action of the fundamental group of C based at p permutes the set of ten branches transitively because D is irreducible. Thus the maps f_i are related to each other by many relations of the form $f_i = g_{ij} \circ f_j \circ \gamma_{ij}$, where $\gamma_{ij} \in \Gamma_C$ and $g_{ij} \in G_C$.

Proposition 3.8.

The inverse of the tangential correspondence on the Fermat quartic is asymptotically stable, with exponential convergence with respect to either the Kantorovich or Prohorov metrics associated to the hyperbolic metric on C.

Proof. Let ν be any Borel probability measure on C. Then we can find a measure $\tilde{\nu}$ supported on a compact fundamental domain L for the group Γ_C, such that $(\pi_C)_* \tilde{\nu} = \nu$. From lemma 3.7, there is a compact set $L' \supset L$ that is forward-invariant under the IFS $\mathcal{F}$. It is now a standard result that there is an invariant law μ for $\mathcal{F}$ and that $\mathcal{F}_*^n\tilde{\nu} \to \mu$ exponentially as $n \to \infty$, with respect to either the Kantorovich metric or the Prohorov metric on the hyperbolic plane. The projection to C does not increase hyperbolic distances and therefore does not increase the Kantorovich or Prohorov metrics.

For an alternative proof not involving the lifts f_i, we could argue as follows. T^{-1} has at least one invariant probability measure μ, as explained in §3.1.5. Carry out the a coupling argument as in our proof of Theorem 3.4, using the Kantorovich metric K_d associated to the hyperbolic metric d on C. Lemma 3.7 leads easily to exponential convergence in the Kantorovich metric. Because the diameter of C is finite, the Kantorovich metric K_d is bi-Lipschitz equivalent to the Prohorov metric by Lemma 1.10. □

Let us point out that the Fermat quartic example is rather surprising. One might conclude that there should be many correspondences on hyperbolic surfaces with no singular points in one direction, by reasoning along the following lines. If R is a Riemann surface of genus $g \geq 2$, then R has many unbranched covers from surfaces of higher genus. If one of these surfaces also has a branched cover of R, then we may compose the inverse of the unbranched cover with the branched cover to obtain an irreducible correspondence T with no forward-singular points. Perhaps we could find such surfaces R by variation within the moduli space of marked surfaces? Unfortunately a dimension count shows that we should not expect to find such surfaces for dimensional reasons alone. Some extra symmetry seems to be required. The Fermat curve has a large automorphism group, of order 96. According to [51], there is only one other smooth plane quartic that has 12

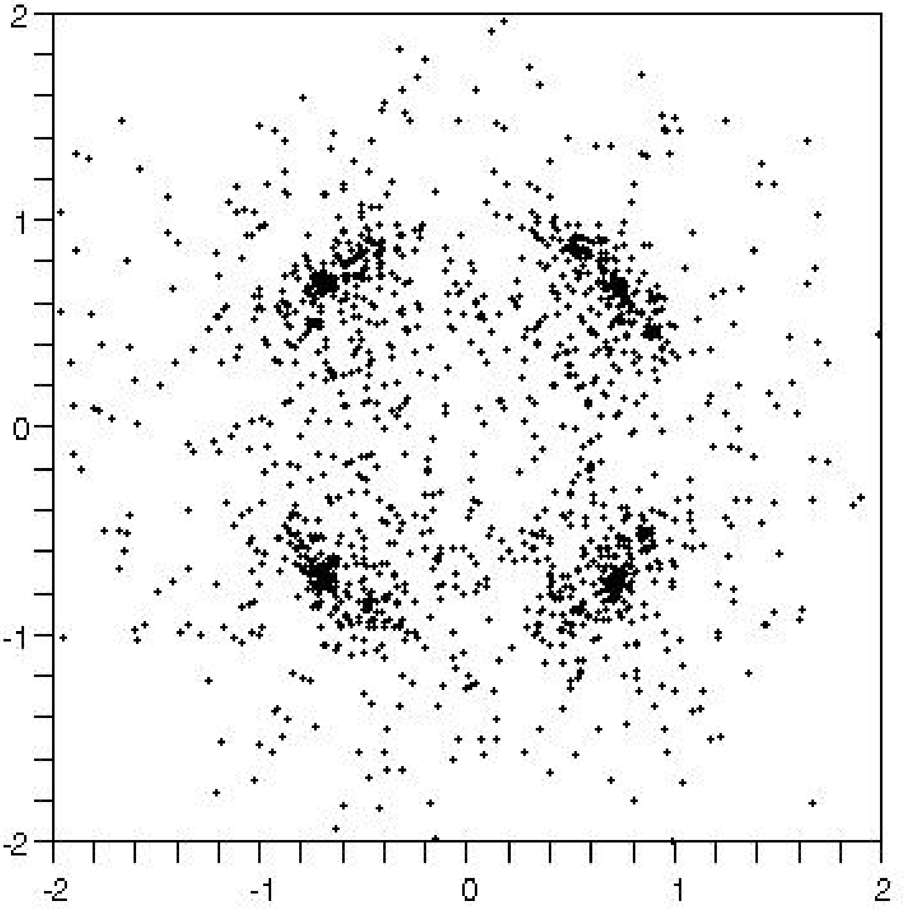

Figure 3.1: A Maple plot of the image under the map $\psi : (X : Y : Z) \mapsto X/Y$ of a random orbit segment of length 7000 for the inverse of the tangential correspondence on the Fermat quartic $X^4+Y^4+Z^4 = 0$. The map $\psi : C \to \widehat{\mathbb{C}}$ is the quotient map for a subgroup of order 4 of the automorphism group of the Fermat quartic, branched with index 4 over each of the fourth roots of -1, which explains why the empirical measure looks dense at those points.

hyperflexes and no simple flexes, namely the curve with 24 automorphisms given by

$$X^4 + Y^4 + Z^4 + 3\left(X^2Y^2 + X^2Z^2 + Y^2Z^2\right) = 0\,.$$

We might also try looking for pairs of Fuchsian groups such that $\mathbb{D}/\Gamma \cong \mathbb{D}/\Delta$, where Γ is free, Δ is not free, and Γ and Δ are commensurable.

3.4.5 Asymptotic stability of the inverse

Theorem 3.9.

Let the correspondence $T : C \to C$ be forward critically finite but not critically finite, and let A be the union of the forward orbits of all backward-singular points. Suppose that the Riemann surface $X = C \setminus A$ is hyperbolic. Then the restriction of T^{-1} to X is asymptotically stable.

Proof. If T has no backward-singular points but there are forward-singular points then the proof is just like the proof we used for the tangential correspondence on the Fermat quartic. If there are no singular points at all but there is an infinite orbit, then C cannot be hyperbolic, because it would have to be weakly arithmetic.

Now assume that there are some backward-singular points. Restrict T^{-1} to X. Lift all branches to the universal cover of X. These give analytic maps which do not increase the hyperbolic distance. There is an invariant measure for T^{-1}, as explained in §3.1.5.

We use the Kantorovich metric coupling argument on X, as in our proof of Theorem 3.4. We need to prove that the Kantorovich distance with respect to modifications of the hyperbolic metric on X is strictly decreased at each step. All we need to prove is that the branches of T^{-1} are not all local hyperbolic isometries. If they were all local isometries then there could be no deficient points for $T^{-1} : X \to X$ (i.e. points in X with at least one T-image in A). From the construction of A, this would imply that A was completely invariant. As X would also contain no forward-singular points, it would follow that the correspondence was critically finite, contrary to the hypothesis. □

3.5 Correspondences on elliptic curves

In this section we consider irreducible correspondences on compact Riemann surfaces of genus 1. We begin by applying the Riemann-Hurwitz formula to relate the numbers of singular points to the bidegree. Let C be any Riemann surface of genus 1, which we will think of also as an elliptic curve. Let T be a correspondence of bidegree (m, n) on C. As usual we write D for the desingularised graph of T and g_D for the genus of D. The forward-singular points of T are the v_1 critical points of the degree n analytic map $\pi_1 : D \to C$, and the backward-singular points of T are the v_2 critical points of the degree m map $\pi_2 : D \to C$ (all counted with multiplicity). The Riemann-Hurwitz formula tells us that

$$v_1 = 2(g_D - 1) = v_2\,.$$

As remarked earlier, if T is separable to $\widehat{\mathbb{C}}$ then the dynamics of T are best studied on $\widehat{\mathbb{C}}$, so in this section we will only consider such correspondences when they arise in families of non-separable ones. However, we will need to consider correspondences which are separable to other surfaces of genus 1.

3.5.1 Correspondences with no singular points

Suppose that T is an irreducible correspondence with no singular points on a genus 1 surface C, of bidegree (m, n). Then the desingularised graph D of T is an unbranched cover of C, so is also a compact surface of genus 1. Make C and D into elliptic curves by giving them a marked point (called 0), in such a way that $\pi_1 : D \to C$ satisfies $\pi_1(0) = 0$, where $T = \pi_2 \circ \pi_1^{-1}$. Then π_1 is an *isogeny*, i.e. both an analytic map and a group homomorphism. We may represent D as $\mathbb{C}/\Lambda_0$, where Λ_0 is a lattice in $\mathbb{C}$. The analysis of section 3.2.3 applies, with the commensurable Fuchsian groups G_1 and G_2 replaced by commensurable lattices Λ_1 and Λ_2 in $\mathbb{C}$. [5] The map M is now an automorphism of $\mathbb{C}$ rather than $\mathbb{H}$, so it is of the form $M : z \mapsto az + b$. If M is a translation, then it commutes with the translations in Λ_1, so $\Lambda_2 = \Lambda_1$ and T is merely an automorphism of C. Henceforth suppose that M is not a translation. After conjugating everything by a suitable translation, (which does not change the groups Λ_i), we may assume that M is of the form $z \mapsto az$. Thinking of the lattices as subsets of $\mathbb{C}$ we have

$$a\Lambda_2 = \Lambda_1 \,.$$

It need not be the case that $m = n$ (the area argument that we used in the hyperbolic case does not apply because branches of the correspondence need not be local Euclidean isometries, merely similarities).

Note that all the covering groups are subgroups of the *abelian* group of translations of $\mathbb{C}$, which we will write as $\mathbb{C}$. This allows us to carry the algebra further than we could in the hyperbolic case. Indeed, the quotient

[5] This application is implicit in [19].

map $\pi_D : \mathbb{C} \to D = \mathbb{C}/\Lambda_0$ is actually a homomorphism of abelian groups; we may identify $\Lambda_0 = \ker \pi_D$. Now $\pi_D(\Lambda_1)$ and $\pi_D(\Lambda_2)$ are subgroups of D of orders n and m respectively, with trivial intersection. These subgroups therefore generate a subgroup $K < D$ of order mn, isomorphic to their direct product.[6] The group generated by Λ_1 and Λ_2 is also a lattice, $\Lambda_3 = \pi_D^{-1}(K)$, and it follows that the correspondence T is separable to the quotient surface $\mathbb{C}/\Lambda_3$. Nevertheless, T need not have valency – for example it could be a non-integer endomorphism of the elliptic curve. Thus we see that Lemma 3.1 no longer applies when $\widehat{\mathbb{C}}$ is replaced by an elliptic curve. In this case, passing to the associated correspondence on $\mathbb{C}/\Lambda_3$ would not help us to study the dynamics because it reduces neither the genus nor the bidegree.

Let $\pi_1^* : C \to D$ be the dual isogeny, so that $\pi_1^* \circ \pi_1$ is multiplication by $n = \deg(\pi_1)$ on D and $\pi_1 \circ \pi_1^*$ is multiplication by n on C. Then

$$T = (\pi_2 \circ \pi_1^*) \circ (\pi_1 \circ \pi_1^*)^{-1} .$$

The factor on the right is the inverse of the isogeny $C \to C$ given by multiplication by n. The factor on the left may be expressed $z \mapsto \alpha(z) + t$, where $\alpha : C \to C$ is an isogeny, t is a point of C and addition is in the group C. In the case when C does not have complex multiplication, α must be multiplication by an integer m'. Thus we may express T as multiplication by a rational m'/n followed by a translation. By changing the choice of $0 \in C$ to a fixed point of T, we can dispense with the translation. Put m'/n in its

[6] In contrast, two commensurable Fuchsian groups may together generate a non-discrete subgroup of $\mathrm{PSL}_2(\mathbb{R})$. For example, Jørgensen's inequality shows that the commensurable groups corresponding to the third power of the AGM correspondence generate a non-discrete group.

lowest form p/q. Then the bidegree of T is actually (p^2, q^2). It follows that if either m or n is not square then C does have complex multiplication.

In the special case of bidegree $(2, 2)$, we must have an isogeny of C of degree 4 which is not simply the composition of an automorphism of C with multiplication by 2. We can analyse these in terms of $\Lambda_0 = \langle 1, \tau\rangle$, where $|\tau| \geq 1$ and $-1/2 < \mathrm{Re}(\tau) \leq 1/2$. In terms of these generators, there are only three possibilities for the unordered pair of lattices Λ_1 and Λ_2. They must be two of $\langle 1, \tau/2\rangle$, $\langle 1/2, \tau\rangle$ and $1,(1+\tau)/2$. On the other hand they are related by $\Lambda_1 = a\Lambda_2$, where we may assume that $|a| = 1$ and $\mathrm{Im}(a) > 0$. Since $1 \in \Lambda_2$, we have $a \in \Lambda_1$, so $2a \in \Lambda_0$ so $2a \in \{\tau - 2, \tau - 1, \tau, \tau + 1, \tau + 2\}$. Each of these possibilities gives us a quadratic equation for a with integer coefficients, and we find after some calculation that a must be i or $(1+\sqrt{-3})/2$, and that $a\Lambda_0 = \Lambda_0$. The elliptic curves involved are $C = \mathbb{C}/\langle 1, 2i\rangle$ or $C = \mathbb{C}/\langle 1, \sqrt{-3}\rangle$ respectively. Since we can dispense with the translation by choosing the zero in C to be a fixed point of T, we get finite dynamics in both cases.

3.5.2 Correspondences of bidegree (2,2) on a torus with two singular points in each direction

We want π_1 and π_2 to be branched covers of degree 2 from a compact Riemann surface R to a torus S, with precisely two critical values each, and therefore precisely two critical points each. The Riemann-Hurwitz formula tells us that R must have genus 2. Then π_1 and π_2 are quotient maps associated to involutions I_1 and I_2 of R, with precisely two fixed points each. Since we want $f = \pi_2 \circ \pi_1^{-1}$ to be a genuine $(2, 2)$ correspondence and not an automorphism of the torus S, the involutions must be distinct. Now I_1 and

I_2 generate a group $G = \langle I_1, I_2 \rangle$ of automorphisms of R, a subgroup of the finite automorphism group $\mathrm{Aut}(R)$, of order at least 4.

Lemma 3.10.

Any two distinct involutions a, b of any compact Riemann surface R generate a dihedral group of automorphisms.

Proof. Every element of $\langle a, b \rangle$ can be expressed as an alternating word in the letters a and b. The alternating words of odd length are all conjugate either to a or to b, hence are not the identity. However, $\mathrm{Aut}(R)$ is finite so some even-length word is the identity. Now ab and its inverse ba have the same order, say m. We know $m \geq 2$ since $a^{-1} = a \neq b$. Then $\langle a, b \rangle$ is the dihedral group of order $2m$. □

We may view the quotient map $\pi_G : R \to R/G$ as an analytic branched cover of degree $|G| \geq 4$, branched at the (finitely many) fixed points of G. This makes R/G into a Riemann surface, which the Riemann-Hurwitz formula tells us must have genus 0. Now the quotient map π_G factors through S in two ways:

$$\pi_G = \pi_3 \circ \pi_1 = \pi_4 \circ \pi_2,$$

where π_3 and π_4 are analytic branched covers of degree $|G|/2$.

If $|G| = 4$, i. e. if I_1 and I_2 commute, then π_3 and π_4 are of degree 2, and in fact $f = \pi_4^{-1} \circ \pi_3$, so f is separable.

