FREE ENERGY AS A DYNAMICAL AND GEOMETRIC INVARIANT (OR CAN YOU HEAR THE SHAPE OF A POTENTIAL?)

MARK POLLICOTT AND HOWARD WEISS

The University of Manchester and The Pennsylvania State University

August 15, 2002 11:12am

ABSTRACT. The lattice gas provides an important and illuminating family of models in statistical physics. An interaction Φ on a lattice $L \subset \mathbb{Z}^d$ determines an idealized lattice gas system with potential A_Φ. The pressure $P(A_\Phi)$ and free energy $F_{A_\Phi}(\beta) = -(1/\beta)P(\beta A_\Phi)$ are fundamental characteristics of the system. However, even for the simplest lattice systems, the information about the potential that the free energy captures is subtle and poorly understood. We study whether, or to what extent, potentials for certain model systems are determined by their free energy. In particular, we show that for a one-dimensional lattice gas, the free energy of finite range interactions typically determines the potential, up to natural equivalence, and there is always at most a finite ambiguity; we exhibit exceptional potentials where uniqueness fails; and we establish deformation rigidity for the free energy. The proofs use a combination of thermodynamic formalism, algebraic geometry, and matrix algebra.

In the language of dynamical systems, we study whether a Hölder continuous potential for a subshift of finite type is naturally determined by its periodic orbit invariants: orbit spectra (Birkhoff sums over periodic orbits with various types of labeling), beta function (essentially the free energy), or zeta function. These rigidity problems have striking analogies to fascinating questions in spectral geometry that Kac adroitly summarized with the question "Can you hear the shape of a drum?".

We also introduce the free energy as a new geometric invariant for negatively curved surfaces and discuss some of its properties. In this case we show that the free energy is intimately related to a Poincaré-type series which encodes both the lengths of closed geodesics and word lengths of the corresponding words in the fundamental group. Thus free energy contains some refined information on the ratio of word length to hyperbolic length of closed geodesics, as studied by Milnor.

Key words and phrases. Pressure, free energy, orbit spectrum, beta function, zeta function, rigidity, thermodynamic formalism, length spectrum, hyperbolic surface, Hadamard product.

The work of the second author was partially supported by a National Science Foundation grant DMS-0100252. This work began during the second author's sabbatical visit at IPST, University of Maryland, and he wishes to thank IPST for their gracious hospitality. This work was completed during a visit by the first author to Penn State as a Shapiro Fellow.

Contents

I. Motivations and Introduction
II. The Zeta Function
III. The Free Energy and Beta Function
  A. Locally Constant Functions
  B. Hölder Continuous Functions
IV. The Orbit Spectrum
V. Geometric Characterization of Free Energy and Beta Function for Hyperbolic Surfaces
App. I. Uncountably Many Hölder Continuous Functions Sharing the Same Unmarked Orbit Spectrum (and Free Energy)
App. II. Locally Constant Functions with the Same Unmarked Orbit Spectrum (and Free Energy)
App. III. Uncountably Many Hölder Continuous Functions Sharing the Same Weak Orbit Spectrum but Having Different Unmarked Orbit Spectra

I.A. Physical Motivation: Does the Free Energy of a Lattice Gas determine the Potential?

The lattice gas provides an important and illuminating family of models in statistical physics. An interaction Φ on $\mathbb{Z}^r$ determines an idealized lattice gas system, and the pressure $P(\Phi)$ and free energy $F_\Phi(\beta)$ are fundamental characteristics of the system. However, even for the simplest lattice systems, the information about the interaction or the potential that the free energy captures is subtle and poorly understood. We study whether, or to what extent, potentials for certain model systems are determined by their free energy.

Following Ruelle [Rue, pp. 36–38], we consider the lattice $\mathbb{Z}^d$ and the full shift space $\Sigma^m \doteq \{1, 2, \cdots, m\}^{\mathbb{Z}^d}$ of configurations equipped with the product topology. For $S \subset \mathbb{Z}^d$ define $\Sigma_S^m \doteq \Sigma_S \doteq \{1, 2, \cdots, m\}^S$. When studying lattice gasses one usually begins with an *interaction* Φ - a function defined on all finite subsets $\Lambda \subset \mathbb{Z}^d$ and which satisfies some regularity condition. For example, the interaction Φ for the general Ising model is defined by[1]

$$\Phi_\Lambda(\xi) = \begin{cases} -h(x)\xi_x & \text{if } \Lambda = \{x\} \\ -J(x,y)\xi_x\xi_y & \text{if } \Lambda = \{x, y\} \\ 0 & \text{otherwise,} \end{cases}$$

[1]The (formal) Hamiltonian for general the Ising model is

$$H(\xi) = -\sum_{(x,y)\in\mathbb{Z}^d} J(x,y)\xi_x\xi_y - \sum_{x\in\mathbb{Z}^d} h(x)\xi_x.$$

To prove convergence of the key quantities of interest, one is essentially forced to first define these quantities on finite sets and then take a (thermodynamic) limit.

where ξ_x denotes the restriction of the configuration ξ to the site x.

One then constructs a *partition function* for each finite subset $\Lambda \subset \mathbb{Z}^d$. Ruelle found that it is often useful to study the associated *potential function* A_Φ on Σ^m defined by

$$A_\Phi(\xi) \doteq -\sum_{0 \in Y} \frac{1}{\#Y} \Phi(\xi|Y),$$

where $\#Y$ denotes the cardinality of the set Y. This potential is essentially the contribution of the lattice site 0 to the energy in the configuration ξ, and the regularity requirement on Φ ensures that A_Φ is well defined. Using this potential function and the $d-$dimensional family of shift maps $\{\sigma^x, x \in \mathbb{Z}^d\}$, one can construct an analogous partition function for each finite subset $\Lambda \subset \mathbb{Z}^d$. More generally, for any continuous function A on Σ^m and finite subset $\Lambda \subset \mathbb{Z}^d$, one can define a partition function

$$Z_\Lambda(A) \doteq \sum_{\xi \in \Sigma^*_\Lambda} \exp\left(\sum_{x \in \Lambda} A(\sigma^x \xi^*)\right),$$

where $\Sigma^*_\Lambda = \{\xi \in \Sigma_\Lambda \,|\, \text{there exists } \xi^* \in \Sigma^m \text{ such that } \xi = \xi^*|\Lambda\}$ and where, for each $\xi \in \Sigma^*_\Lambda$ one makes an arbitrary choice of $\xi^* \in \Sigma_\Lambda$ such that $\xi^*|\Lambda = \xi$.

To define the *pressure* and *free energy* of A one needs to compute the *thermodynamic limit* as $\Lambda \to \infty$:

$$P(A) \doteq \lim_{\text{“}\Lambda \nearrow \infty\text{”}} \frac{1}{\#\Lambda} \log Z_\Lambda(A) \quad \text{and} \quad F_A(\beta) \doteq -\frac{1}{\beta} P(\beta A).$$

The pressure and free energy are the two fundamental objects of study in statistical physics of lattice gasses. For instance, phase transitions correspond to non-differentiability for some derivative of free energy. Establishing the existence of these limits is a non-trivial task and regularity restrictions on the function A are required for the thermodynamic limit to exist. A physically important class of interactions are the finite range interactions. We study to what extent potentials (especially those related to finite range interactions) are determined by their free energy. *In particular, we show that for a one-dimensional lattice gas the free energy of finite range interactions typically determines the potential, up to natural equivalence and there is always at most a finite ambiguity (Theorem 3 and Theorem 4); we exhibit exceptional potentials where uniqueness fails (Proposition 3.1); and we establish deformation rigidity for the free energy (Theorem 5).*

I.B. Mathematical Motivation

Since free energy plays such an essential role in statistical physics, it is natural to study this quantity for dynamical systems and Riemannian manifolds. In particular, one would like to have a geometric/topological interpretation of the free energy. In the special case of locally constant functions for subshifts of finite type, Tuncel [Tun1] introduced a quantity closely related to free energy. This quantity was introduced as an invariant in coding

theory for Markov chains and appears to have been studied only in this context. The term beta function is a bit confusing, since in thermodynamics, beta usually denotes inverse temperature and there is another beta function in dynamics which is the mapping of the unit interval defined by $x \to \beta x \mod 1$. We will follow the established nomenclature in the dynamics literature.

For a general smooth dynamical system $T : X \to X$ one can define the pressure of a continuous function $f : X \to \mathbb{R}$ using the variational principle:

$$P(f) = \sup \left\{ h_\mu(T) + \int_X f d\mu \ : \mu \text{ is a } T - \text{invariant probability measure} \right\},$$

where $h_\mu(T)$ denotes the measure theoretic entropy with respect to the measure μ. By analogy with statistical physics, we define the *free energy* for f by

$$F_f(\beta) = -(1/\beta) \exp P(\beta f).$$

Since the first factor $-(1/\beta)$ in the definition of free energy plays no further role in our analysis, it is notationally convenient to replace the free energy by the beta function for f defined by $\beta_f(t) = \exp P(tf)$.

For certain classes of dynamical systems, e.g., subshifts of finite type and hyperbolic maps, the beta function can be defined using Birkhoff sums (or Birkhoff averages) of the function f over periodic orbits. This is one of a hierarchy of several natural periodic orbit invariants, including the zeta function, the marked orbit spectrum (the set of Birkhoff sums around periodic orbits labeled by the periodic orbit), the unmarked orbit spectrum (the set of Birkhoff sums around periodic orbits labeled by the period), and the orbit spectrum (the unlabeled set of Birkhoff sums around periodic orbits). The weak orbit spectrum seems to be a less natural and less useful invariant than the unmarked orbit spectrum in the context of subshifts of finite type. The main objectives of this paper are to study the relations between these invariants and to show that in many cases the function f can be recovered, up to some unavoidable natural ambiguities, from these various spectrum. Results on subshifts of finite type and Hölder functions can be easily reformulated in terms of one-dimensional Axiom A flows (as we will elaborate at the end of this subsection).

We observe that such rigidity problems in the study of dynamical systems have striking similarities to fascinating questions in length geometry (spectral geometry) which Kac adroitly summarized with the question *"Can you hear the shape of a drum?"*. Given a compact hyperbolic surface, the unmarked length spectrum consists of the set of lengths of all closed geodesics, and the marked length spectrum (the analogue of the marked orbit spectrum) consists of the lengths of closed geodesics labeled by the free homotopy class of the geodesic. The marked length spectrum determines the hyperbolic surface up to isometry [Ota], but the unmarked length spectrum does not [Vig, Sun, Bus]. By analogy, for subshifts of finite type, the marked orbit spectrum determines the function up to a natural equivalence (Lemma 1.2), while the unmarked length spectrum does not (Proposition A.II.1). The unmarked length spectrum for a hyperbolic surface *typically* does determines the surface [Wol] and there is a uniform bound, depending only on the

genus of the surface, on the number of non-isometric hyperbolic surfaces having the same unmarked length spectrum [McK]. By analogy, for subshifts of finite type, the unmarked orbit spectrum for a locally constant function typically determines the function (Theorem 6) and that there is a uniform bound, depending only on the number of coordinates of the locally constant function, on the number of locally constant functions having the same unmarked orbit spectrum (Theorem 6). Finally, Guillemin and Kazhdan [GK] showed that the unmarked length spectrum for hyperbolic surfaces is deformation rigid, i.e., there are no smooth curves of non-isometric surfaces. We show the analogue of this result for the unmarked orbit spectrum for Hölder functions in Theorem 7.

We summarize these analogies in Table 1, where the hyperbolic surfaces (with fixed genus) are determined up to isometry, and the locally constant functions (with fixed number of coordinates) are defined up to coboundary and automorphism of the shift.

Hyperbolic surfaces	**Locally constant functions**
Marked length spectrum determines surface [Otal] (true for negatively curved surfaces)	Marked orbit spectrum determines function [Livsic] (true for Hölder functions)
Unmarked length spectrum does not determine surface [Vigneras/Sunada/Buser]	Unmarked orbit spectrum does not determine function [Proposition A.II.1]
Unmarked length spectrum typically determines surface [Wolpert]	Unmarked orbit spectrum typically determines function [Theorem 6]
Uniform bound on number of surfaces with same unmarked length spectrum [McKean]	Uniform bound on number of functions with same unmarked orbit spectrum [Theorem 6]
No smooth arc with same unmarked length spectrum [Kazdan-Guillemin] (true for negatively curved surfaces)	No smooth arc with same unmarked orbit spectrum [Theorem 7] (true for Hölder functions)
Unmarked length spectrum never simple [Randol]	Unmarked orbit spectrum never simple [Proposition 4.2]

TABLE 1. SUMMARY OF RESULTS ON ORBIT SPECTRUM AND COMPARISONS WITH CORRESPONDING RESULTS FOR HYPERBOLIC SURFACES

In a slightly different direction, we show that there exist uncountably many inequivalent Hölder continuous functions sharing the unmarked orbit spectrum (Proposition A.I.1).

