HIDDEN MARKOV PROCESSES IN THE CONTEXT OF SYMBOLIC DYNAMICS

MIKE BOYLE AND KARL PETERSEN

Abstract. In an effort to aid communication among different fields and perhaps facilitate progress on problems common to all of them, this article discusses hidden Markov processes from several viewpoints, especially that of symbolic dynamics, where they are known as sofic measures, or continuous shift-commuting images of Markov measures. It provides background, describes known tools and methods, surveys some of the literature, and proposes several open problems.

Contents

Received by the editors August 13, 2010.
2010 *Mathematics Subject Classification.* Primary: 60K99, 60-02, 37-02; Secondary: 37B10, 60J10, 37D35, 94A15.

1. Introduction

Symbolic dynamics is the study of shift (and other) transformations on spaces of infinite sequences or arrays of symbols and maps between such systems. A symbolic dynamical system, with a shift-invariant measure, corresponds to a stationary stochastic process. In the setting of information theory, such a system amounts to a collection of messages. Markov measures and hidden Markov measures, also called sofic measures, on symbolic dynamical systems have the desirable property of being determined by a finite set of data. But not all of their properties, for example the entropy, can be determined by finite algorithms. This article surveys some of the known and unknown properties of hidden Markov measures that are of special interest from the viewpoint of symbolic dynamics. To keep the article self contained, necessary background and related concepts are reviewed briefly. More can be found in [66, 78, 77, 96].

We discuss methods and tools that have been useful in the study of symbolic systems, measures supported on them, and maps between them. Throughout we state several problems that we believe to be open and meaningful for further progress. We review a swath of the complicated literature starting around 1960 that deals with the problem of recognizing hidden Markov measures, as closely related ideas were repeatedly rediscovered in varying settings and with varying degrees of generality or practicality. Our focus is on the probability papers that relate most closely to symbolic dynamics. We have left out much of the literature concerning probabilistic and linear automata and control, but we have tried to include the main ideas relevant to our problems. Some of the explanations that we give and connections that we draw are new, as are some results near the end of the article. In Section 5.2 we give bounds on the possible order (memory) if a given sofic measure is in fact a Markov measure, with the consequence that in some situations there is an algorithm for determining whether a hidden Markov measure is Markov. In Section 6.3 we show that every factor map is hidden Markovian, in the sense that every hidden Markov measure on an irreducible sofic subshift lifts to a fully supported hidden Markov measure.

2. Subshift background

2.1. **Subshifts.** Let $\mathcal{A}$ be a set, usually finite or sometimes countable, which we consider to be an alphabet of symbols.

$$\mathcal{A}^* = \bigcup_{k=0}^{\infty} \mathcal{A}^k \tag{2.1}$$

denotes the set of all finite blocks or words with entries from $\mathcal{A}$, including the empty word, ϵ; $\mathcal{A}^+$ denotes the set of all nonempty words in $\mathcal{A}^*$; $\mathbb{Z}$ denotes the integers and $\mathbb{Z}_+$ denotes the nonnegative integers. Let $\Omega(\mathcal{A}) = \mathcal{A}^{\mathbb{Z}}$ and $\Omega^+(\mathcal{A}) = \mathcal{A}^{\mathbb{Z}_+}$ denote the set of all two or one-sided sequences with entries from $\mathcal{A}$. If $\mathcal{A} = \{0, 1, \ldots, d-1\}$ for some integer $d > 1$, we denote $\Omega(\mathcal{A})$ by Ω_d and $\Omega^+(\mathcal{A})$ by Ω_d^+. Each of these spaces is a metric space with respect to the metric defined by setting for $x \neq y$

$$k(x,y) = \min\{|j| : x_j \neq y_j\} \quad \text{and } d(x,y) = e^{-k(x,y)}. \tag{2.2}$$

For $i \leq j$ and $x \in \Omega(\mathcal{A})$ we denote by $x[i,j]$ the block or word $x_i x_{i+1} \dots x_j$. If $\omega = \omega_0 \dots \omega_{n-1}$ is a block of length n, we define

$$\mathcal{C}_0(w) = \{y \in \Omega(\mathcal{A}) : y[0, n-1] = \omega\}, \tag{2.3}$$

and, for $i \in \mathbb{Z}$,

$$\mathcal{C}_i(\omega) = \{y \in \Omega(\mathcal{A}) : y[i, i+n-1] = \omega\}. \tag{2.4}$$

The cylinder sets $\mathcal{C}_i(\omega), \omega \in \mathcal{A}^*, i \in \mathbb{Z}$, are open and closed and form a base for the topology of $\Omega(\mathcal{A})$.

In this paper, a *topological dynamical system* is a continuous self map of a compact metrizable space. The *shift transformation* $\sigma : \Omega_d \to \Omega_d$ is defined by $(\sigma x)_i = x_{i+1}$ for all i. On Ω_d the maps σ and σ^{-1} are one-to-one, onto, and continuous. The pair (Ω_d, σ) forms a topological dynamical system which is called the *full d-shift.*

If X is a closed σ-invariant subset of Ω_d, then the topological dynamical system (X, σ) is called a *subshift.* In this paper, with "σ-invariant" we include the requirement that the restriction of the shift be surjective. Sometimes we denote a subshift (X, σ) by only X, the shift map being understood implicitly. When dealing with several subshifts, their possibly different alphabets will be denoted by $\mathcal{A}(X), \mathcal{A}(Y)$, etc.

The *language* $\mathcal{L}(X)$ of the subshift X is the set of all finite words or blocks that occur as consecutive strings

$$x[i, i+k-1] = x_i x_{i+1} \dots x_{i+k-1} \tag{2.5}$$

in the infinite sequences x which comprise X. Denote by $|w|$ the length of a string w. Then

$$\mathcal{L}(X) = \{w \in \mathcal{A}^* : \text{there are } n \in \mathbb{Z}, y \in X \text{ such that } w = y_n \dots y_{n+|w|-1}\}. \tag{2.6}$$

Languages of (two-sided) subshifts are characterized by being *extractive* (or *factorial*) (which means that every subword of any word in the language is also in the language) and *insertive* (or *extendable*) (which means that every word in the language extends on both sides to a longer word in the language).

For each subshift (X, σ) of (Ω_d, σ) there is a set $\mathcal{F}(X)$ of finite "forbidden" words such that

$$X = \{x \in \Omega_d : \text{for each } i \leq j, x_i x_{i+1} \dots x_j \notin \mathcal{F}(X)\}. \tag{2.7}$$

A *shift of finite type (SFT)* is a subshift (X, σ) of some $(\Omega(\mathcal{A}), \sigma)$ for which it is possible to choose the set $\mathcal{F}(X)$ of forbidden words defining X to be finite. (The choice of set $\mathcal{F}(X)$ is not uniquely determined.) The SFT is *n-step* if it is possible to choose the set of words in $\mathcal{F}(X)$ to have length at most $n+1$. We will sometimes use "SFT" as an adjective describing a dynamical system.

One-step shifts of finite type may be defined by $0, 1$ transition matrices. Let M be a $d \times d$ matrix with rows and columns indexed by $\mathcal{A} = \{0, 1, \dots, d-1\}$ and entries from $\{0, 1\}$. Define

$$\Omega_M = \{\omega \in \mathcal{A}^{\mathbb{Z}} : \text{for all } n \in \mathbb{Z}, M(\omega_n, \omega_{n+1}) = 1\}. \tag{2.8}$$

These were called *topological Markov chains* by Parry [72]. A topological Markov chain Ω_M may be viewed as a *vertex shift*: its alphabet may be identified with the vertex set of a finite directed graph such that there is an edge from vertex i to vertex j if and only if $M(i, j) = 1$. (A square matrix with nonnegative integer entries can similarly be viewed as defining an *edge shift*, but we will not need edge shifts in this paper.) A topological Markov chain with transition matrix M as above is called *irreducible* if for all $i, j \in \mathcal{A}$ there is k such that $M^k(i, j) > 0$. Irreducibility corresponds to the associated graph being strongly connected.

2.2. **Sliding block codes.** Let (X, σ) and (Y, σ) be subshifts on alphabets $\mathcal{A}, \mathcal{A}'$, respectively. For $k \in \mathbb{N}$, a *k-block code* is a map $\pi : X \to Y$ for which there are $m, n \geq 0$ with $k = m + n + 1$ and a function $\pi : \mathcal{A}^k \to \mathcal{A}'$ such that

$$(\pi x)_i = \pi(x_{i-m} \dots x_i \dots x_{i+n}). \tag{2.9}$$

We will say that π is a *block code* if it is a k-block code for some k.

Theorem 2.1 Curtis-Hedlund-Lyndon Theorem**.** *For subshifts (X, σ) and (Y, σ), a map $\psi : X \to Y$ is continuous and commutes with the shift ($\psi\sigma = \sigma\psi$) if and only if it is a block code.*

If (X, T) and (Y, S) are topological dynamical systems, then a *factor map* is a continuous onto map $\pi : X \to Y$ such that $\pi T = S\pi$. (Y, S) is called a *factor* of (X, T), and (X, T) is called an *extension* of (Y, S). A one-to-one factor map is called an *isomorphism* or *topological conjugacy.*

Given a subshift (X, σ), $r \in \mathbb{Z}$ and $k \in \mathbb{Z}_+$, there is a block code $\pi = \pi_{r,k}$ onto the subshift which is the *k-block presentation* of (X, σ), by the rule

$$(\pi x)_i = x[i + r, i + r + 1, \dots, i + r + k - 1] \quad \text{for all } x \in X. \tag{2.10}$$

Here π is a topological conjugacy between (X, σ) and its image $(X^{[k]}, \sigma)$ which is a subshift of the full shift on the alphabet $\mathcal{A}^k$.

Two factor maps ϕ, ψ are *topologically equivalent* if there exist topological conjugacies α, β such that $\alpha\phi\beta = \psi$. In particular, if ϕ is a block code with $(\phi x)_0$ determined by $x[-m, n]$ and $k = m + n + 1$ and ψ is the composition $(\pi_{m,k})^{-1}$ followed by ϕ, then ψ is a 1-block code (i.e. $(\psi x)_0 = \psi(x_0)$) which is topologically equivalent to ϕ.

A *sofic* shift is a subshift which is the image of a shift of finite type under a factor map. A sofic shift Y is *irreducible* if it is the image of an irreducible shift of finite type under a factor map. (Equivalently, Y contains a point with a dense forward orbit. Equivalently, Y contains a point with a dense orbit, and the periodic points of Y are dense.)

2.3. **Measures.** Given a subshift (X, σ), we denote by $\mathcal{M}(X)$ the set of σ-invariant Borel probability measures on X. These are the measures for which the coordinate projections $\pi_n(x) = x_n$ for $x \in X, n \in \mathbb{Z}$, form a two-sided finite-state stationary stochastic process.

Let P be a $d \times d$ stochastic matrix and p a stochastic row vector such that $pP = p$. (If P is irreducible, then p is unique.) Define a $d \times d$ matrix M with entries from $\{0, 1\}$ by $M(i, j) = 1$ if and only if $P(i, j) > 0$. Then P determines a 1-step stationary (σ-invariant) Markov measure μ on the shift of finite type Ω_M by

$$\begin{aligned}\mu(\mathcal{C}_i(\omega[i, j])) &= \mu\{y \in \Omega_M : y[i, j] = \omega_i\omega_{i+1} \dots \omega_j\} \\ &= p(\omega_i)P(\omega_i, \omega_{i+1}) \cdots P(\omega_{j-1}, \omega_j)\end{aligned} \tag{2.11}$$

(by the Kolmogorov Extension Theorem).

For $k \geq 1$, we say that a measure $\mu \in \mathcal{M}(X)$ is *k-step Markov* (or more simply *k-Markov*) if for all $i \geq 0$ and all $j \geq k - 1$ and all x in X,

$$\mu(\mathcal{C}_0(x[0, i])|\mathcal{C}_0(x[-j, -1])) = \mu(\mathcal{C}_0(x[0, i])|\mathcal{C}_0(x[-k, -1])). \tag{2.12}$$

A measure is 1-step Markov if and only if it is determined by a pair (p, P) as above. A measure is k-step Markov if and only if its image under the topological conjugacy taking (X, σ) to its k-block presentation is 1-step Markov. We say that a measure is *Markov* if it is k-step Markov for some k. The set of k-step Markov measures is denoted by $\mathcal{M}_k$ (adding an optional argument to specify the system or transformation if necessary.) *From here on, "Markov" means "shift-invariant Markov with full support"*, that is, every nonempty cylinder subset of X has positive measure. With this convention, a Markov measure with defining matrix P is ergodic if and only if P is irreducible.

A probabilist might ask for motivation for bringing in the machinery of topological and dynamical systems when we want to study a stationary stochastic process. First, looking at $\mathcal{M}(X)$ allows us to consider and compare many measures in a common setting. By relating them to continuous functions ("thermodynamics"—see Section 3.2 below) we may find some distinguished measures, for example maximal ones in terms of some variational problem. Second, by topological conjugacy we might be able to simplify a situation conceptually; for example, many problems involving block codes reduce to problems involving just 1-block codes. And third, with topological and dynamical ideas we might see (and know to look for) some structure or common features, such as invariants of topological conjugacy, behind the complications of a particular example.

2.4. **Hidden Markov (sofic) measures.** If (X, σ) and (Y, σ) are subshifts and $\pi : X \to Y$ is a sliding block code (factor map), then each measure $\mu \in \mathcal{M}(X)$ determines a measure $\pi\mu \in \mathcal{M}(Y)$ by

$$(\pi\mu)(E) = \mu(\pi^{-1}E) \quad \text{for each measurable } E \subset Y. \tag{2.13}$$

(Some authors write $\pi_*\mu$ or $\mu\pi^{-1}$ for $\pi\mu$.)

If X is SFT, μ is a Markov measure on X and $\pi : X \to Y$ is a sliding block code, then $\pi\mu$ on Y is called a *hidden Markov measure* or *sofic measure*. (Various other names, such as "submarkov" and "function of a Markov chain" have also been used for such a measure or the associated stochastic process.) Thus $\pi\mu$ is a convex combination of images of ergodic Markov measures. *From here on, unless otherwise indicated, the domain of a Markov measure is assumed to be an irreducible SFT, and the Markov measure is assumed to have full support (and thus by irreducibility be ergodic). Likewise, unless otherwise indicated, a sofic measure is assumed to have full support and to be the image of an ergodic Markov measure.* Then the sofic measure is ergodic and it is defined on an irreducible sofic subshift. Hidden Markov measures provide a natural way to model systems governed by chance in which dependence on the past of probabilities of future events is limited (or at least decays, so that approximation by Markov measures may be reasonable) and complete knowledge of the state of the system may not be possible.

Hidden Markov processes are often defined as probabilistic functions of Markov chains (see for example [33]), but by enlarging the state space each such process can be represented as a deterministic function of a Markov chain, such as we consider here (see [8]).

The definition of hidden Markov measure raises several questions.

Problem 2.2. Let μ be a 1-step Markov measure on (X, σ) and $\pi : X \to Y$ a 1-block code. The image measure may not be Markov—see Example 2.8. What are necessary and sufficient conditions for $\pi\mu$ to be 1-step Markov?

This problem has been solved, in fact several times. Similarly, given μ and π, it is possible to determine whether $\pi\mu$ is k-step Markov. Further, given π and a Markov measure μ, it is possible to specify k such that either $\pi\mu$ is k-step Markov or else is not Markov of any order. These results are discussed in Section 5.

Problem 2.3. Given a shift-invariant measure ν on (Y, σ), how can one tell whether or not ν is a hidden Markov measure? If it is, how can one construct Markov measures of which it is the image?

The answers to Problem 2.3 provided by various authors are discussed in Section 4. The next problem reverses the viewpoint.

Problem 2.4. Given a sliding block code $\pi : X \to Y$ and a Markov measure ν on (Y, σ), does there exist a Markov measure μ on X such that $\pi\mu = \nu$?

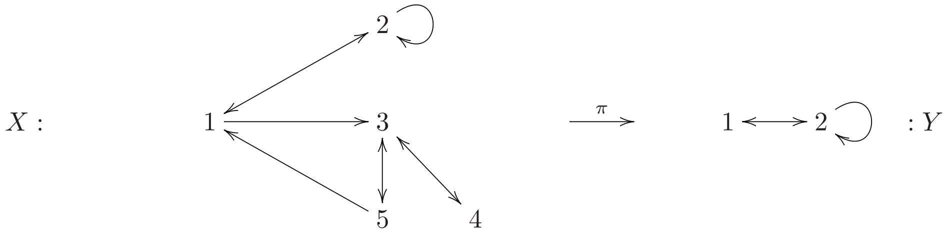
In Section 3, we take up Problem 2.4 (which apart from special cases remains open) and some theoretical background that motivates it.

Recall that a factor map $\pi : X \to Y$ between irreducible sofic shifts has a *degree*, which is the cardinality of the preimage of any doubly transitive point of Y [66]. (If the cardinality is infinite, it can only be the power of the continuum, and we simply write degree$(\pi) = \infty$.) If π has degree $n < \infty$, then an ergodic measure ν with full support on Y can lift to at most n ergodic measures on X. We say that the *degree of a hidden Markov measure* ν, also called its *sofic degree*, is the minimal degree of a factor map which sends some Markov measure to ν.

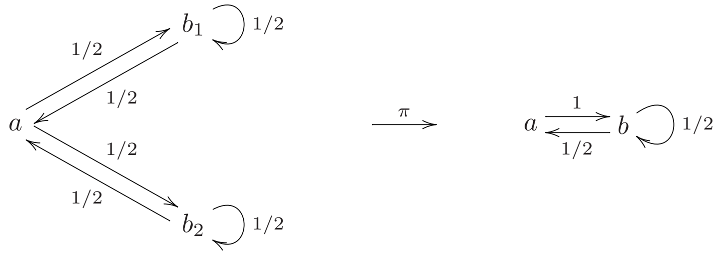
Problem 2.5. Given a hidden Markov measure ν on (Y, σ), how can one determine the degree of ν? If the degree is $n < \infty$, how can one construct Markov measures of which ν is the image under a degree n map?

We conclude this section with examples.

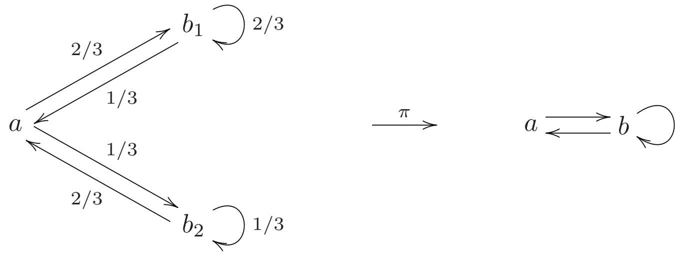
Example 2.6. An example was given in [69] of a code $\pi : X \to Y$ that is *non-Markovian*: some Markov measure on Y does not lift to any Markov measure on X, and hence (see Section 3.1) no Markov measure on Y has a Markov preimage on X. The following diagram presents a simpler example, due to Sujin Shin [91, 93], of such a map. Here π is a 1-block code: $\pi(1) = 1$ and $\pi(j) = 2$ if $j \neq 1$.



Example 2.7. Consider the shifts of finite type given by the graphs below, the 1-block code π given by the rule $\pi(a) = a, \pi(b_1) = \pi(b_2) = b$, and the Markov measures μ, ν defined by the transition probabilities shown on the edges. We have $\pi\mu = \nu$, so the code is *Markovian*—some Markov measure maps to a Markov measure.



Example 2.8. This example uses the same shifts of finite type and 1-block code as in Example 2.7, but we define a new 1-step Markov measure on the upstairs shift of finite type X by assigning transition probabilities as shown.



The entropy of the Markov measure μ (the definition is recalled in Sec. 3.2) is readily obtained from the familiar formula $-\sum p_i P_{ij} \log P_{ij}$, but there is no such simple rule for computing the entropy of ν. If ν were the finite-to-one image of some other Markov measure μ', maybe on some other shift of finite type, then we would have $h(\nu) = h(\mu')$ and the entropy of ν would be easily computed by applying the familiar formula to μ'. But for this example (due to Blackwell [13]) it can be shown [69] that ν is not the finite-to-one image of any Markov measure. Thus Problem 2.5 is relevant to the much-studied problem of estimating the entropy of a hidden Markov measure (see [44, 45] and their references).