This construction gives us the family of $(2, 2)$ correspondences on the sphere with four forward-singular and four backward-singular points. Every one of these is Möbius conjugate to some correspondence given by $M \circ \wp \circ T \circ \wp^{-1}$, where $\wp$ is the Weierstrass $\wp$-function for the (unique) torus branched over the four forward singular points, T is a torus automorphism (which we may always take to be a translation), and M is a Möbius map. Composing the same maps but starting instead at the torus, we get a separable $(2, 2)$ correspondence on the torus: $\wp^{-1} \circ M \circ \wp \circ T$. When this is non-degenerate, its graph is a genus 2 surface, and separability tells us that the induced involutions do commute.

The reader may wonder at this point whether are there *any* non-separable $(2, 2)$ correspondences on any torus. We will show that there are, and attempt to classify all such correspondences.

To show that a correspondence f is non-separable, it suffices to exhibit distinct points $x, y \in S$ such that $f(x)$ and $f(y)$ have distinct but not disjoint sets of values. We have seen above that a non-separable $(2, 2)$ correspondence on a torus S is $\pi_2 \circ \pi_1^{-1}$, where for $j = 1, 2$, π_j is the quotient map for an involution I_j, and where I_1 and I_2 *do not commute*.

Lemma 3.11.
Let R be a compact genus 2 Riemann surface and I a non-trivial involution of R with precisely two fixed points. Write S for the torus R/I obtained by identifying z with $I(z)$ for each $z \in R$, and choose $0 \in S$ so that the branch points of the quotient map $\pi_I : R \to S$ are α and $-\alpha$. Then the involution $w \mapsto -w$ on S has two single-valued lifts to R, say J and K. Both are involutions commuting with I, and $IJ = JI = K$. We can label the lifts so that J has two fixed points and K has six fixed points.

Proof. The involution of S induces an involution of the fundamental group of $S \setminus \{\alpha, -\alpha\}$ which necessarily preserves the holonomy of the cover by R (this could fail if the cover had more than two sheets); hence the involution lifts, and the two lifts are related by the sheet-swapping involution I. One of the lifts has two fixed points over $0 \in S$ while the other swaps the two points over 0. Their squares are both lifts of the identity map on S, with fixed points, and therefore both lifts are involutions. An involution of a genus 2 surface has either 2 or 6 fixed points. The quotient by the group $\{1, I, J, K\}$ is the Riemann sphere and the Riemann-Hurwitz formula tells us that the quotient map has 10 critical points. Each of those critical points is a fixed point of just one of I, J or K (since they cannot share any fixed points). Thus without loss of generality we may take J to have two fixed points and K to have 6. □

Since $\{1, I, J, K\}$ is a subgroup of order four of $\mathrm{Aut}(R)$, we find that whenever $\mathrm{Aut}(R)$ contains an involution with two fixed points, then 4 divides $|\mathrm{Aut}(R)|$. In particular in order for R to be the graph of a non-separable $(2, 2)$ correspondence on a torus, we must have $|\mathrm{Aut}(R)| \geq 8$.

An involution of a compact Riemann surface R of genus $g \geq 2$ is called *hyperelliptic* if it induces a branched covering map of degree 2 to $\widehat{\mathbb{C}}$, which happens if and only if it has $2g + 2$ fixed points. When $g = 2$ then R necessarily has a hyperelliptic involution. The following lemma is standard.

Lemma 3.12.

Let R be a compact Riemann surface of genus $g \geq 2$. Then R has at most one hyperelliptic involution.

Proof. A hyperelliptic involution presents R as a curve $w^2 = f(z)$ where f is a polynomial of degree $2g + 2$ with simple roots. The ratios of holomorphic differentials on R generate the subfield of meromorphic functions $\mathbb{C}(z) \subset$

$\mathbb{C}(R)$, which in turn determines the hyperelliptic involution: two points $x, y \in R$ are related by the involution if and only if every function in this subfield takes the same value at x and y. □

Corollary 3.13.

Let R be a compact genus 2 Riemann surface, and suppose that I_1, I_2 are two non-commuting involutions of R with precisely two fixed points each. For $m = 1, 2$, construct J_m, K_m associated to I_m as in Lemma 3.11, so that K_1, K_2 have six fixed points each. Then $K_1 = K_2$, and I_1 and I_2 descend to give involutions A_1, A_2 of $R/K_1 = \widehat{\mathbb{C}}$.

Note that $I_1 I_2$ is not the identity, nor is it K, since then we would have $I_2 = I_1 K = J_1$, which commutes with I_1. Thus $A_1 A_2$ does not descend to the identity map on R/K, so $A_1 \neq A_2$. Now A_1 and A_2 may or may not commute, so we have two cases to consider:

Lemma 3.14.

Assume the conditions of Corollary 3.13.

1. *If $A_1 A_2 = A_2 A_1$ then $\langle I_1, I_2 \rangle = D_8$. The space of conformal classes of compact genus 2 Riemann surfaces with D_8 symmetry is one-complex dimensional, and there are just three examples in which the quotient tori R/I_1 and R/I_2 are isomorphic, the j-invariants being $(20)^3$ in one example and $-(15)^3$ in the other two (complex conjugate) examples.*

2. *If $A_1 A_2 \neq A_2 A_1$ then $\langle I_1, I_2 \rangle = D_{12}$. The space of conformal classes of compact genus 2 Riemann surfaces with D_{12} symmetry is one-dimensional (over $\mathbb{C}$), and there are just five examples in which the quotient tori*

R/I_1 and R/I_2 are isomorphic, the j-invariants being $2.(2.3.5)^3$ in one example, and also $-(2^5)^3$ and $(20)^3$, there being two complex conjugate cases of each.

Here j is the j-invariant of an elliptic curve, with the normalisation common in number theory, namely $j = 2^8\frac{(\lambda^2-\lambda+1)^3}{\lambda^2(\lambda-1)^2}$ for an elliptic curve branched over $\{0, 1, \infty, \lambda\}$. Note that this is 1728 times Klein's elliptic modular function.

Proof. We treat the two cases separately.

1. Suppose A_1 and A_2 commute. Since I_1 and I_2 do not commute, we must have $I_1I_2 = I_2I_1K$. Note that K commutes with I_1 and with I_2, so $I_1I_2I_1I_2 = K = I_2I_1I_2I_1$, hence $\mathrm{ord}(I_1I_2) = 4$, and we have D_8 as required. The conformal type of R is determined by the conformal type of the orbifold quotient $R/\langle I_1, I_2\rangle$, which is $\widehat{\mathbb{C}}$ with four branch points marked. In fact by a Möbius conjugation we may assume that A_1 acts on $\widehat{\mathbb{C}}$ as $z \mapsto -z$ and that A_2 acts as $z \mapsto 1/z$. We still have a choice of which lift of A_1 to call I_1 and which to call J_1, and of how to label the lifts of A_2 as I_2 and J_2. Since we may conjugate by A_1 and A_2 we lose no generality by taking I_1 to fix the two points of R lying over $0 \in R/K = \widehat{\mathbb{C}}$, and I_2 to fix the two points lying over 1.

   The quotient map $R \mapsto R/K = \widehat{\mathbb{C}}$ is branched over six points, which are permuted with no fixed points by A_1 and permuted with no fixed points by A_2 (because I_1 and I_2 share no fixed points with K). Since an automorphism of $\widehat{\mathbb{C}}$ is determined by its action on three points, A_1 and A_2 must act by different but commuting fixed-point-free involutions of the set of six branch points. These two involutions have one common two-cycle, consisting of the solutions of $z = -1/z$, i. e. $\pm i$. The remaining four branch points are $x, -x, 1/x, -1/x$ for some $x \in \widehat{\mathbb{C}} \setminus \{0, \infty, 1, -1, i, -i\}$. We may write the quotient map $R/K \to R/\langle I_1, I_2\rangle$ as $\phi : z \mapsto \frac{1}{4}(z + 1/z)^2$, which is a degree 4 map with critical points $\{0, \infty, \pm 1, \pm i\}$, and critical values $0, 1, \infty$. Now the torus R/I_1 is determined by the four critical values of the degree 2 quotient map $R/I_1 \to R/\langle I_1, K\rangle = (R/K)/A_1$. The quotient map $R/K \to (R/K)/A_1$ is realised by $z \mapsto z^2$. Now, I_1 fixes the two points over $0 \in R/K$, so these two points are critical for $R \to R/I_1$. Hence they are critical for the degree 4 quotient map $R \to R/\langle I_1, K\rangle = (R/K)/A_1 = \widehat{\mathbb{C}}$, whose critical values are $0, \infty, -1, x^2, 1/x^2$. Thus R/I_1 is the torus obtained as a degree 2 branched cover of the sphere branched over $\infty, -1, x^2, 1/x^2$. Similar considerations apply to R/I_2. This time the quotient map $(R/K) \to (R/K)/A_2$ is represented by $z \mapsto \frac{1}{2}(z+1/z)$ and then the critical values for the degree 4 quotient map $R \to R/\langle I_2, K\rangle = (R/K)/A_2 = \widehat{\mathbb{C}}$ are $1, -1, 0, \frac{1}{2}(x + 1/x), \frac{-1}{2}(x + 1/x)$. However 1 is accounted for by the fact that $R \to R/I_2$ is critical at the two points over $1 \in R/K$.

We obtain a non-separable dynamical correspondence of bidegree $(2, 2)$ when R/I_1 and R/I_2 are isomorphic. This happens precisely when their j-invariants are equal, i. e. when the cross-ratios

$$\left[\infty, -1, x^2, 1/x^2\right], \left[-1, 0, \frac{1}{2}(x + 1/x), \frac{-1}{2}(x + 1/x)\right]$$

lie in the same orbit of the action of S_3 on $\widehat{\mathbb{C}}$ generated by $z \mapsto 1 - z$

and $z \mapsto 1/z$. We write $w = \frac{1}{4}(x + 1/x)^2$ for the co-ordinate on the quotient sphere $R/\langle I_1, I_2\rangle$. Then

$$j(R/I_1) = \frac{64(4w-3)^3}{(w-1)}$$

and

$$j(R/I_2) = \frac{64(3+w)^3}{(w-1)^2}\,.$$

Comparing these shows that $j(R/I_1) = j(R/I_2)$ if and only if

$$64w^3 - 209w^2 + 243w - 162 = 0\,.$$

The roots of this polynomial are

$$w = 2, \quad w = \frac{9(9 \pm 5\sqrt{-7})}{128}\,.$$

The corresponding tori have j-invariants $(20)^3$ and $(-15)^3$ respectively. The torus with j-invariant $(20)^3$ is

$$\mathbb{C}\Big/\left\langle 1, i/\sqrt{2}\right\rangle\,.$$

The torus with j-invariant $(-15)^3$ is

$$\mathbb{C}\Bigg/\left\langle 1, \frac{1+i\sqrt{7}}{2}\right\rangle\,.$$

2. Consider the action of A_1 and A_2 permuting the six branch points of the quotient map $R \to R/K$. Neither fixes any of those points (since I_1 and K share no fixed points, for example). We assume now that A_1 and A_2 do not commute; in particular A_1A_2 cannot fix all six points since it is a non-identity Möbius map. The only possibility is that A_1A_2 has order 3 with no fixed points; hence $(I_1I_2)^3$ is either 1 or K. It can be 1, but then we may replace I_2 with I_2K, which also has two fixed points and does not commute with I_1, so as to get K.

□

The elliptic curves arising here are rather special, being five of the 13 elliptic curves that have complex multiplication and integer j-invariant. These are the elliptic curves whose ring of endomorphisms is bigger than $\mathbb{Z}$ but is nevertheless a principal ideal domain. They arise from quadratic number fields with class number equal to 1. The fact that these curves have complex multiplication may be explained by the action on the Jacobian, if we can check that the correspondences do not have valency. Perhaps an analysis of the actions of I_1 and I_2 on the Jacobian of R would lead to a more conceptual explanation of why we get these particular curves. At the moment we can only remark that it is no surprise because these are in some sense the simplest elliptic curves. The affine linear examples of irreducible $(2,2)$ correspondences with no singular points also had complex multiplication with integer j-invariants: the lattice $\langle 1, 2i\rangle$ gives $j = (66)^3$, while the lattice $\langle 1, \sqrt{-3}\rangle$ gives $j = 2.(30)^3$.

3.6 Dynamics of a D_8 example

This section contains the preparatory work for a numerical exploration of one of the examples computed in the previous section. To study the dynamics numerically we need a fairly precise and rapid means of calculation – repeated calls to functions that evaluate elliptic integrals or theta functions are to be avoided. Instead we represent the genus 2 surface as a hyperelliptic curve, find the involutions I_1 and I_2 explicitly, and represent their quotients and quotient maps explicitly as hyperelliptic curves and algebraic maps. Finally we must represent algebraically the isomorphism between the two tori, using the known Möbius map sending the branch points of one to the branch points of the other. To compose with a translation of the torus we need the elliptic curve addition formula, which again is an algebraic map. For a $(2,2)$ correspondence one should be able to get the hard computational work down to finding one square root per iteration. Of course, the validity of any iterative computations depends on the stability of the system, which is one reason for being interested in results such as Theorem 3.9.

3.6.1 Algebraic equations for the $(2,2)$ correspondences on the torus $\mathbb{C}/\left\langle 1, i/\sqrt{2}\right\rangle$

We follow the notation and normalisation from the proof of Lemma 3.14. The genus 2 surface R has hyperelliptic involution K, with $R \to R/K$ branched over the six points $\pm i, \pm 1 \pm \sqrt{2}$, i. e. R is the compactification of the curve in $\mathbb{C}^2$ given by

$$y^2 = (x^2+1)(x^2-2x-1)(x^2+2x-1).$$

We have involutions

$$\begin{aligned} I_1 &: (x,y) \mapsto (-x,y)\,, \\ I_2 &: (x,y) \mapsto (1/x, y/x^3) \end{aligned}$$

inducing the involutions A_1 and A_2 on $R/K = \widehat{\mathbb{C}}$ by forgetting the second co-ordinate. The quotient map $R \to R/I_1$ is represented as

$$(x,y) \mapsto (x^2, y),$$

mapping onto the torus R/I_1 represented as the curve

$$t^2 = (z+1)(z^2 - 6z + 1),$$

i. e. a double cover of the z-plane branched over $z = -1, 3 \pm 2\sqrt{2}, \infty$. The quotient map $R \to R/I_2$ is represented as

$$(x, y) \mapsto \left(\frac{1}{2}(x + 1/x), \frac{1}{4}(y/x + y/x^2) \right),$$

mapping onto the torus R/I_2 represented as the curve

$$t^2 = z(z^2 - 2)(z + 1),$$

i. e. a double cover of the z-plane branched over $z = 0, \pm\sqrt{2}, -1$.

The Möbius map that sends the latter set of branch points to the former is $z \mapsto \frac{z-1}{z+1}$, which induces a map $R/I_2 \to R/I_1$ given by

$$(z, t) \mapsto \left(\frac{z-1}{z+1}, \frac{t\sqrt{-8}}{(z+1)^2} \right)$$

n. b. The choice of a value for $\sqrt{-8}$ corresponds to a choice of one of the two isomorphisms $R/I_2 \to R/I_1$ covering the Möbius map $z \mapsto \frac{z-1}{z+1}$. It will be convenient to take the (unique) point of R/I_1 over $z = \infty$ to be the zero for the group law on R/I_1, because then $\zeta \mapsto -\zeta$ on the torus is represented by $(z, t) \mapsto (z, -t)$, i. e. by the induced action of K on R/I_1. Then the two isomorphisms are each other's negative.