Our results for the unmarked length spectrum are intimately related to corresponding results for the beta and zeta functions. For example, a knowledge of the unmarked length spectrum is equivalent to a knowledge of the zeta function (Proposition 2.2). In contrast, the beta function is a more subtle invariant.

This paper is organized into separate sections on rigidity for the zeta function (§II), rigidity for the beta function (§III), and rigidity for the unmarked orbit spectrum (§IV). The free energy or beta function seems not to have been previously studied for geodesic flows, and there seems to be no geometric characterization of free energy in the literature. In §V we show that the graph of the beta function has many features which are reminiscent of the Manhattan curve of Burger [Bur] and contains some refined information on the ratio of word length to hyperbolic length of closed geodesics, as studied by Milnor [Mil] (Theorem 8). In the three appendices we construct functions which exhibit various types of degeneracy or (strong) nonrigidity. For instance, in Appendix I we construct an uncountable family of Hölder continuous functions which share the same unmarked orbit spectrum and thus have the same free energy. We also observe that Theorem 5, deformation rigidity of free energy, naturally extends to smooth hyperbolic maps.

Whereas the analogy between length spectrum rigidity for geodesic flows and the problems we consider is a useful guide, it is not possible to translate results directly from one setting to the other. For example, the height functions over subshifts of finite type corresponding to geodesic flows on hyperbolic surfaces form a very small subclass of all Hölder functions. Furthermore, the automorphism group of a Riemann surface is typically trivial (and always finite), while the automorphism group for a subshift of finite type is usually quite large. This crucial disparity arises from the fundamental difference in the topology of the spaces involved.

More generally, we consider the situation we are studying as the Axiom A analogue of the results for geodesic flows, at least in the case where the functions are positive. More precisely, we recall that for any subshift of finite type $\sigma : X \to X$ and any Hölder continuous function $f : X \to \mathbb{R}$ we can construct a space

$$X^f = \{(x,t) \in X \times \mathbb{R} : 0 \le t \le f(x)\},$$

where we identify $(x, r(x)) \sim (\sigma x, 0)$, and a flow ψ defined locally by $\psi(x,t) = (x, t+u)$, subject to the identifications. The following result is a corollary to a theorem of Bowen and shows the relationship with Axiom A flows. Any Axiom A flow (restricted to a basic set) is called one dimensional if the basic set has a cross section which is a Cantor set.

Proposition 1.1 [Bow]**.** *For any one-dimensional Axiom A flow we can associate a subshift and Hölder continuous function for which the length spectrum coincides with the orbit spectrum. Conversely, given any Hölder continuous function and any $r \ge 1$ there is a C^r Axiom A flow with a basic set whose length spectrum coincides with the orbit spectrum.*

In particular, we see that the questions we consider about Hölder continuous functions could be equally well formulated in terms of the properties of one-dimensional Axiom A flows. This helps to reinforce the analogies with the problems for surfaces and geodesic flows.

I.C Subshifts of finite type and their periodic orbit invariants

Let A be a $n \times n$ aperiodic (transition) matrix with entries 0 or 1. We define

$$\Sigma_A^+ = \left\{ x \in \prod_{k=0}^{\infty} \{1, 2, \ldots, n\} : A(x_k, x_{k+1}) = 1 \right\}$$

which is compact, totally disconnected, and zero dimensional in the Tychonoff product topology. Let $\sigma : \Sigma_A^+ \to \Sigma_A^+$ be the subshift of finite type defined by $(\sigma x)_n = x_{n+1}$. Since A is aperiodic, the shift map is topologically mixing and has a dense set of periodic points.

We let $\mathrm{Aut}(\sigma)$ be the group of shift commuting homeomorphisms τ (i.e., $\sigma \circ \tau = \tau \circ \sigma$). This group is always countable and except in cases of small n contains free groups. This is in stark contrast to the situation for hyperbolic surfaces where the group of automorphisms is always finite and typically trivial. This crucial disparity arises from the fundamental difference in the topology of the spaces involved. In particular, it is natural that the zero dimensional space Σ_A^+ allows a much larger space of automorphisms.

We define a metric on Σ_A^+ by $d(x, y) \stackrel{\circ}{=} \sum_{n=0}^{\infty}(1 - \delta_{x_n y_n})/2^n$ that enables us to define the Banach space of Hölder continuous functions on Σ_A^+. An important class of Hölder continuous functions are the *locally constant functions*, i.e., functions that only depend on finitely many coordinates. We let $LC(n)$ be the n-dimensional vector space of locally constant functions which depend on only the first n coordinates. These spaces are nested, i.e., $LC(n) \subset LC(n+1)$ for all $n \in \mathbb{N}$. In the physical nomenclature, locally constant functions correspond to finite range interactions which form an important class of potentials for lattice gasses. If f is a locally constant function, then after recoding if necessary, we can always assume that $f(x) = f(x_0 x_1)$, i.e., f is in $LC(2)$ for some subshift of finite type. For such functions the thermodynamic formalism reduces to matrix algebra.

Let $f : \Sigma_A^+ \to \mathbb{R}$ be a Hölder continuous function and let $S_n f$ denote the sequence of Birkhoff sums

$$S_n f(x) = \sum_{k=0}^{n-1} f(\sigma^k x).$$

We can associate to f the **unmarked orbit spectrum**

$$\mathcal{L}_f = \{(S_n f(x), n) : \sigma^n x = x\}$$

and the **weak orbit spectrum**

$$\mathcal{W}_f = \{S_n f(x) : \sigma^n x = x\}.$$

Since the weak orbit spectrum does not contain the periods of the orbits, it is a weaker invariant than the unmarked orbit spectrum. In Appendix III we construct an uncountably family of pairwise inequivalent Hölder continuous functions with different unmarked orbit spectrum, but all sharing the same weak orbit spectrum.

The following observation shows two natural ways for functions to have the same orbit spectrum. The proof is obvious.

Lemma 1.1.

(i) *If f_1 and f_2 are cohomologous ($f_1 \sim f_2$), i.e., there exists a function $u \in C(\Sigma_A^+, \mathbb{R})$ with $f_1 - f_2 = u \circ \sigma - u$, then $\mathcal{L}_{f_1} = \mathcal{L}_{f_2}$;*

(ii) *If $f_2 = f_1 \circ \tau$ where $\tau \in Aut(\sigma)$ is a shift commuting homeomorphism (i.e. $\sigma \circ \tau = \tau \circ \sigma$), then $\mathcal{L}_{f_1} = \mathcal{L}_{f_2}$.*

Thus when studying the orbit spectrum, it is natural to ignore, or to factor out, these two type of trivial relations.

Definition. We define two functions f_1 and f_2 on Σ_A^+ to be **equivalent**, $f_1 \approx f_2$, if $f_2 \sim f_1 \circ \tau$, where $\tau \in \mathrm{Aut}(\sigma)$. We say two functions are **non-equivalent** if they are not equivalent .

The simplest type of such shift commuting homeomorphism $\tau_0 \in \mathrm{Aut}(\sigma)$, at least in the case of a full shift, is given by some perturbation of the alphabet. For one sided subshifts of finite type these questions were studied by Hedlund [Hed], who showed that the automorphism group of the (one-sided) full shift on two symbols is simply generated by the shift map and permutations of blocks of symbols. In contrast, he showed that for the (one-sided) full shift on three or more symbols the automorphism group is more complex.

We can also associate to f the **marked periodic spectrum**

$$\mathcal{M}_f = \{(S_n f(x), x) : \sigma^n x = x\}.$$

The following observation shows that the marked orbit spectrum essentially determines the function f up to cohomology.

Lemma 1.2. *Two functions f_1 and f_2 are cohomologous if and only if $\mathcal{L}_{f_1} = \mathcal{L}_{f_2}$*

Proof. This is an immediate corollary of Livsic's Theorem [Liv]. ■

Two other important invariants of a function $f\colon \Sigma_A^+ \to \mathbb{R}$ are the **zeta-function**, $\boldsymbol{\zeta_f}$, defined by the power series

$$\zeta_f(z,t) \doteq \exp \sum_{n=1}^{\infty} \frac{z^n}{n} \sum_{\sigma^n x = x} \exp(t S_n f(x)),$$

and the **beta-function**, $\boldsymbol{\beta_f}$, defined by $\beta_f(t)$ being the reciprocal of the radius of convergence of the zeta function $\zeta_f(z,t)$, i.e.,

$$\log \beta_f(t) = P(tf) = \lim_{n\to\infty} \frac{1}{n} \log \left(\sum_{\sigma^n x = x} \exp\left(t S_n f(x)\right) \right), \tag{1}$$

where $P(g)$ denotes thermodynamic pressure of a general Hölder continuous function g defined by

$$P(g) = \lim_{n\to\infty} \frac{1}{n} \log \left(\sum_{\sigma^n x = x} \exp\left(S_n g(x)\right) \right).$$

For subshifts of finite type this definition of pressure is equivalent to the variational definition given in I.B.

If f is a locally constant function, after recoding if necessary we can always assume that $f(x) = f(x_0x_1)$. A routine calculation [PP] shows that

$$\zeta_f(z,t) = \frac{1}{\det(I - zA_{t\cdot f})}, \tag{2}$$

where A_f denotes the $n \times n$ matrix with entries $A_f(i,j) = A(i,j)\exp(f(i,j))$.

The hierarchy of these four invariants is illustrated by the following diagram

$$f_1 \approx f_2 \Longrightarrow \mathcal{L}_{f_1} = \mathcal{L}_{f_2} \Longrightarrow \zeta_{f_1} = \zeta_{f_2} \Longrightarrow \beta_{f_1} = \beta_{f_2}.$$

In this manuscript we investigate under what conditions these arrows can be reversed.

II. The zeta function

The main result in this section is that the zeta function for locally constant functions typically determines the equivalence class of the function. We begin with the following simple lemma for matrix algebra. We recall that P is a *permutation matrix* if exactly one entry in each row and column is equal to 1 and all the others are 0.

Lemma 2.1. *Let $B = (b_{ij})$ be a $n \times n$ matrix with non-negative entries and let $B^{(t)} = (b_{ij}^t)$ denote the Hadamard t-th power of B with characteristic polynomial $q(z,t) = \det(tI - B^{(t)})$.*

(1) *The polynomial q is invariant under conjugation of B by any permutation matrix P, i.e., $q(z,t) = \det(tI - (P^{-1}BP)^{(t)})$ for all $t \in \mathbb{R}$. There are precisely $n!$ permutation matrices of size $n \times n$.*

(2) *For integer values of t, the polynomial q is invariant under conjugation of B by any diagonal matrix D, i.e., $q(z,n) = \det(nI - (D^{-1}BD)^{(n)})$ for $n \in \mathbb{Z}$.*

Proof. The proof is a straightforward calculation. ∎

Since the characteristic polynomial $q(z,n)$ of $B^{(t)}$ is invariant under conjugating B by a permutation matrix, Lemma 2.1 implies there is an inherent finite ambiguity in trying to recover B from $q(z,n)$. At best, one can recover B only up to conjugation by a permutation matrix, and there are $n!$ permutation matrices.

The following is a generalization of a lemma on Newton's identities [Wae] allowing negative terms, and is a special case of a result on Dirichlet series due to Mandelbrojt which will appear in Proposition 2.2.

Lemma 2.2. *Consider an expression of the form*

$$s(t) = \lambda_1^t + \lambda_2^t + \cdots + \lambda_m^t - \lambda_{m+1}^t - \lambda_{m+2}^t - \cdots - \lambda_{m+n}^t,$$

where $\lambda_k > 0$ and $t \in \mathbb{R}$. Then one can obtain the numbers λ_k (up to permutation) from $s(t)$.

Proof. In the special case that $s(t) = \lambda_1^t + \lambda_2^t + \cdots + \lambda_m^t$ is a sum of powers, the lemma follows immediately from Newton's identities, in which case one only needs to know $s(1), s(2), \ldots, s(m)$.