Example 2.9. In this example presented in [97], $X = Y = \Sigma_2 =$ full 2-shift, and the factor map is the 2-block code

$$(\pi x)_0 = x_0 + x_1 \mod 2. \tag{2.14}$$

Suppose $0 < p < 1$ and μ_p is the Bernoulli (product) measure on X, with $\mu(\mathcal{C}_0(1)) = p$. Let ν_p denote the hidden Markov measure $\pi\mu_p = \pi\mu_{1-p}$. If $p \neq 1/2$, then ν_p is a hidden Markov measure strictly of degree 2 (it is not degree 1).

3. Factor maps and thermodynamical concepts

3.1. **Markovian and non-Markovian maps.** We have mentioned (Example 2.8) that the image under a factor map $\pi : X \to Y$ of a Markov measure need not be Markov, and (Example 2.6) that a Markov measure on Y need not have any Markov preimages. In this section we study maps that do not have the latter undesirable property. Recall our convention: a Markov measure is required to have full support.

Definition 3.1. [18] A factor map $\pi : \Omega_A \to \Omega_B$ between irreducible shifts of finite type (A and B are $0, 1$ transition matrices, see (2.8)) is *Markovian* if for every Markov measure ν on Ω_B, there is a Markov measure on Ω_A such that $\pi\mu = \nu$.

Theorem 3.2. [18] *For a factor map $\pi : \Omega_A \to \Omega_B$ between irreducible shifts of finite type, if there exist any fully supported Markov μ and ν with $\pi\mu = \nu$, then π is Markovian.*

Note that if a factor map is Markovian, then so too is every factor map which is topologically equivalent to it, because a topological conjugacy takes Markov measures to Markov measures. We will see a large supply of Markovian maps (the "e-resolving factor maps") in Section 6.1.

These considerations lead to a reformulation of Problem 2.4:

Problem 3.3. Give a procedure to decide, given a factor map $\pi : \Omega_A \to \Omega_B$, whether π is Markovian.

We sketch the proof of Theorem 3.2 for the 1-step Markov case: if any 1-step Markov measure on Ω_B lifts to a 1-step Markov measure, then every 1-step Markov measure on Ω_B lifts to a 1-step Markov measure. For this, recall that if M is an irreducible matrix with spectral radius ρ, with positive right eigenvector r, then the *stochasticization* of M is the stochastic matrix

$$\mathrm{stoch}(M) = \frac{1}{\rho} D^{-1} M D \ , \tag{3.1}$$

where D is the diagonal matrix with diagonal entries $D(i,i) = r(i)$.

Now suppose that $\pi : \Omega_A \to \Omega_B$ is a 1-block factor map, with $\pi(i)$ denoted $\overline{i}$ for all i in the alphabet of Ω_A; that μ, ν are 1-step Markov measures defined by stochastic matrices P, Q; and that $\pi\mu = \nu$. Suppose that $\nu' \in \mathcal{M}(\Omega_B)$ is defined by a stochastic matrix Q'. We will find a stochastic matrix P' defining μ' in $\mathcal{M}(\Omega_A)$ such that $\pi\mu' = \nu'$.

First define a matrix M of size matching P by $M(i,j) = 0$ if $P(i,j) = 0$ and otherwise

$$M(i,j) = Q'(\overline{i}, \overline{j}) P(i,j) / Q(\overline{i}, \overline{j}). \tag{3.2}$$

This matrix M will have spectral radius 1. Now set $P' = \mathrm{stoch}(M)$. The proof that $\pi\mu' = \nu'$ is a straightforward computation that $\pi\mu' = \nu'$ on cylinders $\mathcal{C}_0(y[0,n])$ for all $n \in \mathbb{N}$ and $y \in \Omega_B$. This construction is the germ of a more general thermodynamic result, the background for which we develop in the next section. We finish this section with an example.

Example 3.4. In this example one sees explicitly how being able to lift one Markov measure to a Markov measure, allows one to lift other Markov measures to Markov measures.

Consider the 1-block code π from $\Omega_3 = \{0,1,2\}^{\mathbb{Z}}$ to $\Omega_2 = \{0,1\}^{\mathbb{Z}}$, via $0 \mapsto 0$ and $1, 2 \mapsto 1$. Let ν be the 1-step Markov measure on Ω_2 given by the transition matrix

$$\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \ .$$

Given positive numbers $\alpha, \beta, \gamma < 1$, the stochastic matrix

$$\begin{pmatrix} 1/2 & \alpha(1/2) & (1-\alpha)(1/2) \\ 1/2 & \beta(1/2) & (1-\beta)(1/2) \\ 1/2 & \gamma(1/2) & (1-\gamma)(1/2) \end{pmatrix} \tag{3.3}$$

defines a 1-step Markov measure on Ω_3 which π sends to ν.

Now, if ν' is any other 1-step Markov measure on X_2, given by a stochastic matrix

$$\begin{pmatrix} p & q \\ r & s \end{pmatrix} \ ,$$

then ν' will lift to the 1-step Markov measure defined by the stochastic matrix

$$\begin{pmatrix} p & \alpha q & (1-\alpha)q \\ r & \beta s & (1-\beta)s \\ r & \gamma s & (1-\gamma)s \end{pmatrix} \ . \tag{3.4}$$

3.2. **Thermodynamics on subshifts 001.** We recall the definitions of entropy and pressure and how the thermodynamical approach provides convenient machinery for dealing with Markov measures (and hence eventually, it is hoped, with hidden Markov measures).

Let (X, σ) be a subshift and $\mu \in \mathcal{M}(X)$ a shift-invariant Borel probability measure on X. The *topological entropy* of (X, σ) is

$$h(X) = \lim_{n\to\infty} \frac{1}{n} \log |\{x[0, n-1] : x \in X\}|. \tag{3.5}$$

The *measure-theoretic entropy* of the measure-preserving system (X, σ, μ) is

$$h(\mu) = h_\mu(X) = \lim_{n\to\infty} \frac{1}{n} \sum \{-\mu(\mathcal{C}_0(w)) \log \mu(\mathcal{C}_0(w)) : w \in \{x[0, n-1] : x \in X\}\}. \tag{3.6}$$

(For more background on these concepts, one could consult [78, 96].)

Pressure is a refinement of entropy which takes into account not only the map $\sigma : X \to X$ but also weights coming from a given "potential function" f on X. Given a continuous real-valued function $f \in C(X, \mathbb{R})$, we define the *pressure of f (with respect to σ)* to be

$$P(f, \sigma) = \lim_{n\to\infty} \frac{1}{n} \log \sum \{\exp[S_n(f, w)] : w \in \{x[0, n-1] : x \in X\}\}, \tag{3.7}$$

where

$$S_n(f, w) = \sum_{i=0}^{n-1} f(\sigma^i x) \quad \text{for some } x \in X \quad \text{such that } x[0, n-1] = w. \tag{3.8}$$

(In the limit the choice of x doesn't matter.) Thus,

$$\text{if } f \equiv 0, \text{ then } P(f, \sigma) = h(X). \tag{3.9}$$

The pressure functional satisfies the important *Variational Principle*:

$$P(f, \sigma) = \sup\{h(\mu) + \int f \, d\mu : \mu \in \mathcal{M}(X)\}. \tag{3.10}$$

An *equilibrium state* for f (with respect to σ) is a measure $\mu = \mu_f$ such that

$$P(f, \sigma) = h(\mu) + \int f \, d\mu. \tag{3.11}$$

Often (e.g., when the potential function f is Hölder continuous on an irreducible shift of finite type), there is a unique equilibrium state μ_f which is a *(Bowen) Gibbs measure* for f: i.e., $P(f, \sigma) = \log(\rho)$, and

$$\mu_f(\mathcal{C}_0(x[0, n-1])) \sim \rho^{-n} \exp S_n f(x). \tag{3.12}$$

Here "$\sim$" means the ratio of the two sides is bounded above and away from zero, uniformly in x and n.

If $f \in C(\Omega_A, \mathbb{R})$, depends on only two coordinates, $f(x) = f(x_0 x_1)$ for all $x \in \Omega_A$, then f has a unique equilibrium state μ_f, and $\mu_f \in \mathcal{M}(\Omega_A)$. This measure μ_f is the 1-step Markov measure defined by the stochastic matrix $P = \mathrm{stoch}(Q)$, where

$$Q(i,j) = \begin{cases} 0 & \text{if } A(i,j) = 0, \\ \exp[f(ij)] & \text{otherwise} \quad . \end{cases} \tag{3.13}$$

(For an exposition see [73].)

The pressure of f is $\log \rho$, where ρ is the spectral radius of Q. Conversely, a Markov measure with stochastic transition matrix P is the equilibrium state of the potential function $f[ij] = \log P(i,j)$.

By passage to the k-block presentation, we can generalize to the case of k-step Markov measures: if $f(x) = f(x_0 x_1 \cdots x_k)$, then f has a unique equilibrium state μ, and μ is a k-step Markov measure.

Definition 3.5. We say that a function on a subshift X is *locally constant* if there is $m \in \mathbb{N}$ such that $f(x)$ depends only on $x[-m,m]$. $\mathrm{LC}(X,\mathbb{R})$ is the vector space of locally constant real-valued functions on X. $C_k(X,\mathbb{R})$ is the set of f in $\mathrm{LC}(X,\mathbb{R})$ such that $f(x)$ is determined by $x[0,k-1]$.

We can now express a viewpoint on Markov measures, due to Parry and Tuncel [95, 74], which follows from the previous results.

Theorem 3.6. [74] *Suppose Ω_A is an irreducible shift of finite type; $k \geq 1$; and $f, g \in C_k(X,\mathbb{R})$. Then the following are equivalent.*

(1) $\mu_f = \mu_g$.
(2) *There are $h \in C(X,\mathbb{R})$ and $c \in \mathbb{R}$ such that $f = g + (h - h\circ\sigma) + c$.*
(3) *There are $h \in C_{k-1}(X,\mathbb{R})$ and $c \in \mathbb{R}$ such that $f = g + (h - h\circ\sigma) + c$.*

Proposition 3.7. [74] *Suppose Ω_A is an irreducible shift of finite type. Let*

$$W = \{h - h\circ\sigma + c : h \in LC(\Omega_A,\mathbb{R}), c \in \mathbb{R}\} \ . \tag{3.14}$$

Then the rule $[f] \mapsto \mu_f$ defines maps

$$\begin{aligned} C_k(\Omega_A,\mathbb{R})/W &\rightarrow \mathcal{M}_k(\sigma_A) \\ LC(\Omega_A,\mathbb{R})/W &\rightarrow \cup_k \mathcal{M}_k(\sigma_A), \end{aligned}$$

and these maps are bijections.

3.3. **Compensation functions.** Let $\pi : (X,T) \to (Y,S)$ be a factor map between topological dynamical systems. A *compensation function* for the factor map is a continuous function $\xi : X \to \mathbb{R}$ such that

$$P_Y(V) = P_X(V\circ\pi + \xi) \quad \text{for all } V \in \mathcal{C}(Y,\mathbb{R}). \tag{3.15}$$

Because $h(\pi\mu) \leq h(\mu)$ and $\int V\, d(\pi\mu) = \int V\circ\pi\, d\mu$, we always have

$$P_Y(V) = \sup\{h(\nu) + \int_Y V\, d\nu : \nu \in \mathcal{M}(Y)\} \tag{3.16}$$

$$\leq \sup\{h(\mu) + \int_X V\circ\pi\, d\mu : \mu \in \mathcal{M}(X)\} = P_X(V\circ\pi), \tag{3.17}$$

with possible strict inequality when π is infinite-to-one, in which case a strict inequality $h(\mu) > h(\pi\mu)$ can arise from (informally) the extra information/complexity arising from motion in fibers over points of Y. The pressure equality (3.15) tells us that the addition of a compensation function ξ to the functions $V \circ \pi$ takes into account (and exactly cancels out), for all potential functions V on Y at once, this measure of extra complexity. Compensation functions were introduced in [18] and studied systematically in [97]. A compensation function is a kind of oracle for how entropy can appear in a fiber. The Markovian case is the case in which the oracle has finite range, that is, there is a locally constant compensation function.

A compensation function for a factor map $\pi : X \to Y$ is *saturated* if it has the form $G \circ \pi$ for a continuous function G on Y.

Example 3.8. For the factor map in Examples 2.7 and 2.8, the formula

$$G(y) = \begin{cases} -\log 2 & \text{if } y = .a\dots \\ 0 & \text{if } y = .b\dots \end{cases} \tag{3.18}$$

determines a saturated compensation function $G \circ \pi$ on Ω_A. The sum (or *cocycle*) $S_nG(y) = G(y) + G(\sigma y) + \cdots + G(\sigma^{n-1}y)$ measures the growth of the number of preimages of initial blocks of y:

$$|\pi^{-1}(y_0 \dots y_{n-1})| = 2^{\#\{i: y_i = a, 0 \le i < n\} \pm 1} \sim 2^{\#\{i: y_i = a, 0 \le i < n\}} = e^{-S_nG(y)}. \tag{3.19}$$

Example 3.9. In the situation described at the end of Section 3.1, in which a 1-step Markov measure maps to a 1-step Markov measure under a 1-block map, an associated compensation function is

$$\xi(x) = \log P(i,j) - \log Q(\overline{i}, \overline{j}) \quad \text{when } x_0x_1 = ij. \tag{3.20}$$

Theorem 3.10. [18, 97] *Suppose that $\pi : \Omega_A \to \Omega_B$ is a factor map between irreducible shifts of finite type, with $f \in LC(\Omega_A)$ and $g \in LC(\Omega_B)$, and $\pi\mu_f = \mu_g$. Then there is a constant c such that $f - g \circ \pi + c$ is a compensation function. Conversely, if ξ is a locally constant compensation function, then $\mu_{\xi + g\circ\pi}$ is Markov and $\pi\mu_{\xi+g\circ\pi} = \mu_g$.*

In Theorem 3.10, the locally constant compensation function ξ relates potential functions on Ω_B to their lifts by composition on Ω_A in the same way that the corresponding equilibrium states are related:

$$\begin{aligned} \mathrm{LC}(\Omega_B) \ &\hookrightarrow\ \mathrm{LC}(\Omega_A) \quad \text{via } g \to (g \circ \pi) + \xi \\ \mathcal{M}(\Omega_B) &\hookrightarrow \mathcal{M}(\Omega_A) \quad \text{via } \mu_g \to \mu_{(g\circ\pi)+\xi}. \end{aligned} \tag{3.21}$$

Theorem 3.10 holds if we replace the class of locally constant functions with the class of Hölder (exponentially decaying) functions, or with functions in the larger and more complicated "Walters class" (defined in [97, Section 4]). More generally, the arguments in [97, Theorem 4.1] go through to prove the following.

Theorem 3.11. *Suppose that $\pi : \Omega_A \to \Omega_B$ is a factor map between irreducible shifts of finite type. Let $\mathcal{V}_A, \mathcal{V}_B$ be real vector spaces of functions in $C(\Omega_A, \mathbb{R}), C(\Omega_B, \mathbb{R})$ respectively such that the following hold.*

(1) *$\mathcal{V}_A$ and $\mathcal{V}_B$ contain the locally constant functions.*
(2) *If f is in $\mathcal{V}_A$ or $\mathcal{V}_B$, then f has a unique equilibrium state μ_f, and μ_f is a Gibbs measure.*
(3) *If $f \in \mathcal{V}_B$, then $f \circ \pi \in \mathcal{V}_A$.*

Suppose $f \in \mathcal{V}_A$ and $g \in \mathcal{V}_B$, and $\pi\mu_f = \mu_g$. Then there is a constant C such that $f - g\circ\pi + C$ is a compensation function. Conversely, if ξ in $\mathcal{V}_A$ is a compensation function, then for all $g \in \mathcal{V}_B$ it holds that $\pi\mu_{\xi+g\circ\pi} = \mu_g$.

Moreover, if $G \in \mathcal{V}_B$, then $G \circ \pi$ is a compensation function if and only if there is $c \geq 1$ such that

$$\frac{1}{c} \leq e^{S_nG(y)}\,|\pi^{-1}(y_0 \dots y_{n-1})| \leq c \text{ for all } y, n. \tag{3.22}$$

Problem 3.12. Determine whether there exists a factor map $\pi : X \to Y$ between mixing SFT's and a potential function $F \in \mathcal{C}(X)$ which is *not* a compensation function but has a unique equilibrium state μ_F whose image $\pi\mu_F$ is the measure of maximal entropy on Y. If there were such an example, it would show that the assumptions on function classes in Theorem 3.11 cannot simply be dropped.

We finish this section with some more general statements about compensation functions for factor maps between shifts of finite type.

Proposition 3.13. [97] *Suppose that $\pi : \Omega_A \to \Omega_B$ is a factor map between irreducible shifts of finite type. Then*

(1) *There exists a compensation function.*
(2) *If ξ is a compensation function, $g \in \mathbb{C}(\Omega_B, \mathbb{R})$, and μ is an equilibrium state of $\xi + g \circ \pi$, then $\pi\mu$ is an equilibrium state of g.*
(3) *The map π takes the measure of maximal entropy (see Section 3.5) of Ω_A to that of Ω_B if and only if there is a* constant *compensation function.*

Yuki Yayama [99] has begun the study of compensation functions which are bounded Borel functions.

3.4. **Relative pressure.** When studying factor maps, relativized versions of entropy and pressure are relevant concepts. Given a factor map $\pi : \Omega_A \to \Omega_B$ between shifts of finite type, for each $n = 1, 2, \cdots$ and $y \in Y$, let $D_n(y)$ be a set consisting of exactly one point from each nonempty set $[x_0 \cdots x_{n-1}] \cap \pi^{-1}(y)$. Let $V \in \mathcal{C}(\Omega_A, \mathbb{R})$ be a potential function on Ω_A. For each $y \in \Omega_B$, the *relative pressure of V at y with respect to π* is defined to be

$$P(\pi, V)(y) = \limsup_{n\to\infty} \frac{1}{n} \log \Big[\sum_{x \in D_n(y)} \exp\Big(\sum_{i=0}^{n-1} V(\sigma^i x) \Big) \Big]. \tag{3.23}$$

The *relative topological entropy function* is defined for all $y \in Y$ by

$$P(\pi, 0)(y) = \limsup_{n\to\infty} \frac{1}{n} \log \big| D_n(y) \big|, \tag{3.24}$$

the relative pressure of the potential function $V \equiv 0$.

For the relative pressure function, a *Relative Variational Principle* was proved by Ledrappier and Walters ([64], see also [30]): for all ν in $\mathcal{M}(\Omega_B)$ and all V in $C(\Omega_A)$,

$$\int P(\pi, V)\, d\nu = \sup\Big\{h(\mu) + \int V d\mu : \pi\mu = \nu\Big\} - h(\nu). \tag{3.25}$$

In particular, for a fixed $\nu \in \mathcal{M}(\Omega_B)$, the maximum measure-theoretic entropy of a measure on Ω_A that maps under π to ν is given by

$$\begin{aligned} h(\nu) + \sup\{h_\mu(X|Y) : \pi\mu = \nu\} &= h(\nu) + \sup\{h(\mu) - h(\nu) : \pi\mu = \nu\} \\ &= h(\nu) + \int_Y P(\pi, 0)\, d\nu \ . \end{aligned} \tag{3.26}$$

In [80] a finite-range, combinatorial approach was developed for the relative pressure and entropy, in which instead of examining entire infinite sequences x in each fiber over a given point $y \in \Omega_B$, it is enough to deal just with preimages of finite blocks (which may or may not be extendable to full sequences in the fiber). For each $n = 1, 2, \dots$ and $y \in Y$ let $E_n(y)$ be a set consisting of exactly one point from each nonempty cylinder $x[0, n-1] \subset \pi^{-1}y[0, n-1]$. Then for each $V \in C(\Omega_A)$,

$$P(\pi, V)(y) = \limsup_{n\to\infty} \frac{1}{n} \log \Big[\sum_{x \in E_n(y)} \exp\Big(\sum_{i=0}^{n-1} V(\sigma^i x)\Big)\Big] \tag{3.27}$$

a.e. with respect to every ergodic invariant measure on Y. Thus, we obtain the value of $P(\pi, V)(y)$ a.e. with respect to every ergodic invariant measure on Y if we delete from the definition of $D_n(y)$ the requirement that $x \in \pi^{-1}(y)$.