Now we compose these maps to get a $(2, 2)$ correspondence $f_0 : R/I_1 \to R/I_1$, given by

$$(z, t) \mapsto \left(\left(\frac{1 - \sqrt{z}}{1 + \sqrt{z}} \right)^2, \frac{t\sqrt{-8}}{(1 + \sqrt{z})^3} \right).$$

The correspondence f_0 has two forward singular points, $(0, \pm 1)$. Each of these has only one image, namely $(0,1) \mapsto (1, \sqrt{-8})$ and $(0,-1) \mapsto (1, -\sqrt{-8})$. The point over $z = \infty$ is not forward-singular since $R \mapsto R/I_1$ is branched over ∞. The two points $(0,1)$ and $(0,-1)$ are also the two critical values of f_0 (i.e backward-singular points). They are one of the images of each of $(1, \sqrt{-8})$ and $(1, -\sqrt{-8})$, the other image of each of these being the point over $z = \infty$, which itself maps back to those two points. The two branches of f_0 take the same value at $(-1, 0)$, which is in fact a fixed point. Thus there are three points in the torus where f_0 has only one value.

Finally we make explicit the group law on R/I_1, having chosen the point over $z = \infty$ as the zero. Three points sum to zero if they are collinear. Thus we get the addition law

$$(z_0, t_0) \oplus (z_1, t_1) = (z_2, t_2),$$

where

$$z_2 = 5 + \left(\frac{t_1 - t_0}{z_1 - z_0} \right)^2 - z_1 - z_0,$$
$$t_2 = -t_0 - (z_2 - z_0) \left(\frac{t_1 - t_0}{z_1 - z_0} \right).$$

We will use Greek letters to denote points of R/I_1.

We now have a family of correspondences f_ζ, one for each $\zeta \in R/I_1$. If $\zeta = (z_0, t_0)$, we have

$$f_\zeta : (z,t) \mapsto \left(\left(\frac{1 - \sqrt{z}}{1 + \sqrt{z}} \right)^2, \frac{t\sqrt{-8}}{(1 + \sqrt{z})^3} \right) \oplus (z_0, t_0).$$

We had two choices for the isomorphism $R/I_2 \to R/I_1$; let's call the resulting families f_ζ (taking $\sqrt{-8} = i\sqrt{8}$) and g_ζ (taking $\sqrt{-8} = -i\sqrt{8}$).

Then $f_\zeta = -g_{-\zeta}$. The correspondence f_0 is odd with respect to the group law, so

$$f_\zeta(-\theta) = \zeta - f_0(\theta) = -f_{-\zeta}(\theta),$$

(by which we mean they have the same pairs of values). Thus f_ζ and $f_{-\zeta}$ are conjugate (via K). The same holds for the family g_ζ, so $f_\zeta = g_\zeta \circ K = K \circ g_{-\zeta}$, and

$$f_\zeta \circ f_\zeta = g_\zeta \circ g_{-\zeta}\,.$$

3.6.2 Dynamics of f_0

As is evident from its algebraic formula, f_0 has a factor correspondence obtained by ignoring the second co-ordinate. The factor is the $(2,2)$ correspondence on $\widehat{\mathbb{C}}$ given by

$$z \mapsto \left(\frac{1-\sqrt{z}}{1+\sqrt{z}}\right)^2.$$

This correspondence is separable; in fact it has the Möbius involution $w \mapsto \frac{w+3}{w-1}$ as a factor via the map $z \mapsto \frac{1}{2}(z + 1/z)$. The correspondence f_0^2 has two algebraic components:

$$(z,t) \mapsto (z,-t),$$

occurring in two ways, and

$$(z,t) \mapsto (1/z, \pm t/z\sqrt{z}).$$

So even iterates of (z,t) eventually take just four values, and odd iterates eventually take just four values (possibly fewer for certain points (z,t)). This is rather similar to the situation in §3.5.1. However, inserting a generic translation of the torus destroys the factor, and allows the dynamics to be much more complicated.

Note that f_0 is a critically finite correspondence, so after removing the complete critical orbit we could lift to the hyperbolic plane $\mathbb{H}$ and describe the correspondence by means of a discrete subgroup G of $SL_2(\mathbb{R})$ and a commensurate conjugate group $A^{-1}GA$ (as Bullett did for critically finite correspondences on $\widehat{\mathbb{C}}$ [19]). It would be interesting to describe these groups explicitly.

As a result of numerical experiments we discovered a rather nice geometrical way to describe the action of f_0. Slit the torus by removing all the points (z, t) with $z \in [0, \infty]$. This corresponds to cutting along the boundary of a certain rectangular fundamental domain in the covering plane, then making two short slits part of the way in towards the centre of the rectangle from the midpoints of the long sides. The vertices of this rectangle all correspond to the same point of the curve, namely $(3 + 2\sqrt{2}, 0)$. The centre of the rectangle is the coincident fixed point $(-1, 0)$. The short slits end at the forward-singular points of f_0, which are $(0, -1)$ and $(0, 1)$, and meet the long sides of the rectangle at $(3 - 2\sqrt{2}, 0)$. Now each branch of f_0 consists of an automorphism of this slit-rectangle, conjugate to a rotation of the unit disc; both branches fix the central point of the rectangle, and they rotate by $\pi/4$ or $3\pi/4$ about that point. The two short sides of the rectangle and each half of each long side represent one eighth of the boundary of the domain (in terms of harmonic measure with respect to the central point), and both sides of a slit together represent a further one eighth. It is rather amusing to see how the two branches fit with each other when we cross a boundary segment of the slit fundamental domain.

Consider the equivalence relation whose equivalence classes are grand orbits of f_0. The quotient map for this relation is a degree 8 meromorphic function. In fact it is a Belyi map: we can arrange for it to be branched only over $0, 1, \infty$ and for the slits in the previous paragraph to form some of the edges of the associated dessin d'enfant on the elliptic curve.

f_0 has a coincident fixed point, i.e. a point which is fixed for both branches of f_0; taking this to be the zero of the elliptic curve, we can easily compute that the action on the Jacobian (which is naturally identified the elliptic curve itself) is multiplication by $i\sqrt{2}$. In particular f_0 does not have valency.

The topological description of f_0 in terms of the Belyi map defined above does in fact determine uniquely the complex analytic structure of the elliptic curve. Likewise a topological description of the correspondence f_0 allows us to compute the action on the Jacobian (which is a homological invariant) and this again determines the complex analytic structure uniquely. It is possible to describe *topological correspondences* in terms of pairs of topological branched covering maps. However, these need not always be realised by any holomorphic correspondence. It would be interesting to understand the obstruction.

Now choose the zero in the elliptic curve to be the midpoint of the long side of the slit, which is the point $(3 - 2\sqrt{2}, 0)$ in our algebraic description. The composition of f_0 followed by multiplication by $i\sqrt{2}$ gives us a forward-critically finite correspondence. This is our first explicit example for Theorem 3.9 in which there are singular points in both directions.

Chapter 4

Area distortion of polynomial mappings

This chapter has been accepted for publication in the Bulletin of the London Mathematical Society under the title ‘The area of polynomial images and pre-images’.

4.1 Introduction

A lemniscate is the level set of a complex polynomial, i.e.

$$E(p, r) = \{z \in \mathbb{C} : |p(z)| = r^n\},$$

where p is a polynomial of degree n over $\mathbb{C}$. A well-known example is Bernoulli’s lemniscate, the degree 4 plane curve $|z^2 - 1| = 1$. The region enclosed by a lemniscate is a sublevel set for the modulus of a polynomial, and it is natural to ask how large this set can be. To take account of scaling, we take p to be monic. With this normalisation, Pólya proved in [59] that the area (two-dimensional Lebesgue measure dA) enclosed by the lemniscate $E(p, 1)$ is bounded by a constant that does not depend on the choice of p.

Theorem 4.1 ((Pólya's inequality)). *Let p be a monic complex polynomial of degree n and let D be a closed disc in $\mathbb{C}$. Then the Euclidean area of $p^{-1}(D)$ is at most $\pi\left(\frac{\mathrm{Area}(D)}{\pi}\right)^{1/n}$, with equality only when $p : z \mapsto a(z-b)^n+c$ and the centre of D is the point c, which is the unique critical value of p.*

Here we have made a slight reformulation of Pólya's original statement, intended to suggest a generalisation to an arbitrary measurable set of finite area, in place of the disc D. This generalisation is the main result of this chapter.

Theorem 4.2. *Let p be a monic polynomial of degree n over $\mathbb{C}$. Let K be any measurable subset of the plane. Then*

$$\mathrm{Area}(p^{-1}(K)) \leq \pi \left(\frac{\mathrm{Area}(K)}{\pi}\right)^{1/n},$$

with equality if and only if K is a disc, up to a set of measure zero, and p has a unique critical value at the centre of that disc.

In fact we will establish the following bound on the area of the image of a set under a monic polynomial mapping.

Theorem 4.3. *Let p be a monic polynomial of degree n over $\mathbb{C}$, and L be any measurable subset of the plane. Define the multiplicity $n(z, p, L)$ to be the number of p-pre-images of z in L, counted according to their valency. Then the area of $p(L)$ counted with multiplicity satisfies*

$$\int_{\mathbb{C}} n(z, p, L)\, d\mathrm{A}(z) \;=\; \int_L |p'(z)|^2\, dA(z) \;\geq\; n\pi \left(\frac{\mathrm{Area}(L)}{\pi}\right)^n,$$

with equality if and only if L is a disc (up to a set of measure zero) and p has a unique critical value at the centre of that disc.

To deduce Theorem 4.2 from Theorem 4.3, take $L = p^{-1}(K)$, which maps onto K with multiplicity n everywhere, so that

$$\mathrm{Area}(K) = \frac{1}{n} \int_{p^{-1}(K)} |p'(w)|^2 \, d\mathrm{A}(w) \,.$$

A survey of area estimates for lemniscates has recently been given by Lubinsky [48], with a view towards applications in the convergence theory of Padé approximation. In [31], Eremenko and Hayman address the related problem of bounding the length of $E(p, r)$. Fryntov and Rossi [33] have obtained a hyperbolic analogue of Pólya's inequality, giving a sharp upper bound for the hyperbolic area of the pre-image of a hyperbolic disc under a finite Blaschke product. This raises the question of finding analogues for finite Blaschke products of theorems 4.2 and 4.3.

4.2 Logarithmic capacity

In this section and the next we outline the machinery that we will use to prove Theorem 4.3. Logarithmic capacity and condenser capacity are powerful tools of potential theory with many alternative descriptions. In order to make this paper as accessible as possible, we describe them only in terms of polynomials and rational functions. The results in this section are all classical, but the author does not know of one comprehensive reference. For a clear introduction to potential theory in $\mathbb{C}$, including proofs of our statements about logarithmic capacity, see [62]. Further information about the capacity and modulus of a condenser may be found in [4, 5, 39, 65], and the appendix of [67]. Ahlfors [2, Ch. 4] casts condenser capacity as a conformal invariant of extremal length type, called *extremal distance*.

Given a compact subset $K \subset \mathbb{C}$ and a polynomial q, we write

$$\|q\|_K = \max_{z \in K} |q(z)| \,.$$

Let $\mathcal{M}$ be the set of all monic complex polynomials. The *logarithmic capacity* of K is

$$\operatorname{cap}(K) = \inf\{(\|q\|_K)^{1/\deg q} : q \in \mathcal{M}\}$$

Logarithmic capacity may also be characterised as the unique non-negative real function of compact sets satisfying the following four conditions.

1. $\operatorname{cap}\left(\overline{\mathbb{D}}\right) = 1$, where $\overline{\mathbb{D}}$ is the closed unit disc in $\mathbb{C}$.
2. (monotonicity) If $K_1 \subset K_2$ then $\operatorname{cap}(K_1) \leq \operatorname{cap}(K_2)$.
3. (outer continuity) If $K_n \downarrow K$ then $\operatorname{cap}(K_n) \downarrow \operatorname{cap}(K)$.
4. If $p(z) = a_n z^n + \ldots a_0$ is a complex polynomial with $a_n \neq 0$, then

$$\operatorname{cap}(p^{-1}(K)) = \left(\frac{\operatorname{cap}(K)}{|a_n|}\right)^{1/n} \,.$$

Thus for any monic polynomial p, the capacity of the lemniscate $E(p, r)$ is r. Using a linear polynomial in condition (4) we see that

$$\operatorname{cap}(aK + b) = a \operatorname{cap}(K) \,,$$

so logarithmic capacity is a one-dimensional measurement of the size of K. If d is the diameter of K then $\operatorname{cap}(K) \leq d/2$, and if K is connected then

$\operatorname{cap}(K) \geq d/4$. Thus it is not surprising that there is an 'isoperimetric' inequality comparing logarithmic capacity and area. A compact set $K \subset \mathbb{C}$ is called *polar* when $\operatorname{cap}(K) = 0$.

Theorem 4.4. *For any compact set $K \subset \mathbb{C}$,*

$$\operatorname{Area}(K) \;\leq\; \pi \operatorname{cap}(K)^2 \,,$$

with equality if and only if K is the union of a closed disc and a polar set.

For a proof using a simple kind of symmetrization, see [62, Thm. 5.3.5]. Because of condition (4) above, Theorem 4.1 is the special case of this isoperimetric inequality in which K is restricted to be a lemniscate. In fact Pólya first proved Theorem 4.1 by a clever use of Gronwall's area formula then used it to deduce Theorem 4.4 [48, 59]. To make this deduction, consider the filled-in set $\tilde{K} = \mathbb{C} \setminus U$, where U is the unbounded component of $\mathbb{C} \setminus K$. The maximum modulus theorem shows that $\operatorname{cap}(\tilde{K}) = \operatorname{cap}(K)$. On the other hand the following result shows that we can estimate $\operatorname{cap}(\tilde{K})$ by approximating $\tilde{K}$ from outside by filled-in lemniscates, whose areas are bounded above in Theorem 4.1.

Theorem 4.5 ((Hilbert's Lemniscate Theorem)). *Let K be a compact subset of $\mathbb{C}$ such that $\mathbb{C} \setminus K$ is connected, and let V be any open neighbourhood of K. Then there exists a complex polynomial q such that*

$$\frac{|q(z)|}{\|q\|_K} > 1 \qquad \text{for all } z \in \mathbb{C} \setminus V.$$

We can use the scaling properties of logarithmic capacity to formulate a scale-invariant version of Theorem 4.2. For a non-polar compact set K, define

$$\rho(K) = \frac{\text{Area}(K)}{\pi \operatorname{cap}(K)^2} \,.$$

Then $\rho(K)$ is a measure of the *roundness* of K: from Theorem 4.4 we find that $\rho(K) \in [0,1]$ and that $\rho(K) = 1$ if and only if K is a full-measure subset of a disc. Divide both sides of the inequality of Theorem 4.2 by $\operatorname{cap}(K)^{2/n} = \operatorname{cap}(p^{-1}(K))^2$ to obtain the following theorem.

Theorem 4.6. *If p is any complex polynomial of degree n, not necessarily monic, and K is any non-polar compact subset of the plane, then*

$$\rho(p^{-1}(K)) \leq \rho(K)^{1/n} \,.$$

This is sharp for each value of $\rho(K)$. Indeed, any value of $\rho(K)$ in $[0,1]$ can be achieved by taking $K = [0,1] \cup \{z \in \mathbb{C} : |z| \leq \alpha\}$ and $p : z \mapsto z^n$. This gives equality in Theorem 4.2 and hence also in Theorem 4.6.

4.3 Condenser capacity

Definition. A *plane condenser* is a pair (K, L) of disjoint compact subsets of the Riemann sphere $\widehat{\mathbb{C}}$.

We call a continuous function $f : \widehat{\mathbb{C}} \to \mathbb{R}$ admissible for the condenser (K, L) if $f = 0$ on K, $f = 1$ on L and f is continuously diffferentiable on $\widehat{\mathbb{C}} \backslash (K \cup L)$. The condenser capacity $\operatorname{cap}(K, L)$ is the infimum over admissible f of the Dirichlet integral

$$\mathcal{D}(f) = \frac{1}{2\pi} \int_{\mathbb{C}} |\nabla f(z)|^2 dA(z) \,.$$

Caveat: in some works the condenser capacity is given as half of this. Condenser capacity has the following properties:

1. cap $(\overline{D(0,R)}, \widehat{\mathbb{C}} \setminus D(0,S)) = 1/(\log S - \log R)$, where $D(0,R)$ is the open disc of radius R about 0.