Now assume that in $s(t)$ no two of the numbers λ_k coincide; the most general case requires a trivial modification to this argument. Then

$$\lim_{t\to\infty} r^t s(-t) = \lim_{t\to\infty} \left\{ \sum_{k=1}^{m} \left(\frac{r}{\lambda_k}\right)^t - \sum_{k=m+1}^{m+n} \left(\frac{r}{\lambda_k}\right)^t \right\} = \left\{ \begin{array}{c} 0, \text{ if } 0 \le r < \min\limits_k \lambda_k \\ \pm\infty, \text{ if } r \ge \min\limits_k \lambda_k \end{array} \right\},$$

where the minimizing λ_k belongs to $\{\lambda_1, \ldots, \lambda_m\}$ if the limit is $+\infty$ and the minimizing λ_k belongs to $\{\lambda_{m+1}, \ldots, \lambda_{m+n}\}$ if the limit is $-\infty$. Hence one can first detect the smallest λ_k as the jump point of the limit, remove that λ_k from the sum, then detect the next smallest λ_k, remove it from the sum, and so on. ∎

When studying the zeta function of a function $f \in LC(2)$ it is notationally convenient to work with the characteristic polynomial $q_f(z,t) = \det(zI - A_f^{(t)})$ of $A_f^{(t)}$ instead of $\zeta_f(z,t)$. Using the expression for the zeta function in (2) one can easily see the relationship

$$q_f(z,t) = z^n \zeta_f(1/z,t)^{-1}.$$

Proposition 2.1. *There exists an explicit uniform bound $C = C(n) > 0$ on the number of $n \times n$ aperiodic matrices with non-negative entries (up to conjugation by diagonal and permutation matrices) with the same characteristic polynomial $q(z,t)$. Furthermore, the typical such $n \times n$ matrix is actually determined by its characteristic polynomial $q(z,t)$ (up to conjugation by diagonal and permutation matrices).*

Proof. Let us first consider the case where the entries of B are positive. We only need to work with $E_1(B^{(t)}), E_2(B^{(t)})$, and $E_3(B^{(t)})$, the first three principal minors for the matrix $B^{(t)}$. It is well known [HJ] that these three minors are the coefficients of the first three terms (in z) in q_f, i.e., $q(z,t) = z^n - E_1(B^{(t)})z^{n-1} + E_2(B^{(t)})z^{n-2} - \cdots \pm E_n(B^{(t)})$, where $E_i(B^{(t)})$ denotes the i-th principal minor of $B^{(t)}$ and that the principal minors of B are themselves invariant under conjugation by diagonal and permutation matrices.

By conjugating B by a suitable diagonal matrix, we can assume that each entry in the first column of B is 1 except (possibly) the entry b_{11}. The first principal minor, $E_1(B^{(t)}) = \text{trace}(B^{(t)})$, is the sum $\sum_{i=1}^n b_{ii}^t$, which by Lemma 2.2 determines the unordered list of diagonal entries $\{b_{11}, b_{22}, \ldots, b_{nn}\}$. Since we are only interested in recovering the matrix B up to conjugation by permutation matrices, we can assume we know the ordered list of diagonal entries.

The general term in the second principal minor corresponds to the determinant of the special principal matrix $B^{(t)}\{i,j\}$ and is of the form $b_{ii}^t b_{jj}^t - b_{ij}^t b_{ji}^t$. In particular, when $i = 1$ and $j \ge 2$, the terms are of the form $b_{11}^t b_{jj}^t - b_{1j}^t$. Let us assume the genericity

condition $b^t_{ii}b^t_{jj} \neq b^t_{kl}b^t_{lk}$ for all i,j,k,l. Then Lemma 2.2 allows us to obtain the unordered list of all double products $\{b_{ij}b_{ji}\}$. Included in this unordered list are all the entries from the first row $\{b_{1j}b_{j1}\} = \{b_{1j}\}$.

The general term in the third principal minor corresponds to the determinant of the special principal matrix $B^{(t)}\{i,j,k\}$ and is of the form $b^t_{ii}b^t_{jj}b^t_{kk} - b^t_{ii}b^t_{jk}b^t_{kj} - b^t_{ji}b^t_{ij}b^t_{kk} - b^t_{ki}b^t_{ik}b^t_{jj} + b^t_{ji}b^t_{ik}b^t_{kj} + b^t_{ki}b^t_{ij}b^t_{jk}$. Let us also assume the degeneracy condition that for all $\{i,j,k\}$ all six terms are distinct and do not cancel with any terms in the determinant of other special principal matrices $B^{(t)}\{q,r,s\}$. The matrices which do not satisfy all our nondegeneracy assumptions are easily seen to form an algebraic variety of codimension one.

Included in $E_3(B^{(t)})$ are terms of the form $b^t_{kk}b^t_{ij}b^t_{jk}$, where we already know the diagonal elements b^t_{kk}. This observation, along with our genericity assumptions and Lemma 2.2, allows us to obtain, without any ambiguity, all double products $\{b_{ij}b_{ji}\}$, including all the entries of the first row $\{b_{1j}\}$.

We now show how to recover the general entry b_{jk}. The components of $E_3(B)$ include all terms $b_{j1}b_{1k}b_{kj}$ for $j \geq 2$, and thus terms $b_{1k}b_{kj}$. Let τ be one of the terms in $E_3(B)$, and suppose that

$$\tau = b_{1k}b_{kj} = b_{1k}\frac{b_{jk}b_{kj}}{b_{jk}}.$$

Inverting this expression, we obtain

$$b_{jk} = b_{1k}\frac{b_{jk}b_{kj}}{\tau},$$

where we know the terms b_{1k} and $b_{jk}b_{kj}$. Since $E_3(B)$ contains terms of the form $b_{ki}b_{ij}b_{jk}$, it contains the special terms of the form $b_{k1}b_{1j}b_{jk}$, and thus terms $b_{1j}b_{jk}$ for $k \geq 2$. Thus if we multiply the expression obtained just above for b_{jk} by the known term b_{1j}, we must obtain a term in $E_3(B)$. If we do not, then our genericity assumption implies that $\tau \neq b_{1k}b_{kj}$. This lets us determine the product $b_{1j}b_{jk}$, and since we know the element b_{1j}, it lets us determine the entry b_{jk}.

If B is non-generic, it may a priori happen that the product $b_{1j}b_{jk}$ (in the previous paragraph) is a term in $E_3(B)$, even though $\tau \neq b_{1k}b_{kj}$. Dividing by b_{1j} would give the *wrong value* of b_{jk}. However, very crudely, one can not obtain more than $\#E_3(B) \leq 6n^3$ *wrong values* for b_{jk}. Thus, again very crudely, there are at most $C(n) = (6n^3)^{n^2}$ such matrices with the same characteristic polynomial $q(z,t)$.

For a general matrix B with non-negative entries, by assumption, there exists an integer $M \geq 1$ such that B^{M} has all positive entries. Since the characteristic polynomial for B determines the characteristic polynomial for B^{M}, we can apply the preceding argument to the characteristic polynomial for B^{M}, to obtain the matrix B^{M} up to conjugation by diagonal and permutation matrices. By extracting the (unique) M-th root of B^{M} we can recover the matrix B up to conjugation by diagonal and permutation matrices. ■

We remark that Proposition 2.1 need not hold for a general non-aperiodic matrix with

non-negative entries. Consider the matrix B defined by

$$B = \begin{pmatrix} 1 & 0 & e \\ 0 & 1 & f \\ 0 & 0 & 1-e-f \end{pmatrix}.$$

Since the matrix B and hence $B^{(t)}$ are upper triangular, the characteristic polynomial $q(z,t) = (z-1)^2(z-(1-e-f)^t)$. Hence one can obtain $e+f$ from $q(z,t)$ but there is no way to obtain e and f separately.

We also remark that $C(2) = 1$. After conjugating by a diagonal matrix we can assume that

$$B = \begin{pmatrix} a & b \\ 1 & d \end{pmatrix}.$$

The entry b is positive since B is aperiodic. Knowledge of the trace of $B^{(t)}$ together with Lemma 2 gives the diagonal entries a and d, up to permutation. Since we can conjugate B by a permutation matrix, we can assume that we know a and d precisely. If $\det B \neq 0$, then the knowledge of the determinant of $B^{(t)}$ together with Lemma 2 gives b. Since the only way for $\det B = 0$ is for $b = ad$, we can obtain b precisely when $\det B = 0$.

We now show that conjugating the matrix representing a locally constant function by a permutation matrix results in a new function which differs from the original function by an automorphism of the shift.

Lemma 2.3. *Consider a locally constant function f represented by the $n \times n$ matrix A_f. Let P be a $n \times n$ permutation matrix and let g denote the locally constant function represented by the matrix $B_g = P^{-1}A_fP = P^TA_fP$, where P^T denotes the transpose of P (we are using the fact that $P^{-1} = P^T$ for any permutation matrix). Then the functions f and g are related by an automorphism of the shift determined by permuting the n letters of the alphabet by the permutation defined by P.*

Proof. Let us write the permutation matrix

$$P = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \cdots & r_{n,n} \end{pmatrix}.$$

The matrix P induces a permutation $\tau : \{1,\ldots,n\} \to \{1,\ldots,n\}$ by defining $r_k = e_{\tau(k)}$, where e_k denotes the row vector $(0,0,\ldots,0,1,0,\ldots 0)$ which contains a single 1 in the k-th place and 0 in all other places. This permutation defines an automorphism σ of the subshift Σ_A^+ simply by replacing every occurrence of the letter k in a word by the letter $\tau(k)$.

The (i,j) entry of the matrix $B_g = P^{-1}A_fP$ is $\sum_{k=1}^n \sum_{l=1}^n r_{k,i}r_{l,j} \exp(f(k,l))$, and thus the (i,j) entry of the matrix for the function $g \circ \sigma$ is $\sum_{k=1}^n \sum_{l=1}^n r_{k,i}r_{l,j} \exp(f(\tau(k),\tau(l)))$.

From definitions, we see that all of the products $r_{k,i}r_{l,j}$ vanish unless $i = \tau(k)$ and $j = \tau(l)$, in which case the (i,j) entry of the matrix for $g \circ \sigma$ is $\exp(f(i,j))$. ■

It is implicit in the hypothesis of Lemma 2.3 that conjugation by the permutation matrix P preserves the transition matrix of the subshift of finite type. If this is not the case then the two functions f and g are defined on different subshifts of finite type.

The next lemma says that in all but the trivial case the matrix obtained by conjugating a column stochastic matrix by a diagonal matrix is not column stochastic.

Lemma 2.4. *Let A denote an aperiodic $n \times n$ non-negative column stochastic matrix and D a diagonal matrix such that $D^{-1}AD$ is a column stochastic matrix. Then D is the identity matrix.*

Proof. If $A = \{A_{ij}\}$ and $D = \text{diag}\{d_1, \ldots d_n\}$, then $D^{-1}AD = \{(d_j/d_i)a_{ij}\}$. Since the matrix $D^{-1}AD$ is invariant under multiplication of D by a scaler, we can assume that $d_1 = 1$. Since A is column stochastic, the column sum must satisfy $\sum_{i=1}^{n} a_{ij} = 1$ for $j = 1, \ldots n$, and thus $(1, \ldots, 1)$ is a left eigenvector of A with eigenvalue 1. For $D^{-1}AD$ to be column stochastic, the columns must satisfy $\sum_{i=1}^{n} d_i a_{ij} = d_j$ for $j = 1, \ldots n$, and thus $(d_1, \ldots, d_n)$ is a left eigenvector of A with eigenvalue 1. By simplicity of the maximal eigenvalue (we are assuming that the transition matrix is aperiodic) we deduce that $(d_1, \ldots, d_n) = (1, \ldots, 1)$. ■

From the above results we conclude the main result of this section.