In particular, the relative topological entropy is given by

$$P(\pi, 0)(y) = \limsup_{n\to\infty} \frac{1}{n} \log |\pi^{-1} y[0, n-1]| \tag{3.28}$$

a.e. with respect to every ergodic invariant measure on Y.

And if μ is relatively maximal over ν, in the sense that it achieves the supremum in (3.26), then

$$h_\mu(X|Y) = \int_Y \lim_{n\to\infty} \frac{1}{n} \log |\pi^{-1} y[0, n-1]|\, d\nu(y). \tag{3.29}$$

3.5. **Measures of maximal and relatively maximal entropy.** Already Shannon [90] constructed the measures of maximal entropy on irreducible shifts of finite type. Parry [72] independently and from the dynamical viewpoint rediscovered the construction and proved uniqueness. For an irreducible shift of finite type the unique measure of maximal entropy is a 1-step Markov measure whose transition probability matrix is the stochasticization, as in (3.1), of the 0, 1 matrix that defines the subshift. When studying factor maps $\pi : \Omega_A \to \Omega_B$ it is natural to look for *measures of maximal relative entropy*, which we also call *relatively maximal measures* : for fixed ν on Ω_B, look for the $\mu \in \pi^{-1}\nu$ which have maximal entropy in that fiber. Such measures always exist by compactness and upper semicontinuity, but, in contrast to the Shannon-Parry case (when Ω_B consists of a single point), they need not be unique. E.g., in Example 2.9, the two-to-one map π respects entropy, and for $p \neq 1/2$ there are exactly two ergodic measures (the Bernoulli measures μ_p and μ_{1-p}) which π sends to ν_p. Moreover, there exists some $V_p \in \mathcal{C}(Y)$ which has ν_p as a unique equilibrium state [52, 81], and $V_p \circ \pi$ has exactly two ergodic equilibrium states, μ_p and μ_{1-p}.

Here is a useful characterization of relatively maximal measures due to Shin.

Theorem 3.14 [92]**.** *Suppose that $\pi : X \to Y$ is a factor map of shifts of finite type, $\nu \in \mathcal{M}(Y)$ is ergodic, and $\pi\mu = \nu$. Then μ is relatively maximal over ν if and only if there is $V \in \mathcal{C}(Y, \mathbb{R})$ such that μ is an equilibrium state of $V \circ \pi$.*

If there is a *locally constant* saturated compensation function $G \circ \pi$, then every Markov measure on Y has a unique relatively maximal lift, which is Markov, because then the relatively maximal measures over an equilibrium state of $V \in \mathcal{C}(Y, \mathbb{R})$ are the equilibrium states of $V \circ \pi + G \circ \pi$ [97]. Further, the measure of maximal entropy $\max_X$ is the unique equilibrium state of the potential function 0 on X; and the relatively maximal measures over $\max_Y$ are the equilibrium states of $G \circ \pi$.

It was proved in [79] that for each ergodic ν on Y, there are only a finite number of relatively maximal measures over ν. In fact, for a 1-block factor map π between 1-step shifts of finite type X, Y, the number of ergodic invariant measures of maximal entropy in the fiber $\pi^{-1}\{\nu\}$ is at most

$$N_\nu(\pi) = \min\{|\pi^{-1}\{b\}| : b \in \mathcal{A}(Y), \nu[b] > 0\}. \tag{3.30}$$

This follows from the theorem in [79] that for each ergodic ν on Y, any two distinct ergodic measures on X of maximal entropy in the fiber $\pi^{-1}\{\nu\}$ are *relatively orthogonal*. This concept is defined as follows.

For $\mu_1, \dots, \mu_n \in \mathcal{M}(X)$ with $\pi\mu_i = \nu$ for all i, their *relatively independent joining* $\hat{\mu}$ over ν is defined by:

if $A_1, \dots, A_n$ are measurable subsets of X and $\mathcal{F}$ is the σ-algebra of Y, then

$$\hat{\mu}(A_1 \times \ldots \times A_n) = \int_Y \prod_{i=1}^{n} \mathbb{E}_{\mu_i}(\mathbf{1}_{A_i} | \pi^{-1}\mathcal{F}) \circ \pi^{-1} \, d\nu \tag{3.31}$$

in which $\mathbb{E}$ denotes conditional expectation. Two ergodic measures μ_1, μ_2 with $\pi\mu_1 = \pi\mu_2 = \nu$ are *relatively orthogonal* (over ν), $\mu_1 \perp_\nu \mu_2$, if

$$(\mu_1 \otimes_\nu \mu_2)\{(u, v) \in X \times X : u_0 = v_0\} = 0. \tag{3.32}$$

This means that with respect to the relatively independent joining or coupling, there is zero probability of coincidence of symbols in the two coordinates.

That the second theorem (distinct ergodic relatively maximal measures in the same fiber are relatively orthogonal) implies the first (no more than $N_\nu(\pi)$ relatively maximal measures over ν) follows from the Pigeonhole Principle. If we have $n > N_\nu(\pi)$ ergodic measures $\mu_1, \dots, \mu_n$ on X, each projecting to ν and each of maximal entropy in the fiber $\pi^{-1}\{\nu\}$, we form the relatively independent joining $\hat{\mu}$ on X^n of the measures μ_i as above. Write p_i for the projection $X^n \to X$ onto the i'th coordinate. For $\hat{\mu}$-almost every $\hat{x}$ in X^n, $\pi(p_i(\hat{x}))$ is independent of i; abusing notation for simplicity, denote it by $\pi(\hat{x})$. Let b be a symbol in the alphabet of Y such that b has $N_\nu(\pi)$ preimages $a_1, \dots, a_{N_\nu(\pi)}$ under the block map π. Since $n > N_\nu(\pi)$, for every $\hat{x} \in \pi^{-1}[b]$ there are $i \neq j$ with $(p_i\hat{x})_0 = (p_j\hat{x})_0$. At least one of the sets $S_{i,j} = \{\hat{x} \in X^n : (p_i\hat{x})_0 = (p_j\hat{x})_0\}$ must have positive $\hat{\mu}$-measure, and then also

$$(\mu_i \otimes_\nu \mu_j)\{(u,v) \in X \times X : \pi u = \pi v, u_0 = v_0\} > 0, \tag{3.33}$$

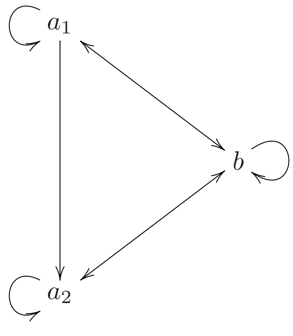
contradicting relative orthogonality. (Briefly, if you have more measures than preimage symbols, two of those measures have to coincide on one of the symbols: with respect to each measure, that symbol a.s. appears infinitely many times in the same place.)

The second theorem is proved by "interleaving" measures to increase entropy. If there are two relatively maximal measures over ν which are not relatively orthogonal, then the measures can be 'mixed' to give a measure with greater entropy. We concatenate words from the two processes, using the fact that the two measures are supported on sequences that agree infinitely often. Since X is a 1-step SFT, we can switch over whenever a coincidence occurs. That the switching increases entropy is seen by using the strict concavity of the function $-t \log t$ and lots of calculations with conditional expectations.

Example 3.15. Here is an example (also discussed in [79, Example 1]) showing that to find relatively maximal measures over a Markov measure it is not enough to consider only sofic measures which map to it. We describe a factor map π which is both left and right e-resolving (see section 6.1) and such that there is a unique relatively maximal measure μ above any fully-supported Markov measure ν, but the measure μ is not Markov, and it is not even sofic.

We use vertex shifts of finite type. The alphabet for the domain subshift is $\{a_1, a_2, b\}$ (in that order for indexing purposes), and the factor map (onto the 2-shift (Ω_2, σ)) is the 1-block code π which erases subscripts. The transition diagram and matrix A for the domain shift of finite type (Ω_A, σ) are

(3.34)



$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Above the word ba^nb in Ω_2 there are $n+1$ words in Ω_A: above a^n we see k a_1's followed by $n-k$ a_2's, where $0 \le k \le n$. Let us for simplicity consider the maximal measure ν on (Ω_2, T); so, $\nu(\mathcal{C}_0(ba^nb)) = 2^{-n-2}$. Now the maximal entropy lift μ of ν will assign equal measure $2^{-(n+2)}/(n+1)$ to each of the preimage blocks of ba^nb. If μ is sofic, then (as in Sec. 4.1.4) there are vectors u, v and a square matrix Q such that $\mu(\mathcal{C}_0(b(a_1)^nb) = uQ^nv$ for all $n > 0$. Then the function $n \mapsto uQ^nv$ is some finite sum of terms of the form $rn^j(\lambda^n)$ where $j \in \mathbb{Z}_+$ and r, λ are constants. The function $n \mapsto 2^{-(n+2)}/(n+1)$ is not a function of this type.

Problem 3.16. Is it true that for every factor map $\pi : \Omega_A \to \Omega_B$ every (fully supported) Markov measure ν on Ω_B has a unique relatively maximal measure that maps to it, and this is also a measure with full support?

Remark 3.17. After the original version of this paper was posted on the Math Arxiv and submitted for review, we received the preprint [100] of Jisang Yoo containing the following result: "Given a factor map from an irreducible SFT X to a sofic shift Y and an invariant measure ν on Y with full support, every measure on X of maximal relative entropy over ν is fully supported." This solves half of Problem 3.16.

3.6. **Finite-to-one codes.** Suppose $\pi : \Omega_A \to \Omega_B$ is a finite-to-one factor map of irreducible shifts of finite type. There are some special features of this case which we collect here for mention. Without loss of generality, after recoding we assume that π is a 1-block code. Given a Markov measure μ and a periodic point x we define the *weight-per-symbol* of x (with respect to μ) to be

$$\mathrm{wps}_\mu(x) := \lim_{n\to\infty} \frac{1}{n} \log \mu\{y : x_i = y_i, 0 \le i < n\} \ . \tag{3.35}$$

Proposition 3.18. *Suppose $\pi : \Omega_A \to \Omega_B$ is a finite-to-one factor map of irreducible shifts of finite type. Then*

(1) *The measure of maximal entropy on Ω_B lifts to the measure of maximal entropy on Ω_A.*
(2) *Every Markov measure on Ω_B lifts to a unique Markov measure of equal order on Ω_A.*
(3) *If μ, ν are Markov measures on Ω_A, Ω_B respectively, then the following are equivalent:*
   (a) $\pi\mu = \nu$
   (b) *for every periodic point x in Ω_A,* $\mathrm{wps}_\mu(x) = \mathrm{wps}_\nu(\pi x)$.

Proofs can be found in, for example, [56]. For infinite-to-one codes, we do not know an analogue of Prop. 3.18 (3).

3.7. **The semigroup measures of Kitchens and Tuncel.** There is a hierarchy of sofic measures according to their sofic degree. Among the degree-1 sofic measures, there is a distinguished and very well behaved subclass, properly containing the Markov measures. These are the *semigroup measures* introduced and studied by Kitchens and Tuncel in their memoir [57]. Roughly speaking, semigroup measures are to Markov measures as sofic subshifts are to SFT's.

A sofic subshift can be presented by a semigroup [98, 57]. Associated to this are nonnegative transition matrices R_0, L_0. A semigroup measure (for the semigroup presentation) is defined by a state probability vector and a pair of stochastic matrices R, L with 0/+ pattern matching R_0, L_0 and satisfying certain consistency conditions. These matrices can be multiplied to compute measures of cylinders. A measure is a semigroup measure if there exist a semigroup and apparatus as above which can present it. We will not review this constructive part of the theory, but we mention some alternate characterizations of these measures.

For a sofic measure μ on X and a periodic point x in X, the weight-per-symbol of x with respect to μ is still well defined by (3.35). Let us say a factor map π *respects* μ*-weights* if whenever x, y are periodic points with the same image we have $\mathrm{wps}_\mu(x) = \mathrm{wps}_\mu(y)$. Given a word $U = U[-n \dots 0]$ and a measure μ, let μ_U denote the conditional measure on the future, i.e. if UW is an allowed word then $\mu_U(W) = \mu(UW)/\mu(U)$.

Theorem 3.19. [57] *Let ν be a shift-invariant measure on an irreducible sofic subshift Y. Then the following are equivalent:*

(1) *ν is a semigroup measure.*
(2) *ν is the image of a Markov measure μ under a finite-to-one factor map which respects μ-weights.*
(3) *ν is the image of a Markov measure μ under a degree 1 resolving factor map which respects μ-weights.*
(4) *The collection of conditional measures μ_U, as U ranges over all Y-words, is finite.*

There is also a thermodynamic characterization of these measures as unique equilibrium states of bounded Borel functions which are locally constant on doubly transitive points, very analogous to the characterization of Markov measures as unique equilibrium states of continuous locally constant functions. The semigroup measures satisfy other nice properties as well.

Theorem 3.20. [57] *Suppose $\pi : X \to Y$ is a finite-to-one factor map of irreducible sofic subshifts and μ and ν are semigroup measures on X and Y respectively. Then*

(1) *ν lifts by π to a unique semigroup measure on X, and this is the unique ergodic measure on X which maps to ν;*
(2) *$\pi\mu$ is a semigroup measure if and only if π respects μ-weights;*
(3) *there is an irreducible sofic subshift X' of X such that π maps X' finite-to-one onto X* [69]*, and therefore ν lifts to a semigroup measure on X'.*

In contrast to the last statement, it can happen for an infinite-to-one factor map between irreducible SFTs that there is a Markov measure on the range which cannot lift to a Markov measure on any subshift of the domain [69].

We finish here with an example. There are others in [57].

Example 3.21. This is an example of a finite-to-one, one-to-one a.e. 1-block code $\pi : \Omega_A \to \Omega_B$ between mixing vertex shifts of finite type, with a 1-step Markov measure μ on Ω_A, such that the following hold:

(1) For all periodic points x, y in Ω_A, $\pi x = \pi y$ implies that $\mathrm{wps}_\mu(x) = \mathrm{wps}_\mu(y)$.
(2) $\pi\mu$ is not Markov on Ω_B.

Here the alphabet of Ω_A is $\{1, 2, 3\}$; the alphabet of Ω_B is $\{1, 2\}$;

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

and π is the 1-block code sending 1 to 1 and sending 2 and 3 to 2. The map π collapses the points in the orbit of $(23)^*$ to a fixed point and collapses no other periodic points. (Given a block B, we let B^* denote a periodic point obtained by infinite concatenation of the block B.)

Let f be the function on Ω_A such that $f(x) = \log 2$ if $x_0x_1 = 23$, $f(x) = \log(1/2)$ if $x_0x_1 = 32$ and $f(x) = 0$ otherwise. Let μ be the 1-step Markov measure which is the unique equilibrium state for f, defined by the stochasticization P of the matrix

$$M = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 2 \\ 1 & 1/2 & 0 \end{pmatrix}.$$

Let λ denote the spectral radius of M. Suppose that $\nu = \pi\mu$ is Markov, of any order. Then $\mathrm{wps}_\nu(2^*) = \mathrm{wps}_\mu((23)^*) = -\log\lambda$. Also, there must be a constant c such that for all large n,

$$\mathrm{wps}_\nu((12^n)^*) = \frac{1}{n+1}(c + (n+1)\mathrm{wps}_\nu(2^*)) = \frac{c}{n+1} - \log\lambda\ . \tag{3.36}$$

So, for all large n,
(3.37)

$$\frac{c}{2n+1} - \log\lambda = \mathrm{wps}_\nu((12^{2n})^*) = \mathrm{wps}_\mu((1(23)^n)^*) = \frac{1}{2n+1}\log(2\lambda^{-(2n+1)})$$

and

$$\frac{c}{2n+2} - \log\lambda = \mathrm{wps}_\nu((12^{2n+1})^*) = \mathrm{wps}_\mu((1(23)^n2)^*) = \frac{1}{2n+2}\log(\lambda^{-(2n+2)})\ .$$

Thus $c = \log 2$ and $c = 0$, a contradiction. Therefore $\pi\mu$ is not Markov.

4. Identification of hidden Markov measures

Given a finite-state stationary process, how can we tell whether it is a hidden Markov process? If it is, how can we construct some Markov process of which it is a factor by means of a sliding block code? When is the image of a Markov measure under a factor map again a Markov measure? These questions are of practical importance, since scientific measurements often capture only partial information about systems under study, and in order to construct useful models the significant hidden variables must be identified and included. Beginning in the 1960's some criteria were developed for recognizing a hidden Markov process: loosely speaking, an abstract algebraic object constructed from knowing the measures of cylinder sets should be in some sense finitely generated. Theorem 4.20 below gives equivalent conditions, in terms of formal languages and series (the series is "rational"), linear algebra (the measure is "linearly representable"), and abstract algebra (some module is finitely generated), that a shift-invariant probability measure be the image under a 1-block map of a shift-invariant 1-step Markov measure. In the following we briefly explain this result, including the terminology involved.

Kleene [59] characterized rational languages as the linearly representable ones, and this was generalized to formal series by Schützenberger [89]. In the study of stochastic processes, functions of Markov chains were analyzed by Gilbert [40], Furstenberg [39], Dharmadhikari [23, 24, 25, 26, 27, 28], Heller [48, 49], and others. For the connection between rational series and continuous images of Markov chains, we follow Berstel-Reutenauer [9] and Hansel-Perrin [46], with an addition to explain how to handle zero entries. Subsequent sections describe the approaches of Furstenberg and Heller and related work.

Various problems around these ideas were (and continue to be) explored and solved. In particular, it is natural to ask when is the image of a Markov measure μ under a continuous factor map π a Gibbs measure (see (3.12), or when is the image of a Gibbs measure again a Gibbs measure? Chazottes and Ugalde [21] showed that if μ is k-step Markov on a full shift Ω_d and π maps Ω_d onto another full shift Ω_D, then the image $\pi\mu$ is a Gibbs measure which is the unique equilibrium state of a Hölder continuous potential which can be explicitly described in terms of a limit of matrix products and computed at periodic points. They also gave sufficient conditions in the more general case when the factor map is between SFT's. The case when μ is Gibbs but not necessarily Markov is considered in [22]. For higher-dimensional versions see for example [63, 68, 43].

Among the extensive literature that we do not cite elsewhere, we can mention in addition [47, 70, 35, 10, 88].

4.1. **Formal series and formal languages.**

4.1.1. *Basic definitions.* As in Section 2.1, continue to let $\mathcal{A}$ be a finite alphabet, $\mathcal{A}^*$ the set of all finite words on $\mathcal{A}$, and $\mathcal{A}^+$ the set of all finite nonempty words on $\mathcal{A}$. Let ϵ denote the empty word. A *language* on $\mathcal{A}$ is any subset $\mathcal{L} \subset \mathcal{A}^*$.

Recall that a *monoid* is a set S with a binary operation $S \times S \to S$ which is associative and has a neutral element (identity). This means we can think of $\mathcal{A}^*$ as the multiplicative free monoid generated by $\mathcal{A}$, where the operation is concatenation and the neutral element is ϵ.

A *formal series* (nonnegative real-valued, based on $\mathcal{A}$) is a function $s : \mathcal{A}^* \to \mathbb{R}_+$. For all $w \in \mathcal{A}^*$, $s(w) = (s, w) \in \mathbb{R}_+$, which can be thought of as the coefficient of w in the series s. We will think of this s as $\sum_{w\in\mathcal{A}^*} s(w)w$, and this will be justified later. If $v \in \mathcal{A}^*$ and s is the series such that $s(v) = 1$ and $s(w) = 0$ otherwise, then we sometimes use simply v to denote s.

Associated with any language $\mathcal{L}$ on $\mathcal{A}$ is its *characteristic series* $F_{\mathcal{L}} : \mathcal{A}^* \to \mathbb{R}_+$ which assigns 1 to each word in $\mathcal{L}$ and 0 to each word in $\mathcal{A}^* \setminus \mathcal{L}$. Associated to any Borel measure μ on $\mathcal{A}^{\mathbb{Z}_+}$ is its *corresponding series* F_μ defined by

$$F_\mu(w) = \mu(\mathcal{C}_0(w)) = \mu\{x \in \mathcal{A}^{\mathbb{Z}_+} : x[0, |w| - 1] = w\}. \tag{4.1}$$

It is sometimes useful to consider formal series with values in any *semiring* K, which is just a ring without subtraction. That is, K is a set with operations $+$ and $\cdot$ such that $(K, +)$ is a commutative monoid with identity element 0, $(K, \cdot)$ is a monoid with identity element 1; the product distributes over the sum; and for $k \in K$, $0k = k0 = 0$.