2. (monotonicity) If $K_1 \subset K_2$ and $L_1 \subset L_2$ then cap $(K_1, L_1) \leq$ cap (K_2, L_2).

3. (outer continuity) If (K_n, L_n) is a sequence of condensers such that $K_n \downarrow K$ and $L_n \downarrow L$ as $n \to \infty$, then cap $(K_n, L_n) \to$ cap (K, L).

4. (conformal invariance) Suppose that (E, F) and (K, L) are condensers, with $U = \widehat{\mathbb{C}} \setminus (E \cup F)$ and $V = \widehat{\mathbb{C}} \setminus (K \cup L)$. Suppose that $\varphi : U \to V$ is analytic and n-valent, i.e. every point in V has precisely n pre-images in U (counted with multiplicity). Finally suppose that $\varphi(z) \to K$ as $z \to E$ in U and $\varphi(z) \to L$ as $z \to F$ in U. Then

$$\operatorname{cap}(E, F) = n \operatorname{cap}(K, L)\,.$$

To prove (4), approximate K and L by their closed ϵ-neighbourhoods K_ϵ and L_ϵ, whose boundaries are regular for the Dirichlet problem. Then there is a unique extremal function f in the definition of cap (K_ϵ, L_ϵ). It is the unique admissible function that is harmonic on $\widehat{\mathbb{C}} \setminus (K_\epsilon \cup L_\epsilon)$. Consider the pulled-back condenser $(E', F') = (E \cup \varphi^{-1}(K_\epsilon), F \cup \varphi^{-1}(L_\epsilon))$. The pullback $f \circ \varphi$ extended by 0 on E and by 1 on F is admissible for (E', F') and harmonic on $\widehat{\mathbb{C}} \setminus (E' \cup F')$. Therefore it is the unique extremal function for (E', F'). We have

$$\begin{aligned}
\mathcal{D}(f \circ \varphi) &= \int_{\widehat{\mathbb{C}} \setminus (E' \cup F')} |\nabla (f \circ \varphi)(z)|^2 \, dA(z) \\
&= \int_{\widehat{\mathbb{C}} \setminus (E' \cup F')} |(\nabla f)(\varphi(z))|^2 \, |\varphi'(z)|^2 \, dA(z) \\
&= n \int_{\widehat{\mathbb{C}} \setminus (K_\epsilon \cup L_\epsilon)} |\nabla f(w)|^2 dA(w) \quad = \quad n \, \mathcal{D}(f)\,.
\end{aligned}$$

Finally take the limit as $\epsilon \searrow 0$ and apply outer continuity of condenser capacity.

An easy consequence of the conformal invariance is that for any rational function R of degree n, we have

$$\left(\frac{\min_L |R|}{\max_K |R|}\right)^{1/n} \leq \exp(1/\mathrm{cap}\,(K, L))\,.$$

In fact the right-hand side is the supremum of the left-hand side as R ranges over all rational functions (see [34]). This is the analogue for condensers of our initial definition of logarithmic capacity.

Whereas logarithmic capacity scales one-dimensionally, the condenser capacity is invariant under scaling, so we cannot use it to estimate the area of K or L. However, in the case that $\infty \in L$ we can estimate the ratio of the Euclidean areas of K and $\widehat{\mathbb{C}} \setminus L$:

Theorem 4.7. *[Carleman, 1918] Let (K, L) be a condenser with $\infty \in L$. Then*

$$\frac{2}{\mathrm{cap}\,(K, L)} \leq \log\left(\frac{\mathrm{Area}(\widehat{\mathbb{C}} \setminus L)}{\mathrm{Area}(K)}\right)\,,$$

with equality if and only if $\widehat{\mathbb{C}} \setminus L$ and K are concentric discs in the plane, up to the addition of closed sets of logarithmic capacity zero to K and L.

The proof of Carleman's inequality uses the fact that the Dirichlet integral $\mathcal{D}(f)$ does not increase when f is replaced by its Schwarz symmetrization. This is the function $S(f)$ whose superlevel sets are concentric discs with the same Euclidean area as the corresponding superlevel sets of f. For details, see the classic book of Pólya and Szegö, [60], or [5] for a more recent account and a generalisation to variable metrics.

The logarithmic capacity can be recovered from condenser capacities:

$$-\log \operatorname{cap}(K) = \lim_{R\to\infty}\left[\frac{1}{\operatorname{cap}(K, \widehat{\mathbb{C}} \setminus D(0,R))} - \log R\right].$$

This is easy to prove using the characterisations of logarithmic capacity and condenser capacity in terms of minimal energy of Borel measures. It reveals the isoperimetric theorem for logarithmic capacity as a limiting case of Carleman's isoperimetric inequality for condenser capacity.

4.4 Proof of Theorem 4.2

Lemma 4.8. *For any complex polynomial g of degree d,*

$$\int_{\mathbb{C}} |g(w)|\, \mathbf{1}_{\{|g(w)|\le x\}}\, d\mathrm{A}(w) \;\ge\; \frac{2x}{d+2}\,\mathrm{Area}(\{w\in\mathbb{C} : |g(w)|\le x\}).$$

Proof. By conformal invariance of condenser capacity, we have

$$\operatorname{cap}\left(g^{-1}\left(\widehat{\mathbb{C}}\setminus B(0,x)\right), g^{-1}(B(0,s))\right) = \frac{d}{(\log x - \log s)}.$$

Theorem 4.7 gives

$$\frac{\mathrm{Area}\,(\{w\in\mathbb{C} : s\le|g(w)|\le x\})}{\mathrm{Area}\,(\{w\in\mathbb{C} : |g(w)|\le x\})} \;\ge\; 1-\left(\frac{s}{x}\right)^{2/d},$$

so

$$\begin{aligned}\int_{\mathbb{C}} |g(w)|\, \mathbf{1}_{\{|g(w)|\le x\}}\, d\mathrm{A}(w) \;&=\; \int_0^x \mathrm{Area}\,(\{w\in\mathbb{C} : s\le|g(w)|\le x\})\, ds\\ &\ge\; \mathrm{Area}\,(\{w\in\mathbb{C} : |g(w)|\le x\})\int_0^x \left(1-\left(\frac{s}{x}\right)^{2/d}\right) ds\\ &=\; \frac{2x}{d+2}\,\mathrm{Area}(\{w\in\mathbb{C} : |g(w)|\le x\}).\end{aligned}$$

□

Now fix a monic polynomial p and $A > 0$. Among all measurable sets K satisfying $\mathrm{Area}(K) = A$, the Dirichlet integral

$$\int_K |p'(w)|^2 d\mathrm{A}(w)$$

is minimised when K is the sublevel set

$$K_t = \{w \in \mathbb{C} : |p'(w)|^2 \leq t\}.$$

Here t is determined uniquely by the condition $\mathrm{Area}(K_t) = A$. The polynomial $z \mapsto (p'(z)/n)^2$ is monic, with degree $2n-2$, so Theorem 4.1 gives

$$A = \mathrm{Area}(K_t) \leq \pi \left(\frac{\pi (t/n^2)^2}{\pi}\right)^{1/(2n-2)} .$$

Rearranging this we have

$$t \geq n^2 \left(\frac{A}{\pi}\right)^{n-1} .$$

Now we apply Lemma 4.8 to the polynomial $g = (p')^2$ to obtain

$$\begin{aligned} \int_{K_t} |p'(w)|^2 d\mathrm{A}(w) &= \int_{\mathbb{C}} |p'(w)|^2 \mathbf{1}_{\{|p'(w)|^2 \leq t\}} \, d\mathrm{A}(w) \\ &\geq \frac{2t}{2n} \,\mathrm{Area}\,(K_t) = \frac{tA}{n} \\ &\geq n\pi \left(\frac{A}{\pi}\right)^n . \end{aligned}$$

For equality, we must have equality in our application of Pólya's inequality, so p must be $p : z \mapsto (z-b)^n + c$, and K can differ from disc K_t at most by a set of 2–dimensional Lebesgue measure zero. This completes the proof of Theorem 4.3.

Chapter 5

Smale's Mean Value Conjecture

Smale's mean value conjecture is a well-known inequality constraining the location of critical points and critical values of a polynomial mapping the complex plane into itself. We give an algebraic proof that for any isolated local extremum for Smale's mean value conjecture, all the objective values are equal. We also generalise Smale's conjecture to rational maps of the Riemann sphere, proving a version which only fails to be best possible by a constant multiplicative factor. We also discuss the special case of rational maps all of whose critical points are fixed, giving a construction based on the Newton–Raphson method.

5.1 Smale's Mean Value Conjecture

Let p be any polynomial with coefficients in $\mathbb{C}$. We say that $\zeta \in \mathbb{C}$ is a *critical point* of p if $p'(\zeta) = 0$. Then $p(\zeta)$ is the corresponding *critical value.* In 1981 Stephen Smale proved the following result about critical points and critical values of polynomials.

Theorem 5.1.

Let p be a non-linear polynomial and z any given complex number. Then there exists a critical point ζ of p such that

$$\left|\frac{p(\zeta)-p(z)}{\zeta-z}\right| \leq 4\left|p'(z)\right|.$$

The values of $\left|\frac{p(\zeta)-p(z)}{(\zeta-z)p'(z)}\right|$ as ζ ranges over all critical points of p are called the *objective values*, and any critical point ζ which minimises $\left|\frac{p(\zeta)-p(z)}{\zeta-z}\right|$ among all critical points is called an *essential* critical point with respect to z.

Smale asked whether the constant 4 could be reduced to 1, or in the case of a polynomial of degree n, to $1-1/n$, which would be best possible in view of the example $p(z) = z^n + z$. Thus Smale's mean value conjecture is the statement that

$$\min\left|\frac{p(\zeta_i)-p(z)}{(\zeta_i-z)p'(z)}\right| \leq \frac{n-1}{n}.$$

Various special cases have been solved, but at present the best known result that applies to all polynomials replaces 4 by $4^{(n-2)/(n-1)}$ [15]. It is widely assumed that for each $n \geq 2$ there exists at least one polynomial of degree n that achieves the best possible constant for polynomials of degree n. We will refer to such extremal polynomials as *globally maximal* polynomials. In fact it is conjectured that the polynomial $p(z) = z^n + z$ is (after normalisation) the unique globally maximal polynomial. David Tischler has shown that this polynomial is an isolated local maximum (see [72] and [73]).

5.1.1 Algebraic results on Smale's Conjecture

We will work in the family

$$\mathcal{G}_n = \{\text{monic polynomials } p \text{ of degree } n \text{ with } p(0) = 0 \text{ and } p'(0) = 1\}.$$

Each such polynomial is of the form

$$p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_2 z^2 + z,$$

and the vector $(a_2, \ldots, a_{n-1}) \in \mathbb{C}^{n-2}$ gives co-ordinates for the parameter space. Note that by pre- and post-composition with affine maps, every polynomial of degree n can be put into a unique such form, and that these transformations do not change the objective values in Smale's mean value conjecture.

Proposition 5.2.
For each $n \geq 2$ there exists a polynomial Q_n such that $(\lambda_1, \ldots, \lambda_{n-1})$ is the vector of objective values

$$\lambda_i = \frac{p(\zeta_i)}{\zeta_i\, p'(0)}$$

for some polynomial p of degree n if and only if

$$(\lambda_1, \ldots, \lambda_{n-1}) \in X \setminus T\,,$$

where X is the affine hypersurface in $\mathbb{C}^{n-1}$ defined by $Q_n = 0$ and T is some proper subvariety of X. In particular the set of possible vectors of objective values is an open subset of X in the usual topology.

Remark: the related problems of constructing a polynomial with given critical points or with given critical values do not have this feature: in those problems, every vector of objective values is possible (see [14]).

Proof. Let q be the polynomial given by $q(z) = \frac{p(z)}{z}$, where $p \in \mathcal{G}_n$. The objective values $\lambda_i = \frac{p(\zeta_i)}{\zeta_i}$ are the values of t for which $q(z)-t$ and $p'(z)$ have a common root in z, so they are the roots of the resultant

$$R(t) = \operatorname{Res}\left(q(z) - t, \frac{p'(z)}{n}\right) = \prod_{i=1}^{n-1}\prod_{j=1}^{n-1}(\alpha_i - \zeta_j),$$

where the α_i are the roots of $q(z) = t$ as a polynomial in z, and the ζ_j are the roots of $p'(z) = 0$, both repeated according to multiplicity. Note that the multiplicity of roots of $R(t)$ counts the λ_i appropriately. We can calculate the resultant $R(t)$ from Bezout's or Sylvester's determinant, so the coefficients of $R(t)$ are polynomials in the coefficients a_i. One can easily check that the leading term of $R(t)$ is $(-1)^{n-1}t^{n-1}$, independent of the choice of p. (This is a useful consequence of our normalisation of p). Let s_k be the sum of the products of the λ_i taken k at a time (repeating the λ_i according to their multiplicity). These elementary symmetric functions are given by the coefficients of $R(t)$ taken with appropriate signs. Thus for $k = 1, \ldots, n-1$ there are polynomials σ_k in $n-2$ variables such that

$$s_k = \sigma_k(a_2, \ldots, a_{n-1}).$$

Now there must be an algebraic relation $R_n(\sigma_1, \ldots, \sigma_{n-1}) = 0$ between the polynomials σ_k in the polynomial ring $\mathbb{C}[a_2, \ldots a_{n-1}]$, since there are $n-1$ of them. (If not, the polynomials σ_i would generate a subring of $\mathbb{C}[a_2, \ldots, a_{n-1}]$ of (Krull) dimension $n-1$, which is impossible.) In fact we can choose R_n to be irreducible; later we will check that this implies that the map $\varphi : (a_2, \ldots, a_{n-1}) \mapsto (\sigma_1(a_2, \ldots a_{n-1}), \ldots, \sigma_{n-1}(a_2, \ldots, a_{n-1})$ is a dominant morphism. On the other hand there is no non-trivial algebraic relation between the s_k in the polynomial ring generated by the variables λ_i. Therefore substituting the expansions of the s_k gives the required non-trivial algebraic relation between the λ_i, say

$$Q_n(\lambda_1, \ldots \lambda_{n-1}) = 0.$$

For the converse, we will show in Lemma 5.3 that the image of φ is $(n-2)$-dimensional, then the irreducibility of R_n implies that φ is dominant. Therefore the image of φ is $W \setminus S$, where W is the affine hypersurface defined by $R_n = 0$ and S is some proper subvariety of W. In particular the image of φ is an open subset of W in the usual topology (as well as in the Zariski topology). Now the map that sends the vector of roots of a monic polynomial of degree $(n-1)$ to the corresponding vector of elementary symmetric functions is a finite surjective morphism $\psi : \mathbb{C}^{n-1} \to \mathbb{C}^{n-1}$, of degree $(n-1)!$, so the quasi-affine variety $\psi^{-1}(W \setminus S) = X \setminus T$ has the properties described by the theorem.

We may also wish to include in S the image under φ of the vectors $(a_2, \ldots, a_{n-1})$ corresponding to polynomials with repeated roots; this is contained in the subvariety of W defined by $\prod_{i=1}^{n-1} \lambda_i = 0$. □

Lemma 5.3.
The image of φ is $(n-2)$-dimensional.

Proof. We consider the vector of values $c_i = p(\zeta_i)/\zeta_i$ as functions of the critical points $\zeta_1, \ldots, \zeta_{n-1}$, after dropping the requirement that $p'(0) = 1$. We claim that the mapping $\chi : (\zeta_1, \ldots, \zeta_{n-1}) \to (c_1, \ldots, c_{n-1})$ is a dominant morphism from $\mathbb{C}^{n-1}$ to $\mathbb{C}^{n-1}$. In fact we will show that there exists a polynomial p_0 at which the derivative of χ has rank $(n-1)$, such that the product function $\pi : (\zeta_1, \ldots, \zeta_{n-1}) \mapsto \prod_{i=1}^{n-1} \zeta_i$ takes the value 1 and has non-zero derivative at p_0. The image of φ is the image of χ restricted to the level set $(\pi = 1)$, so the implicit function theorem yields the lemma.