Theorem 1. *Let A be a $n \times n$ transition matrix, (Σ_A^+, σ) be a mixing one-sided subshift of finite type, and $f \in LC(m)$ be a locally constant function. Then there are at most $C(n^m)$ non-equivalent locally constant functions in $LC(m)$ with the same zeta function. Furthermore, for generic $f \in LC(m)$ (i.e., on the complement of a codimension one algebraic set) the zeta function determines the function (up to conjugation by diagonal and permutation matrices).*

Proof. By recoding, we can assume that $f \in LC(2)$ for a new subshift on n^m symbols. We apply Proposition 2.1 to the matrix A_f and use Lemmas 2.3 and 2.4 to show that the ambiguity in the conjugacy corresponds to equivalence of functions. ■

Corollary 1.1. *Given any locally constant function f there are at most countably many non-equivalent locally constant functions g with the same zeta function.*

Proof. It suffices to observe that if $f \in LC(m)$ and $g \in LC(m+l)$ for $l \geq 0$, and f and g share the same zeta function, then Theorem 1 shows that there are finitely many non-equivalent classes of such functions g. By considering the union over l the result follows. ■

In Appendix II we construct examples of non-equivalent locally constant functions with the same zeta function. In particular, this shows that $C(m) \geq 2$ for some m.

In the more general context of Hölder continuous functions, it is easy to see that knowledge of the zeta function is equivalent to knowing the unmarked orbit spectrum.

Proposition 2.2. *The zeta function for a Hölder continuous function determines the unmarked orbit spectrum, i.e., $\zeta_{f_1} = \zeta_{f_2}$ implies that $\mathcal{L}_{f_1} = \mathcal{L}_{f_2}$. Thus $\zeta_{f_1} = \zeta_{f_2}$ if and only if $\mathcal{L}_{f_1} = \mathcal{L}_{f_2}$.*

Proof. Let us write $\zeta(z,t) = \sum_{n=1}^{\infty}(z^n/n)\, a_n(t)$, where $a_n(t) = \sum_{\sigma^n x = x} \exp(tS_n f(x))$. The power series $\zeta(z,t)$ defines a holomorphic function in z in the disk of radius $\exp\beta(t)$. This implies that the functions $a_n(t)$ are all uniquely determined and can be obtained by differentiating the power series for fixed t. For each n we can apply Lemma 2.2 to the sum of exponentials to obtain the set of numbers $\{S_n f(x)\colon \sigma^n(x) = x\}$, and thus we can recover the entire unmarked orbit spectrum. ■

The zeta function, being a function of two variables, seems to contain a great deal of information about the function. Below we show that if we fix the variable $z = 1$, the zeta function still captures the weak unmarked orbit spectrum.

Proposition 2.3. *The restricted zeta function $\zeta_f(1,t)$ determines the weak orbit spectrum, i.e., $\zeta_{f_1}(1,t) = \zeta_{f_2}(1,t)$ implies that $\mathcal{W}_{f_1} = \mathcal{W}_{f_2}$.*

Proof. We can expand a restricted zeta function as a Dirichlet series

$$\zeta_f(1,t) = \sum_{n=1}^{\infty} a_n \exp(-\mu_n t),$$

where $\mu_1 \le \mu_2 \le \ldots \to \infty$ are real numbers corresponding to the unmarked orbit spectrum and $a_n \in \mathbb{C}$. This series converges on a half-plane containing some point $c \in \mathbb{R}$. Then by a result of Mandelbrojt [Man, p.388] we have for each $\nu \in \mathbb{R}$

$$\sum_{\mu_n < \nu} (\nu - \mu_n) = \begin{cases} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\zeta_f(1, c+it)}{(c+it)^2} \exp \nu(c+it)dt & \text{if } \nu > \mu_1 \\ 0 & \text{if } \nu \le \mu_1. \end{cases}$$

In particular, this allows us to recover the coefficients a_n and the exponents μ_n uniquely (up to permutation). ■

Finally, we show that the zeta function is deformation rigid. We call this property isozetal rigidity.

Theorem 2. *Let (Σ_A^+, σ) be a mixing one-sided subshift of finite type. Assume that $(-\epsilon, \epsilon) \ni \lambda \to f_\lambda \in C^\alpha(\Sigma_A^+, \mathbb{R})$ is a C^2 family of Hölder continuous functions with identical zeta functions (i.e., $\zeta_{f_\lambda} = \zeta_{f_0}$ for all $-\epsilon < \lambda < \epsilon$). Then the deformation is trivial, i.e., $f_\lambda \sim f_0$ for all $-\epsilon < \lambda < \epsilon$.*

Since we will prove a stronger result on beta functions in Section III.B, we omit the proof of this theorem.

III.A The free energy and beta functions: Locally constant functions

We remind the reader that we shall use the terms free energy and beta function interchangeably, since they are trivially related. Our formulation will be in terms of beta functions. We begin by showing that it is not always true that the beta function determines the zeta function.

Proposition 3.1. *There exist locally constant functions f_1, f_2 with $\beta_{f_1} = \beta_{f_2}$ but $\zeta_{f_1} \neq \zeta_{f_1}$.*

Proof. Let $\Sigma_2^+ = \prod_{n=0}^{\infty}\{1,2\}$. We can choose two rationally independent numbers $0 < a < b$ and define locally constant functions

$$f_1(x) = \begin{cases} a \text{ if } x_0 = 1 \\ b \text{ if } x_0 = 2 \end{cases} \quad \text{and} \quad f_2(x) = \begin{cases} a \text{ if } (x_0, x_1) = (1,1) \text{ or } (2,2) \\ b \text{ if } (x_0, x_1) = (1,2) \text{ or } (2,1) \end{cases}.$$

We easily see that $\beta_{f_1}(t) = \beta_{f_2}(t) = \exp(ta) + \exp(tb)$. Using (2), the associated zeta functions are given by

$$\zeta_{f_1}(z,t)^{-1} = \det \begin{pmatrix} 1 - z\exp(ta) & -z\exp(ta) \\ -z\exp(tb) & 1 - z\exp(tb) \end{pmatrix} = 1 - z(\exp(ta) + \exp(tb))$$

and

$$\begin{aligned} \zeta_{f_2}(z,t)^{-1} &= \det \begin{pmatrix} 1 - z\exp(ta) & -z\exp(tb) \\ -z\exp(tb) & 1 - z\exp(ta) \end{pmatrix} \\ &= 1 - 2z\exp(ta) + z^2(\exp(2ta) - \exp(2tb)), \end{aligned}$$

which are clearly different provided $a \neq b$. ∎

We now discuss the ambiguity with which the beta function determines the function. There do exist functions f for which the beta function β_f *does* determine the equivalence class of the function f. A trivial example is a coboundary $f = u \circ \sigma - u$. In this case, the beta function $\beta_f = 0$. Ruelle observed (see Lemma 3.3) that the lack of strict convexity of the pressure function implies that the function must be a coboundary.

We now show that *typically* the beta function for a locally constant function f determines the equivalence class of the locally constant function.

Theorem 3. *Let (Σ_A^+, σ) be a mixing one-sided subshift of finite type with incidence matrix A. For each $m \in \mathbb{N}$, for generic locally constant functions $f \in LC(m)$ the β-function determines the equivalence class of the function in $LC(m)$.*

The strategy for the proof of Theorem 3 is as follows. Proposition 3.2 below says that typically the beta function determines the characteristic polynomial. Proposition 2.1 says that typically the characteristic polynomial determines the function. Combining these results then gives Theorem 3. More precisely, we show that typically the characteristic polynomial $q_f(z,t)$ is minimal and thus the beta function determines $q_f(z,t)$.

We begin with a few preliminaries. Suppose that $\beta_{f_1} = \beta_{f_2}$, for $f_1, f_2 \in LC(m)$. By recoding if necessary, we can assume that $f_1, f_2 \in LC(2)$. Both functions have the same pressure $P = \log \beta(1)$. Let A be the associated $n \times n$ transition matrix.

If μ_i is the unique equilibrium state for f_i, let g_i denote the the Jacobian function defined by $g_i(x_0, x_1) = d\sigma^* \mu_i / d\mu_i(x_0, x_1)$, whenever $A(x_0, x_1) = 1$, and zero otherwise. There exists $u(x) = u(x_0)$ such that $f_i = g_i + u \circ \sigma - u + P$. By construction $\beta_{g_1} = \beta_{g_2}$ and $P(g_1) = P(g_2) = 0$. Thus the function g_i can be viewed as a canonical representative in the cohomology class of f_i. It suffices to show that generically g_1 is cohomologous to g_2.

We can associate to g_1 and g_2 non-negative column *stochastic* matrices $P_1 = (P_1(r,s))_{ij}$ and $P_2 = (P_2(r,s))_{ij}$, where $P_i(r,s) = A(r,s) \exp g_i(r,s)$ for $1 \leq r, s \leq n$. The function $\beta_{g_i}(t)$ is the maximal eigenvalue for $P_i^{(t)} \stackrel{\circ}{=} (P_i(r,s)^t)$, the matrix given by raising all of the entries to the power t. The matrices $P_i^{(t)} \in M_n(\mathcal{R})$, where $M_N n(\mathcal{R})$ denotes the ring of $n \times n$ matrices with entries in the ring

$$\mathcal{R} = \left\{ \sum_{i=1}^{k} n_i a_i^t : n_i \in \mathbb{Z}, a_i > 0 \right\},$$

and the beta function $\beta_{g_i}(t)$ is a zero of the characteristic polynomial $q_i(z,t) = \det(zI - P_i^{(t)}) \in \mathbb{Z}[\mathcal{R}]$.

Proposition 3.2. *The beta function $\beta_f(t)$ for a locally constant function defined by a $n \times n$ non-negative column stochastic matrix A typically determines the characteristic polynomial $q_f(z,t)$.*

The *minimal polynomial* $p(z)$ for $\beta_f(t)$ is the (unique) monic polynomial in $\mathbb{Z}[\mathcal{R}]$ of smallest degree for which $p(\beta_f(t)) = 0$. Clearly the beta function always determines its minimal polynomial. We begin by recalling the following result.

Lemma 3.1. *The minimal polynomial $p(z)$ for $\beta_f(t)$ divides the characteristic polynomial $q_f(z,t)$. In particular, if $q_f(z,t)$ is minimal, then it is determined by the beta function.*

Proof. We include a sketch of the proof (due to Tuncel cf.[Tun2]) for completeness. Let us consider the field $\mathcal{F}$ of rational fractions on $\mathcal{R}$. We define an ideal $\mathcal{I} \subset \mathcal{F}[z]$ by

$$\mathcal{I} = \{ f(z,t) \in \mathcal{F}[z] : f(\beta(t), t) = 0 \}.$$

Since $\mathcal{F}[z]$ is a principal ideal domain [Fra, p. 282] there exists an element $p(z,t) \in \mathcal{F}[z]$ such that $\mathcal{I} = p(z,t)\mathcal{F}[z]$. The element $p(z,t)$ must have minimal non-zero degree in $\mathcal{I}$ since it is generating. In particular, we can write $q_f(z,t) = p(z,t) \cdot s(z,t)$, for some $s(z,t) \in \mathcal{F}[z]$. Suppose that $q_f(z,t) = z^l - q_1 z^{l-1} - \ldots - q_{l-1} z - q_l$, $p(z,t) = z^d - p_1 z^{d-1} - \ldots - p_{d-1} z - p_d$ and $s(z,t) = z^m - s_1 z^{m-1} - \ldots - s_{m-1} z - s_m$, where the coefficients $q_i \in \mathcal{R}$ and $p_i, s_i \in \mathcal{F}$.

Let $\mathcal{S}$ be the group ring over $\mathbb{Z}$ generated using all the exponentials from the p_i, and exponentials from the numerators and denominators of p_i and s_i. We can assume that

$p_i = \tilde{p}_i/p_0$ and $s_i = \tilde{s}_i/s_0$, say, where $\tilde{p}_i, p_0, \tilde{s}_i, s_0 \in \mathcal{S}$. We can then rewrite $q_f(z,t) = p(z,t) \cdot s(z,t)$ as an identity in $\mathcal{S}$:

$$\begin{aligned} & s_0 p_0 \left[z^l - q_1 z^{l-1} - \ldots - q_{l-1} z - q_l \right] \\ & = \left[\tilde{p}_0 z^d - \tilde{p}_1 z^{d-1} - \ldots - \tilde{p}_{d-1} z - \tilde{p}_d \right] \left[s_0 z^m - \tilde{s}_1 z^{m-1} - \ldots - \tilde{s}_{m-1} z - \tilde{s}_m \right] \end{aligned}$$

However, since $\mathcal{S}$ is a unique factorization domain each irreducible factor of $p_0 s_0$ must divide one of the two terms multiplied on the right hand side. For example, if it divides the first term, then it divides each term $\tilde{p}_0, \ldots, \tilde{p}_d$, and is thus invertible (since they can be assumed to be coprime). Thus p_0 is a monomial and $p(z,t) \in \mathcal{R}[z]$. We deduce that p is a minimal polynomial of $\beta(t)$. ■

By a *typical matrix* A we mean that no non-trivial product of integer powers of entries a_{ij} is equal to 1 (or equivalently, the numbers $\log a_{ij}$ are rationally independent).