We denote the set of all K-valued formal series based on $\mathcal{A}$ by $K\langle\langle \mathcal{A} \rangle\rangle$ or $\mathcal{F}_K(\mathcal{A})$. We further abbreviate $\mathbb{R}_+\langle\langle \mathcal{A} \rangle\rangle = \mathcal{F}(\mathcal{A})$.

Then $\mathcal{F}(\mathcal{A})$ is a semiring in a natural way: For $f_1, f_2 \in \mathcal{F}(\mathcal{A})$, define

(1) $(f_1 + f_2)(w) = f_1(w) + f_2(w)$
(2) $(f_1 f_2)(w) = \sum f_1(u) f_2(v)$, where the sum is over all $u, v \in \mathcal{A}^*$ such that $uv = w$, a finite sum.

The neutral element for multiplication in $\mathcal{F}(\mathcal{A})$ is

$$s_1(w) = \begin{cases} 1 & \text{if } w = \epsilon \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

As discussed above, we will usually write simply ϵ for s_1. There is a natural injection $\mathbb{R}_+ \hookrightarrow \mathcal{F}(\mathcal{A})$ defined by $t \mapsto t\epsilon$ for all $t \in \mathbb{R}_+$.

Note that:

- $\mathbb{R}_+$ acts on $\mathcal{F}(\mathcal{A})$ on both sides:
  $(ts)(w) = ts(w),\ (st)(w) = s(w)t,$ for all $w \in \mathcal{A}^*$, for all $t \in \mathbb{R}_+$.
- There is a natural injection $\mathcal{A}^* \hookrightarrow \mathcal{F}(\mathcal{A})$ as a multiplicative submonoid: For $w \in \mathcal{A}^*$ and $v \in \mathcal{A}^*$, define
  $$w(v) = \delta_{wv} = \begin{cases} 1 & \text{if } w = v \\ 0 & \text{otherwise.} \end{cases}$$
  This is a 1-term series.

Definition 4.1. The *support* of a formal series $s \in \mathcal{F}(\mathcal{A})$ is

$$\operatorname{supp}(s) = \{w \in \mathcal{A}^* : s(w) \neq 0\}.$$

Note that $\operatorname{supp}(s)$ is a language. A language corresponds to a series with coefficients 0 and 1, namely its characteristic series.

Definition 4.2. A *polynomial* is an element of $\mathcal{F}(\mathcal{A})$ whose support is a finite subset of $\mathcal{A}^*$. Denote the K-valued polynomials based on $\mathcal{A}$ by $\wp_K(\mathcal{A}) = K\langle\mathcal{A}\rangle$. The *degree* of a polynomial p is $\deg(p) = \max\{|w| : p(w) \neq 0\}$ and is $-\infty$ if $p \equiv 0$.

Definition 4.3. A family $\{f_\lambda : \lambda \in \Lambda\} \subset \mathcal{F}(\mathcal{A})$ of series is called *locally finite* if for all $w \in \mathcal{A}^*$ there are only finitely many $\lambda \in \Lambda$ for which $f_\lambda(w) \neq 0$. A series $f \in \mathcal{F}(\mathcal{A})$ is called *proper* if $f(\epsilon) = 0$.

Proposition 4.4. *If $f \in \mathcal{F}(\mathcal{A})$ is proper, then $\{f^n : n = 0, 1, 2, \dots\}$ is locally finite.*

Proof. If $n > |w|$, then $f^n(w) = 0$, because

$$f^n(w) = \sum_{\substack{u_1 \dots u_n = w \\ u_i \in \mathcal{A}^*,\ i=1,\dots,n}} f(u_1) \dots f(u_n)$$

and at least one u_i is ϵ. □

Definition 4.5. If $f \in \mathcal{F}(\mathcal{A})$ is proper, define

$$f^* = \sum_{n=0}^{\infty} f^n \text{ and } f^+ = \sum_{n=1}^{\infty} f^n \text{ (a pointwise finite sum),}$$

with $f^0 = 1 = 1 \cdot \epsilon = \epsilon$.

4.1.2. *Rational series and languages.*

Definition 4.6. The *rational operations* in $\mathcal{F}(\mathcal{A})$ are sum $(+)$, product $(\cdot)$, multiplication by real numbers (tw), and $* : f \to f^*$. The family of *rational series* consists of those $f \in \mathcal{F}(\mathcal{A})$ that can be obtained by starting with a finite set of polynomials in $\mathcal{F}(\mathcal{A})$ and applying a finite number of rational operations.

Definition 4.7. A language $\mathcal{L} \subset \mathcal{A}^*$ is *rational* if and only if its characteristic series

$$F(w) = \begin{cases} 1 & \text{if } w \in \mathcal{L} \\ 0 & \text{if } w \notin \mathcal{L} \end{cases} \tag{4.3}$$

is rational.

Recall that regular languages correspond to *regular expressions*: The set of regular expressions includes $\mathcal{A}$, ϵ, $\emptyset$ and is closed under $+$, $\cdot$, $*$. A language recognizable by a finite-state automaton, or consisting of words obtained by reading off sequences of edge labels on a finite labeled directed graph, is regular.

Proposition 4.8. *A language $\mathcal{L}$ is rational if and only if it is* regular. *Thus a nonempty insertive and extractive language is rational if and only if it is the language of a sofic subshift.*

4.1.3. *Distance and topology in $\mathcal{F}(\mathcal{A})$.* If $f_1, f_2 \in \mathcal{F}(\mathcal{A})$, define

$$D(f_1, f_2) = \inf\{n \geq 0 : \text{ there is } w \in \mathcal{A}^n \text{ such that } f_1(w) \neq f_2(w)\} \tag{4.4}$$

and

$$d(f_1, f_2) = \frac{1}{2^{D(f_1,f_2)}}. \tag{4.5}$$

Note that $d(f_1, f_2)$ defines an *ultrametric* on $\mathcal{F}(\mathcal{A})$:

$$d(f, h) \leq \max\{d(f, g), d(g, h)\} \leq d(f, g) + d(g, h). \tag{4.6}$$

With respect to the metric d, $f_k \to f$ if and only if for each $w \in \mathcal{A}^*$, $f_k(w) \to f(w)$ in the discrete topology on $\mathbb{R}$, i.e. $f_k(w)$ eventually equals $f(w)$.

Proposition 4.9. *$\mathcal{F}(\mathcal{A})$ is* complete *with respect to the metric d and is a* topological semiring *with respect to the metric d (that is, $+$ and $\cdot$ are continuous as functions of two variables).*

Definition 4.10. A family $\{F_\lambda : \lambda \in \Lambda\}$ of formal series is called *summable* if there is a series $F \in \mathcal{F}(\mathcal{A})$ such that for every $\delta > 0$ there is a finite set $\Lambda_\delta \subset \Lambda$ such that for each finite set $I \subset \Lambda$ with $\Lambda_\delta \subset I$, $d(\sum_{i \in I} F_i, F) < \delta$. Then F is called the *sum* of the series and we write $F = \sum_{\lambda \in \Lambda} F_\lambda$.

Proposition 4.11. *If $\{F_\lambda : \lambda \in \Lambda\}$ is locally finite, then it is summable, and conversely.*

Thus any $F \in \mathcal{F}(\mathcal{A})$ can be written as $F = \sum_{w \in \mathcal{A}^*} F(w)w$, where the formal series is a convergent infinite series of polynomials in the metric of $\mathcal{F}(\mathcal{A})$. Recall that

$$(F(w)w)(v) = \begin{cases} F(w) & \text{if } w = v \\ 0 & \text{if } w \neq v, \end{cases}$$

where $F(w)w \in \mathcal{F}(\mathcal{A})$ and $w \in \mathcal{A}^*$, so that $\{F(w)w : w \in \mathcal{A}^*\}$ is a locally finite, and hence summable, subfamily of $\mathcal{F}(\mathcal{A})$.

We note here that the set $\wp(\mathcal{A})$ of all polynomials is dense in $\mathcal{F}(\mathcal{A})$.

4.1.4. *Recognizable (linearly representable) series.*

Definition 4.12. $F \in \mathcal{F}(\mathcal{A})$ is *linearly representable* if there exists an $n \geq 1$ (the *dimension* of the representation) such that there are a $1 \times n$ nonnegative row vector $x \in \mathbb{R}_+^n$, an $n \times 1$ nonnegative column vector $y \in \mathbb{R}_+^n$, and a morphism of multiplicative monoids $\phi : \mathcal{A}^* \to \mathbb{R}_+^{n \times n}$ (the multiplicative monoid of nonnegative $n \times n$ matrices) such that for all $w \in \mathcal{A}^*$, $F(w) = x\phi(w)y$ (matrix multiplication). A *linearly representable measure* is one whose associated series is linearly representable. The triple (x, ϕ, y) is called the *linear representation* of the series (or measure).

Example 4.13. Consider a Bernoulli measure $\mathcal{B}(p_0, p_1, \dots, p_{d-1})$ on $\Omega_+(\mathcal{A}) = \mathcal{A}^{\mathbb{Z}_+}$ where $\mathcal{A} = \{a_0, a_1, \dots, a_{d-1}\}$, and $p = (p_0, p_1, \dots, p_{d-1})$ is a probability vector. Let $f = \sum_{i=0}^{d-1} p_i a_i \in \mathcal{F}(\mathcal{A})$. Then

$$f(w) = \begin{cases} p_i & \text{if } w = a_i \\ 0 & \text{if } w \neq a_i. \end{cases}$$

Define $F_p = f^* = \sum_{n \geq 0} f^n$. Note that f is proper since we have $f(\epsilon) = 0$. Consider the particular word $w = a_2 a_0$. Then $f^0(w) = f(w) = 0$, and for $n \geq 3$, we have $f^n(w) = 0$ because any factorization $w = u_1 u_2 u_3$ includes ϵ and $f(\epsilon) = 0$. Thus $F_p(w) = f^*(w) = f^2(w) = \sum_{uv=w} f(u)f(v) = f(a_2)f(a_0) = p_2 p_0$. Continuing in this way, we see that for $w_i \in \mathcal{A}$, $F_p(w_1 w_2 \dots w_n) = p_{w_1} p_{w_2} \dots p_{w_n}$.

Example 4.14. Consider a Markov measure μ on $\Omega_+(\mathcal{A})$ defined by a $d \times d$ stochastic matrix P and a d-dimensional probability row vector $p = (p_0, p_1, \cdots, p_{d-1})$. Define $F_{p,P} \in \mathcal{F}(\mathcal{A})$ by $F_{p,P}(w_1 \dots w_n) = \mu(\mathcal{C}_0(w_1 \dots w_n))$ for all $w_1, \dots, w_n \in \mathcal{A}$. Put $y = (1, \dots, 1)^{\text{tr}} \in \mathbb{R}^d_+$, $x = p \in \mathbb{R}^d_+$, and let ϕ be generated by $\phi(a_j)$, $j = 0, 1, ..., d-1$, where

$$\phi(a_j) = \begin{pmatrix} 0 & \cdots & P_{0j} & 0 & \cdots & 0 \\ 0 & \cdots & P_{1j} & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & \cdots & P_{d-1,j} & 0 & \cdots & 0 \end{pmatrix} \quad \text{for each } a_j \in \mathcal{A}. \tag{4.7}$$

Then the triple (x, ϕ, y) represents the given Markov measure μ. In this Markov case each matrix $\phi(a_j)$ has at most one nonzero column and thus has rank at most 1.

Example 4.15. Now we show how to obtain a linear representation of a sofic measure that is the image under a 1-block map π of a 1-step Markov measure. Let μ be a 1-step Markov measure determined by a $d \times d$ stochastic matrix P and fixed vector p as in Example 4.14. Let $\pi : X \to Y$ be a 1-block map from the SFT X to a subshift Y. For each a in the alphabet $B = \mathcal{A}(Y)$ let P_a be the $d \times d$ matrix such that

$$P_a(i', j') = \begin{cases} P(i', j') & \text{if } \pi(j') = a \\ 0 & \text{otherwise.} \end{cases} \tag{4.8}$$

Thus P_a just zeroes out all the columns of P except the ones corresponding to indices in the π-preimage of the symbol a in the alphabet of Y. Again let $y = (1, \dots, 1)^{\text{tr}}$. For each $a \in B$ define $\phi(a) = P_a$. That the ν-measure of each cylinder in Y is the sum of the μ-measures of its preimages under π says that the triple (x, ϕ, y) represents $\nu = \pi\mu$.

In working with linearly representable measures, it is useful to know that the nature of the vectors and matrix involved in the representation can be assumed to have a particular restricted form. Below, we say a matrix P is a direct sum of irreducible stochastic matrices if the index set for the rows and columns of P is the disjoint union of sets for which the associated principal submatrices of P are irreducible stochastic matrices. (Equivalently, there are irreducible stochastic matrices $P_1, \dots, P_k$ and a permutation matrix Q such that QPQ^{-1} is the block diagonal matrix whose successive diagonal blocks are $P_1, \dots, P_k$.)

Proposition 4.16. *A formal series $F \in \mathcal{F}(\mathcal{A})$ corresponds to a linearly representable shift-invariant probability measure μ on $\Omega_+(\mathcal{A})$ if and only if F has a linear representation (x, ϕ, y) with $P = \sum_{a \in \mathcal{A}} \phi(a)$ a stochastic matrix, y a column vector of all 1's, and $xP = x$. Moreover, in this case the vector x can be chosen to be positive with the matrix P a direct sum of irreducible stochastic matrices.*

Proof. It is straightforward to check that any (x, ϕ, y) of the specified form linearly represents a shift-invariant measure. Conversely, given a linear representation (x, ϕ, y) as in Definition 4.12 of a shift-invariant probability measure μ, define $P = \sum_{a \in \mathcal{A}} \phi(a)$ and note that, by induction, for all $w \in \mathcal{A}^*, \mu(\mathcal{C}_0(w)) = x\phi(w)P^k y = xP^k\phi(w)y$ for all natural numbers k.

Next, one shows that it is possible to reduce to a linear representation (x, ϕ, y) of μ such that each entry of x and y is nonzero, and, with P defined as $P = \sum_{a \epsilon \mathcal{A}} \phi(a)$, $xP = x$ and $Py = y$. This requires some care. If indices corresponding to 0 entries in x or y, or to 0 rows or columns in P, are jettisoned nonchalantly, the resulting new ϕ may no longer be a morphism.

Definition 4.17. A triple (x', ϕ', y') is obtained from (x, ϕ, y) by *deleting a set I of indices* if the following holds: the indices for (x, ϕ, y) are the disjoint union of the set I and the indices for (x', ϕ', y'); and for every symbol a and all indices i, j not in I we have $x'_i = x_i, y'_i = y_i$ and $\phi'(a)(i,j) = \phi(a)(i,j)$. Then we let ϕ' also denote the morphism determined by the map on generators $a \mapsto \phi'(a)$.

First, suppose that j is an index such that column j of P (and therefore column j of every $\phi(a) := M_a$) is zero. By shift invariance of the measure, (xP, ϕ, y) is still a representation, so we may assume without loss of generality that $x_j = 0$. Let (x', ϕ', y) be obtained from (x, ϕ, y) by deleting the index j. We claim that (x', ϕ', y) still gives a linear representation of μ. This is because for any word $a_1 \dots a_m$, the difference $[x\phi(a_1)\cdots\phi(a_m)y] - [x'\phi'(a_1)\cdots\phi'(a_m)y']$ is a sum of terms of the form

$$x(i_0)M_{a_1}(i_0, i_1)M_{a_2}(i_1, i_2)\cdots M_{a_m}(i_{m-1}, i_m)y(i_m) \tag{4.9}$$

in which at least one index i_t equals j. If $i_0 = j$, then $x(i_0) = 0$; if $i_t = j$ with $t > 0$, then $M_{a_t}(i_{t-1}, i_t) = 0$. In either case, the product is zero.

By the analogous argument involving y rather than x, we may pass to a new representation by deleting the index of any zero row of P. We repeat until we arrive at a representation in which no row or column of P is zero.

An *irreducible component* of P is a maximal principal submatrix C which is an irreducible matrix. C is an *initial component* if for every index j of a column through C, $P(i,j) > 0$ implies that (i,j) indexes an entry of C. C is a *terminal component* if for every index i of a row through C, $P(i,j) > 0$ implies that (i,j) indexes an entry of C.

Now suppose that $\mathcal{I}$ is the index set of an initial irreducible component of P, and $x(i) = 0$ for every i in $\mathcal{I}$. Define (x', ϕ', y) by deleting the index set $\mathcal{I}$. By an argument very similar to the argument for deleting the index of a zero column, the triple (x', ϕ', y') still gives a linear representation of μ. Similarly, if $\mathcal{J}$ is the index set of a terminal irreducible component of P, and $y(j) = 0$ for every j in $\mathcal{J}$, we may pass to a new representation by deleting the index set $\mathcal{J}$.

Iterating these moves, we arrive at a representation for which P has no zero row and no zero column; every initial component has an index i with $x(i) > 0$; and every terminal component has an index j with $y(j) > 0$. We now claim that for this representation the set of matrices $\{P^n\}$ is bounded. Suppose not. Then there is a pair of indices i, j for which the entries $P^n(i, j)$ are unbounded. There is some initial component index i_0, and some $k \geq 0$, such that $x(i_0) > 0$ and $P^k(i_0, i) > 0$. Likewise there is a terminal component index j_0 and an $m \geq 0$ such that $y(j_0) > 0$ and $P^m(j, j_0) > 0$. Appealing to shift invariance of μ, for all $n > 0$ we have

$$1 = xP^{n+k+m}y \geq x(i_0)P^k(i_0, i)P^n(i, j)P^m(j, j_0)y(j_0), \tag{4.10}$$

which is a contradiction to the unboundedness of the entries $P^n(i, j)$. This proves the family of matrices P_n is bounded.

Next let Q_n be the Cesàro sum, $(1/n)(P + ... + P^n)$. Let Q be a limit of a subsequence of the bounded sequence $\{Q_n\}$. Then $PQ = Q = QP$; xQ and Qy are fixed vectors of P; and (xQ, ϕ, Qy) is a linear representation of μ. It could be that xQ vanishes on all indices through some initial component, or that Qy vanishes on all indices through some terminal component. In this case we simply cycle through our reductions until finally arriving a linear representation (x, ϕ, y) of μ such that $xP = x$; $Py = y$; the set of matrices $\{P^n\}$ is bounded; P has no zero row or column; x does not vanish on all indices of any initial component; and y does not vanish on all indices of any terminal component.

If C is an initial component of P, then the restriction of x to the indices of C is a nontrivial fixed vector of C. Thus this restriction is positive, and the spectral radius of C is at least 1. The spectral radius of C must then be exactly 1, because the set $\{P^n\}$ is bounded.

We are almost done. Suppose P is not the direct sum of irreducible matrices. Then there must be an initial component with index set $\mathcal{I}$ and a terminal component with index set $\mathcal{J} \neq \mathcal{I}$, with some $i \in \mathcal{I}$, $j \in \mathcal{J}$ and m minimal in $\mathbb{N}$ such that $P^m(i, j) > 0$. Because $\mathcal{I}$ indexes an initial component, for any $k \in \mathbb{N}$ we have that $(xP^k)_i$ is the sum of the terms $x_{i_0}P(i_0, i_1) \cdots P(i_{k-1}, i)$ such that $i_t \in \mathcal{I}$, $0 \leq t \leq k - 1$. Because $\mathcal{J}$ indexes an terminal component, for any $k \in \mathbb{N}$ we have that $(P^k y)_j$ is the sum of the terms $P(j, i_1) \cdots P(i_{k-1}, i_k)y(i_k)$ such that $i_t \in \mathcal{J}$, $1 \leq t \leq k$. Because $\mathcal{I} \neq \mathcal{J}$, by the minimality of m we have for all $n \in \mathbb{N}$ that

$$xy = xP^{m+n}y \geq \sum_{k=0}^{n} (xP^k)_i P^m(i, j)(P^{n-k}y)_j = (n + 1)x_i P^m(i, j)y_j \ , \tag{4.11}$$

a contradiction.