Let r_k be the k-th elementary symmetric function of the critical points. We may take p to be monic, so that

$$p'(z) = n\prod_{i=1}^{n-1}(z-\zeta_i) = n\left(z^{n-1} - r_1 z^{n-2} + r_2 z^{n-3} - \cdots + (-1)^{n-1} r_{n-1}\right) ,$$

$$\frac{p(z)}{nz} = \frac{z^{n-1}}{n} - r_1\frac{z^{n-2}}{n-1} + \cdots + (-1)^{n-1} r_{n-1}\frac{z^0}{1} .$$

In particular, $p(\zeta_i)/(n\zeta_i)$ is a polynomial function of the critical points. At the point $Z_0 = (1, 1, , \ldots, 1)$ we can compute the derivative matrix of χ as follows.

For $i \neq j$,

$$\begin{aligned} \frac{\partial}{\partial\zeta_j}\left(\frac{p(\zeta_i)}{n\zeta_i}\right) &= -\frac{\binom{n-2}{0}}{n-1} + \frac{\binom{n-2}{1}}{n-2} - \frac{\binom{n-2}{2}}{n-3} + \cdots + (-1)^{n-1}\frac{\binom{n-2}{n-2}}{1} \\ &= \frac{(-1)^{n-1}}{n-1} . \end{aligned}$$

For $i = j$,

$$\begin{aligned} \frac{\partial}{\partial\zeta_i}\left(\frac{p(\zeta_i)}{n\zeta_i}\right) &= \left[-\frac{\binom{n-2}{0}}{n-1} + \frac{\binom{n-2}{1}}{n-2} - \frac{\binom{n-2}{2}}{n-3} + \cdots + (-1)^{n-1}\frac{\binom{n-2}{n-2}}{1}\right] + \\ &\quad \left[\left(1-\frac{1}{n}\right)\binom{n-1}{0} - \left(1-\frac{1}{n-1}\right)\binom{n-1}{1} + \ldots\right. \\ &\quad \left.\cdots + (-1)^{n-1}\left(1-\frac{1}{1}\right)\binom{n-1}{n-1}\right] \\ &= \frac{(-1)^{n-1}}{n-1} + \frac{(-1)^{n-1}}{n} . \end{aligned}$$

It follows that the determinant of the derivative of χ at Z_0 is non-zero (because $\frac{1-n}{n}$ is not a root of the characteristic polynomial of the $(n-1)$ by $(n-1)$ matrix each of whose entries is 1. □

As a consequence of Proposition 5.2, we obtain are able to obtain Theorem 5.4, which may be a useful step towards a proof of Smale's conjecture. Bear in mind that Tischler has shown that for each degree n there is at least one polynomial *locally strictly maximal* in $\mathcal{G}_n$ for Smale's conjecture, namely $z^n + z$, (see [72] and [73]) (i. e. it is an isolated local maximum for min $\left|\frac{p(\zeta_i)}{\zeta_i p'(0)}\right|$). We know that the vector of objective values for any polynomial p lies inside a relatively open subset of the analytic set $X \setminus T$. Recall that we call a critical point ζ of p *essential* when $\left|\frac{p(\zeta)}{\zeta}\right|$ is minimal among all critical points of p.

We are not aware of any proof in the literature of the following highly plausible conjecture, although it is likely that the quasiconformal deformation method of Tischler could be adapted to prove it, if it were known that the best possible constant for Smale's mean value conjecture is a strictly increasing function of the degree.

Conjecture 2.

The (globally) maximal polynomials of degree n are contained in a compact region of the parameter space $\mathcal{G}_n$.

Theorem 5.4.

For any isolated local maximal polynomial of degree n, all the objective values in Smale's mean value conjecture have equal modulus, i. e. every critical point is essential. If conjecture 2 is true then the same applies to any globally maximal polynomial. Even without conjecture 2, if there exists a globally maximal polynomial of degree n, then there exists a globally maximal polynomial for which all the objective values have equal modulus.

The usefulness of this theorem is that (subject to proving conjecture 2) it reduces proving Smale's mean value conjecture to proving it for the special case where all the objective values are equal in modulus. This is a condition that might be suitable for attack by results of geometric function theory.

Proof. Suppose that $\hat{p}$ is a maximal polynomial for which the moduli of the objective values $\hat{\lambda}_i$ are not all equal; in fact suppose that the essential critical points are $\zeta_1, \dots \zeta_k$, (repeated according to multiplicity), where $k \leq n-2$. Write $v_0 = (\hat{\lambda}_1, \dots \hat{\lambda}_{n-1}) \in \mathbb{C}^{n-1}$. Consider the complex line L in $\mathbb{C}^{n-1}$ given by $x_1 = \hat{\lambda}_1, \dots, x_{n-2} = \hat{\lambda}_2$. Suppose that L were contained in X, (the analytic hypersurface $Q_n = 0$). Then $L \cap T$ must be finite since $v_0 \in L \setminus T$. But $L \cap (X \setminus T)$ cannot be all of $L \setminus T$ because then all points of $L \setminus T$ outside a disc in L (and including v_0) would come from polynomials achieving the same minimum objective modulus in Smale's conjecture as v_0, contrary to the hypothesis that $\hat{p}$ is a locally maximal polynomial. If conjecture 2 is true, then because φ is continuous and ψ is proper, this implies that the vectors of objective values for globally maximal polynomials of degree n are also contained in a compact set, which again shows the impossibility of $L \subset X$.

The Weierstrass Preparation Theorem has the following consequence.

> The zero locus of an analytic function $f(z_1, \dots, z_{n-2}, w)$, not vanishing identically on the w-axis, projects locally onto the hyperplane $(w = 0)$ as a finite-sheeted cover branched over the zero locus of an analytic function.

(See [37], chapter 0.1, for details).

Applying this to $f = Q_n$, it follows that there is an subset $U \subset X$, open in the usual topology, close to v_0, on which $|x_1| > |\hat{\lambda}_1|$, $|x_2| > |\hat{\lambda}_2|$, $\dots$, $|x_{n-2}| > |\hat{\lambda}_{n-2}|$, but $|x_n|$ is as close as we like to $|\hat{\lambda}_{n-1}|$. In U there must be a point v_1

of $X \subset T$ because T is a proper subvariety of X. Then v_1 shows that v_0 was not maximal for Smale's mean value conjecture after all. This contradiction shows that it is not possible for the objective values of a locally maximal polynomial not to be all equal.

Finally, if conjecture 2 is false and $\hat{p}$ is a globally maximal polynomial then we cannot use the Weierstrass Preparation Theorem, but we can find a point of the complex line L with $|x_{n-1}| = |\hat{\lambda}_1|$; indeed we can choose such a point to avoid T because T cannot contain a circle in L. By permuting the co-ordinates and repeating the argument we eventually obtain a globally maximal polynomial all of whose objective values are equal in modulus. □

Lemma 5.5.

For any p in $\mathcal{G}_n$, we have

$$\prod_{i=1}^{n} |\lambda_i| = \frac{\operatorname{disc}(p)}{n^{n-1}}.$$

In particular for a maximal polynomial p in $\mathcal{G}_n$, we have for all i

$$(n-1)/n \leq \min_i |\lambda_i| = \frac{(\operatorname{disc}(p))^{1/(n-1)}}{n} = \frac{(\operatorname{disc}(q))^{1/(n-1)}}{n}.$$

Hence any maximal polynomial p satisfies $\operatorname{disc}(p) \geq (n-1)^{(n-1)}$.

Proof. Suppose that p is a maximal polynomial in $\mathcal{G}_n$. Let $z_0 = 0, z_1, \ldots, z_{n-1}$ be the (distinct) roots of p, let $q(z) = p(z)/z$ and let $\zeta_1, \ldots \zeta_{n-1}$ be the roots of p', repeated according to multiplicity. Then the constant coefficient of

$R(t)$ is

$$\begin{aligned}\mathrm{Res}(q, \frac{p'}{n}) &= \prod_{j=1}^{n-1}\prod_{i=1}^{n-1}(z_j - \zeta_i) = \prod_{j=1}^{n-1}\frac{p'(z_j)}{n} \\ &= n\prod_{j=0}^{n-1}\frac{p'(z_j)}{n} = n\prod_{j=0}^{n-1}\frac{1}{n}\prod_{i\neq j}(z_j - z_i) \\ &= \frac{(-1)^{n(n-1)/2}}{n^{n-1}}\,\mathrm{disc}(p),\end{aligned}$$

where $\mathrm{disc}(p)$ stands for the discriminant of p. The condition $p'(0) = 1$ tells us that $\mathrm{disc}(p) = \pm\mathrm{disc}(q)$. The leading term of $R(t)$ is $(-1)^{n-1}t^{n-1}$. Recalling that the λ_i are the roots of $R(t)$, we find that

$$\prod_{i=1}^{n-1}|\lambda_i| = \frac{\mathrm{disc}(p)}{n^{n-1}}.$$

Since p is maximal for Smale's mean value conjecture, we know from Theorem 5.4 that the $|\lambda_i|$ are all equal, which gives the first part of the corollary. The second part arises from the fact that we know $|\lambda_i| \geq \frac{n-1}{n}$ for a maximal polynomial, because of the example $z^n + z$. □

This immediately gives us as a corollary a result originally proved by Tischler using a different method.

Theorem 5.6 (Tischler).
Among polynomials p in $\mathcal{F}_n$ such that the roots of $q(z) = p(z)/z$ all lie on the unit circle, $p(z) = z^n + z$ is maximal for Smale's mean value conjecture.

Proof. It is well-known that among all monic polynomials q whose roots all lie on the unit circle, the modulus of the discriminant is maximised uniquely when those roots are equally spaced. Observe that the minimum of the $|\lambda_i|$

is at most their geometric mean, with equality only if they are all equal. The further condition that $q(0) = p'(0) = 1$ leaves us with a unique polynomial maximal among those whose roots lie on the unit circle, namely $q(z) = z^{n-1} + 1$. □

Any explicit information that we can obtain about the polynomial Q_n may help us to pin down the extremal polynomials. For example, in the case $n = 3$ we have

$$Q_3(\lambda_1, \lambda_2) = (\lambda_1 - 1/2)(\lambda_2 - 1/2) - 1/36.$$

Setting $Q_3 = 0$ forces either λ_1 or λ_2 to lie in the closed disc of radius 1/4 centred on 1/2, therefore to have modulus at most 3/4. The only case of equality here is $\lambda_1 = \lambda_2 = 3/4$, which corresponds (uniquely, after normalisation) to the polynomial $p(z) = z^3 + z$.

The relation Q_4 is already formidable: Let $u = 4(\lambda_1 - 1/2)$, $v = 4(\lambda_2 - 1/2)$, and $w = 4(\lambda_3 - 1/2)$. Then

$$\begin{aligned} &Q_4(\lambda_1, \lambda_2, \lambda_3) = \\ &\quad 3 - 4\,(u + v + w) - 4\,(uv + vw + wu) \\ &\quad + 8\,(u^2v + v^2u + v^2w + w^2v + w^2u + u^2w) + 33\,uvw \\ &\quad + 4\,uvw(u + v + w) - 16\,uvw(u^2 + v^2 + w^2) \\ &\quad - 52\,uvw(uv + vw + wu) - 3u^2v^2w^2 \\ &\qquad + 72\,u^2v^2w^2(u + v + w) - 81\,u^3v^3w^3. \end{aligned}$$

Unfortunately it is not possible to separate the variables to rewrite the equation $Q_4 = 0$ as

$$R(\lambda_1) \cdot R(\lambda_2) \cdot R(\lambda_3) = \text{constant},$$

where R is any rational function.

As for Q_5, a crude estimate to bound its degree shows that the degree is at most 3420; it is probably rather smaller, but nevertheless Q_5 is likely to be impossible to calculate explicitly.

One further thing we can do is apply calculus to variation of the arguments of λ_i. This gives the condition for a maximal polynomial that $\arg(\lambda_i \frac{\partial Q_n}{\partial \lambda_i})$ must be equal for all $i = 1, .., n-1$. Together with the condition that the λ_i have equal moduli and satisfy Q_n, this gives in theory enough constraints to leave only a finite set of candidate maxima to examine. However since we are unable to make effective use of this information, we omit the details.

5.2 The mean value conjecture for rational maps

In Smale's original proof (to obtain the constant 4), the only fact about polynomials as special examples of rational maps that is used is that $P(\infty) = \infty$. The proof does not use the fact that ∞ is a critical point for P or that there are no other pre-images of ∞. This allows us to use the Smale's proof to prove a version of the mean value conjecture for rational maps:

Theorem 5.7.

Let R be any rational map of degree at least 2, and let x, y be points of $\widehat{\mathbb{C}}$ with $R(x) \neq R(y)$, such that x is not a critical point of R. Then there exists a critical point ζ of R and a Möbius map M such that $M \circ R$ fixes each of x, y and ζ, and with respect to any local co-ordinate at x we have $|(M \circ R)'(x)| \geq 1/4$.

Remark: The special case with $y = \infty$ and R a polynomial is a simple reformulation of Smale's result. Because x is fixed by $M \circ R$, the derivative at x of $M \circ R$ does not depend on the choice of local co-ordinate.

Proof. The statement is certainly true for R if it is true for $\hat{R} = T \circ R \circ S$, where T and S are Möbius maps. Choose a Möbius map S such that $S(\infty) = y$ and $S(0) = x$, and a Möbius map T such that $T(R(y)) = \infty$ and $T(R(x)) = 0$. Then $\hat{R} = T \circ R \circ S$ is a rational map fixing 0 and ∞. Let D be the largest open disc centred on 0 that carries a single-valued branch β of $\hat{R}^{-1}$ mapping 0 to 0; let $\tilde{D}$ be the image of β, i. e. the component of $\hat{R}^{-1}(D)$ containing 0. There is a critical point ζ of $\hat{R}$ on the boundary $\partial\tilde{D}$, with $\hat{R}(\zeta) \in \partial D$. Now β is a univalent map from the disc about 0 of radius $|\hat{R}(\zeta)|$ to the domain $\tilde{D}$, which omits ζ, so Koebe's $\frac{1}{4}$-Theorem gives

$$|(\hat{R}^{-1})'(0)| \leq 4\frac{|\zeta|}{|\hat{R}(\zeta)|}.$$

□

We now ask what is the best possible constant in Theorem 5.7 to replace 1/4, (perhaps in terms of the degree of R). The example $R(z) = z^n + z$, $y = \infty$ shows that $n/(n-1)$ would be best possible. The proof precludes the case of equality, because then the branch β of R^{-1} would cover the whole of the Riemann sphere except for a slit, and the complement of D could not be covered by R. A degree 2 rational map must have exactly two simple critical points; after Möbius maps at both ends we may assume that the map is $z \mapsto z^2$, and that $x = 1$ in the theorem; in this case the derivative at x is 2. So the situation for degree 2 is no different from the polynomial case.

The reader may think that this generalised mean value conjecture looks a little unnatural due to the unequal roles of x and y. We can make it more symmetrical by letting y be a variable *critical* point, giving it the same status as ζ. Here is the analogue of Smale's Theorem:

Theorem 5.8.
Let R be a rational map of degree at least 2, and $x \in \widehat{\mathbb{C}}$ be given, such that $R(x)$ is not a critical value of R. Then there exist two critical points ζ and κ of R such that if M is the Möbius map that makes x, ζ and κ fixed points of $M \circ R$, then with respect to any local co-ordinate at x we have $|(M \circ R)'(x)| \geq 1/2$.