Lemma 3.2. *A characteristic polynomial $q_f(z,t)$ for a typical $n \times n$ non-negative column stochastic matrix A is minimal, i.e., $q_f(z,t) \neq a(z,t) \cdot b(z,t)$, where $a(z,t), b(z,t) \in \mathbb{Z}[\mathcal{R}]$ are non-constant functions.*

Proof. If we assume for a contraction that the characteristic polynomial is not minimal and we write it as a product $q_f(z,t) = a(z,t) \cdot b(z,t)$, then we can multiply out the coefficients and obtain a contradiction by comparing equations from the coefficients of the two polynomials multiplied together.

First, consider for the purposes of illustration the case of degree 2. The general case is similar. Assume for a contradiction that we can write

$$\begin{aligned} q_f(z,t) &= z^2 - \mathrm{tr}(A^{(t)}) z - \det(A^{(t)}) \\ &= (z - \sum_i l_i \lambda_i^t)(z - \sum_j m_j \mu_j^t) \\ &= z^2 - \left(\sum_i l_i \lambda_i^t + \sum_j m_j \mu_j^t \right) z + \sum_{i,j} l_i m_j \lambda_i^t \mu_j^t. \end{aligned}$$

Let $A = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}$. Then $\mathrm{tr}(A^{(t)}) = p^t + q^t$ and $\det(A^{(t)}) = p^t q^t - (1-p)^t (1-1)^t$. Comparing the z coefficients we see that $\{\lambda_i, \mu_j\} = \{p, q\}$. However, comparing the constant terms we see that we have $\{\lambda_i \mu_j\} = \{pq, (1-p)(1-q)\}$. However, this means that $pq = (1-p)(1-q)$, $p^2 = (1-p)(1-q)$ or $q^2 = (1-p)(1-q)$, which imposes relations on p and q. For typical functions this condition fails. This contradiction shows that $q_f(z,t)$ is minimal in the case $n = 2$.

For the general case, let us consider a typical matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

Then we can write

$$q_f(z,t) = z^n - \left(\sum_{i=1}^{n} a_{ii}^t\right) z^{n-1} + \left(\sum_{i\neq j} a_{ij}^t a_{ji}^t\right) z^{n-2} - \ldots \pm \sum_{\tau\in S_n} a_{1\tau(1)}^t a_{2\tau(2)}^t \cdots a_{n\tau(n)}^t,$$

where S_n denotes the set of permutations on n symbols. Let us assume for a contradiction that $q_f(z,t) = a(z,t)\cdot b(z,t)$ where $a(z,t), b(z,t) \in \mathbb{Z}[\mathcal{R}]$. If we write

$$\begin{aligned} a(z,t) &= z^d + \left(\sum_i l_i \lambda_i^s\right) z^{d-1} + \ldots + \left(\sum_r m_i \mu_r^s\right) \\ b(z,t) &= z^{n-d} + \left(\sum_j k_j \alpha_i^s\right) z^{n-d-1} + \ldots + \left(\sum_l p_l \eta_l^s\right), \end{aligned} \tag{3}$$

then we see that

$$q_f(z,t) = z^n - \left(\sum_i l_i \lambda_i^t + \sum_i k_i \alpha_i^t\right) z^{n-1} + \ldots + \left(\sum_{r,l} m_i p_l \mu_r^t \eta_l^t\right). \tag{4}$$

Comparing the coefficients of (3) and (4) we see the following:

(1) By comparing the constant terms we have that

$$\pm \sum_{\tau\in S_n} a_{1\tau(1)}^t a_{2\tau(2)}^t \cdots a_{n\tau(n)}^t = \sum_{r,l} m_r p_l \mu_r^t \eta_l^t.$$

In particular, there is a correspondence between the terms $\mu_r\eta_l$ and the terms $a_{1\tau(1)}a_{2\tau(2)}\cdots a_{n\tau(n)}$ for some $\tau \in S_n$.

(2) By comparing the z^d term we have that each η_r must be of the form

$$\frac{a_{1\tau(1)}a_{2\tau(2)}\cdots a_{n\tau(n)}}{a_{i_1 i_1}\ldots a_{i_{n-d} i_{n-d}}}, \tag{5}$$

for some permutation $\tau \in S_n$, which also fixes *precisely* d-terms $i_1, \ldots, i_d$ (i.e., (5) represents the terms for all the fixed points of the permutation). This is easily seen since in the expansion of $\det(zI - P^s)$ the z^d contribution comes from d entries on the diagonal corresponding to rows (and columns) $i_1, \ldots, i_d$, say.

(3) Similarly, by comparing the z^{n-d} term we have that each μ_r must be of the form

$$\frac{a_{1\tau(1)}a_{2\tau(2)}\cdots a_{n\tau(n)}}{a_{i_1 i_1}\ldots a_{i_{n-d} i_{N-d}}}, \tag{6}$$

for some permutation $\tau \in S_n$, which additionally fixes *precisely* $(n-d)$-terms $i_1, \ldots, i_{n-d}$.

Consider any term $a_{1\tau(1)}a_{2\tau(2)}\cdots a_{n\tau(n)}$ where $\tau \in S_n$ has no fixed point. This must occur in the constant term described in (1). However, this cannot be written as a product of a term η_l of the form (5) and a term μ_r of the form (6).[2] This contradiction shows that $q_f(z,t)$ is irreducible. ■

Remark. For $n \leq 4$ one can also show these results using the quadratic, cubic, and quartic formulae. In the case $n = 2$ let

$$A = \begin{pmatrix} a & 1-b \\ 1-a & b \end{pmatrix}.$$

The characteristic polynomial of $A^{(t)}$ is $q(z,t) = z^2 - \mathrm{tr}A^{(t)}\, z + \det A^{(t)}$. For this expression to be factorized into two linear non-trivial polynomials over $\mathcal{R}$ we require that the square root of the discriminant lie in $\mathcal{R}$, i.e.,

$$\sqrt{\mathrm{tr}^2 A^{(t)} - 4\det A^{(t)}} = \sqrt{(a^t + b^t)^2 - 4(a^t b^t - (1-a)^t(1-b)^t)} \in \mathcal{R}.$$

This can only hold when $a + b = 1$.

Since $\beta(t)$ is the maximal eigenvalue of $A^{(t)}$ it follows that

$$2\beta_f(t) = \mathrm{tr}A^{(t)} + \sqrt{\mathrm{tr}^2 A^{(t)} - 4\det A^{(t)}}.$$

If $\mathrm{tr}^2 A^{(t)} \neq 4\det A^{(t)}$, the beta function determines both $\mathrm{tr}A^{(t)}$ and $\sqrt{\mathrm{tr}^2 A^{(t)} - 4\det A^{(t)}}$. Substituting $\mathrm{tr}A^{(t)}$ into the term in the square root, one sees that the beta function also determines $\det A^{(t)}$, and thus $q(z,t)$. The nondegeneracy condition is equivalent to the condition $a + b \neq 1$.

The following theorem shows that for *any* $f \in LC(m)$ there are only finitely many functions in $LC(n)$ with the same beta function. The proof uses an interesting interplay between ring theory and thermodynamic formalism.

Theorem 4. *Let (Σ_A^+, σ) be a mixing one-sided subshift of finite type. For each $m \in \mathbb{N}$ and every locally constant function $f \in LC(m)$, there are at most finitely many non-equivalent locally constant functions in $LC(m)$ with the same beta function.*

Proof. Given the beta function associated to f we can consider the family of stochastic matrices given by

$$\mathcal{S} = \left\{ P : \sum_{i=1}^{n} P(i,j) = 1, \text{ for all } j \text{ and } \det(\beta(t)I - P^{(t)}) = 1, \text{ for all t} \in \mathbb{R} \right\}.$$

[2] To illustrate this, consider the case of a full shift on 4 symbols and $n = 4$. The "factors" of degrees $d = n - d = 2$. The constant term of $\det(zI - A^t)$ is a sum of 24 terms of the form $\pm a_{1\tau(1)}a_{2\tau(2)}a_{3\tau(3)}a_{4\tau(4)}$, where $\tau \in S_4$ is a permutation on 4 symbols. This includes $a_{12}a_{23}a_{34}a_{41}$ corresponding to the cyclic permutation (1234) The coefficient of z^2 must be a sum of the terms $a_{ij}a_{ji}$ and $a_{ii}a_{jj}$, since each z^2 contribution eliminates 2 rows (and columns) and the corresponding coefficient is the determinant of the remaining 2×2 matrix. In particular, we need that $a_{12}a_{23}a_{34}a_{41} = a_{ij}a_{ji}a_{rs}a_{sr}$, say, which is impossible for a typical matrix.

By analyticity considerations this is equivalent to

$$\mathcal{S} = \left\{ P : \sum_{i=1}^{n} P(i,j) = 1, \quad \det(\beta(t)I - P^{(k)}) = 1, \text{ for all } j \text{ and for all } k \in \mathbb{N} \right\}.$$

It is convenient to use this latter formulation to think of $\mathcal{S}$ as being an algebraic set in $\mathbb{R}^{n^2}$ given by an infinite set of polynomials. However, any such algebraic set can always be defined by only finitely many polynomials by the Hilbert Basis Theorem [Ful, p.13]. In particular, there are two possibilities:

(i) $\mathcal{S}$ is a finite set; or
(ii) $\mathcal{S}$ contains non-trivial connected components.

However, case (ii) cannot occur, since it contradicts Theorem 5 (Deformation Rigidity). We therefore conclude that $\mathcal{S}$ is finite, as claimed. ■

The proof of the following corollary is very similar to the proof of Corollary 1.1.

Corollary 4.1. *Given any locally constant function f there are at most countably many non-equivalent locally constant functions g with the same beta function.*

III.A The free energy and beta functions: Hölder continuous functions

In this section we prove the beta function is deformation rigid. In particular, this implies that there are no connected sets of isobetal functions. We call this phenomena isobetal rigidity. We observe that our proof easily extends to the beta function for smooth hyperbolic maps and Hölder continuous functions.

Let us briefly recall Ruelle's derivative formulas for pressure [PP, Rue].

Lemma 3.3. *Let f and g be Hölder continuous functions on Σ_A^+.*

(a) *The first derivative of pressure can be expressed as*

$$\left.\frac{\partial}{\partial s}\right|_{s=0} P(f+sg) = \int_{\Sigma_A^+} g d\mu_f,$$

where μ_f is Gibbs measure for potential f.

(b) *The second derivative of pressure can be expressed as*

$$\left.\frac{\partial^2}{\partial s^2}\right|_{s=0} P(f+sg) = var(g),$$

where μ_f is the Gibbs measure for potential f (i.e., $P(f) = h_{\mu_f}(\sigma) + \int_{\Sigma_A^+} f d\mu_f$), and

$$var_{\mu_f}(g) = \lim_{n\to\infty} \frac{1}{n} \int_{\Sigma_A} \left(\sum_{k=0}^{n-1} g(\sigma^i x) - \int_{\Sigma_A^+} g d\mu_f \right)^2 d\mu_f(x) \geq 0.$$

(c) *The expression $var_{\mu_f}(g) = 0$ if and only if $g \sim c$, where c is a constant.*

This brings us to the statement of isobetal rigidity.