Consequently, P is now a direct sum of irreducible matrices, each of which has spectral radius 1. The eigenvectors x, y are now positive. Let D be the diagonal matrix with $D(i, i) = y(i)$. Define $(x', \phi', y) = (xD, D^{-1}\phi D, D^{-1}y)$. Then (x', ϕ', y) is the linear representation satisfying all the conditions of the theorem.

$\square$

Example 4.18. The conclusion of the Proposition does not follow without the hypothesis of stationarity: there need *not* be any linear representation with positive vectors x, y, and there need not be any linear representation in which the nonnegative vectors x, y are fixed vectors of P. For example, consider the nonstationary Markov measure μ on two states a, b with initial vector $p = (1, 0)$ and transition matrix

$$T = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = N_a + N_b \ . \tag{4.12}$$

If q is the column vector $(1,1)^{\mathrm{tr}}$, then p, N_a, N_b, q generate a linear representation of μ, e.g. $1 = \mu(\mathcal{C}_0(a)) = pN_aq$, and $(1/2)^k = \mu(\mathcal{C}_0(a^kb^m)) = p(N_a)^k(N_b)^mq$ when $k, m > 0$.

Now suppose that there is a linear representation of μ generated by positive vectors x, y and nonnegative matrices M_a, M_b. Then

$$\begin{aligned} 1 &= \mu(\mathcal{C}_0(a)) = xM_ay, \\ 0 &= \mu(\mathcal{C}_0(b)) = xM_by. \end{aligned} \tag{4.13}$$

From the second of these equations, $M_b = 0$, since $x > 0$ and $y > 0$. But this contradicts $0 < \mu(\mathcal{C}_0(ab)) = xM_aM_by$.

Next suppose there is a linear representation for which x, y could be chosen eigenvectors of $P = M_a + M_b$ (necessarily with eigenvalue 1, since $xP^ny = 1$ for all $n > 0$). Then

$$\frac{1}{2} = \mu(\mathcal{C}_0(ab)) = xM_aM_by \leq xPM_by = xM_by = \mu(\mathcal{C}_0(b)) = 0, \tag{4.14}$$

which is a contradiction.

4.2. **Equivalent characterizations of hidden Markov measures.**

4.2.1. *Sofic measures—formal series approach.*

The semiring $\mathcal{F}(\mathcal{A})$ of formal series on the alphabet $\mathcal{A}$ is an $\mathbb{R}_+$*-module* in a natural way. On this module we have a *(linear) action of* $\mathcal{A}^*$ defined as follows:

For $F \in \mathcal{F}(\mathcal{A})$ and $w \in \mathcal{A}^*$, define $(w, F) \to w^{-1}F$ by

$$(w^{-1}F)(v) = F(wv) \text{ for all } v \in \mathcal{A}^*.$$

Thus

$$w^{-1}F = \sum_{v \in \mathcal{A}^*} F(wv)v.$$

If $F = u \in \mathcal{A}^*$, then

$$(w^{-1}F)(v) = u(wv) = \begin{cases} 1 & \text{if } wv = u \\ 0 & \text{if } wv \neq u. \end{cases}$$

Thus $w^{-1}u \neq 0$ if and only if $u = wv$ for some $v \in \mathcal{A}^*$, and then $w^{-1}u = v$ (in the sense that they are the same function on $\mathcal{A}^*$): $w^{-1}v$ erases w from v if v has w as a prefix, otherwise $w^{-1}v$ gives 0. Note also that this is a *monoid action* :

$$(vw)^{-1}F = w^{-1}(v^{-1}F)\ . \tag{4.15}$$

Definition 4.19. A submodule M of $\mathcal{F}(\mathcal{A})$ is called *stable* if $w^{-1}F \in M$ for all $F \in M$, i.e. $w^{-1}M \subset M$, for all $w \in \mathcal{A}^*$.

Theorem 4.20. *Let $\mathcal{A}$ be a finite alphabet. For a formal series $F \in \mathcal{F}_{\mathbb{R}_+}(\mathcal{A})$ that corresponds to a shift-invariant probability measure ν in $\Omega_+(\mathcal{A})$, the following are equivalent:*

1. *F is linearly representable.*
2. *F is a member of a stable finitely generated submodule of $\mathcal{F}_{\mathbb{R}_+}(\mathcal{A})$.*
3. *F is rational.*
4. *The measure ν is the image under a 1-block map of a shift-invariant 1-step Markov probability measure μ.*

In the latter case, the measure ν is ergodic if and only if it is possible to choose μ ergodic.

In the next few sections we sketch the proof of this theorem

4.2.2. *Proof that a series is linearly representable if and only if it is a member of a stable finitely generated submodule of $\mathcal{F}(\mathcal{A})$.* Suppose that F is linearly representable by (x, ϕ, y). For each $i = 1, 2, \cdots, n$ (where n is the dimension of the representation) and each $w \in \mathcal{A}^*$, define

$$F_i(w) = [\phi(w)y]_i.$$

Let $M = \langle F_1, \cdots, F_n \rangle$ be the span of the F_i with coefficients in $\mathbb{R}_+$, which is a submodule of $\mathcal{F}(\mathcal{A})$. Since

$$F(w) = x\phi(w)y = \sum_{i=1}^{n} x_i[\phi(w)y]_i = \sum_{i=1}^{n} x_i F_i(w),$$

we have that $F = \sum_{i=1}^{n} x_i F_i$, which means $F \in M$.

We next show that M is stable. Let $w \in \mathcal{A}^*$. Then for $u \in \mathcal{A}^*$,

$$\begin{aligned}(w^{-1}F_i)(u) &= F_i(wu) = [\phi(wu)y]_i = [\phi(w)\phi(u)y]_i \\ &= \sum_{j=1}^{n} \phi(w)_{ij}[\phi(u)y]_j = \sum_{j=1}^{n} \phi(w)_{ij} F_j(u).\end{aligned}$$

Since $\phi(w)_{ij} \in \mathbb{R}_+$, we have $\sum_{j=1}^{n} \phi(w)_{ij} F_j(u) \in M$, so

$$w^{-1}F_i = \sum_{j=1}^{n} x_i \phi(w)_{ij} F_j \in \langle F_1, ...F_n \rangle = M.$$

Conversely, let M be a stable finitely generated left submodule, and assume that $F \in \langle F_1, \cdots, F_n \rangle = M$. Then there are $x_1, \cdots, x_n \in \mathbb{R}_+$ such that $F = \sum_{i=1}^n x_i F_i$. Since M is stable, for each $a \in \mathcal{A}$ and each $i = 1, 2, \cdots, n$, we have that $a^{-1} F_i \in \langle F_1, ...F_n \rangle$. So there exist $c_{ij} \in \mathbb{R}_+, j = 1, 2, \cdots, n$, such that $a^{-1} F_i = \sum_{j=1}^n c_{ij} F_j$. Define $\phi(a)_{ij} = c_{ij}$ for $i, j = 1, 2, \cdots, n$. Note by linearity that for any nonnegative row vector $(t_1, \ldots, t_n)$ we have

$$a^{-1}(\sum_{i=1}^n t_i F_i) = \sum_{j=1}^n \Big((t_1, \ldots, t_n)\phi(a) \Big)_j F_j \ . \tag{4.16}$$

Extend ϕ to a monoid morphism $\phi : \mathcal{A}^* \to \mathbb{R}_+^{n \times n}$ by defining $\phi(a_1 \cdots a_n) = \phi(a_1) \cdots \phi(a_n)$. Because the action of $\mathcal{A}^*$ on $\mathcal{F}(\mathcal{A})$ satisfies the monoidal condition (4.15), we have from (4.16) that for any $w = a_1 a_2 \cdots a_n \in \mathcal{A}^*$,

$$\begin{aligned} w^{-1}(\sum_{i=1}^n t_i F_i) &= (a_1 \cdots a_n)^{-1}(\sum_{i=1}^n t_i F_i) = (a_n^{-1} \cdots (a_1^{-1} \sum_{i=1}^n t_i F_i) \cdots) \\ &= \sum_j \Big((t_1, \ldots, t_n)\phi(a_1) \cdots \phi(a_n) \Big)_j F_j = \sum_j \Big((t_1, \ldots, t_n)\phi(w) \Big)_j F_j \ . \end{aligned}$$

Define the column vector y by $y_j = F_j(1)$ for $j = 1, 2, \cdots, n$ and let x be the row vector $(x_1, \ldots, x_n)$. Then

$$F(w) = w^{-1}F(1) = \left(\sum_j \Big(x\phi(w) \Big)_j F_j \right)(1) = \sum_j \Big(x\phi(w) \Big)_j F_j(1) = x\phi(w)y \ , \tag{4.17}$$

showing that (x, ϕ, y) is a linear representation for F.

4.2.3. *Proof that a formal series is linearly representable if and only if it is rational.* This equivalence is from [59, 89]. Recall that a series is rational if and only if it is in the closure of the polynomials under the rational operations $+$ (union), $\cdot$ (concatenation), $*$, and multiplication by elements of $\mathbb{R}_+$.

First we prove by a series of steps that every rational series F is linearly representable.

Proposition 4.21. *Every polynomial is linearly representable.*

Proof. If $w \in \mathcal{A}$ and $|w|$ is greater than the degree of the polynomial F, then $w^{-1} \equiv 0$. Let $S = \{w^{-1}F : w \in \mathcal{A}^*\}$. Then S is finite and stable, hence S spans a finitely generated stable submodule M to which F belongs. (Take $\epsilon^{-1}F = F$). By Section 4.2.2, F is linearly representable. □

The next observation follows immediately from the definition of stability. The proof of the Lemma is included for practice.

Proposition 4.22. *If F_1 and F_2 are in stable finitely generated submodules of $\mathcal{F}(\mathcal{A})$ and $t \in \mathbb{R}_+$, then $(F_1 + F_2)$ and (tF_1) are in stable finitely generated submodules of $\mathcal{F}(\mathcal{A})$.*

Lemma 4.23. *For $F, G \in \mathcal{F}(\mathcal{A})$ and $a \in \mathcal{A}$, $a^{-1}(FG) = (a^{-1}F)G + F(\epsilon)a^{-1}G$.*

Proof. For any $w \in \mathcal{A}^*$,

$$\begin{aligned}(a^{-1}(FG))(w) = (FG)(aw) &= \sum_{uv=aw} F(u)G(v) \\ &= F(\epsilon)G(aw) + \sum_{u'v'=w} F(au')G(v') \\ &= F(\epsilon)G(aw) + \sum_{u'v'=w} (a^{-1}F)(u')G(v') \\ &= F(\epsilon)(a^{-1}G)(w) + ((a^{-1}F)(G))(w).\end{aligned} \tag{4.18}$$

□

Proposition 4.24. *Suppose that for $i = 1, 2$, $F_i \in M_i$, where each M_i is a stable, finitely generated submodule. Let $M = M_1F_2 + M_2$. Then M is finitely generated and stable and contains F_1F_2.*

Proof. The facts that $F_1F_2 \in M$ and M is finitely generated are immediate. The proof that M is stable is a consequence of the Lemma. For if $f_1F_2+f_2$ is an element of M and $a \in \mathcal{A}$, then

$$a^{-1}(f_1F_2 + f_2) = (a^{-1}f_1)F_2 + f_1(\epsilon)(a^{-1}F_2) + a^{-1}f_2. \tag{4.19}$$

Note that $a^{-1}f_1 \in M_1$ and $a^{-1}f_2$, $a^{-1}F_2 \in M_2$. Thus $f_1(\epsilon)(a^{-1}F_2) + f_2 \in M_2$, so we conclude that M is stable.

□

Lemma 4.25. *If F is proper (that is $F_1(\epsilon) = 0$) and $a \in \mathcal{A}$, then $a^{-1}(F^*) = (a^{-1}F)F^*$.*

Proof. Recall that $F_1^* = \sum_{n\geq 0} F_1^n$. Thus $a^{-1}(F^*) = a^{-1}(1 + FF^*) = a^{-1}(\epsilon + FF^*) = a^{-1}\epsilon + (a^{-1}F)F^* + F(\epsilon)a^{-1}(F^*)$.

Because $(a^{-1}\epsilon)(w) = \epsilon(aw) = 0$ for all $w \in \mathcal{A}^*$ and $F(\epsilon) = 0$, we get that $a^{-1}F^* = (a^{-1}F)F^*$. □

Proposition 4.26. *Suppose M_1 is finitely generated and stable, and that $F_1 \in M_1$ is proper. Then F_1^* is in a finitely generated stable submodule.*

Proof. Define $M = \mathbb{R}_+ + M_1F_1^*$. We have

$$F_1^* = 1 + \sum_{n\geq 1} F_1^n = (1 + F_1F_1^*) \in M.$$

Also M is finitely generated (by 1 and the $f_iF_1^*$ if the f_i generate M_1).

To show that M is stable, suppose that $t \in \mathbb{R}_+$ and $a \in \mathcal{A}$. Then for any $u \in \mathcal{A}^*$ we have $(a^{-1}t)(u) = t(au) = 0$, so $a^{-1}t = 0 \in \mathbb{R}_+$. And for any $f_1 \in M_1$ and $a \in \mathcal{A}$, $a^{-1}(f_1F_1^*) = (a^{-1}f_1)F_1^* + f_1(\epsilon)a^{-1}(F_1^*)$. Since M_1 is stable, $a^{-1}f_1 \in M_1$ and the first term is in $M_1F_1^*$. By the Lemma, the second term is $f_1(\epsilon)(a^{-1}F_1)F_1^*$, which is again in $M_1F_1^*$. $\square$

These observations show that if F is rational, then F lies in a finitely generated stable submodule, so by Section 4.2.2 F is linearly representable.

Now we turn our attention to proving the statement in the title of this section in the other direction. So assume that $F \in \mathcal{F}(\mathcal{A})$ is linearly representable. Then $F(w) = x\phi(w)y$ for all $w \in \mathcal{A}$ for some (x, ϕ, y). Consider the semiring of formal series $\mathcal{F}_K(\mathcal{A}) = K^{\mathcal{A}^*}$, where K is the semiring $\mathbb{R}_+^{n\times n}$ of $n \times n$ nonnegative real matrices and n is the dimension of the representation. Let $D = \sum_{a\in\mathcal{A}} \phi(a)a \in \mathcal{F}_K(\mathcal{A})$. The series D is proper, so we can form

$$D^* = \sum_{h\geq 0} D^h = \sum_{h\geq 0}\Big(\sum_{a\in\mathcal{A}} \phi(a)a\Big)^h = \sum_{h\geq 0}\Big(\sum_{w\in\mathcal{A}^h} \phi(w)w\Big) = \sum_{w\in\mathcal{A}} \phi(w)w. \tag{4.20}$$

This series D^* is a rational element of $\mathcal{F}_K(\mathcal{A})$, since we started with a polynomial and formed its *. By Lemma 4.27 below, each entry $(D^*)_{ij}$ is rational in $\mathcal{F}_{\mathbb{R}_+}(\mathcal{A})$.

With D and D^* now defined, we have that

$$F(w) = x\phi(w)y = \sum_{i,j} x_i\phi(w)_{ij}y_j = \sum_{i,j} x_iD^*(w)_{ij}y_j, \tag{4.21}$$

and each $D^*(w)_{ij}$ is a rational series applied to w. Thus $F(w)$ is a finite linear combination of rational series D^*_{ij} applied to w and hence is rational.

Lemma 4.27. *Suppose D is an $n \times n$ matrix whose entries are proper rational formal series (e.g., polynomials). Then the entries of D^* are also rational.*

Proof. We use induction on n. The case $n = 1$ is trivial. Suppose the lemma holds for $n - 1$, and D is $n \times n$ with block form $D = \begin{pmatrix} a & u \\ v & Y \end{pmatrix}$, with a a rational series. The entries of D can be thought of as labels on a directed graph; a path in the graph has a label which is the product of the labels of its edges; and then $D^*(i, j)$ represents the sum of the labels of all paths from i to j (interpret the term "1" in $D(i, i)$ as the label of a path of length zero). With this view, one can see that $D^* = \begin{pmatrix} b & w \\ x & Z \end{pmatrix}$, where

(1) $b = (a + uY^*v)^*$,
(2) $Z = (Y + va^*u)^*$,
(3) $w = buY^*$,
(4) $x = Y^*vb$.

Now Y^* and Z have rational entries by the induction hypothesis, and consequently all entries of D^* are rational. $\square$

4.2.4. *Linearly representable series correspond to sofic measures.* The (topological) support of a measure is the smallest closed set of full measure. Recall our convention (Sec. 2.4) that Markov and sofic measures are ergodic with full support.

Theorem 4.28 [39, 46, 48]**.** *A shift-invariant probability measure ν on $\Omega_+(\mathcal{A})$ corresponds to a linearly representable (equivalently, rational) formal series $F = F_\nu \in \mathcal{F}_{\mathbb{R}_+}(\mathcal{A})$ if and only if it is a convex combination of measures which (restricted to their supports) are sofic measures. Moreover, if (x, ϕ, y) is a representation of F_ν such that x and y are positive and the matrix $\sum_{i \in B} \phi(i)$ is irreducible, then ν is a sofic measure.*

Proof. Suppose that ν is the image under a 1-block map (determined by a map $\pi : \mathcal{A} \to \mathcal{B}$ between the alphabets) of a 1-step Markov measure μ. Then ν is linearly representable by the construction in Example 4.15.

Alternatively, if F_μ is represented by (x, ϕ, y) then for each $w \in \mathcal{A}^*$ we have

$$F_\mu(w) = \sum_{i,j} x_i \phi(w)_{ij} y_j = \sum_{i,j} x_i \Big([\sum_{a \in \mathcal{A}} \phi(a) a]^*(w) \Big)_{ij} y_j. \tag{4.22}$$

For $u \in \mathcal{B}^*$ define

$$F_\nu(u) = \sum_{i,j} x_i \left(\Big[\sum_{b \in \mathcal{B}} (\sum_{a \in \mathcal{A}, \phi(a) = b} \phi(a)) b \Big]^* (u) \right)_{ij} y_j \tag{4.23}$$

to see that F_ν is a linear combination of rational series and to see its linear representation.

Conversely, suppose that ν corresponds to a rational (and hence linearly representable) formal series $F = F_\nu \in \mathcal{F}_{\mathbb{R}_+}(\mathcal{B})$ with dimension n. Let (x, ϕ, y) represent F. To indicate an ordering of the alphabet $\mathcal{B}$, we use notation $\mathcal{B} = \{1, 2, \dots, k\}$ and $\phi(i) = P_i$. First assume that the $n \times n$ matrix P is irreducible and the vectors x and y are positive. We will construct a Markov measure μ and a 1-block map π such that $\nu = \pi\mu$.

Applying the standard stochasticization trick as in the last paragraph of the proof of Proposition 4.16, we may assume that the irreducible matrix P is stochastic, every entry of y is 1, and x is stochastic. Define matrices with block forms,

$$M = \begin{pmatrix} P_1 & P_2 & \cdots & P_k \\ P_1 & P_2 & \cdots & P_k \\ \cdots & \cdots & \cdots & \cdots \\ P_1 & P_2 & \cdots & P_k \end{pmatrix}, \qquad R = \begin{pmatrix} I \\ I \\ \cdots \\ I \end{pmatrix},$$

$$C = \begin{pmatrix} P_1 & P_2 & \cdots & P_k \end{pmatrix}, \quad M_i = \begin{pmatrix} 0 & \cdots & P_i & \cdots & 0 \\ 0 & \cdots & P_i & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & P_i & \cdots & 0 \end{pmatrix}$$

where each P_i is $n \times n$; R is $nk \times k$; I is the $n \times n$ identity matrix; C and the M_i are $nk \times nk$; and M_i is zero except in the i'th block column, where it is RP_i. The matrix M is stochastic, but it can have zero columns. (We thank Uijin Jung for pointing this out.) Let M' be the largest principal submatrix of M with no zero column or row.