Proof. The conclusion is unchanged if we replace R by $M \circ R \circ N$, where M and N are any Möbius maps, so w. l. o. g. we may assume $x = \infty = R(x)$; then all the critical values of R are finite and there are at least two of them. Let K be the convex hull of the critical values. Then there exist two critical points ζ and κ such that $\operatorname{diam}(K) = |R(\zeta) - R(\kappa)| > 0$. Now there exists a single-valued branch β of R^{-1} defined on $\widehat{\mathbb{C}} \setminus K$, taking ∞ to ∞, and omitting ζ and κ. Let $f : \widehat{\mathbb{C}} \setminus \overline{\mathbb{D}} \to \widehat{\mathbb{C}} \setminus K$ be a Riemann map fixing ∞. The logarithmic capacity of a compact subset of $\mathbb{C}$ of diameter d is at most $d/2$ [62, Theorem 5.3.4] so $f^{\#}(\infty) \geq 2/|R(\zeta) - R(\kappa)|$. The Koebe $\frac{1}{4}$-Theorem tells us that $(\beta \circ f)^{\#}(\infty) \leq 4/|\zeta - \kappa|$, so we have $\frac{1}{R^{\#}(\infty)} = \beta^{\#}(\infty) \leq 2\frac{|R(\zeta)-R(\kappa)|}{|\zeta-\kappa|}$. Take M to be the affine map that makes $M \circ R$ fix ζ, κ and ∞. Then $M^{\#}(\infty) = \frac{|R(\zeta)-R(\kappa)|}{|\zeta-\kappa|}$, which gives the result. □

As before, we ask what is the best possible constant to replace 1/2 in Theorem 5.8. Taking $R(z) = (z^n - nz)/(1 - n)$, for which zero and all the critical points including ∞ are fixed, there is only one choice for M, namely the identity map; in this case we have $(M \circ R)'(0) = (n/n-1)$. So the constant $n/(n-1)$ to replace $1/2$ in Theorem 5.8 would be best possible, and the constant 1 would be the best possible bound independent of the degree of R.

5.2.1 Rational maps with all critical points fixed

If we can arrange for every critical point to be fixed and for there to be at least one non-critical fixed point left over, then we get an example for Theorem 5.8 in which there is only one choice of Möbius map M, making it easy to compute an upper bound on the best possible constant for that theorem. A rational map of degree $n \geq 2$ has $2n-2$ critical points and $n+1$ fixed points, both counted with multiplicity, so there will have to be some multiple critical points. In this section we will present a method of constructing rational maps all of whose critical points are fixed, and which have only one further fixed point (which after conjugation we may take to be ∞). Note that there are rational maps all of whose critical points are fixed but which do not have this additional property, for example $z \mapsto z^n$ for $n \geq 3$.

Suppose that g is a polynomial of degree n with no repeated roots, such that all the roots of g'' are also roots of g; it follows that g' has no multiple roots. Then let R_g be the associated Newton-Raphson map

$$R_g(z) = z - \frac{g(z)}{g'(z)}.$$

R_g has a superattracting fixed point (i. e. a critical fixed point) at each of the roots of g. In fact

$$R_g'(z) = \frac{g(z)g''(z)}{g'(z)^2},$$

so the only critical points of R_g are the roots of g. Observe that R_g has no multiple poles since g' has no multiple roots. In particular, *every critical point of R_g is a fixed point.* Note that R_g has a fixed point at ∞ with multiplier $R^{\#}(\infty) = \frac{n}{n-1}$. Any such R_g will show that the best possible constant for Theorem 5.8 is at most $n/(n-1)$, just as the polynomial $p(z) = (z^n + nz)/(n-1)$ does. In fact, when we take $h(z) = a(z-c)^n + b(z-c)$, we find that R_h is Möbius-conjugate to p, via $z \mapsto 1/z$; moreover, this is the only case of this construction for which R_g can be conjugate to a polynomial!

Here is an example of an R_g which is not conjugate to a polynomial, but all of whose critical points are fixed. Take

$$h(z) = z^4 + 2z^3 + 6z^2 + 5z + 4 = (z^2 + z + 4)(z^2 + z + 1)\,.$$

For this h, we have

$$R_h(z) = \frac{3z^4 + 4z^3 + 6z^2 - 4}{4z^3 + 6z^2 + 12z + 5}$$

$$h''(z) = 12(z^2 + z + 1)\,.$$

Since h has no repeated roots, the roots of h'' are roots of h but not of h'. Thus R_h has two finite fixed points of valency three, two finite fixed points of valency two, and no other critical points. R_h is not conjugate to a polynomial because it does not have a fixed point of valency equal to its degree.

In Theorem 5.8, it seems likely that there are several extremal rational maps for each degree. This would be interesting for the following reason: any method that solves Smale's mean value conjecture might be expected also to apply to give the best possible constant for Theorem 5.8, so it should not rely on the uniqueness of the extremum!

The following lemma characterises those rational maps that occur as Newton-Raphson maps of rational functions in terms of their fixed points and the multipliers at those fixed points. Buff and Henriksen [22] characterise the rational maps that occur as König's methods for polynomials, and this includes Newton's method for polynomials as a special case for which they give priority to [40, Prop. 2.1.2]. The extension here to Newton maps of rational functions may possibly be new.

Lemma 5.9 (Characterisation of Newton-Raphson maps of rational functions).
A rational map R is the Newton-Raphson map associated to some non-linear rational function g if and only if all the fixed points of R are simple and each finite fixed point χ of R satisfies $\frac{1}{1-R'(\chi)} \in \mathbb{Z}$. Moreover, g is a polynomial if and only if these integers are all positive.

The condition that all fixed points be simple applies even to any fixed point at ∞.

Proof. Suppose that $g(z) = p(z)/q(z)$ in lowest terms is a rational function of degree $n \geq 2$ and $R = R_g(z) := z - \frac{g(z)}{g'(z)}$. The fixed points of R are precisely the roots and poles χ of g. If g has order $m \neq 0$ at $\chi \in \mathbb{C}$ then

$$1 - R'(\chi) = \frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{g(z)}{g'(z)}\right)\Bigg|_{z=\chi} = \frac{1}{m}\,.$$

For the converse, consider

$$g(z) := \exp \int^z \frac{\mathrm{d}z}{z - R(z)}\,.$$

This function is certainly locally defined away from fixed points of R and satisfies the Newton-Raphson equation $R(z) = z - \frac{g(z)}{g'(z)}$. The integrand $\frac{1}{z-R(z)}$ has no multiple poles since at each fixed point of R we are told $\frac{\mathrm{d}}{\mathrm{d}z}(z - R(z)) \neq 0$. We can therefore express the integrand in partial fractions as

$$\frac{1}{z - R(z)} = q(z) + \sum_i \frac{A_i}{z - \chi_i},$$

where q is a polynomial. R does not have a multiple fixed point at ∞, so $z - R(z) \to \infty$ as $z \to \infty$, hence $q = 0$. Near a fixed point χ_i of R we know that $z - R(z) = (z - \chi_i)/m + O((z - \chi_i)^2)$ for some positive integer m_i, so $A_i = m_i$. Now we can perform the integration explicitly:

$$\int^z \frac{\mathrm{d}z}{z - R(z)} = \sum_i m_i \log(z - \chi_i) + c,$$

so

$$g(z) = \exp(c). \prod_i (z - \chi_i)^{m_i},$$

which is a rational function; it is a polynomial when all $m_i \geq 1$. □

The following special case is also a special case of [22, Prop. 4].

Corollary 5.10 (Characterisation of Newton-Raphson maps associated to polynomials without repeated roots).
The following are equivalent for a rational map R:

1. *R has a simple fixed point at ∞ and all the finite fixed points of R are also critical points of R;*

2. *$R(z) = z - \frac{g(z)}{g'(z)}$ for some non-linear polynomial g with no repeated roots.*

In this situation, the fixed point of R at ∞ is not critical; indeed its multiplier is $n/(n-1)$, where $n = \deg g = \deg R$.

If a Newton-Raphson map R_g associated to a polynomial g has all its critical points fixed, then all of its finite fixed points must be critical, and hence g cannot have any repeated root. Indeed, R_g would otherwise have a non-critical fixed point of multiplier less than 1 in modulus, and then a suitable iterate of R_g would violate Theorem 5.8.

Note that for a Newton-Raphson map associated to a rational function g of degree n, we could have repelling fixed points associated to poles of g, but the multiplier would be $k/(k-1)$ for a pole of g of order k. It is possible for the degree of R_g to be less than the degree of g, so we might still hope to produce better examples for Theorem 5.8 using the present construction. However in the next section we will show that the multiplier $n/(n-1)$ is the smallest possible multiplier of a non-critical fixed point of any R_g of degree n whose critical points are all fixed.

A final observation: R has a simple non-critical fixed point at ∞ precisely when

$$\lim_{z\to\infty} \frac{R(z)}{z} \in \widehat{\mathbb{C}} \setminus \{0, 1, \infty\}.$$

5.2.2 Forbidden multipliers

The special case of Smale's mean value conjecture for polynomials in which the critical points are all fixed is often referred to as Kostrikin's conjecture. It was observed by Shub that if the critical points of a polynomial are all fixed, then the multiplier at each remaining fixed point must be greater than or equal to 1 in modulus (just plug a suitable iterate of the given map into

Smale's Theorem). We will give a different argument which in fact applies in general to rational maps whose critical points are all fixed, and shows that the multiplier must be strictly greater than 1 in modulus.

To see this, consider first the case in which R has only two critical values. Then R pulls back a complete Euclidean metric from $\widehat{\mathbb{C}} \setminus \{\text{critical values}\}$ to $\widehat{\mathbb{C}} \setminus R^{-1}(\{\text{critical values}\})$, so there can only be two critical points, and the map is Möbius-conjugate to $z \mapsto z^n$, in which the multiplier at the remaining fixed points is n. Otherwise, there are at least three critical values, and the multivalued map $R^{-1} : \widehat{\mathbb{C}} \setminus \{\text{critical values}\} \to \widehat{\mathbb{C}} \setminus R^{-1}(\{\text{critical values}\})$ lifts to a conformal isomorphism between the universal covers of these two domains. Thus any branch of R^{-1} is locally a hyperbolic isometry between the natural hyperbolic metrics on $\widehat{\mathbb{C}}\setminus\{\text{critical values}\}$ to $\widehat{\mathbb{C}}\setminus R^{-1}(\{\text{critical values}\})$. However, the inclusion map $I : \widehat{\mathbb{C}}\setminus\{\text{critical values}\} \to \widehat{\mathbb{C}}\setminus R^{-1}(\{\text{critical values}\})$ omits some points because it is impossible for each of the critical values to have only one pre-image. Therefore the inclusion map is everywhere a strict contraction between the hyperbolic metrics. At any non-critical fixed point this shows that the multiplier is greater than 1 in modulus.

The following theorem excludes further values of the multiplier at any non-critical fixed point, thus making a small amount of progress on Kostrikin's Conjecture.

Theorem 5.11.

Suppose that R is a rational map all of whose critical points are fixed. Suppose that R has degree n and has m critical points (not counting multiplicity). Then R has exactly $n+1$ fixed points, of which $n+1-m$ are non-critical. The multiplier at any non-critical point does not lie in the closed disc whose *diameter is the interval* $[1, 1+\frac{2}{n+m-2}]$*, except in the case where* $m = n$ *and the multiplier of the remaining fixed point is* $n/(n-1)$*, which is on the boundary of this disc.*

Proof. We use the notion of residue fixed point index for holomorphic maps. If $f : U \to \mathbb{C}$ is holomorphic on an open set $U \subset \mathbb{C}$ and z_0 is an isolated fixed point of f, then the residue fixed point index of f at z_0 is defined as

$$\iota(f, z_0) = \frac{1}{2\pi i} \int \frac{dz}{z - f(z)} \ ,$$

where the integral is taken around a positively-oriented circle around z_0 so small that it contains no other fixed points of f. If the multiplier of f at z_0 is $\lambda \neq 1$ then $\iota(f, z_0) = \frac{1}{1-\lambda}$. If the fixed point is simple then the index is still well-defined and finite, but this formula does not apply.

Theorem 5.12 (Rational Fixed Point Theorem). *[54, Theorem 12.4] For any rational map* $f : \widehat{\mathbb{C}} \to \widehat{\mathbb{C}}$ *which is not the identity map, the sum over all the fixed points of the residue fixed point index is* 1.

The first part of Theorem 5.11 is a consequence of Shub's observation that any non-critical fixed points must be repelling; in particular they are simple, so there are no multiple fixed points of R. The residue fixed point index of a fixed critical point is 1. For a non-critical fixed point, the multiplier has modulus strictly greater than one, so the residue fixed point index has real part strictly less than 1/2. In fact in the case where there are n critical fixed points, the rational fixed point theorem shows that the remaining fixed point must have residue fixed point index $1-n$, so must have multiplier $n/(n-1)$. In the general case, select a particular non-critical fixed point z_0. We get a contribution of less than $m + \frac{(n-m)}{2}$ to the real part of the total index from all the other fixed points, so the residue fixed point index $\iota(R, z_0)$ has real part greater than $1 - (m + \frac{(n-m)}{2}) = 1 - \frac{m+n}{2}$, and hence the multiplier does not lie in the disc whose diameter is the interval $[1, 1 + \frac{2}{m+n-2}]$, as required. In fact, further small regions of excluded values of the multiplier may be found by considering iterates of R. □

Note that in the case $m < n$, this disc of excluded values contains the multiplier $\frac{n}{n-1}$ in its interior. Since the multipliers at the fixed points of a Newton-Raphson map associated to a rational function are real and non-negative, Theorem 5.11 implies that one cannot improve on the multiplier $n/(n-1)$ by using Newton-Raphson maps associated to rational functions.

Appendix A

Applications of IFS

The motivation in this thesis for studying IFSs arises from conformal dynamical systems. To put our results in context, we discuss here various other applications of IFSs.

An IFS is often used to model a discrete dynamical system perturbed by random noise, perhaps as a discrete approximation to a stochastic differential equation. Typically the unperturbed system will be stable, say with a locally attracting fixed point. The IFS corresponding to a small perturbation of this system may have an invariant measure concentrated near the original fixed point. However, it may also happen for a dynamical system with many distinct locally attracting fixed points that even small perturbations are asymptotically stable.

A well-known application of N-map IFSs is in computer graphics, for generating natural-looking landscapes and textures. Michael Barnsley's fern [7] is a famous example. These pictures are produced by plotting the points of a random orbit of an N-map continuous IFS. In order to be sure that this will produce the required picture, we would like to know that with probability one the distribution of the first n points of the orbit will converge (as $n \to \infty$) to some non-random probability measure, which represents the picture. As we will see (particularly in section 2.3.2), this requirement is closely related to another property of IFSs called *asymptotic stability*, which will be of central importance to us.

For *image compression* one must solve the inverse problem of finding an IFS with a reasonably small number of simple maps whose invariant measure approximates a given probability measure. One approach is to look for an IFS of affine maps, using moments of the target measure to determine parameters. Although there are commercial image compression systems which encode digital images in terms of collections of IFSs, they have not been successful because they are computationally intensive in comparison to compression methods based on Fourier or wavelet transforms (such as JPEG) yet they do not typically achieve much better compression ratios.

Another common reason for introducing an IFS is to represent a given Markov chain. Consider a Markov Chain with state-space Y and transition probabilities $P(x, A)$, where for each $x \in Y$, $P(x, \cdot)$ is a Borel probability measure on Y, and for each Borel set $A \subset Y$, $P(\cdot, A)$ is a Borel-measurable function. Any such Markov Chain is representable by an IFS (see chapter 1 of [50]). However not every such Markov Chain can be represented by a continuous IFS. Indeed, for a continuous IFS, the map from Y to $\mathcal{P}(Y)$ (with the weak topology) given by $x \mapsto P(x, \cdot)$ is necessarily continuous. Blumenthal and Corson [17] gave a partial converse: they showed that any Markov chain on a connected, locally connected and compact space Y such that the transition map $Y \to \mathcal{P}(Y)$ is continuous, and such that for each $x \in Y$ the support of $P(x, \cdot)$ is the whole of Y, is representable by an IFS of continuous maps. Another restrictive condition on a Markov chain is the existence of a representation by an IFS with finitely many maps (not necessarily continuous). Such a representation exists if and only if the transition probabilities $P(x, A)$ take only finitely many distinct values. Finally, representability by a continuous N-map IFS is a more subtle question. For example, consider the Markov Chain $(x_i)_{i=0}^{\infty}$ with state space $\mathbb{C}$ such that $\mathbb{P}(x_{n+1}^2 = x_n) = 1$, where the two square roots are taken with equal probability when $x_n \neq 0$. The corresponding map $\mathbb{C} \to \mathcal{P}(\mathbb{C})$ is continuous and the transition probabilities take only finitely many values, yet this chain cannot be represented using any continuous IFS.