Theorem 5. *Let (Σ_A^+, σ) be a mixing one-sided subshift of finite type. Assume that $(-\epsilon, \epsilon) \ni \lambda \to f_\lambda \in C^\alpha(\Sigma_A^+, \mathbb{R})$ is a C^2 family of Hölder continuous functions with identical beta functions (i.e., $\beta_{f_\lambda} = \beta_{f_0}$ for all $-\epsilon < \lambda < \epsilon$). Then $f_\lambda \sim f_0$ for all $-\epsilon < \lambda < \epsilon$.*

Proof. For any $s_0 \in (-\epsilon, \epsilon)$ we can use the C^2 assumption to make the expansion

$$f_s = f_{s_0} + (s - s_0) f_{s_0}^{(1)} + (s - s_0)^2 f_{s_0}^{(2)} + O((s - s_0)^3),$$

where $f_{s_0}^{(1)}$, $f_{s_0}^{(2)} \in C^\alpha(\Sigma_A^+)$. The hypothesis implies that the beta functions for this one parameter family all coincide. Thus $P(-tf_s) = 0$ for all $t \in \mathbb{R}$. From Lemma 3.3 we obtain that

$$0 = \left.\frac{\partial}{\partial s}\right|_{s=s_0} P(-tf_s) = -t \int_{\Sigma_A^+} f_{s_0}^{(1)} d\mu \tag{7}$$

and

$$0 = \left.\frac{\partial^2}{\partial^2 s}\right|_{s=s_0} P(-tf_s) = -t \int_{\Sigma_A^+} f_{s_0}^{(2)} d\mu + t^2 \mathrm{var}_\mu(f_{s_0}^{(1)}), \tag{8}$$

where $\mu = \mu_{-tf_{s_0}}$ denotes the Gibbs measure for the potential $-tf_{s_0}$.

Choose arbitrarily small t having the opposite sign of $\int_{\Sigma_A^+} f_{s_0}^{(2)} d\mu$. The expression on the right hand side of (8) is thus non-negative, and it follows that $\mathrm{var}(f_{s_0}^{(1)}) = 0$. Lemma 3.3(c) allows us to conclude that $f_{s_0}^{(1)} \sim c$, and (7) implies that $c = 0$. Thus for all $s_0 \in (-\epsilon, \epsilon)$ there exists $h_{s_0} \in C^\alpha(\Sigma_A^+, \mathbb{R})$ such that

$$\left.\frac{\partial f_s}{\partial s}\right|_{s=s_0} = h_{s_0} \circ \sigma - h_{s_0}.$$

We integrate both sides with respect to s, sum over each periodic orbit, and apply the Livsic theorem to conclude that for each s the function $f_s \sim f_0$. ∎

For a smooth hyperbolic map we can apply the proof of Theorem 5 directly on the manifold to obtain the following extension of isobetal deformation rigidity to Axiom-A diffeomorphisms.

Corollary IV.1. *Let M be a smooth manifold and suppose that $\Lambda \subset M$ is a basic set for an Axiom-A diffeomorphism $T: M \to M$. Assume that $(-\epsilon, \epsilon) \ni \lambda \to f_\lambda \in C^\alpha(\Lambda, \mathbb{R})$ is a C^2 family of Hölder continuous functions on Λ with identical beta functions (i.e., $\beta_{f_\lambda} = \beta_{f_0}$ for all $-\epsilon < \lambda < \epsilon$). Then $f_\lambda \sim f_0$ for all $-\epsilon < \lambda < \epsilon$.*

IV. The Unmarked Orbit Spectrum

We now turn to the final type of periodic orbit invariant we shall consider. This is the analogue of the unmarked length spectrum for hyperbolic surfaces where one labels the closed geodesics by word length.

Recall that the length spectrum for a hyperbolic surface is never simple, and in fact has unbounded multiplicity [Ran]. By contrast, a transversality argument easily shows that

the length spectrum of a non-constant negatively curved manifold is generically simple [Abr].

We say that the orbit spectrum for the function f is *simple* if the elements of the set $\mathcal{W}_f$ are all distinct. The next result shows that the orbit spectrum is generically simple, as analogous with the case of negatively curved surfaces.

Proposition 4.1. *Fix $\alpha > 0$. There exists a dense $\mathcal{G}_\delta$ subset of $\alpha-$Hölder continuous functions for which the unmarked orbit spectrum is simple.*

Proof. The space $C^\alpha(\Sigma_A^+, \mathbb{R})$ of $\alpha-$Hölder continuous functions is a complete metric space, and hence a Baire space. Let B_n denote the set of $\alpha-$Hölder continuous functions which give distinct weights to periodic orbits up to period n. This is clearly an open dense set. It follows from the Baire category theorem that the intersection of all B_n will be a dense G_δ set. ■

Even through we know by Proposition 4.1 that generically the spectrum is simple, every locally constant functions has non-simple spectrum, by analogy with the case of hyperbolic surfaces.

Proposition 4.2. *Every locally constant function has non-simple unmarked orbit spectrum with unbounded multiplicity, i.e., for each m sufficiently large, there exists M periodic points of the same period with the same Birkhoff sum.*

Proof. Let $f \in LC(n)$, say, have range $\{\alpha_1, \ldots, \alpha_k\}$ then the values $\{S_m f(x) : \text{ for } \sigma^m x = x\}$ are contained in the set

$$\left\{ \sum_{i=1}^k l_i \alpha_i : \sum_{i=1}^k l_i = m \right\},$$

which has cardinality at most m^k. However, since the number of periodic orbit of period m grows exponentially fast the result easily follows. ■

We now consider the extent to which the unmarked orbit spectrum determines the equivalence class of the function and prove the analogs of the theorems of Wolpert and McKean for hyperbolic surfaces (i.e., that typically the unmarked length spectrum determines the surface and there are uniform bounds depending on the genus of non-isometric surfaces having the same unmarked length spectrum).

Theorem 6. *Let (Σ_A^+, σ) be a mixing one-sided subshift of finite type. If $f \in LC(n)$ is a locally constant function then there are at most $C(n)$ non-equivalent locally constant functions in $LC(n)$ with the same unmarked orbit spectrum. Furthermore, for generic $f \in LC(n)$ (i.e., on the complement of a codimension one algebraic set) the unmarked orbit spectrum determines the function in $LC(n)$ (up to conjugation by diagonal and permutation matrices).*

Proof. This follows immediately from Theorem 1 and Proposition 2.2. ■

We now observe that the unmarked orbit spectrum is deformation rigid.

Theorem 7. *Let (Σ_A^+, σ) be a mixing one-sided subshift of finite type. Assume that $(-\epsilon, \epsilon) \ni \lambda \to f_\lambda \in C^\alpha(\Sigma_A^+, \mathbb{R})$ is a C^2 family of Hölder continuous functions with identical orbit spectra. Then the deformation is trivial, i.e., $f_\lambda \sim f_0$ for all $-\epsilon < \lambda < \epsilon$.*

We have already proved a stronger result on beta functions in Section III.B.

Finally, although for generic $f \in LC(n)$ we know that the beta function does determine the unmarked orbit spectrum, the following result shows that exceptional examples exist.

Proposition 4.6. *There exist non-equivalent $f_1, f_2 \in LC(2)$ with $\beta_{f_1} = \beta_{f_2}$, but $\mathcal{L}_{f_1} \neq \mathcal{L}_{f_2}$.*

Proof. Let $\Sigma_2^+ = \prod_{n=0}^{\infty}\{1,2\}$. Choose $a < b$ such that $\exp[a] + \exp[b] = 1$ and define locally constant functions

$$f_1(x) = \begin{cases} a \text{ if } (x_0, x_1) = (1,2) \text{ or } (2,1) \\ b \text{ if } (x_0, x_1) = (1,1) \text{ or } (2,2) \end{cases}$$

$$f_2(x) = \begin{cases} a \text{ if } (x_0, x_1) = (1,1) \text{ or } (2,2) \\ b \text{ if } (x_0, x_1) = (1,2) \text{ or } (2,1) \end{cases}.$$

We can easily compute

$$\begin{aligned} \sum_{\sigma^n x = x} \exp t(S_n f_1(x)) &= \operatorname{trace} \begin{pmatrix} \exp tb & \exp ta \\ \exp ta & \exp tb \end{pmatrix}^n \\ &= (\exp ta + \exp tb)^n + (\exp tb - \exp ta)^n \\ &\asymp (\exp ta + \exp tb)^n \end{aligned}$$

as $n \to \infty$ and

$$\begin{aligned} \sum_{\sigma^n x = x} \exp tS_n f_2 x)) &= \operatorname{tr} \begin{pmatrix} \exp ta & \exp tb \\ \exp tb & \exp ta \end{pmatrix}^n \\ &= (\exp ta + \exp ta)^n + (\exp ta - \exp tb)^n \\ &\asymp (\exp ta + \exp tb)^n \end{aligned}$$

as $n \to \infty$. It immediately follows that $\beta_{f_1} = \beta_{f_2}$. By considering periodic points of period 3, we see that $3\beta \in \mathcal{L}_{f_1}$ and $3\alpha \notin \mathcal{L}_{f_1}$, while $3\alpha \in \mathcal{L}_{f_2}$ and $3\beta \notin \mathcal{L}_{f_2}$. Thus $f_2 \not\sim f_1 \circ \tau$, for any shift automorphism τ. ∎

V. Geometric Characterization of Free Energy and Beta Function for Hyperbolic Surfaces

In this section we introduce natural notions of free energy and beta function for a hyperbolic surface, or more generally, a negatively curved surface. All results apply in the more general case, but for ease of exposition we only state them for hyperbolic surfaces. We show that it has many features which are reminiscent of the Manhattan curve of Burger

[Bur] and contains some refined information on the ratio of word length to hyperbolic length of closed geodesics, as studied by Milnor [Mil].

The link with subshifts of finite type is via special Markov partitions for the geodesic flow [BS]. In particular, we can always associate to the geodesic flow $\phi_t : S_1V \to S_1V$ on the unit tangent bundle to a hyperbolic surface V a subshift of finite type $\sigma : \Sigma_A^+ \to \Sigma_A^+$ and a function $f : \Sigma_A^+ \to \mathbb{R}$, so that the associated suspended flow models the geodesic flow. Moreover, given two hyperbolic surfaces with the same genus, the underlying subshifts $\sigma : \Sigma_A^+ \to \Sigma_A^+$ are the same. Let Γ_0 be a standard set of generators for the fundamental group, denoted by $\pi_1(V)$. We can assume that Γ_0 is symmetric, i.e., $g \in \Gamma_0$, then $g^{-1} \in \Gamma_0$. Given $g \in \pi_1(V) - \{e\}$ we can define the word length $|g|$ to be the smallest number of generators from Γ_0 needed to represent g.

Using the Bowen-Series coding, we can associate a subshift $\sigma : \Sigma_A^+ \to \Sigma_A^+$ and a Hölder continuous function $r : \Sigma_A^+ \to \mathbb{R}$. In particular, every closed geodesic γ corresponds to periodic orbit $\{x, \sigma x, \dots, \sigma^{n-1}x\}$ with $\sigma^n x = x$ and we can identify $r^n(x) = l(\gamma)$ and $n = |\gamma|$, where l denotes the hyperbolic length in V. This property of the coding is crucial for our analysis, and we are not aware of any other construction of Markov partitions where this property has been verified.

We can associate to the hyperbolic surface V and generating set Γ_0 the *beta function*

$$\beta(t) = \beta_{\Gamma_0,V}(t) = \lim_{n\to+\infty} \frac{1}{n} \log \left(\sum_{|\gamma|=n} \exp\left(-tl(\gamma)\right) \right)$$

It is easy to see that this function is *always* strictly convex. This is closely related to a *modified Poincaré series* for V and Γ_0 by

$$\rho(a,b) = \rho_{\Gamma_0,V}(a,b) \stackrel{\circ}{=} \sum_{\gamma} \exp\left(-al(\gamma) - b|\gamma|\right).$$

This infinite sum converges provided $a, b > 0$ are sufficiently large. We denote by $R = R(\Gamma_0, V) = \{(a,b) \in \mathbb{R}^2 : \rho(a,b) < \infty\}$ the domain of convergence of the function ρ. We define $L = L(\Gamma_0, V)$ to be the boundary curve for this set (see Figure 1).

Lemma 5.1. *The curve L is parameterized by $(a, \beta(a))$*

Proof. Using symbolic dynamics we can write

$$\rho(a,b) = \sum_{n=1}^{\infty} \frac{1}{n} \sum_{\sigma^n x = x} \exp\left(-aS_n r(x) - bn\right).$$

By the root test, this series converges if

$$\exp\left(P(-ar-b)\right) = \lim_{n\to\infty} \left(\sum_{\sigma^n x = x} \exp\left(-aS_n r(x) - bn\right) \right)^{1/n} < 1.$$

In particular, we see that

$$R = \{(a,b) \in \mathbb{R}^2 \,:\, P(-ar-b) < 0\} \text{ and } L = \{(a,b) \in \mathbb{R}^2 \,:\, P(-ar-b) = 0\}.$$

Since $P(-ar-b) = 0$ we see that $b = P(-ar) = \beta(a)$. ■

We recall that by a classical result of Milnor that for any hyperbolic surface there exist $A, B > 0$ such that $A \leq l(\gamma)/|\gamma| \leq B$ for all closed geodesics γ. If we choose $A = \inf_\gamma l(\gamma)/|\gamma|$ and $B = \sup_\gamma l(\gamma)/|\gamma|$ then using the Anosov closing lemma it is easy to show that the ratios $\{l(\gamma)/|\gamma| \,:\, \gamma = \text{closed geodesic}\}$ are dense in the interval $[A, B]$.