We have a strong shift equivalence $M = RC$, $P = CR$, and it then follows from the irreducibility of P that M' is irreducible. Therefore, there is a unique left stochastic fixed vector X for M. Let Y be the $nk \times 1$ column vector with every entry 1. We have $MR = RP$, and consequently $XR = x$. Also, $M_i R = RP_i$ for each i. So, for any word $i_1 \cdots i_j$, we have

$$\begin{aligned} xP_{i_1} \cdots P_{i_j} y &= XRP_{i_1} \cdots P_{i_j} y \\ &= XM_{i_1} \cdots M_{i_j} Ry = XM_{i_1} \cdots M_{i_j} Y \ . \end{aligned}$$

This shows that (X, Φ, Y) is also a representation of F_ν, where $\Phi(i) = M_i$. Let $X', \Phi'(i) = M_i', Y'$ be the restrictions of $X, \Phi(i), Y$ to the vectors/matrices on the indices of M'. Then (X', Φ', Y') is also a representation of F_ν. Let A' be the $0,1$ matrix of size matching M' whose zero entries are the same as for M'. Then (X', M', Y') defines an ergodic Markov measure μ on $\Omega_{A'}$ and there is a 1-block code π such that $\pi\mu = \nu$. Explicitly, π is the restriction of the code which sends $\{1, 2, \dots n\}$ to 1; $\{n+1, n+2, \dots 2n\}$ to 2; and so on. Thus ν is a sofic measure.

Now, for the representation (x, ϕ, y) of F_ν, we drop the assumption that the matrix P is irreducible. However, by Proposition 4.16, without loss of generality we may assume that P is the direct sum of irreducible stochastic matrices $P^{(j)}$; x is a positive stochastic left fixed vector of P; and y is the column vector with every entry 1. Restricted to the indices through $P^{(j)}$, x is a fixed vector of $P^{(j)}$ and therefore is a multiple $c_j x^{(j)}$ of the stochastic left fixed vector $x^{(j)}$ of $P^{(j)}$. Note, $\sum_j c_j = 1$. If $y^{(j)}$ denotes the column vector with every entry 1 such that $P^{(j)} y^{(j)} = y^{(j)}$, then

$$(x, \phi, y) = \sum_j c_j (x_j, P^{(j)}, y^{(j)}) \ .$$

If follows from the irreducible case that μ is a convex combination of sofic measures.

□

4.3. **Sofic measures—Furstenberg's approach.** Below we are extracting from [39, Secs. 18–19] only what we need to describe Furstenberg's approach to the identification of sofic measures and compare it to the others. This leaves out a lot. We follow Furstenberg's notation, apart from change of symbols, except that we refer to shift-invariant measures as well as finite-state stationary processses.

Furstenberg begins with the following definition.

Definition 4.29. [39, Definition 18.1] A *stochastic semigroup of order* r is a semigroup $\mathcal{S}$ having an identity e (i.e., a monoid), together with a set of r elements $\mathcal{A} = \{a_1, a_2, \dots, a_r\}$ generating $\mathcal{S}$, and a real-valued function F defined on S satisfying

(1) $F(e) = 1$,
(2) $F(s) \geq 0$ for each $s \in S$ and $F(a_i) > 0$, $i = 1, 2, \dots, r$,
(3) $\Sigma_{i=1}^r F(a_i s) = \Sigma_{i=1}^r F(s a_i) = F(s)$ for each $s \in S$.

Given a subshift X on an alphabet $\{a_1, a_2, \dots, a_r\}$ with shift-invariant Borel probability μ and $\mu(a_i) > 0$ for every i, let S be the free semigroup of all formal products of the a_i, with the empty product taken as the identity e. Define F on S by $F(e) = 1$ and $F(a_{i_1}a_{i_2}\dots a_{i_k}) = \mu(\mathcal{C}_0(a_{i_1}a_{i_2}\dots a_{i_k}))$. Clearly the triple $(\{a_1, a_2, \dots, a_r\}, S, F)$ is a stochastic semigroup, which we denote $S(X)$.

Conversely, any stochastic semigroup $(\{a_1, a_2, \dots, a_r\}, S, F)$ determines a unique shift-invariant Borel probability μ for which $F(a_{i_1}a_{i_2}\dots a_{i_k}) = \mu(\mathcal{C}_0((a_{i_1}a_{i_2}\dots a_{i_k})))$ for all $a_{i_1}a_{i_2}\dots a_{i_k}$. We denote by $X(S)$ this finite-state stationary process (equivalently the full shift on r symbols with invariant measure μ). Two stochastic semigroups are called *equivalent* if they define the same finite-state stationary process modulo a bijection of their alphabets. A *cone* in a linear space is a subset closed under addition and multiplication by positive real numbers [39, Sec. 15.1].

Definition 4.30. [39, Definition 19.1] Let D be a linear space, D^* its dual, and let $\mathcal{C}$ be a cone in D such that for all x, y in D, if $x + \lambda y \in \mathcal{C}$ for all real λ, then $y = 0$. Let $\theta \in \mathcal{C}$ and $\theta^* \in D^*$, and suppose that θ^* is nonnegative on $\mathcal{C}$. A *linear stochastic semigroup* S on $(\mathcal{C}, \theta, \theta^*)$ is a stochastic semigroup $(\{a_1, \dots, a_r\}, S, F)$ whose elements are linear transformations *from* $\mathcal{C}$ *to* $\mathcal{C}$ satisfying

(1) $\sum a_i\theta = \theta$;
(2) $\sum a_i^*\theta^* = \theta^*$ (where L^* denotes the transformation of D^* adjoint to a transformation L of D);
(3) $F(s) = (\theta^*, s\theta)$ for $s \in S$, where $(\cdot, \cdot)$ denotes the dual pairing of D^* and D;
(4) $(\theta^*, a_i\theta) > 0$, $i = 1, 2, \dots, r$.

$(S, D, \mathcal{C}, \theta, \theta^*)$ was called *finite dimensional* by Furstenberg if there is $m \in \mathbb{N}$ such that $D = \mathbb{R}^m$, $\mathcal{C}$ is the cone of vectors in $\mathbb{R}^m$ with all entries nonnegative, and each element of S is an $m \times m$ matrix with nonnegative entries.

A semigroup S of transformations satisfying (1) to (4) does define a stochastic semigroup if $(\theta^*, \theta) = 1$.

Theorem 4.31. [39, Theorem 19.1] *Every stochastic semigroup S is equivalent to some linear stochastic semigroup.*

Proof. Let $A_0(S)$ be the real semigroup algebra of S, i.e., the real vector space with basis S and multiplication determined by the semigroup multiplication in S and the distributive property,

$$\Big(\sum \alpha_s s\Big)\Big(\sum \beta_t t\Big) = \sum \alpha_s\beta_t st. \tag{4.24}$$

(Each sum above has finitely many terms.)

If S is the free monoid generated by r symbols, then $A_0(S)$ is isomorphic to the set $\wp_{\mathbb{R}}(\mathcal{A})$ of real-valued polynomials, i.e. finitely supported formal series $\mathcal{A}^* \to \mathbb{R}$ (see Definition 4.2).

Extend F from S to a linear functional on $A_0(S)$, i.e. $F(\sum \alpha_s s) = \sum \alpha_s F(s)$. Define $I = \{u \in A_0(S) : F(u) = 0\}$, an ideal in $A_0(S)$, and the algebra $A = A(S) = A_0(S)/I$. Define the element $\tau = a_1 + a_2 + \cdots + a_r$ in $A(S)$ (here a_i abbreviates $a_i + I$) and set $D = A/A(e - \tau)$.

The elements of A and in particular those of S operate on D by left multiplication. Let a_i' denote the operator induced by left multiplication by $a_i \in S$. Take V to be the image in D of the set of elements of A that can be represented as positive linear combinations of elements in S. Denote by $\overline{u}$ the image in D of an element u in A. Set $\theta = \overline{e}$ and let θ^* be the functional induced on D by F on A (F vanishes on $A(e - \tau)$).

Then the four conditions in the definition of linear stochastic semigroup are satisfied. This linear stochastic semigroup given by

$$(\{a_1', \dots, a_r'\}, D, V, \theta, \theta^*) \tag{4.25}$$

is equivalent to the given S because $F(s') = (\theta^*, s'\theta) = F(s)$. (We will see later that this construction is closely related to Heller's "stochastic module" construction.) □

Given a shift-invariant sofic measure on the set of two-sided sequences on the alphabet $\{1, \dots, r\}$ which assigns positive measure to each symbol, it is possible to associate an explicit finite-dimensional linear stochastic semigroup to μ in the same way that we attached a linear representation in Example 4.15. Here μ is the image under some 1-block code π of a Markov measure defined from some $m \times m$ stochastic matrix P. For $1 \le i \le r$, let P_i be the $m \times m$ matrix such that $P_i(i', j') = P(i', j')$ if $\pi(j') = i$ and otherwise $P_i(i', j') = 0$. Let θ^* be a stochastic (probability) left fixed vector for P and let θ be the column vector with every entry 1. Let C be the cone of all nonnegative vectors in $D = \mathbb{R}^m$. If we identify P_i with the symbol i, then these data give a finite-dimensional linear stochastic semigroup equivalent to $S(X)$. Along with this observation, Furstenberg established the converse.

Theorem 4.32. [39, Theorem 19.2] *A linear stochastic semigroup S is finite dimensional if and only if the stochastic process that it determines is a 1-block factor of a 1-step stationary finite-state Markov process.*

In the statement of Theorem 4.32, "Markov" does not presume ergodic. The construction for the theorem is essentially the one given in Theorem 4.28, with a simplification. Because of the definition of linear stochastic semigroup (Definition 4.30), Furstenberg can begin with θ^*, θ actual fixed vectors of $P := \sum_i P_i$. The triple (P, θ^*, θ) corresponds to (P, x, y) in Theorem 4.16, where x, y need not be fixed vectors. Thus Furstenberg can reduce more quickly to the form where θ^* and θ are *positive* fixed vectors of P. Note that "finite dimensional" in Theorem 4.32 means more than having the cone C of the linear stochastic semigroup generating a finite-dimensional space D: here C is a cone in $\mathbb{R}^m$ with exactly m (in particular, finitely many) extreme rays.

4.4. **Sofic measures—Heller's approach.** Repeating some problems already stated, but with some refinements, here are the natural questions about sofic measures which we are currently discussing, in subshift language.

Problem 4.33. Let $\pi : \Omega_A \to Y$ be a 1-block map from a shift of finite type to a (sofic) subshift and let μ be a (fully supported) 1-step Markov measure on Ω_A. When is $\pi\mu$ Markov? Can one determine what the *order* (a k such that the measure is k-step Markov) of the image measure might be?

Problem 4.34. Given a shift-invariant probability measure ν on a subshift Y, when are there a shift of finite type Ω_A, a factor map $\pi : \Omega_A \to Y$, and a 1-step shift-invariant fully supported Markov measure μ on Ω_A such that $\pi\mu = \nu$?

Problem 4.35. If ν is a sofic measure, how can one explicitly construct Markov measures of which ν is a factor? Are there procedures for constructing Markov measures that map to ν which have a minimal number of states or minimal entropy?

Problem 4.33 was discussed in [20], for the reversible case. Later complete solutions depend on Heller's solution of Problem 4.34, so we discuss that first. Effective answers to the first part of Problem 4.35 are given by Furstenberg and in the proof of Theorem 4.28.

Problem 4.34 goes back at least to a 1959 paper of Gilbert [40]. Following Gilbert and Dharmadhikari [23, 24, 25, 26], Heller (1965) created his stochastic module theory and within this gave a characterization [48, 49] of sofic measures (1965). We describe this next.

4.4.1. *Stochastic module.* We describe the stochastic module machinery setup of Heller [48] (with some differences in notation). Let $S = \{1, 2, ..., s\}$ be a finite state space for a stochastic process. Let A_S be the associative real algebra with free generating set S. An *A_S-module* is a real vector space V on which A_S acts by linear transformations, such that for each $i \in S$ there is a linear transformation $M_i : V \to V$ such that a word $u_1...u_k$ sends $v \in V$ to $M_{u_1}(M_{u_2}(...(M_{u_k}(v))..))$. We denote an A_S-module as $(\{M_i\}, V)$ or for brevity just $\{M_i\}$, where the M_i are the associated generating linear transformations $V \to V$ as above.

Definition 4.36. A *stochastic S-module* for a stochastic process with state space S is a triple $(l, \{M_i\}, r)$, where $(\{M_i\}, V)$ is an A_S-module, $r \in V$, $l \in V^*$, and for every word $u = u_1...u_t$ on S its probability $\mathrm{Prob}(u) = \mathrm{Prob}(\mathcal{C}_0(u))$ is given by

$$\mathrm{Prob}(u) = lM_{u_1}M_{u_2}...M_{u_t}r. \tag{4.26}$$

Given an A_S-module M, an $l \in V^*$ and $r \in V$, a few axioms are required to guarantee that they define a stochastic process with state space S. Define $\sigma = \sum\{a_i : a_i \in S\}$ and denote by $\mathcal{C}_S$ the cone of polynomials in A_S with nonnegative coefficients. Then the axioms are that

(1) $lr = 1$;
(2) $l(\mathcal{C}_S r) \subset [0, \infty)$;
(3) for all $f \in A_S$, $l(f(\sigma - 1)\,r) = 0$.

Example 4.37. *A stochastic module for a sofic measure.* As we saw in Section 4.3, this setup of a stochastic module arises naturally when a 1-block map π is applied to a 1-step Markov measure μ with state space S given by an $s \times s$ stochastic transition matrix P and row probability vector l. For each $i \in S$, let M_i be the matrix whose j'th column equals column j of P if $\pi(j) = i$ and whose other columns are zero. The probability of an S-word $u = u_1 ... u_t$ is $lM_{u_1}M_{u_2}...M_{u_t}r$, where r is the vector of all 1's. With $V = \mathbb{R}^s$, presented as column vectors, $(l, \{M_i\}, r)$ is a stochastic module for the process given by $\pi\mu$.

4.4.2. *The reduced stochastic module.* A stochastic module $(l, (\{M_i\}, V), r)$ is *reduced* if (i) V is the smallest invariant (under the operators M_i) vector space containing r and (ii) l annihilates no nonzero invariant subspace of V. Given a stochastic module $(l, \{M_i\}, r)$ for a stochastic process, with its operators M_i operating on the real vector space V, a smallest stochastic module $(l', \{M_i'\}, r')$ describing the stochastic process may be defined as follows. Let R_1 be the cyclic submodule of V generated by the action on r; let L_1 be the cyclic submodule of V^* generated by the (dual) action on l; let V' be R_1 modulo the subspace annihilated by L_1; for each $i \in S$ let M_i' be the (well defined) transformation of V' induced by M_i; let r', l' be the elements of V' and $(V')^{\perp}$ determined by r, l. Now (l', M', r') is *the reduced stochastic module* of the process. V' is the subspace generated by the action of the M_i' on r', and no nontrivial submodule of V' is annihilated by l'. The reduced stochastic module is still a stochastic module for the original stochastic process. We say "the" reduced stochastic module because any stochastic modules describing the same stochastic process have isomorphic reduced stochastic modules.

4.4.3. *Heller's answer to Problem 4.34.* We give some preliminary notation. A process is "induced from a Markov chain" if its states are lumpings of states of a finite state Markov process, that is, there is a 1-block code which sends the associated Markov measure to the measure associated to the stochastic process. Let $(A_S)_+$ be the subset of A_S consisting of linear combinations of words with all coefficients nonnegative. A *cone* in a real vector space V is a union of rays from the origin. A convex cone $\mathcal{C}$ is *strongly convex* if it contains no line through the origin. It is *polyhedral* if it is the convex hull of finitely many rays.

Theorem 4.38. *Let $(l, (\{M_i\}, V), r)$ be a reduced stochastic module. The associated stochastic process is induced from a Markov chain if and only if there is a cone $\mathcal{C}$ contained in the vector space V such that the following hold:*

(1) $r \in \mathcal{C}$,
(2) $l\mathcal{C} \subset [0, \infty)$,
(3) $(A_S)_+\mathcal{C} \subset \mathcal{C}$,
(4) $\mathcal{C}$ *is strongly convex and polyhedral.*

Heller stated this result in [48, Theorem 1]. The proof there contained a minor error which was corrected in [49]. Heller defined a process to be *finitary* if its associated reduced stochastic module is finite dimensional. (We will call the corresponding measure finitary.) A consequence of Theorem 4.38 is the (obvious) fact that the reduced stochastic module of a sofic measure must be finitary. Heller gave an example [48] of a finitary process which is not a 1-block factor of a 1-step Markov measure, and therefore is not a factor of any Markov measure. (However, a subshift with a weakly mixing finitary measure is measure theoretically isomorphic to a Bernoulli shift [12].)

4.5. **Linear automata and the reduced stochastic module for a finitary measure.** The 1960's and 1970's saw the development of the theory of probabilistic automata and linear automata. We have not thoroughly reviewed this literature, and we may be missing from it significant points of contact with and independent invention of the ideas under review. However, we mention at least one. A finite dimensional stochastic module is a special case of a linear space automaton, as developed in [51] by Inagaki, Fukumura and Matuura, following earlier work on probabilistic automata (e.g. [76, 83]. They associated to each linear space automaton its canonical (up to isomorphism) equivalent irreducible linear space automaton. When the linear space automaton is a stochastic module, its irreducible linear space automaton corresponds exactly to Heller's canonical (up to isomorphism) reduced stochastic module. Following [51] and Nasu's paper [70], we will give some concrete results on the reduced stochastic module.

We continue the Example 4.37 and produce a concrete version of the reduced stochastic module in the case that a measure on a subshift is presented by a stochastic module which is finite dimensional as a real vector space (for example, in the case of a sofic measure). Our presentation follows a construction of Nasu [70] (another is in [51]). Correspondingly, in this section we will reverse Heller's roles for row and column vectors and regard the stochastic module as generated by row vectors.

So, let $(u, \{M_i\}, v)$ be a finite dimensional stochastic module on finite alphabet $\mathcal{A}$. We take the presentation so that there is a positive integer n such that the M_i are $n \times n$ matrices; u and v are n-dimensional row and column vectors; and the map $a \mapsto M_a$ induces a monoid homomorphism ϕ from $\mathcal{A}^*$, sending a word $w = a_1 \cdots a_j$ to the matrix $\phi(w) = M_{a_1} \cdots M_{a_j}$.

Let $\mathcal{U}$ be the vector space generated by vectors of the form $u\phi(w)$, $w \in \mathcal{A}^*$. Similarly define $\mathcal{V}$ as the vector space generated by vectors of the form $\phi(w)v$, $w \in \mathcal{A}^*$. Let $k = \dim(\mathcal{U})$. If $k < n$, then construct a smaller module (presenting the same measure) as follows. Let L be a $k \times n$ matrix whose rows form a basis of $\mathcal{U}$. For each symbol a there exists a $k \times k$ matrix $\widehat{M_a}$ such that $LM_a = \widehat{M_a}L$. Define $\widehat{u}$ to be the k dimensional row vector such that $\widehat{u}L = u$ and set $\widehat{v} = Lv$. Let $a \to \widehat{M_a}$ induce a monoid homomorphism $\widehat{\phi}$ from $\mathcal{A}^*$, sending a word $w = a_1 \cdots a_j$ to $\widehat{\phi}(w) = \widehat{M}_{a_1} \cdots \widehat{M}_{a_j}$. The subspace $\widehat{\mathcal{U}}$ of $\mathbb{R}^k$ generated by vectors of the form $\widehat{u}\widehat{\phi}(w)$ is equal to $\mathbb{R}^k$ because $\widehat{\mathcal{U}}L = \mathcal{U}$ and $\dim(\mathcal{U}) = k$. It is easily checked that $\widehat{u}\widehat{\phi}(w)\widehat{v} = u\phi(w)v$, for every w in $\mathcal{A}^*$. Let $\widehat{\mathcal{V}}$ be the subspace of $\mathbb{R}^k$ generated by column vectors $\widehat{\phi}(w)\widehat{v}$. We have for each a that $LM_a v = \widehat{M}_a Lv = \widehat{M}_a \widehat{v}$, so L maps $\mathcal{V}$ onto $\widehat{\mathcal{V}}$. Also L maps the space of n-dimensional column vectors onto $\mathbb{R}^k$. It follows that if $\dim(\mathcal{V}) = n$, then $\dim(\widehat{\mathcal{V}}) = k$.