The survey article [27] gives an example of the the use of an IFS to represent the waiting time process in the G/G/1 queue, and gives references to further applications in queuing theory. The same article describes the recent method of Propp and Wilson which uses backward iteration of a suitably contractive IFS to simulate exactly from a distribution on a very large finite state space, discussing in particular the Ising model on a large but finite grid in two dimensions. That method is based on Letac's principle, among other ideas, and in this connection theorem 2.24 may be of interest.

Appendix B

Known stability results for IFS

There are many papers in which IFSs are proved to be asymptotically stable as a consequence of contractivity conditions on the defining maps. [27] is a very readable recent survey article on applications of contractive IFSs. It includes a complete proof of the following theorem (which was known before) and gives references to papers containing stronger results.

Theorem B.1.

Let $\mathcal{F}$ be an IFS on a complete separable metric space (Y, d). Suppose that f is a.s. Lipschitz, with $\mathbb{E} \log \mathrm{Lip}(f) < 0$, and that there exist $\alpha > 0$ and $\beta > 0$ such that $\mathbb{P}(\mathrm{Lip}(f) > u) < \alpha/u^{\beta}$ for all $u > 0$. Then $\mathcal{F}$ is asymptotically stable and there exists a point $x_0 \in X$ and constants $a > 0, b > 0$ and $0 < r < 1$ such that

$$\pi_d \left((\mathcal{F}_*)^n \delta_x, \mu \right) \leq (a + bd(x, x_0)) r^n \, ,$$

where π_d is the Prohorov metric and μ is the invariant law of $\mathcal{F}$.

The idea of using the natural extension of the skew product to relate the forward and reverse iterates was used by Elton [30] to prove stability results and ergodic theorems in the situation of the above theorem, but allowing the sequence of maps (F_n) to be a more general stationary process than a Bernoulli process. Then the sequence $F_n \circ \cdots \circ F_1(Z_0)$ no longer forms a Markov chain, but what has been called a Markov chain in a random environment. This raises the question of whether our results on non-uniformly contracting IFSs can be extended to deal with stationary sequences of maps.

For IFSs on a complete separable metric space (X, d), Stenflo [70] supposes that there exists $c < 1$ such that for all $x, y \in X$,

$$\mathbb{E}(d(f(x), f(y)) \leq cd(x, y)\,,$$

and

$$\mathbb{E}(d(x, f(x)) < \infty$$

He proves that there exists a unique invariant law μ and that for any $x \in X$,

$$K_d\left(\mathcal{F}_*^n(\delta_x), \mu\right) = O(c^n)\,,$$

uniformly on bounded subsets. For non-uniformly contracting IFSs, one cannot hope to give any similar bound on the rate of convergence. In the same situation, Stenflo proves the continuous dependence of the invariant law on the parameters defining the IFS.

A more common assumption of average contractivity is a spatially uniform bound

$$\mathbb{E}\left(\log d(f(x), f(y)) - \log d(x, y)\right) < -\epsilon < 0\,.$$

Barnsley and Elton [10] prove asymptotic stability of IFSs using a slightly extended version of this. [9] shows asymptotic stability using this assumption and allowing *place-dependent probabilities*; in the same situation [74] proves strong results (including a law of the iterated logarithm) about the behaviour of the average of a continuous function h over the first n steps of the orbit, approximating a suitable normalisation of this sequence by a Brownian motion.

The subject of iterated function systems driven by stationary sequences of maps was also developed in Romania under the name of dependence with complete connections; a paper from that school relevant to the present discussion is [38], which generalises the result of [10].

Define the local Lipschitz constant of a map $h : \mathbb{X} \to \mathbb{X}$ to be

$$D_x h = \limsup_{y \to x} \frac{d(f(x), f(y))}{d(x,y)} .$$

Steinsaltz [68] studies IFSs that he calls locally contractive. Here X is a convex subset of $\mathbb{R}^n$ with the Euclidean metric, and it is assumed that there is a function $\phi : X \to [1, \infty)$ and a constant $c < 1$ such that for all $n \in \mathbb{N}$ and all $x \in X$,

$$\mathbb{E}\left(D_x(F_1 \circ \cdots \circ F_n)\right) \leq \phi(x) c^n . \tag{B.1}$$

The main result is that $\mathcal{F}$ is asymptotically stable, with an explicit rate of convergence given for the sequence of reverse iterates. A sufficient condition for (B.1) to hold is given that does not involve iteration. Note that the brief literature survey in [68] repeatedly omits the crucial condition of completeness. It would be interesting to see whether Steinsaltz' results can be applied to simplify any of the results about analytic IFSs on $\mathbb{D}$ in Chapter 2.

There are non-uniformly contracting IFSs which do not fall into any of the classes discussed above, and of course each of these classes contains IFSs which are not non-uniformly contracting.

In the setting of IFSs of maps that do not increase distance, two recent papers [1, 53] go even beyond non-uniform contractivity. They prove the stability of the IFS on $\mathbb{R}^+$ defined by

$$f_\omega(x) = |x - \omega| \,,$$

where $\omega \in \Omega = \mathbb{R}^+$ and the law $\mathbb{P}$ of ω is compactly supported but not supported on any lattice in $\mathbb{R}$. Note that in this case each map is Lipschitz with constant 1, but none of the maps is strictly distance-decreasing. Moreover, the stability result is very delicate indeed: in the N-map case, stability fails when $\mathbb{P}$ is supported on a lattice, so the unstable IFSs are dense in the parameter space.

Appendix C

Dependence of stability on p

In this appendix we give some examples of analytic IFS on $\mathbb{D}$. They show that no condition on the maps alone can be both necessary and sufficient for stability. In all three examples, the maps are fixed but the existence of an invariant probability measure depends on the choice of the associated positive probability vector. From the point of view of random walks this is not surprising, but in the context of IFSs it is interesting – Lasota and Yorke [46] showed that for the closely related class of non-expansive IFSs on *compact* metric spaces, stability depends only on the maps and not on the probabilities. The examples also demonstrate that for a fixed set of maps, neither the existence nor the absence of an invariant measure is necessarily a convex condition on the probability vector. In particular the condition for stability is not necessarily linear.

Example C.1. Define a 2-map IFS $\mathcal{F}$ by

$$f_1 : z \mapsto z^2$$

and

$$f_2 : z \mapsto \frac{z+\tau}{1+\tau z}\,,$$

where $\tau > 0$ is chosen so that f_2 is the hyperbolic isometry of $\mathbb{D}$ that translates the geodesic along the real axis by hyperbolic distance $\log 2$ towards 1. The interval $[0,1)$ is a forward-invariant set for both maps. If $\mathcal{F}$ is asymptotically stable then its restriction to $[0,1)$ is asymptotically stable, and conversely if the restriction is asymptotically stable then the unrestricted system has an invariant probability distribution, so by Ambroladze's Theorem is asymptotically stable. The restriction to $[0,1)$ behaves rather like a random walk with a reflecting barrier at 0, since on the real axis near 1, $z \mapsto z^2$ displaces points towards 0 by a hyperbolic distance that approaches $\log 2$. When $p_1 > \frac{1}{2}$ we will show how to bound the Markov chain $H_n = F_n \circ \cdots \circ F_1(0)$ between two simpler random walks with retaining barriers, both coupled to H_n. Each of these bounding walks has a unique invariant probability distribution, to which its empirical distribution almost surely converges. Letting δ_0 be the unit mass at 0, it follows that $\mathcal{F}^n_*(\delta_0) \nrightarrow 0$ weakly; hence by Ambroladze's Theorem $\mathcal{F}$ has an invariant probability measure and is asymptotically stable. The invariant measure is supported on $[0,1)$ because each $\mathcal{F}^n_*(\delta_0)$ is supported on $[0,1)$.

Lower bound: Consider a random walk Z_n on the points x_m of $[0,1)$ such that $d(0,x_m) = m\log 2$, $m \in \mathbb{N}$, defined as follows. $Z_0 = 0$; the transition from Z_{n-1} to Z_n is a step to the right when $F_n = f_2$ and a step to the left when $F_n = f_1$, with the exception that if $Z_{n-1} = 0$ and $F_n = f_1$ then $Z_n = 0$ again. Since for all $x \in (0,1)$, we have $d(x^2, x) < \log 2$, we obtain $Z_n \leq H_n$ for all n, as required. It is well known that the chain Z_n is recurrent (persistent) when $p_1 > p_2$.

Upper bound: The upper bound is slightly more complicated. Let $R \in (0,1)$. Consider a random walk Y_n on $(0,1)$ such that $Y_0 = 0$ and the transition from Y_{n-1} to Y_n is a step to the right by a hyperbolic distance $\log 2$ when $F_n = f_2$ but is a step to the left by a hyperbolic distance $\frac{p_2+p_1}{2p_1}\log 2$ when $F_n = f_1$, except when this would make $Y_n < R$, in which case $Y_n = Y_{n-1}$. (We have introduced a retaining barrier at R). Except when Y_n is close to R, the expected increment in $d(0,Y_n)$ is $-p_1\frac{p_1+p_2}{2p_1}\log 2 + p_2 \log 2 = \frac{p_2-p_1}{2}\log 2$, so when $p_1 > p_2$, the chain (Y_n) has an invariant distribution. To ensure that Y_n is an upper bound for H_n, we need only choose R large enough that for all $1 > t \geq R$, we have $d\left(t,\sqrt{t}\right) > \frac{p_2+p_1}{2p_1}\log 2$, and since $d\left(t,\sqrt{t}\right) \to \log 2$ as $t \to 1$, there does exist such an R.

When $p_1 \geq p_2$, the chain Z_n is still a lower bound for H_n, but it has no invariant distribution (even though in the case $p_1 = p_2 = \frac{1}{2}$ it is recurrent). In particular, for any $t \in (0,1)$, $\mathbb{P}(Z_n \in [0,t]) \to 0$ as $n \to \infty$, and since $Z_n \leq H_n$, we have $\mathbb{P}(H_n \in [0,t]) \leq \mathbb{P}(Z_n \in [0,t])$. Therefore $\mathcal{F}_*^n(\delta_0) \to 0$ weakly, so by Ambroladze's Theorem $\mathcal{F}$ has no invariant measure.

Example C.2. In this example, stability is not a convex condition on the probability vector. The details are somewhat technical (using hyperbolic trigonometry) but the basic idea is as in example C.1: we find some Liapunov function ϕ on $\mathbb{D}$ (i. e. a continuous real-valued function with $\phi(z) \to \infty$ as $z \to \partial\mathbb{D}$) and consider the process $\phi(F_n \circ \cdots \circ F_1(0))$. Then we construct a Markov chain with values in $\mathbb{R}$ that is a lower or upper bound for this process and which we can prove is transient or positively recurrent, respectively. This general idea appears a great deal in the literature on IFSs.

The maps for this example will all be conjugate to $g : z \mapsto z^2$, so individually they are easy to understand. Pick two points x_1, x_2 of $\mathbb{D}$ with $d(x_1, x_2) = t$, and fix a large positive integer N. (In due course we will specify how large N and t must be). For $i = 1, 2$, let σ_i be a hyperbolic isometry that carries x_i to 0. Let h be the rotation $h : z \mapsto e^{\frac{2\pi i}{N}} z$. Then the maps of our example are the following, as j runs over $\{0, \ldots, N-1\}$ and i runs over $\{1, 2\}$:

$$f_{i,j} = \sigma_i^{-1} \circ h^j \circ g \circ \sigma_i.$$

Note that $h^j \circ g$ is in fact a conjugate of g.

To specify the probabilities associated to these maps, choose $\alpha \in (0, 1)$. Set the probability $p_{1,j}$ associated to each map $f_{1,j}$ equal to α/N, and the probability $p_{2,j}$ associated to each map $f_{2,j}$ equal to $(1-\alpha)/N$. We have now defined an IFS; call it $\mathcal{F}$.

Claim.

1. For α sufficiently close to 0 or to 1, $\mathcal{F}$ is asymptotically stable.

2. For $\alpha = \frac{1}{2}$, $\mathcal{F}$ has no invariant measure.

Proof. 1. Swapping σ_1 and σ_2 gives another system satisfying the above description, but with α and $1 - \alpha$ swapped; thus it suffices to consider the case where α is close to 0, so that the $f_{1,j}$ occur only rarely. The maps $f_{2,j}$ have fixed point x_2, so we will look at $r_n = d(x_2, H_n)$, where $H_n := F_n \circ \cdots \circ F_1(x_2)$. When $F_n = f_{1,j}$ for some j, we apply the triangle inequality twice to obtain

$$\begin{aligned} r_n - r_{n-1} &\leq d(x_2, x_1) + d(x_1, H_n) - d(x_2, H_{n-1}) \\ &\leq d(x_2, x_1) + d(x_1, H_{n-1}) - d(x_2, H_{n-1}) \\ &\leq d(x_1, x_2) + d(x_1, x_2) = 2t. \end{aligned}$$

If $F_n = f_{2,j}$ for some j, and r_{n-1} is large, then $r_n - r_{n-1}$ is approximately $-\log 2$. Just as in example 1, we can bound r_n above by a random walk on $\mathbb{R}$ with a retaining lower barrier, which has an invariant distribution when $2\alpha t < (1-\alpha)\log 2$. This shows that for sufficiently small α, $\mathcal{F}$ itself is asymptotically stable.

2. To deal with the case $\alpha = \frac{1}{2}$, we introduce the random sequences (i_n) and (j_n) such that $F_n = f_{i_n, j_n}$, and define

$$a_n = d(H_n, x_{i_n}), \qquad b_n = d(H_n, x_{3-i_n}).$$

Our aim is to show that with probability 1, $r_n \to \infty$ as $n \to \infty$. We begin with some hyperbolic trigonometry.

$$\cosh d(0, z) = \frac{1}{1 - |z|^2}$$

$$\therefore \quad \cosh d(0, z^2) = \frac{1}{1 - |z|^4} = \frac{1}{1 - |z|^2} \frac{1}{1 + |z|^2}.$$

Hence if $i_n = i_{n-1}$, we have $\cosh r_n \geq \frac{1}{2} \cosh r_{n-1}$.

Now suppose that $i_n \neq i_{n-1}$. We shall use the analogue for hyperbolic triangles of the cosine rule. Let θ be the angle at x_{3-i_n} in the geodesic triangle with vertices at x_1, x_2 and H_{n-1}. The side lengths are $d(x_1, x_2) = t$, $d(x_{i_n}, H_{n-1}) = s_{n-1}$ and $d(x_{3-i_n}, H_{n-1}) = r_{n-1}$. The hyperbolic cosine rule now gives

$$\cosh s_{n-1} = \cosh t \cosh r_{n-1} - \cos\theta \sinh t \sinh r_{n-1}.$$

Dividing through by $\cosh r_{n-1}$, we have

$$\frac{\cosh s_{n-1}}{\cosh r_{n-1}} = \cosh t - \cos\theta \sinh t \tanh r_{n-1}.$$

The angle θ takes one of N possible values evenly spaced around the circle, and which of those values it takes is independent of r_{n-1} and s_{n-1}. So if we require N to be even, with probability at least $\frac{1}{2}$, the contribution of the final term in the above equation is non-negative, so $\frac{\cosh s_{n-1}}{\cosh r_{n-1}} \geq \cosh t$. When can the ratio $\frac{\cosh s_{n-1}}{\cosh r_{n-1}}$ be less than 1? Precisely when $\cos\theta > \frac{\cosh t - 1}{\sinh t} = \tanh(\frac{t}{2})$. We can require t to be sufficiently large that $\cos\frac{\pi}{N} < \tanh(\frac{t}{2})$, and then the probability that $\frac{\cosh s_{n-1}}{\cosh r_{n-1}} < 1$ is at most $1/N$. In this bad case, we resort to the triangle inequality: $s_{n-1} \geq r_{n-1} - t$. In the remaining case we content ourselves with $s_{n-1} \geq r_{n-1}$.