Theorem 8.

(a) *The curve L is real analytic and strictly convex.*

(b) *The points $(0,1)$ and $(h,0)$ lie on L, where*

$$h = \lim_{n \to +\infty} \frac{1}{n} \log Card\{\gamma \,:\, |\gamma| = n\}.$$

(c) *The asymptotic slope of L as $a \to \pm\infty$ are $-A$ and $-B$, respectively.*

Proof. Parts (a) and (b) follow easily from standard properties of pressure [PP]. In particular, h is the topological entropy of the subshift $\sigma : \Sigma_A^+ \to \Sigma_A^+$.

For part (c) we recall that the slope of the curve at $(a, \beta_f(a))$ is $\beta_f'(a) = -\int_{\Sigma_A^+} f d\mu_{-af}$, where μ_{-af} is the Gibbs measure for $-af$. By the variational principle we know that

$$h_{\mu_{-af}}(\sigma) - a \int_{\Sigma_A^+} f d\mu_{-af} \geq h_\mu(\sigma) - a \int_{\Sigma_A^+} f d\mu,$$

for all σ-invariant probability measures μ. Thus $\int_{\Sigma_A^+} f d\mu_{-af} \leq \inf_\mu \int_{\Sigma_A^+} f d\mu + 2h/a$. In particular, letting $a \to \infty$ we see that

$$\lim_{a \to \infty} \beta_f'(a) = -\lim_{a \to \infty} \int_{\Sigma_A^+} f d\mu_{-af} = -\inf_\mu \int_{\Sigma_A^+} f d\mu = -\inf_\gamma l(\gamma)/|\gamma| = -A.$$

The proof of the second part of (c) is similar. ■

Remark: Comparison with the Manhattan curve. We recall that Burger introduced the Manhattan curve in association with two length functions l_1, l_2 on the same hyperbolic surface. More precisely, one can define a Poincaré-type function of two variables by

$$\eta(a,b) = \sum_\gamma \exp\left(-a l_1(\gamma) - b l_2(\gamma)\right),$$

which converges providing a, b are sufficiently large. We denote by $R_M(l_1, l_2) = \{(a,b) \in \mathbb{R}^2 \,:\, \eta(a,b) < \infty\}$, and the *Manhattan curve* $L_M(l_1, l_2)$ is the boundary curve of R_M. Burger showed the following result.

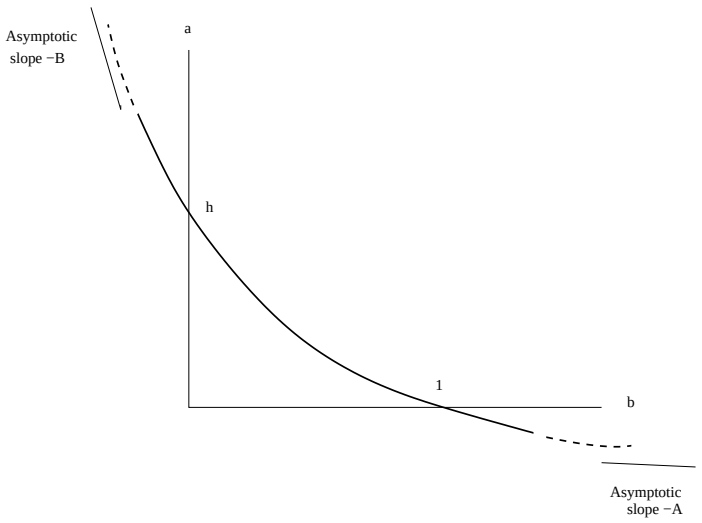

FIGURE 1. THE CURVE L

Proposition 5.1 [Bur]**.**

- **(a)** *The curve $L_M(l_1, l_2)$ is analytic and convex, and strictly convex except when $l_1 = l_2$;*
- **(b)** *The normal to $L_M(l_1, l_2)$ tends to the Thurston stretch $dil_+(l_1, l_2)$ and $dil_+(l_1, l_2)$ as $a \to \pm\infty$;*
- **(c)** *The normals as $L_M(l_1, l_2)$ crosses the axes are the intersection numbers $i(l_1, l_2)$ that are intimately related to the Weil-Petersson distance between the two hyperbolic metrics in Teichmüller space.*

APPENDIX I: UNCOUNTABLY MANY HÖLDER CONTINUOUS FUNCTIONS SHARING THE SAME UNMARKED ORBIT SPECTRUM (AND FREE ENERGY)

In contrast to the case of locally constant functions, the following result shows that for some Hölder continuous functions there are uncountably many mutually non-equivalent Hölder continuous functions with the same unmarked orbit spectrum.

Proposition A.I.1. *There exists an uncountable family of mutually inequivalent Hölder continuous functions which share the same unmarked orbit spectrum.*

Proof. Let Σ_3^+ denote the full one-sided shift on the three symbols 0, 1 and 2. To construct uncountably many functions with the same unmarked length spectrum we choose as our index sequences $\underline{i} = (i_n)_{n=1}^{\infty} \in \{0, 2\}^{\mathbb{N}}$, and define the Hölder continuous functions

$$f_{\underline{i}}(x) = \begin{cases} \theta^n & \text{if } x \in [i_n i_{n-1} \dots i_1 1] \\ 0 & \text{otherwise,} \end{cases}$$

where $[i_n i_{n-1} \dots i_1 1]$ denotes the obvious cylinder set. It is easy to see that for a fixed $\underline{i}$ the function $f_{\underline{i}}(y) = 0$ for every periodic point y having period n that is not of the form

$$y = y_{\underline{i}} = y_0 \dots y_{n_1} i_{k_1} \dots i_1 1 y_{m_2} \dots y_{n_2} i_{k_2} \dots i_1 1 \dots y_{m_r} \dots y_{n_r} i_{k_r} \dots i_1 1,$$

where

(i) $y_{n_j} \neq i_{k_j+1}$, for $j = 1, \ldots, k_r$;
(ii) $y_l \in \{0, 2\}$ for $m_j \leq l \leq n_j$ and $j = 1, \ldots, r$ (where $m_1 = 0$);
(iii) $0 \leq k_1, \ldots, k_r$ and $k_1 + \ldots + k_r + r \leq n$.

If one considers the Birkhoff sum over the periodic orbit for such a point y, an easy calculation shows that

$$S_n f_{\underline{i}}(y) = \frac{1 - \theta^{k_1}}{1 - \theta} + \ldots + \frac{1 - \theta^{k_r}}{1 - \theta}.$$

Let us observe that for the two periodic points (for different maps)

$$y_{\underline{i}} = y_0 \ldots y_{n_1} i_{k_1} \ldots i_1 1 y_{m_2} \ldots y_{n_2} i_{k_2} \ldots i_1 1 \ldots y_{m_r} \ldots y_{n_r} i_{k_r} \ldots i_1 1$$
$$y_{\underline{j}} = y_0 \ldots y_{n_1} j_{k_1} \ldots j_1 1 y_{m_2} \ldots y_{n_2} j_{k_2} \ldots j_1 1 \ldots y_{m_r} \ldots y_{n_r} j_{k_r} \ldots j_1 1,$$

the functions $f_{\underline{i}}(y_{\underline{i}}) = f_{\underline{j}}(y_{\underline{j}})$ and $S_n f_{\underline{i}}(y_{\underline{i}}) = S_n f_{\underline{j}}(y_{\underline{j}})$. This observation allows us to conclude that for each $n \geq 1$ the sets $\{S_n f_{\underline{i}}(y) : \sigma^n y = y\}$ coincide for all $\underline{i} \in \{0, 2\}^{\mathbb{N}}$. It immediately follows that for all $\underline{i} \in \{0, 2\}^{\mathbb{N}}$ the unmarked length spectra for $f_{\underline{i}}$ coincide.

Since different functions must necessarily have different marked length spectra, none of these functions differ by a coboundary. To see that there are uncountably many non-equivalent functions we need only recall that the space of automorphisms is countable. A simple cardinality argument completes the proof. ■

Corollary A.I.II. *There exists an uncountable family of mutually inequivalent Hölder continuous functions which share the same beta function.*

Appendix II: Locally Constant Functions with the Same Unmarked Orbit Spectrum (and Free Energy)

We continue our study of the ambiguity in recovering functions from their unmarked orbit spectrum. Here we work by close analogy with a basic construction of isospectral non-isometric hyperbolic surfaces using a construction of Buser [Bus] involving cospectral graphs. Buser's construction is really a reinterpretation of Sunada's common covering surface construction [Sun]. The goal of this appendix is to prove the following result.

Proposition A.II.1. *There exist non-equivalent locally constant functions f_1 and f_2 on the one-sided full shift on six symbols with $\mathcal{L}_{f_1} = \mathcal{L}_{f_2}$.*

To make the proof of this result as self-contained as possible, we present some preparatory material. If we construct distinct cospectral graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, i.e., two directed graphs with the same number of closed loops (or cycles) of any given period, then it would immediately follow that the corresponding subshifts of finite type have the same number of periodic points of any given period.

Consider the shift on n symbols $\sigma : \Sigma_A^+ \to \Sigma_A^+$, and let G be a finite group with a finite set of n generators $G_0 = \{A_1, \ldots, A_n\}$. We can associate to $G_0 \subset G$ a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, called the *Cayley graph* [Bus, Bol] having the following properties:

(a) There is a 1-1 correspondence between the vertices $\mathcal{V}$ of $\mathcal{G}$ and the elements of G;

(b) The edge set $\mathcal{E}$ contains the edge from vertex g_1 to vertex g_2 provided there is a $g \in G_0$ such that $g_1 = gg_2$.

There is a natural (right) action on the graph, by which $g \in G$ carries a vertex g_1 to the vertex $g_1 g$, and the edge g_1 to g_2 is carried to an edge from $g_1 g$ to $g_2 g$. Given a subgroup $H \subset G$ we can also consider the quotient graph $\mathcal{G}/H$. A trivial example is $\mathcal{G}_0 = \mathcal{G}/G$, which consists of both a single vertex and a directed edge for each element (and inverse) in G_0. In particular, the corresponding subshift of finite type is the full shift.

Given any directed graph $\mathcal{G}$, let $N_n(\mathcal{G})$ be the number of closed loops of length k ($k \geq 1$). Two directed graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are called *cospectral* if $N_k(\mathcal{G}_1) = N_k(\mathcal{G}_2)$, for all $k \geq 1$. Two subgroups $H_1, H_2 \subset G$ are called *almost conjugate* if

$$\#\{[g] \cap H_1\} = \#\{[g] \cap H_2\}$$

for every conjugacy class $[g]$, for every $g \in G$.