If $\dim(\widehat{\mathcal{V}}) < k$, then repeat the reduction move, but applying it to v (column vectors) rather than to u. This will give a stochastic module $(\overline{u}, \{\overline{M}_a\}, \overline{v})$, say with $m \times m$ matrices M_a and invariant subspaces $\overline{\mathcal{U}}, \overline{\mathcal{V}}$ generated by the action on $\overline{u}, \overline{v}$. By construction we have $\dim(\overline{\mathcal{V}}) = m$. And because $\widehat{\mathcal{U}}$ had full dimension, we have $\dim(\overline{\mathcal{U}}) = m$ also. Regarding $\mathcal{V}$ as a space of functionals on $\mathcal{U}$, and letting $\ker(\mathcal{V})$ denote the subspace of $\mathcal{U}$ annihilated by all elements of $\mathcal{V}$, we see that $u \mapsto \overline{u}$ is a presentation of the map $\pi : \mathcal{U} \to \mathcal{U}/\ker(\mathcal{V})$. Thus $(\overline{u}, \{\overline{M}_a\}, \overline{v})$ is a presentation of the reduced stochastic module. Also, for all a, $\pi M_a = \overline{M}_a \pi$, and therefore the surjection π (acting from the right) also satisfies

$$\Big(\sum_a M_a\Big)\pi = \pi\Big(\sum_a \overline{M}_a\Big) . \tag{4.27}$$

If $(\widetilde{u}, \{\widetilde{M_a}\}, \widetilde{v})$ is another such presentation of the reduced stochastic module, then it must have the same (minimal) dimension m, and there will be an invertible matrix G (giving the isomorphism of the two presentations) such that for all a,

$$\Big(\widetilde{u},\ \{\widetilde{M_a}\},\ \widetilde{v}\Big) = \Big(\overline{u}G,\ \{G^{-1}\overline{M}_a G\},\ G^{-1}\overline{v}\Big) . \tag{4.28}$$

To find G, simply take m words w such that the vectors $\overline{u}\overline{\phi}(w)$ are a basis for $\overline{\mathcal{U}}$, and let G be the matrix such that for each of these w,

$$\overline{u}\overline{\phi}G = \widetilde{u}\widetilde{\phi} . \tag{4.29}$$

The rows of the matrix L above (a basis for the space $\mathcal{U}$) may be obtained by examining vectors $u\phi(w)$ in some order, with the length of w nondecreasing, and including as a row any vector not in the span of previous vectors. Let $\mathcal{U}_m$ denote the space spanned by vectors $u\phi(w)$ with w of length at most m. If for some m it holds that $\mathcal{U}_m = \mathcal{U}_{m+1}$, then $\mathcal{U}_m = \mathcal{U}$. In particular, if n is the dimension of the original stochastic module, then the matrix L can be found by considering words of length at most $n-1$.

One can check that if two equivalent stochastic modules have dimensions n_1 and n_2, then they are equivalent (define the same measure) if and only if they assign the same measure to words of length n_1+n_2-1. (This is a special case of [51, Theorem 5.2].) If the reduced stochastic module of a measure has dimension at most n, then one can also construct the reduced stochastic module from the measures of words of length at most $2n-1$ (one construction is given in [51, Theorem 6.2]). However, without additional information about the measure, this forces the examination of a number of words which for a fixed alphabet can grow exponentially as a function of n, as indicated by the following example.

Example 4.39. Let X be the full shift on the three symbols $0, 1, 2$. Given $k \in \mathbb{N}$, define a stochastic matrix P indexed by X-words of length $k+1$ by $P(10^k, 0^k1) = 1/6 = P(20^k, 0^k2)$; $P(10^k, 0^k2) = 1/2 = P(20^k, 0^k1)$; $P(a_0 \cdots a_k, a_1 \cdots a_{k+1}) = 1/3$ otherwise; and all other entries of P are zero. This matrix defines a $(k+1)$-step Markov measure μ on X which agrees with the Bernoulli $(1/3, 1/3, 1/3)$ measure on all words of length at most $k+2$ except the four words $10^k1, 10^k2, 20^k1, 10^k2$. The reduced stochastic module has dimension at most $2k+1$, because for any word U the conditional probabilty function on X-words defined by $\rho_U : W \mapsto \mu(UW|U)$

will be a constant multiple of ρ_V for one of the words $V = 0^{k+1}, 10^j, 20^j$, with $0 \leq j \leq k$. The number of X-words of length $k+2$ is 3^{k+2}.

4.6. **Topological factors of finitary measures, and Nasu's core matrix.** The content of this section is essentially taken from Nasu's paper [70], as we explain in more detail below. Given a square matrix M, in this section we let M^* denote any square matrix similar to one giving the action of M on the maximal invariant subspace on which the action of M is nonsingular.

Adapting terminology from [70], we define the *core matrix* of a finite dimensional stochastic module give by matrices, $(l, \{M_i\}, r)$, to be $\sum_i M_i$. A core matrix for a finitary measure μ is any matrix which is the core matrix of a reduced stochastic module for μ. This matrix is well defined only up to similarity, but for simplicity of language we refer to *the core matrix of* μ, denoted $\text{Core}(\mu)$. Similarly, we define *the eventual core matrix of* μ to be $\text{Core}(\mu)^*$, denoted $\text{Core}^*(\mu)$. E.g., if $\text{Core}(\mu)$ is $\begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$, then $\text{Core}^*(\mu)$ is $\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}$.

Considering square matrices M and N as linear endomorphisms, we say N is a quotient of M if there is a linear surjection π such that, writing action from the right, $M\pi = \pi N$. (Equivalently, by duality, the action of N is isomorphic to the action of M on some invariant subspace.) In this case, the characteristic polynomial of M divides that of N (but, e.g. $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ is a principal submatrix of but not a quotient of $\begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$).

Theorem 4.40. *Suppose ϕ is a continuous factor map from a subshift X onto a subshift Y, $\mu \in \mathcal{M}(X)$ and $\phi\mu = \nu \in \mathcal{M}(Y)$. Suppose μ is finitary. Then ν is finitary, and Core$^*(\nu)$ is a quotient of Core$^*(\mu)$. In particular, if ϕ is a topological conjugacy, then Core$^*(\nu) =$ Core$^*(\mu)$.*

The key to the topological invariance in Theorem 4.40 is the following lemma (a measure version of [70, Lemma 5.2]).

Lemma 4.41. *Suppose μ is a finitary measure on a subshift X and $n \in \mathbb{N}$. Let $X^{[n]}$ be the n-block presentation of X; let $\psi : X^{[n]} \to X$ the 1-block factor map defined on symbols by $[a_1 \cdots a_n] \mapsto a_1$; let $\mu^{[n]} \in \mathcal{M}(X^{[n]})$ be the measure such that $\psi\mu^{[n]} = \mu$. Then $\mu^{[n]}$ is finitary and Core$^*(\mu^{[n]})$ is a quotient of Core$^*(\mu)$.*

Proof of Lemma 4.41. For $n > 1$, the n-block presentation of X is (after a renaming of the alphabet) equal to the 2-block presentation of $X^{[n-1]}$. So, by induction it suffices to prove the lemma for $n = 2$.

Let $(l, \{P_i\}, r)$ be a reduced stochastic module for μ, where the P_i are $k \times k$ and $\mathcal{A}(X) = \{1, 2, \dots, m\}$. For each symbol ij of $\mathcal{A}(X^{[2]})$, define an $mk \times mk$ matrix

P'_{ij} as an $m \times m$ system of $k \times k$ blocks, in which the i, j block is P_i and the other entries are zero. Define $l' = (l, \ldots, l)$ (m copies of l) and define $r' = \begin{pmatrix} P_1 r \\ \vdots \\ P_m r \end{pmatrix}$. Then $(l', \{P'_{ij}\}, r')$ is a stochastic module for $\mu^{[2]}$, which is therefore finitary. Also, we have an elementary strong shift equivalence of the core matrices P and P',

$$P' = \begin{pmatrix} P_1 \\ \vdots \\ P_m \end{pmatrix} \begin{pmatrix} I & \cdots & I \end{pmatrix} \qquad , \qquad P = \begin{pmatrix} I & \cdots & I \end{pmatrix} \begin{pmatrix} P_1 \\ \vdots \\ P_m \end{pmatrix},$$

and therefore $P^* = (P')^*$. Because Core$(\mu^{[2]})$ is a quotient of P', it follows that Core$^*(\mu^{[2]})$ is a quotient of $(P')^* = P^* = \text{Core}^*(\mu)$. $\square$

If $\phi : X \to Y$ is a factor map of irreducible sofic shifts of equal entropy, then ϕ must send the unique measure of maximal entropy of X, μ_X, to that for Y. These are sofic measures, and consequently Theorem 4.40 gives computable obstructions to the existence of such a factor map between given X and Y. In his work, Nasu associated to given X a certain linear (not stochastic) automaton. If we denote it $(l, \{M_i\}, r)$, and let $\log(\lambda)$ denote the topological entropy of X, then $(l, \{(1/\lambda)M_i\}, r)$ would be a stochastic module for μ_X. In the end Nasu's core matrix is $\lambda\text{Core}(\mu_X)$. Nasu remarked in [70] that his arguments could as well be carried out with respect to measures to obtain his results, and that is what we have done here.

Eigenvalue relations between core matrices (not so named) of equivalent linear automata already appear in [51, Sec.7]. Also, Kitchens [55] earlier used the (Markov) measure of maximal entropy for an irreducible shift of finite type in a similar way to show that the existence of a factor map of equal-entropy irreducible SFTs, $\Omega_A \to \Omega_B$, implies (in our terminology) that B^* is a quotient of A^*. This is a special case of Nasu's constraint.

5. When is a sofic measure Markov?

5.1. **When is the image of a 1-step Markov measure under a 1-block map 1-step Markov?**

We return to considering Problem 4.33. In this subsection, suppose μ is a 1-step Markov measure, that is, a 1-step fully supported shift-invariant Markov measure on an irreducible shift of finite type Ω_A. Suppose that π is a 1-block code with domain Ω_A. How does one characterize the case when the measure $\pi\mu$ is again 1-step Markov?

To our knowledge, this problem was introduced, in the language of Markov processes, by Burke and Rosenblatt (1958) [20], who solved it in the reversible case [20, Theorem 1]. Kemeny and Snell [54, Theorems 6.4.8 and 6.3.2] gave another exposition and introduced the "lumpability" terminology. Kemeny and Snell defined a (not necessarily stationary) finite-state Markov process X to be *lumpable* with respect to a partition of its states if for every initial distribution for X the corresponding quotient process is Markov. They defined X to be *weakly lumpable* with respect to the partition if there exists an initial distribution for X for which the quotient process Y is Markov. In all of this, by Markov they mean 1-step Markov. Various problems around these ideas were (and continue to be) explored and solved. For now we restrict our attention to the question of the title of this subsection and describe three answers.

5.1.1. *Stochastic module answer.*

Theorem 5.1. *Let (l, M, r) be a presentation of the reduced stochastic module of a sofic measure ν on Y, in which M_i denotes the matrix by which a symbol i of $\mathcal{A}(Y)$ acts on the module. Suppose $k \in \mathbb{N}$. Then the sofic measure ν is k-step Markov if and only if every product $M_{i(1)} \cdots M_{i(k)}$ of length k has rank at most 1.*

The case $k = 1$ of Theorem 5.1 was proved by Heller [48, Prop.3.2] An equivalent characterization was given a good deal later, evidently without awareness of Heller's work, by Bosch [15], who worked from the papers of Gilbert [40] and Dharmadhikari [23]. The case of general k in Theorem 5.1 was proved by Holland [50, Theorem 4], following Heller.

5.1.2. *Linear algebra answer.* One can approach the problem of deciding whether a sofic measure is Markov with straight linear algebra. There is a large literature using such ideas in the context of automata, control theory and the "lumpability" strand of literature emanating from Kemeny and Snell (see e.g. [41] and its references). Propositions 5.2 and 5.3 and Theorem 5.4 are taken from Gurvits and Ledoux [41]. As with previous references, we are considering only a fragment of this one.

Let N be the size of the alphabet of the irreducible shift of finite type Ω_A. Let π be a 1-block code mapping Ω_A onto a subshift Y. Let P be an $N \times N$ irreducible stochastic matrix defining a 1-step Markov measure μ on Ω_A. Let p be the positive stochastic row fixed vector of P. Let U be the matrix such that $U(i, j) = 1$ if π maps the state i to the state j, and $U(i, j) = 0$ otherwise. Given $i \in \mathcal{A}(\Omega_A)$, let $\bar{i}$ be its image symbol in Y. Given $j \in \mathcal{A}(Y)$, let P_j be the matrix of size P which equals P in columns i such that $\bar{i} = j$, and is zero in other entries. Likewise define p_j. Given a Y-word $w = j_1 \cdots j_k$, we let $P_w = P_{j_1} \cdots P_{j_k}$.

Alert: We are using parenthetical notation for matrix and vector entries and subscripts for lists. If $\pi\mu$ is a 1-step Markov measure on Y, then it is defined using a stochastic row vector q and stochastic matrix Q. The vector q can only be pU, and the entries of Q are determined by $Q(j, k) = (p_j P_k U)(k)/q(j)$. Let ν denote the Markov measure defined using q, Q. Define q_j, Q_j by replacing entries of q, Q with zero in columns not indexed by j. For a word $w = j_0 \dots j_k$ on symbols from $\mathcal{A}(Y)$, we have $(\pi\mu)(\mathcal{C}_0(w)) = \nu(\mathcal{C}_0(w))$ if and only if

$$p_{j_0} P_{j_1} \cdots P_{j_k} U = p_{j_0} U Q_{j_1} \cdots Q_{j_k} \tag{5.1}$$

(since $q_{j_0} = p_{j_0} U$). Thus $\pi\mu = \nu$ if and only if (5.1) holds for all Y-words w. This remark is already more or less in Kemeny and Snell [54, Theorem 6.4.1].

For the additional argument which produces a finite procedure, we define certain vector spaces (an idea already in [31, 56, 86, 87, 41] and elsewhere).

Let $\mathcal{V}_k$ denote the real vector space generated by the row vectors $p_{j_0}P_{j_1}\cdots P_{j_k}$ such that $j_0j_1\cdots j_t$ is a Y-word and $0\le t\le k$. So, $\mathcal{V}_0$ is the vector space generated by the vectors p_{j_0}, and $\mathcal{V}_{k+1}$ is the subspace generated by $\mathcal{V}_k\cup\{vP_j : v\in\mathcal{V}_k, j\in\mathcal{A}(Y)\}$. In fact, for $k\ge 0$, we claim that

$$\mathcal{V}_k = \langle\ \{p_{j_0}P_{j_1}\cdots P_{j_k} : j_0\cdots j_k\in\mathcal{A}(Y)^{k+1}\}\ \rangle\ , \quad \text{and} \tag{5.2}$$

$$\mathcal{V}_{k+1} = \langle\ \{vP_j : v\in\mathcal{V}_k, j\in\mathcal{A}(Y)\}\ \rangle\ , \tag{5.3}$$

where $\langle\quad\rangle$ is used to denote span. Clearly (5.3) follows from (5.2), which is a consequence of stationarity, as follows. Because $\sum_j p_j = p = pP = \sum_j pP_j$, and for $i\ne j$ the vectors p_i and pP_j cannot both be nonzero in any coordinate, we $p_j = pP_j$. So, given t and $j_1\cdots j_t$, we have

$$\begin{aligned} p_{j_1}P_{j_2}\cdots P_{j_t} &= pP_{j_1}P_{j_2}\cdots P_{j_t} \\ &= \sum_{j_0} p_{j_0}P_{j_1}P_{j_2}\cdots P_{j_t}, \end{aligned}$$

from which (5.3) easily follows. Let $\mathcal{V} = \langle\cup_{k\ge 0}\mathcal{V}_k\rangle$.

Proposition 5.2. *Suppose P is an $N\times N$ irreducible stochastic matrix and ϕ is a 1-block code. Let the vector spaces $\mathcal{V}_k$ be defined as above, and let n be the smallest positive integer such that $\mathcal{V}_n=\mathcal{V}_{n+1}$. Then $n\le N-|\mathcal{A}(Y)|$, $\mathcal{V}_n=\mathcal{V}$, and the following are equivalent:*

1. *$\phi\mu$ is a 1-step Markov measure on the image of ϕ.*
2. *$p_{j_0}P_{j_1}\cdots P_{j_n}U = p_{j_0}UQ_{j_1}\cdots Q_{j_n}$, for all $j_0\cdots j_n\in\mathcal{A}(Y)^{n+1}$.*

Proof. For $k\ge 1$, we have $\mathcal{V}_k\subset\mathcal{V}_{k+1}$, and also

$$\mathcal{V}_k=\mathcal{V}_{k+1} \quad \text{implies} \quad \mathcal{V}_k=\mathcal{V}_t=\mathcal{V} \quad \text{for all } t\ge k\ . \tag{5.4}$$

Because $\dim(\mathcal{V}_0)=|\mathcal{A}(Y)|$, it follows that $n\le N-|\mathcal{A}(Y)|$.

Because (1) is equivalent to (5.1) holding for all Y-words $j_0j_1\cdots j_k,\ k\ge 0$, we have that (1) implies (2).

Now suppose (2) holds. For $K\ge 1$, the linear condition (5.1) holds for all Y-words of length k less than or equal to K if and only if $vUQ_j = vP_jU$ for all j in $\mathcal{A}(Y)$ and all v in $\mathcal{V}_K$. (U is the matrix defined above.) Because $\mathcal{V}_K=\mathcal{V}_n$ for $K\ge n$, we conclude from (2) and (5.2) that (5.1) holds for all Y-words $j(0)j(1)\cdots j(k),\ k\ge 0$, and therefore (1) holds. $\square$

Next we consider an irrreducible $N\times N$ matrix P defining a 1-step Markov measure μ on Ω_A and a 1-block code ϕ from Ω_A onto a subshift Y. Given a positive integer $k\ge 1$, we are interested in understanding when $\phi\mu$ is k-step Markov. We use notations $U, p, p_j, P_j, \mathcal{V}_t$ and $\mathcal{V}_n=\mathcal{V}$ as above. Define a stochastic row vector q indexed by Y-words of length k, with $q(j_0\cdots j_{k-1}) = (p_{j_0}P_{j_1}\cdots P_{j_{k-1}}U)(j_{k-1})$.

Let Q be the square matrix indexed by Y-words of length k whose nonzero entries are defined by

$$Q(j_0 \cdots j_{k-1}, j_1 \dots j_k) = \frac{\Big(p_{j_0} P_{j_1} \cdots P_{j_k} U\Big)(j_k)}{q(j_0 \cdots j_{k-1})} \ .$$

Then Q is an irreducible stochastic matrix and q is a positive stochastic vector such that $qQ = q$. Let ν be the k-step Markov measure defined on Y by (q, Q). The measures ν and $\phi\mu$ agree on cylinders $\mathcal{C}_0(j_0 \cdots j_k)$ and therefore on all cylinders $\mathcal{C}_0(j_0 \cdots j_t)$ with $0 \leq t \leq k$. Clearly, if $\phi\mu$ is k-step Markov then $\phi\mu$ must equal ν.