Now, the action of the map F_n does not contract too much: $\cosh r_n \geq \frac{1}{2}\cosh s_{n-1}$, as explained in the proof of part 1. Now let us catalogue the possible multiplicative increments of the sequence $(\cosh r_n)$:

- With probability $\frac{1}{2}$ we have $i_n = i_{n-1}$ so $\cosh r_n \geq \frac{1}{2}\cosh r_{n-1}$.
- With probability $1/4$ we have $i_n \neq i_{n-1}$ and $\cos\theta \leq 0$, so $\cosh r_n \geq \frac{1}{2}\cosh s_{n-1} \geq \frac{1}{2}\cosh t \, \cosh r_{n-1}$.
- With probability at most $1/(2N)$ we have $i_n \neq i_{n-1}$ and $\cos\theta \geq$

$\tanh\frac{t}{2}$, in which case we still have (from the triangle inequality) $\cosh r_n \geq \frac{1}{2}\cosh s_{n-1} \geq \frac{1}{2}\cosh(r_{n-1}-t) \geq \frac{1}{2}e^{-t}\cosh r_{n-1}$.

- With probability at least $1/4 - 1/(2N)$ we have $i_n \neq i_{n-1}$ and $0 < \cos\theta < \tanh\frac{t}{2}$, in which case we have $\cosh r_n \geq \frac{1}{2}\cosh s_{n-1} \geq \frac{1}{2}\cosh r_{n-1}$.

The corresponding events for different values of n are independent. We will now bound $\log\cosh r_n$ below by a transient random walk W_n on $\mathbb{R}$, coupled to the maps F_n. Start with $W_0 = 0$, and let the increment $W_n - W_{n-1}$ be $-\log 2$ in the first and last cases above, $\log\cosh t - \log 2$ in the second case, and $-\log 2 - t$ in the third case. Then

$$\begin{aligned}\mathbb{E}(W_n - W_{n-1}) &\geq \frac{\log\cosh t}{4} - \frac{t}{2N} - \log 2\\ &\geq t\left(\frac{1}{4} - \frac{1}{2N}\right) - \frac{5}{4}\log 2.\end{aligned}$$

If we take $N = 6$ and $t = 6$, then the condition $\cos\frac{\pi}{N} < \tanh(\frac{t}{2})$ is satisfied and the expected increment of (W_n) is positive. (W_n) has independent increments with finite variance, so a.s. $W_n \to \infty$ as $n \to \infty$ by the law of large numbers. Since $W_n \leq \log\cosh r_n$, we also have $r_n \to \infty$ as $n \to \infty$ a.s., so indeed $\mathcal{F}$ has no invariant probability measure when $\alpha = \frac{1}{2}$.

□

Example C.3. This example shows that the property of not having an invariant measure may fail to be a convex condition on the probability vector, for fixed maps. Define an IFS of three analytic self-maps of $\mathbb{D}$:

$$f_1 : z \mapsto z^3\,;$$

$$f_2 : z \mapsto \frac{z+\tau}{1+\tau z}\,,$$

where τ is chosen so that f_2 is a hyperbolic isometry that translates the real axis towards 1 by a distance $\log 3$, and

$$f_3 = f_2^{-1}\,.$$

The real interval $(-1, 1)$ is a forward-invariant set. Now fix $0 < p_1 < \frac{1}{2}$ and vary the other two probabilities. Near the extremes, we can bound the process by a random walk with a drift, so there is no invariant probability measure, but at the midpoint $p_2 = p_3$, we can bound $d(0, H_n)$ above by a random walk with a drift towards 0 and a retaining lower barrier, just as in example 1, so with this probability vector, $\mathcal{F}$ does have an invariant probability measure.

Bibliography

[1] A. Abrams, H. Landau, Z. Landau, J. Pommersheim and E. Zaslow. An iterated random function with Lipschitz number one. *Theory Probab. Appl.* **47** no. 2 (2003), 190–201.

[2] L. V. Ahlfors. *Conformal Invariants.* McGraw-Hill (1973).

[3] A. Ambroladze. Ergodic properties of random iterations of analytic functions. *Ergodic Thy. & Dyn. Sys.* (1999) **19**, 1379–1388.

[4] T. Bagby. The modulus of a plane condenser. *Journal of Mathematics and Mechanics* Vol. 17, No. 4 (1967), 315–329.

[5] C. Bandle. *Isoperimetric Inequalities and Applications*, (Pitman, 1980).

[6] M. Barge and R. Swanson. Pseudo-orbits and Topological Entropy. *Proc. AMS*, Vol. 109, No. 2, (June 1990), 559–566.

[7] M. F. Barnsley. *Fractals Everywhere*, 2nd ed. (Academic Press Professional, Boston, 1993).

[8] M. F. Barnsley and S. Demko. Global construction of fractals. *Proc. Roy. Soc. London*, Series A, **399** , (1985) 243–275.

[9] M. F. Barnsley, S. Demko, J. H. Elton and J. S. Geronimo. Invariant measures for Markov processes arising from iterated function systems with place-dependent probabilities. *Ann. Inst. H. Poincaré Probab. Statist.* **24** (1988), 367–394.

[10] M. F. Barnsley and J. H. Elton. A new class of Markov processes for image encoding. *Adv. in Appl. Probab.* **20** (1988), no. 1, 14–32.

[11] A. F. Beardon. The Schwarz-Pick lemma for Derivatives. *Proc. Amer. Math. Soc.* vol. 125, no. 11, (1997), 3255–3256.

[12] A. F. Beardon. Iteration of contractions and analytic maps. *J. London Math. Soc.*, **41**, (1991), 141–150.

[13] A. F. Beardon, T. K. Carne, D. Minda and T. W. Ng. Random iteration of analytic maps, to appear in *Ergodic Thy. & Dyn. Sys.*

[14] A. F. Beardon, T. K. Carne and T. W. Ng. The Critical Values of a Polynomial. *Constructive Approximation*, **18** (2002), 343–354.

[15] A. F. Beardon, D. Minda and T. W. Ng. Smale's mean value conjecture and the hyperbolic metric. *Mathematische Annalen*, **322**, (2002), 623-632.

[16] P. Billingsley. *Convergence of Probability Measures* 2nd edition, Wiley series in probability and statistics (1999).

[17] R. M. Blumenthal and H. K. Corson. On continuous collections of measures. *Proc.* 6^{th} *Berkeley Symposium on Mathematical Statistics and Probability*, vol. 2 (1972), 33–40.

[18] T. Breuer, *Characters and Automorphism Groups of Compact Riemann Surfaces*, LMS Lecture Note Series 280.

[19] S. R. Bullett. Critically finite correspondences and subgroups of the modular group. *Proc. London Math. Soc.* (3) 65 (1992), 423–448.

[20] S. R. Bullett. Dynamics of the arithmetic-geometric mean. *Topology* **30** No. 2 (1991), 171–190.

[21] S. R. Bullett and C. Penrose, Regular and limit sets for holomorphic correspondences. *Fundamenta Mathematica* 167 (2001), 111–171.

[22] X. Buff and C. Henriksen. On König's root-finding algorithms. *Nonlinearity* **16** (2003) 989–1015.

[23] L. Carleson and T.W. Gamelin. *Complex Dynamics*. Springer (1993).

[24] P.M. Centore and E.R. Vrscay. Continuity of Attractors and Invariant Measures for Iterated Function Systems. *Canad. Math. Bull.* Vol. **37** (3), (1994), 315–329.

[25] L. Clozel and J.-P. Otal. Unique ergodicité des correspondances modulaires.

[26] L. Clozel and E. Ullmo. Correspondances modulaires et mesures invariantes. *J. reine angew. Math.* **558** (2003), 47–83.

[27] P. Diaconis and D. Freedman. Iterated Random Functions. *SIAM Review* **41** No. 1, 45–76.

[28] J. Dieudonné. Recherches sur quelques problèmes relatifs aux polynômes et aux fonctions bornées d'une variables complexe. *Ann. Sci. Ecole Norm. Sup.* **48** (1931), 247–358.

[29] R. M. Dudley. *Real Analysis and Probability.* 2nd edition, Cambridge University Press, (2002).

[30] J. H. Elton. A multiplicative ergodic theorem for Lipschitz maps. *Stochastic Process. Appl.* **34** no. 1 (1990), 39–47.

[31] A. E. Eremenko and W. K. Hayman. On the length of lemniscates. *Michigan J. Math.* 46 (1999) 409–415.

[32] A. Freire, A. Lopes and R. Mañé. An invariant measure for rational maps. *Bol. Soc. Bras. Mat.* **14** No. 1 (1983), 45–62.

[33] A. Fryntov and J. Rossi. Hyperbolic symmetrization and an inequality of Dyn'kin. *Entire functions in modern analysis (Tel–Aviv, 1997)*, Israel Math. Conf. Proc. 15, eds. Y. Lyubich at al. (Bar-Ilan Univ., Ramat Gan, 2001), 103–115.

[34] T. Ganelius. Rational functions, capacities and approximation. *Aspects of contemporary complex analysis (Proc. NATO Adv. Study Inst., Univ. Durham, Durham, 1979)*, eds. D. A. Brannan and J. G. Clunie (Academic Press, London-New York, 1980), 409–414.

[35] L. Greenberg. Commensurable groups of Moebius Transformations, *Discontinuous groups and Riemann surfaces*, Annals of Math. studies **79**, 1974, 227–237.

[36] L. Greenberg. Discrete subgroups of the Lorentz group, Math. Scanf. **10** (1962), 85–107.

[37] P. Griffiths and J. Harris. *Principles of Algebraic Geometry*, John Wiley and Sons, 1978.

[38] Ş. Grigorescu and G. Popescu. Random systems with complete connections as a framework for fractals. *Studii Cercetari Matematici* **41** No. 6 (1989), 27–43.

[39] W. K. Hayman. *Multivalent Functions*. 2nd ed. (Cambridge University Press, 1994).

[40] J. Head. The combinatorics of Newton's method for cubic polynomials. *PhD thesis*, Cornell University, 1989.

[41] A. Hinkkanen and G. J. Martin. The dynamics of semigroups of rational functions I. *Proc. London Math. Soc.* (3) 73 (1996), 358–384.

[42] M. Hurley. On topological entropy of maps. *Ergod. Th. & Dynam. Sys.*, **15** (1995), 557–568.

[43] M. Iancu and S. Williams. Notes on the topology of metric spaces. Preprint, to appear in *Topology and Applications.* Online: www.math.buffalo.edu/˜sww/0papers/Topology-of-Metric-Spaces.pdf

[44] R. Langevin and P. Walczak. Entropie d'une dynamique. *C. R. Acad. Sci. Paris*, tome 312, Série 1, (1991), 141-144.

[45] R. Langevin and F. Przytycki. Entropie de l'image inverse d'une application. *Bull. Soc. Math. France*, **120**, (1992), 237–250.

[46] A. Lasota and J. A. Yorke. Lower Bound Technique for Markov Operators and Iterated Function Systems. *Random and Computational Dynamics* 2(1), (1994), 41–77.

[47] G. Letac. A contraction principle for certain Markov chains and its applications. *Random Matrices and their applications*, Contemp. Math **50** (AMS, 1986), 263–273.

[48] D. S. Lubinsky. Small values of polynomials: Cartan, Pólya and others. *Journal of Inequalities and Applications* 1, (1997), no. 3, 199–222.

[49] M. Y. Lyubich, Entropy properties of rational endomorphisms. *Ergodic Thy. & Dyn. Sys.* (1983) **3**, 351–385.

[50] Y. Kifer, *Ergodic Theory of Random Transformations*, Progress in Probability and Statistics Vol. 10, Birkhaüser, 1986.

[51] A. Kuribayashi and K. Komiya. On Weierstrass points and automorphisms of curves of genus three. *Algebraic Geometry* (Proc. summer meeting, Univ. Copenhagen, 1978), Lecture Notes in Math. 732, (Springer, Berlin, 1979), 253–299.

[52] R. Mañé. On the uniqueness of the maximizing measure for rational maps. *Bo. Soc. Bras. Mat.* **14** No. 1 (1983), 27–43.

[53] J. C. Mattingly. Contractivity and ergodicity of the random map $x \to |z - \theta|$. *Theory Probab. Appl.* **47** no.2 (2003), 333–343.

[54] J. Milnor. *Dynamics in One Complex Variable*, Vieweg 1999.

[55] Z. Nehari. A generalization of Schwarz' lemma, Duke Math. J. **14** (1947) 1035–1049.

[56] T. Ohno. Asymptotic behaviours of dynamical systems with random parameters, *Publ. R.I.M.S. Kyoto Univ.*, **19** (1983), 83–98.

[57] K. R. Parthasarathy. *Probability Measures on Metric Spaces*, Academic Press, New York, (1967).

[58] K. Petersen. *Ergodic Theory*, Cambridge studies in advanced mathematics 2, Cambridge University Press, 1983.

[59] G. Pólya. Beitrag zur Verallgemeinerung des Verzerrungssatzes auf mehrfach zusammenhängedende Gebiete. *S-B. Akad. Wiss, Berlin* (1928), 228–232 & 280–282. Reprinted in G. Pólya, *Collected Papers, Vol. 1: Singularities of Analytic Functions*, ed. R. P. Boas (MIT press, 1974), 352–362.

[60] G. Pólya and G. Szegö. *Isoperimetric Inequalities in Mathematical Physics*, Annals of Mathematics Studies **27**, (Princeton University Press, 1951).

[61] S.T. Rachev, *Probability metrics and the stability of stochastic models*, Wiley series in probability and mathematical statistics, (1991).

[62] T. Ransford, *Potential Theory in the Complex Plane*, London Mathematical Society Student Texts **28**, (Cambridge University Press, 1995).

[63] W. Rudin. *Real and Complex Analysis*, International edition, McGraw-Hill, 1987.

[64] W. Rudin. *Functional Analysis*, 2nd edition, McGraw-Hill, 1991.

[65] E. B. Saff and V. Totik. *Logarithmic Potentials with External Fields*, Grundlehren der mathematischen Wissenschaften, vol 316, (Springer, 1997).

[66] B. Simon. *Representations of Finite and Compact Groups* , AMS Graduate Studies in Mathematics, vol 10, (1996).

[67] H. Stahl and V. Totik, *General Orthogonal Polynomials*, Encyclopedia of mathematics and its applications, vol. 43, (Cambridge University Press, 1992).

[68] D. Steinsaltz, Locally contractive iterated function systems. *Annals of Probability*, vol. 27 No. 4, (1999) 1952–1979.

[69] Ö. Stenflo. Introduction to PhD thesis, Dept. of Mathematics, Umeå University, Sweden (1998). Currently available at www.math.su.se/ stenflo/introduction.pdf.

[70] Ö. Stenflo. Ergodic theorems for Markov Chains represented by iterated function systems. *Bull. Polish. Acad. Sci. Math.* **49** (2001), no. 1, 27–43.

[71] H. Sumi. Skew product maps related to finitely generated rational semigroups. *Nonlinearity* **13** (2000), 995–1019.

[72] D. Tischler. Critical Points and Values of Complex Polynomials. *Journal of Complexity*, **5**, 438–456 (1989).

[73] D. Tischler. Perturbations of Critical Fixed Points of Analytic Maps. *Astérisque* no. 222 (1994), 407–422.

[74] C. P. Walkden. Invariance principles for iterated maps that contract on average. *preprint*, University of Manchester (2002).