Lemma A.II.1. *The quotient graphs $\mathcal{G}/H_1$ and $\mathcal{G}/H_2$ of two almost conjugate subgroups are cospectral.*

Proof. The argument can easily be extracted from the proof of Sunada's theorem [Bus, Sun]. Let G_0^k denotes words of length k in elements of G_0. Observe that

$$\begin{aligned} N_k(\mathcal{G}_i) &= \#\{(g, g_0) \in G \times G_0^k \colon g_0 g H_i = g H_i\} \\ &= \sum_{g_0 \in G_0^k} \#\{g \in G \colon g^{-1} g_0 g H_i = H_i\} \\ &= \frac{1}{\#H_i} \sum_{g_0 \in G_0^k} \#\{g \in G \colon g^{-1} g_0 g \in H_i\} \end{aligned}$$

For $g, h \in G$, the expression $g^{-1} g_0 g = h^{-1} g_0 h$ holds if and only $g_0 = g h^{-1} g_0 h g^{-1} = (g h^{-1}) g_0 (g h^{-1})^{-1}$, which holds if and only if $g h^{-1} \in C_{g_0}$ (or equivalently $g \in C_{g_0} h$), where $C_{g_0} = \{g \in G \colon g g_0 g^{-1} = g_0\}$ denotes the centralizer of g_0 in G. Hence for fixed $g_1 = g^{-1} g_0 g$, one has that $\#\{h \in G \colon h^{-1} g_0 h = g_1\} = \#\{g C_{g_0}\} = \#C_{g_0}$. Thus $\#\{[g_0] \cap H_i\} \cdot \#C_{g_0} = \#\{g \in G \colon g^{-1} g_0 g \in H_i\}$, and we obtain

$$N_k(\mathcal{G}_i) = \frac{1}{\#H_i} \sum_{g_0 \in G_0^n} \#\{[g_0] \cap H_i\} \cdot \#C_{g_0}.$$

It immediately follows that $N_k(\mathcal{G}_1) = N_k(\mathcal{G}_2)$ if and only if the subgroups H_1 and H_2 are almost conjugate. ∎

Example [Bus, Lub]. Consider the group $G = SL(3, \mathbb{Z}_2)$ and $G_0 = \{A, B\}$ where

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

There are two almost conjugate subgroups

$$H_1 = \left\{ \begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \end{pmatrix} \in SL(2,\mathbb{Z}_2) \right\} \text{ and } H_2 = \left\{ \begin{pmatrix} * & 0 & 0 \\ * & * & * \\ * & * & * \end{pmatrix} \in SL(2,\mathbb{Z}_2) \right\},$$

which are not conjugate. The quotient graphs $\mathcal{G}/H_1$ and $\mathcal{G}/H_2$ are *non-isomorphic*[3] cospectral graphs each with seven vertices. See Figure 2.

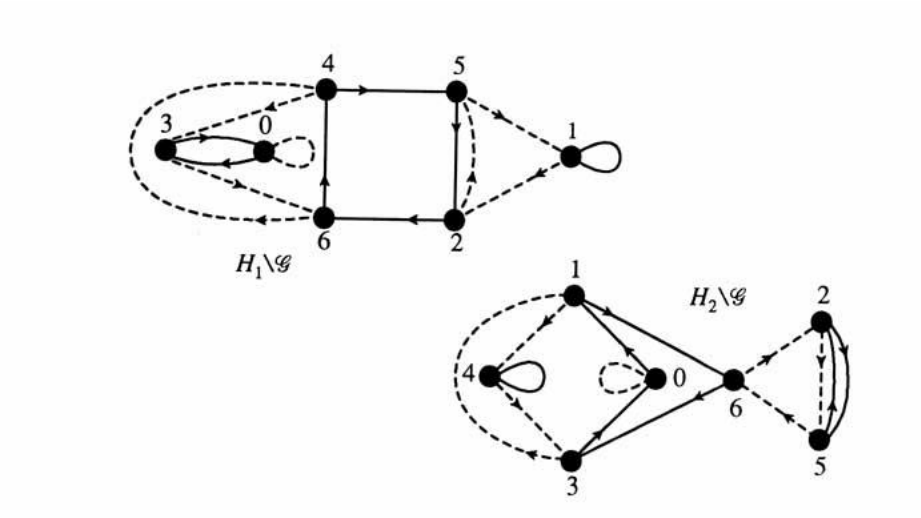

FIGURE 2. CAYLEY GRAPHS FOR H_1 AND H_2

Remark. Suppose G is a finite group that acts freely and isometrically on a compact Riemannian manifold M. Sunada showed that if H has two almost conjugate subgroups H_1 and H_2, then the quotient manifolds M/H_1 and M/H_2 are isospectral (for both lengths of closed geodesics and eigenvalues of the Laplacian).

We now prove Proposition A.II.1.

Proof of Proposition A.II.1. We can colour the edges of the two non-isomorphic quotient Cayley graphs $\mathcal{G}/H_1$ and $\mathcal{G}/H_2$ in the above example according to the generator A (solid edges in Figure 2) or B (dashed edges in Figure 2) which corresponds to that edge. In Lemma A.II.1 we have seen that for each n, the number of closed loops of length n, coincide for these two graphs, and it is easily seen that the number of closed loops of length n *which have the same colouring*, also coincides for each n.

We can extend each of these two graphs $\mathcal{G}/H_1$ by adding additional edges so that every vertex is connected to every other vertex. The corresponding subshift of finite type is then the full shift (on 6 symbols). For the larger graphs, we declare that the edges labeled by A

[3]Two graphs $\mathcal{A}$ and $\mathcal{B}$ are isomorphic if there exists a one-to-one mapping ϕ from the vertex set of $\mathcal{A}$ onto the vertex set of $\mathcal{B}$ such that whenever vertices P and Q of $\mathcal{A}$ are connected by exactly k edges, then $\phi(P)$ and $\phi(Q)$ are also connected by exactly k edges.

have length α, the edges labeled by B have length β, and the additional edges have length 1, where 1, α and β are rationally independent. We can define $f_1, f_2 \in LC(2)$ by declaring that $f_i(x_0x_1)$ equals the length of the edge connecting vertices x_0 to x_1 for the extension of $\mathcal{G}/H_i$. Since the graphs $\mathcal{G}/H_1$ and $\mathcal{G}/H_2$ are non-isomorphic, the functions f_1, f_2 are non-equivalent, however, by design, $\mathcal{L}_{f_1} = \mathcal{L}_{f_2}$. ■

Appendix III: Uncountably Many Hölder Continuous Functions Sharing the Same Weak Orbit Spectrum but Having Different Unmarked Orbit Spectra

In the following proposition we construct an uncountably family of pairwise inequivalent Hölder continuous functions with different unmarked orbit spectrum, but all sharing the same weak orbit spectrum. The construction is in the same spirit as the construction on Appendix I.

Proposition A.III. *There exists an uncountably family of pairwise inequivalent Hölder continuous functions with different unmarked orbit spectra, but all sharing the same weak orbit spectrum.*

Proof. Let Σ_2^+ denote a full shift on the two symbols 0 and 1. for any $n \geq 0$ we denote

$$[0^n 1] := \{x \in \Sigma_2^+ : x_i = 0, 0 \leq i \leq n-1 \text{ and } x_n = 1\}.$$

Let $0 < \theta < 1$ and then we can define a function

$$f(x) = \begin{cases} \theta^n & \text{if } x \in [0^n 1] \\ 0 & y \text{ if } x = (0,0,0,\dots). \end{cases}$$

To write down the orbit spectrum it is convenient to relate this to the shift on infinitely many symbols $[0^n 1]$ (with allowed transitions $[0^n 1] \to [0^{n-1} 1]$ and $[1] \mapsto [0^n 1]$). The representation of the locally constant f on this shift with infinitely many symbols is given by $f([0^n 1]) = \theta^n$. By a simple calculation we see that the length spectrum is the semi-group generated by the values $1 + \theta + \theta^2 + \ldots + \theta^n = (1 - \theta^{n+1})/(1 - \theta)$.

As a first step consider for any $k \geq 1$ the function defined by

$$f_k(x) = \begin{cases} \theta^n & \text{if } x \in [0^n 1], n \notin \{k, k+1, k+2\} \\ \theta^k + \theta^{k+1} & \text{if } x \in [0^k 1] \\ -\theta^{k+1} & \text{if } x \in [0^{k+1} 1] \\ \theta^{k+1} + \theta^{k+2} & \text{if } x \in [0^{k+2} 1] \\ 0 & \text{if } x = (0,0,0,\dots). \end{cases}$$

It is easy to see that for $k \geq 2$ the weak orbit spectra agree (i.e., $\mathcal{W}_{f_k} = \mathcal{W}_{f_1}$) although the unmarked orbit spectrums differ (i.e., $\mathcal{L}_{f_k} \neq \mathcal{L}_{f_1}$). It immediately follows from the latter observation that these functions are not mutually cohomologous. Moreover, it is a simple

matter to modify this construction so that the function is either changed in a similar way, or left unchanged, for $k = 3, 6, 9, \ldots$. In this way we can construct an uncountable family of functions having the same weak orbit spectrum as the original function. This is easy checked to be Hölder.

First observe that none of these functions differ by a coboundary. This is easily seen by observing that different functions must necessarily have different marked length spectra. For example, the weighting for f and f_k of the closed orbit of period $k+1$ in the cylinder $[0^k 1]$ are $1+\theta+\ldots+\theta^k+\theta^k$ and $1+\theta+\ldots+\theta^k+\theta^{k+1}$, respectively. A similar observation applies in other cases. Secondly, observe that there are uncountably many non-equivalent functions we need only recall that the space of automorphisms is countable. A simple cardinality argument completes the proof. ■

References

[Abr] R. Abraham, *Bumpy Metrics*, Proceedings Sym. Pure Math. **14** (1966), 1–3.

[BS] R. Bowen and C. Series, *Markov Maps Associated with Fuchsian Groups*, Inst. Hautes tudes Sci. Publ. Math. **50** (1979), 153–170.

[BPS] L. Barreira, Y. Pesin and J. Schmeling, *On A General Concept of Multifractal Rigidity: Multifractal Spectra for Dimensions, Entropies, and Lyapunov Exponents. Multifractal Rigidity*, Chaos **7** (1999), 27–38.

[Bol] Bollobás, *Modern Graph Theory*, Graduate Texts in Mathematics 184, Springer Verlag, 1998.

[Bow] R. Bowen, *One-dimensional hyperbolic sets for flows*, J. Diff. Equat. **12** (1972), 173–179..

[Bur] M. Burger, *Intersection, the Manhattan Curve, and Patterson-Sullivan Theory in Rank 2*, Internat. Math. Res. Notices **7** (1993), 217–225..

[Bus] P. Buser, *Geometry and Spectra of Compact Riemann Surfaces*, Birkhäuser, 1992.

[Fra] J. Fraleigh, *A First Course in Abstract Algebra*, 8th ed, Addison Wesley Longman,, 1998.

[Ful] W. Fulton, *Algebraic Curves*, Benjamin/Cummings, 1969.

[GK] V. Guillemin and D. Kazhdan, *Some Inverse Spectral Results for Negatively Curved 2-Manifolds*, Topology **19** (1980), 301–312.

[Hed] G. Hedlund, *Endomorphisms and Automorphisms of the Shift Dynamical System*, Math. Systems Th. **3** (1969), 320–375.

[HJ] R. Horn and C. Johnson, *Matrix Analysis*, CUP, 1985.

[Liv] A. Livsic, *Cohomology Properties of Dynamical Systems*, Math. USSR-Izv **6** (1972), 1278–1301.

[Lub] A. Lubotzky, *Discrete groups, Expanding Graphs and Invariant Measures*, Progress in Mathematics 125, Birkhauser, 1994.

[Man] S. Mandelbrojt, *Selecta*, Gauthier-villars, 1981.

[McK] H. McKean, *Selberg's Trace Formula as Applied to a Compact Riemann Surface*, Comm. Pure Appl. Math. **25** (1972), 225–246.

[Mil] J. Milnor, *A note on curvature and fundamental group*, J. Diff. Geom. **2** (1968), 1–7.

[Ota] J. P. Otal, *Le Spectre Marqué des Longueurs des Surfaces à Courbure Negative.*, Ann. of Math. **131** (1990), 151–162.

[PP] W. Parry and M. Pollicott, *Zeta Functions and the Periodic Orbit Structures of Hyperbolic Dynamics*, Astérisque **187–188** (1990).

[Ran] B. Randol, *The Length Spectrum of a Riemann Surface is Always of Unbounded Multiplicity.*, Proc. Amer. Math. Soc. **78** (1980), 455–456.

[Rue] D. Ruelle, *Thermodynamic Formalism*, Addison-Wesley, 1978.

[Sun] T. Sunada, *Riemannian Coverings and Isospectral Manifolds.*, Ann. of Math. **121** (1985), 69–186.

[Tun1] S. Tuncel, *Conditional Pressure and Coding*, Israel J. Math **39** (1981), 101-112.

[Tun2] S. Tuncel, *Coefficient Rings for Beta Function Classes of Markov Chains*, Erg. Th. & Dyn. Sys. **20** (2000), 1477-1493.

[Vig] M. F. Vigneras, *Variétés Riemanniennes Isospectrales et non Isometriques.*, Ann. of Math. **112** (1980), 21–32.

[Wae] B. L. van der Waerden, *Algebra*, vol 1, Ungar, 1970.

[Wol] S. Wolpert, *The Length Spectra as Moduli for Compact Riemann Surfaces.*, Ann. of Math. **109** (1979), 323–351.