Proposition 5.3. [41] *Suppose P is an $N \times N$ irreducible stochastic matrix defining a 1-step Markov measure μ on Ω_A and $\phi : \Omega_A \to Y$ is a 1-block code. Let k be a fixed positive integer. With the notations above, the following are equivalent.*

(1) *$\phi\mu$ is a k-step Markov measure (i.e., $\phi\mu = \nu$).*
(2) *For every Y-word $w = w_0 \cdots w_{k-1}$ of length k and every $v \in \mathcal{V}$,*

$$vP_w(PU - \mathbf{1}Q^w) = 0, \tag{5.5}$$

where $P_w = P_{w_0} \cdots P_{w_{k-1}}$; $\mathbf{1}$ is the size N column vector with every entry 1; and Q^w is the stochastic row vector defined by

$$Q^w(j) = Q(w_0 \cdots w_{k-1}, w_1 \cdots w_{k-1} j) \ , \quad j \in \mathcal{A}(Y) \ . \tag{5.6}$$

Proof. We continue to denote by $z(j)$ the entry in the j'th coordinate of a row vector z. By construction of ν we have for $t = 0$ that

$$(\pi\mu)\mathcal{C}_0(j_0 \cdots j_{t+k}) = \nu\mathcal{C}_0(j_0 \cdots j_{t+k}) \quad \text{for all } j_0 \cdots j_{t+k} \in \mathcal{A}^{t+k+1} \ . \tag{5.7}$$

Now suppose t is a nonnegative integer and (5.7) holds for t. Given $j_0 \cdots j_{t+k}$, let w be its terminal word of length k. Then for $j \in \mathcal{A}(Y)$,

$$\begin{aligned}
&(\pi\mu)\mathcal{C}_0(j_0 \cdots j_{t+k} j) - \nu\mathcal{C}_0(j_0 \cdots j_{t+k} j) \\
&= \Big(p_{j_0} P_{j_1} \cdots P_{j_{t+k}} P_j U\Big)(j) - \Big(\nu\mathcal{C}_0(j_0 \cdots j_{t+k}) Q^w\Big)(j) \\
&= \Big(p_{j_0} P_{j_1} \cdots P_{j_{t+k}} P_j U\Big)(j) - \Big((p_{j_0} P_{j_1} \cdots P_{j_{t+k}} \mathbf{1}) Q^w\Big)(j) \\
&= \Big(p_{j_0} P_{j_1} \cdots P_{j_{t+k}} [P_j U - \mathbf{1}Q^w]\Big)(j) \\
&= \Big(p_{j_0} P_{j_1} \cdots P_{j_t} P_w [PU - \mathbf{1}Q^w]\Big)(j),
\end{aligned}$$

where the term $P_{j_1} \cdots P_{j_t}$ is included only if $t > 0$, and the last equality holds because the jth columns of PU and $P_j U$ are equal. Thus, given (5.7) for t, by (5.2) we have (5.7) for $t+1$ if and only $vP_w[PU - \mathbf{1}Q^w] = 0$ for all $v \in \mathcal{V}_t$ and all w of length k. It follows from induction that (5.7) holds for all $t \geq 0$ (i.e. $\pi\mu = \nu$) if and only if (5.5) holds for all $v \in \mathcal{V}$. ∎

Because $\mathcal{V}$ can be computed, Proposition 5.3 gives an algorithm, given k, for determining whether the image of a 1-step Markov measure is a k-step Markov measure. The next result gives a criterion which does not require computation of the matrix Q.

Theorem 5.4. [41] *Let notations be as in Proposition 5.3. Then $\phi\mu$ is a k-step Markov measure on Y if and only if for every Y-word w of length k,*

$$\Big((\mathcal{V}P_w)\cap\ker(U)\Big)P\subset\ker(U)\ . \tag{5.8}$$

Proof. Let $w = w_0\cdots w_{k-1}$ be a Y-word of length k. Using the computations of the proof of Proposition 5.3, we obtain for $j\in\mathcal{A}(Y)$ that

$$\begin{aligned}0 &= \pi\mu\mathcal{C}_0(w_0\cdots w_{k-1}j) - \nu\mathcal{C}_0(w_0\cdots w_{k-1}j)\\ &= \Big(p_{w_0}P_{w_1}\cdots P_{w_{k-1}}[PU-\mathbf{1}Q^w]\Big)(j)\\ &= \Big(pP_{w_0}P_{w_1}\cdots P_{w_{k-1}}[PU-\mathbf{1}Q^w]\Big)(j)\\ &= \Big(pP_w[PU-\mathbf{1}Q^w]\Big)(j)\ .\end{aligned}$$

Consequently, the vector $v = p$ satisfies (5.5). Moreover,

$$(pP_wU)(w_{k-1}) = (p_{w_0}P_{w_1}\cdots P_{w_{k-1}}U)(w_{k-1}) = \pi\mu\mathcal{C}_0(w) > 0,$$

and therefore $pP_w\notin\ker(U)$. Because $vP_w = 0$ if and only if $vP_w\mathbf{1} = 0$, the space $\mathcal{V}P_w$ is spanned by pP_w and $(\mathcal{V}P_w)\cap\ker(U)$. Thus (5.5) holds for all $v\in\mathcal{V}$ if and only if (5.5) holds for all $v\in\mathcal{V}$ such that $vP_w\in\ker(U)$, which is equivalent to (5.8). □

Gurvits and Ledoux [41, Sec. 2.2.2] explain how Theorem 5.4 can be used to produce an algorithm, polynomial in the number N of states, for deciding whether $\pi\mu$ is a 1-step Markov measure.

5.2. **Orders of Markov measures under codes.** This section includes items relevant to the second part of Problem 4.33.

Definition 5.5. Given positive integers m, n, k with $1\le k\le n$, recursively define integers $N(k,m,n)$ by setting

$$N(n,m,n) = 1 \tag{5.9}$$

$$N(k,m,n) = (1+m^{N(k+1,m,n)})N(k+1,m,n)\ ,\quad \text{if } 1\le k<n\ . \tag{5.10}$$

Proposition 5.6. *Suppose $\pi:\Omega_A\to Y$ is a 1-block code and μ is a 1-step Markov measure on Ω_A. Let n be the dimension of the reduced stochastic module of $\pi\mu$ and let $m = |\mathcal{A}(Y)|$. Suppose $n\ge 2$. (In the case $n = 1$, $\pi\mu$ is Bernoulli.) Let $K = N(2,m,n)$. If $\pi\mu$ is not K-step Markov, then it is not k-step Markov for any k.*

Before proving Proposition 5.6, we state our main interest in it.

Corollary 5.7. *Suppose μ is a 1-step Markov measure on an irreducible SFT Ω_A determined by a stochastic matrix P, and that there are algorithms for doing arithmetic in the field generated by the entries of P. Suppose ϕ is a block code on Ω_A. Then there is an algorithm for deciding whether the measure $\phi\mu$ is Markov.*

Proof. The corollary is an easy consequence of Propositions 5.2 and 5.6. □

The proof of Proposition 5.6 uses two lemmas.

Lemma 5.8. *Suppose $P_1, \dots, P_t$ are $n \times n$ matrices such that* $\mathrm{rank}(P_1 ... P_t P_1) = \mathrm{rank}(P_1) = r$. *Then for all positive integers m,* $\mathrm{rank}(P_1 ... P_t)^m P_1 = r$.

Proof. It follows from the rank equality that $(P_1 ... P_k)$ defines an isomorphism from the image of P_1 (a vector space of column vectors) to itself. □

Lemma 5.9. *Suppose k, m, n are positive integers and $1 \leq k \leq n$. Suppose $\mathcal{Q}$ is a collection of m matrices of size $n \times n$, and there exists a product of $N(k, m, n)$ matrices from $\mathcal{Q}$ with rank at least k. Then there are arbitrarily long products of matrices from $\mathcal{Q}$ with rank at least k.*

Proof. We prove the proposition by induction on k, for k decreasing from n. The case $k = n$ is clear. Suppose now $1 \leq k < n$ and the lemma holds for $k+1$. Suppose a matrix M is a product $Q_{i(1)} \cdots Q_{i(N(k,m,n))}$ of $N(k, m, n)$ matrices from $\mathcal{Q}$ and has rank at least k. We must show there are arbitrarily long products from $\mathcal{Q}$ with rank at least k.

The given product is a concatenation of products of length $N(k + 1, m, n)$, and we define corresponding matrices,

$$(5.11) \quad P_j = Q_{1+(j-1)(N(k+1,m,n))} \cdots Q_{j(N(k+1,m,n))}, \quad 1 \leq j \leq 1 + m^{N(k+1,m,n)} .$$

If any P_j has rank at least $k + 1$, then by the induction hypothesis there are arbitrarily long products with rank at least $k + 1$, and we are done. So, suppose every P_j has rank at most k. Because $\mathrm{rank}(P_j) \geq \mathrm{rank}(M) \geq k$, it follows that M, and every P_j, and every subproduct of consecutive P_j's, has rank k.

There are only $m^{N(k+1,m,n)}$ words of length $N(k + 1, m, n)$ on m symbols, so two of the matrices P_j must be equal. The conclusion now follows from Lemma 5.8. □

Proof of Proposition 5.6. As described in Examples 4.37 and 4.5, there are algorithms for producing the reduced stochastic module for $\pi\mu$ as a set of matrices M_a (one for each symbol from $\mathcal{A}(Y)$) and a pair of vectors u, v such that for any Y-word $a_1 \cdots a_t$, $(\pi\mu)\mathcal{C}_0(a_1 \cdots a_t) = uM_{a_1} \cdots M_{a_t} v$. By Theorem 5.1, $\pi\mu$ is k-step Markov if and only every product $M_{a_1} \cdots M_{a_k}$ has rank at most 1. Let $K = N(2, m, n)$. If $\pi\mu$ is not K-step Markov, then some matrix $\prod_{i=1}^{K} M_{a(i)}$ has rank at least 2, and by Lemma 5.9 there are then arbitrarily long products of M_a's with rank at least 2. By Theorem 5.1, this shows that $\pi\mu$ is not k-step Markov for any k. □

Remark 5.10. Given m and n, the numbers $N(k, m, n)$ grow very rapidly as k decreases. Consequently, the bound K in Proposition 5.6 (and consequently the algorithm of Corollary 5.7) is not practical. However, in an analogous case (Problem 5.13 below) we don't even know the existence of an algorithm.

Problem 5.11. Find a reasonable bound K for Proposition 5.6.

Example 5.12. This is an example to show that the cardinality of the domain alphabet cannot be used as the bound K in Proposition 5.6. Given $n > 1$ in $\mathbb{N}$, let A be the adjacency matrix of the directed graph $\mathcal{G}$ which is the union of two cycles, $a_1b_1b_2\cdots b_{n+4}a_1$ and $a_2b_3b_4\cdots b_{n+3}a_2$. The vertex set $\{a_1, a_2, b_1, \ldots, b_{n+4}\}$ is the alphabet $\mathcal{A}$ of Ω_A. Let ϕ be the 1-block code defined by erasing subscripts, and let Y be the subshift which is the image of ϕ, with alphabet $\{a, b\}$. Let μ be any 1-step Markov measure on Ω_A. In $\mathcal{G}$, there are exactly four first return paths from $\{a_1, a_2\}$ to $\{a_1, a_2\}$: $a_1b_1\cdots b_{n+4}a_1$, $a_1b_1\cdots b_{n+3}a_2$, $a_2b_3\cdots b_{n+4}a_1$ and $a_2b_3\cdots b_{n+3}a_2$. Thus, in a point of Y, successive occurrences of the symbol a must correspondingly be separated by m b's, with $m \in \{n+4, n+3, n+2, n+1\}$. Each Y-word ab^ma has a unique preimage word, so $\phi : \Omega_A \to Y$ is a topological conjugacy. Thus $\phi\mu$ is k-step Markov for some k. We have

$$\begin{aligned}
\phi(b_1\cdots b_{n+3}a_2b_3\cdots b_{n+3}a_2) &= \big(b^{n+3}ab^{n+1}\big)a\ , \quad \text{and} \\
\phi(a_1b_1\cdots b_{n+4}a_1b_1\cdots b_{n+1}) &= ab\big(b^{n+3}ab^{n+1}\big)\ .
\end{aligned}$$

So, $\big(b^{n+3}ab^{n+1}\big)a$ and $ab\big(b^{n+3}ab^{n+1}\big)$ are Y-words, but $ab\big(b^{n+3}ab^{n+1}\big)a$ is not a Y-word. Consequently, we have conditional probabilities,

$$\begin{aligned}
\phi\mu[y_0 = a \mid y_{-(2n+5)}\cdots y_{-1} &= \quad \big(b^{n+3}ab^{n+1}\big)] > 0\ , \\
\phi\mu[y_0 = a \mid y_{-(2n+7)}\cdots y_{-1} &= ab\big(b^{n+3}ab^{n+1}\big)] = 0\ ,
\end{aligned}$$

which shows that $\phi\mu$ cannot be $(2n+5)$-Markov. In contrast, $|\mathcal{A}| = n+6 < 2n+5$.

With regard to the problem (3.3) of determining whether a given factor map is Markovian, the analogue of Proposition 5.6 is the following open problem.

Problem 5.13. Find (or prove there does not exist) an algorithm for attaching to any 1-block code ϕ from an irreducible shift of finite type a number N with the following property: if a 1-step Markov measure μ on the range of ϕ has no preimage measure which is N-step Markov, then μ has no preimage measure which is Markov.

Remark 5.14. (The persistence of memory) Suppose $\phi : \Omega_A \to \Omega_B$ is a 1-block code from one irreducible 1-step SFT onto another. We collect some facts on how the memory of a Markov measure and a Markov image must or can be related.

(1) The image of a 1-step Markov measure can be Markov but not 1-step Markov. (E.g. the standard map from the k-block presentation to the 1-block presentation takes the 1-step Markov measures onto the k-step Markov measures.)
(2) If ϕ is finite-to-one and ν is k-step Markov on Ω_B, then there is a unique Markov measure μ on Ω_A such that $\phi\mu = \nu$, and μ is also k-step Markov (Proposition 3.18).
(3) If any 1-step Markov measure on Ω_B lifts to a k-step Markov measure on Ω_A, then for every n, every n-step Markov measure on Ω_B lifts to an $(n+k)$-step Markov measure on Ω_A. (This follows from the explicit construction (3.2) and passage as needed to a higher block presentation.)

(4) If ϕ is infinite-to-one then it can happen [18, Section 2] ("peculiar memory example") that every 1-step Markov measure on Ω_B lifts to a 2-step Markov measure on Ω_A but not to a 1-step Markov measure, while every 1-step Markov on Ω_A maps to a 2-step Markov measure on Ω_B.

6. Resolving maps and Markovian maps

In this section, Ω_A denotes an irreducible 1-step shift of finite type defined by an irreducible matrix A.

6.1. **Resolving maps.** In this section, $\pi : \Omega_A \to Y$ is a 1-block code onto a subshift Y, with Y not necessarily a shift of finite type, unless specified. U denotes the $0,1$, $|\mathcal{A}(\Omega_A)| \times |\mathcal{A}(Y)|$ matrix such that $U(i,j) = 1$ iff $\pi(i) = j$. Denote a symbol $(\pi x)_0$ by $\overline{x_0}$.

Definition 6.1. The factor map π as above is *right resolving* if for all symbols $i, \overline{i}, k$ such that $\overline{i}k$ occurs in Y, there is at most one j such that ij occurs in Ω_A and $\overline{j} = k$. In other words, for any diagram

$$\begin{array}{ccc} i & & \\ \big\downarrow & & \\ \overline{i} & \longrightarrow & k \end{array} \tag{6.1}$$

there is at most one j such that

$$\begin{array}{ccc} i & \longrightarrow & j \\ \big\downarrow & & \big\downarrow \\ \overline{i} & \longrightarrow & k \end{array} \tag{6.2}$$

Definition 6.2. A factor map π as above is *right e-resolving* if it satisfies the definition above, with "at most one" replaced by "at least one".

Reverse the roles of i and j above to define *left resolving* and *left e-resolving.* A map π is *resolving (e-resolving)* if it is left or right resolving (e-resolving).

Proposition 6.3. (1) *If π is resolving, then $h(\Omega_A) = h(Y)$.*
(2) *If $Y = \Omega_B$ and $h(\Omega_A) = h(\Omega_B)$, then π is e-resolving iff π is resolving.*
(3) *If π is e-resolving, then Y is a 1-step shift of finite type, Ω_B.*
(4) *If π is e-resolving and $k \in \mathbb{N}$, then every k-step Markov measure on $Y = \Omega_B$ lifts to a k-step Markov measure on Ω_A.*

Proof. (1) This holds because a resolving map must be finite-to-one [66, 58].

(2) We argue as in [66, 58]. Suppose π is right-resolving. This means precisely that $AU \le UB$. If $AU \ne UB$, then it would be possible to increase some entry of A by one and have a resolving map onto Ω_B from some irreducible SFT Ω_C properly containing Ω_A. But now $h(\Omega_C) > h(\Omega_A)$, while $h(\Omega_C) = h(\Omega_B) = h(\Omega_A)$ because

the resolving maps respect entropy. This is a contradiction. The other direction holds by a similar argument.

(3) This is an easy exercise [18].

(4) We consider $k = 1$ (the general case follows by passage to the higher block presentation). Suppose π is right e-resolving. This means that $AU \geq UB$. Suppose Q is a stochastic matrix defining a 1-step Markov measure μ on Ω_B. For each positive entry $B(k, \ell)$ of B and i such that $\pi(i) = k$, let $\mathcal{J}(i, k, l)$ be the set of indices j such that $A(i, j) > 0$ and $\pi(j) = \ell$. Now simply choose P to be any nonnegative matrix of size and zero/positive pattern matching A such that for each i, k, l, $\sum_{j \in \mathcal{J}(i,k,l)} P(i, j) = Q(k, \ell)$. Then $PU = UQ$, and this guarantees that $\pi\mu = \nu$. The condition on the $+/0$ pattern guarantees that μ has full support on Ω_A. (The code π in Example 3.4 is right e-resolving, and (3.4) gives an example of this construction.) $\square$

The resolving maps, and the maps which are topologically equivalent to them (the *closing* maps), form the only class of finite-to-one maps between nonconjugate irreducible shifts of finite type which we know how to construct in significant generality [5, 6, 66, 58, 17]. The e-resolving maps, and the maps topologically equivalent to them (the *continuing* maps), are similarly the Markovian maps we know how to construct in significant generality [18]. If Ω_A, Ω_B are mixing shifts of finite type with $h(\Omega_A) > h(\Omega_B)$ and there exists any factor map from Ω_A to Ω_B (as there will given a trivially necessary condition), then there will exist infinitely many continuing (hence Markovian) factor maps from Ω_A to Ω_B. However, the most obvious hope, that the factor map send the maximal entropy measure of Ω_A to that of Ω_B, can rarely be realized. Given Ω_A, there are only finitely many possible values of topological entropy for Ω_B for which such a map can exist [18].

6.2. **All factor maps lift 1-1 a.e. to Markovian maps.** Here “all factor maps” means “all factor maps between irreducible sofic subshifts”. Factor maps between irreducible SFTs need not be Markovian, but they are in the following strong sense close to being Markovian, even if the subshifts X and Y are only sofic.

Theorem 6.4. [17] *Suppose $\pi : X \to Y$ is a factor map of irreducible sofic subshifts. Then there are irreducible SFT's Ω_A, Ω_B and a commuting diagram of factor maps*

$$\begin{array}{ccc} \Omega_A & \xrightarrow{\gamma} & \Omega_B \\ {\scriptstyle\alpha}\downarrow & & \downarrow{\scriptstyle\beta} \\ X & \xrightarrow[\pi]{} & Y \end{array} \tag{6.3}$$

such that α, β are degree 1 right resolving and γ is e-resolving. In particular, γ is Markovian. If Y is SFT, then the composition $\beta\gamma$ is also Markovian.

The Markovian claims in Theorem 6.4 hold because finite-to-one maps are Markovian (Proposition 3.18), e-resolving maps are Markovian (Proposition 6.3), and a composition of Markovian maps is Markovian. In the case when π is degree 1 between irreducible SFTs, the "Putnam diagram" (6.3) is a special case of Putnam's work in [82], which was the stimulus for [17].

6.3. **Every factor map between SFT's is hidden Markovian.** A factor map $\pi : \Omega_A \to \Omega_B$ is Markovian if some (and therefore every) Markov measure on Ω_B lifts to a Markov measure on Ω_A. There exist factor maps between irreducible SFTs which are not Markovian. In this section we will show in contrast that all factor maps between irreducible SFTs (and more generally between irreducible sofic subshifts) are *hidden Markovian*: every sofic (i.e., hidden Markov) measure lifts to a sofic measure. The terms Markov measure and sofic measure continue to include the requirement of full topological support.

Theorem 6.5. *Let $\pi : X \to Y$ be a factor map between irreducible sofic subshifts and suppose that ν is a sofic measure on Y. Then ν lifts to a sofic measure μ on X. Moreover, μ can be chosen to satisfy* $\mathrm{degree}(\mu) \leq \mathrm{degree}(\nu)$.

Proof. We consider two cases.

Case I: ν is a Markov measure on Y. Consider the Putnam diagram (6.3) associated to π in Theorem 6.4. The measure ν lifts to a Markov measure μ^* on Ω_A. Set $\mu = \alpha\mu^*$. Then $\pi\mu = \nu$, and $\mathrm{degree}(\mu) = 1 \leq \mathrm{degree}(\nu)$.

Case II: ν is a degree n sofic measure on Y. (Possibly $n = \infty$.) Then there are an irreducible SFT Ω_C with a Markov measure μ' and a degree n factor map $g : \Omega_C \to Y$ which sends μ' to ν. By Lemma 6.8 below, there exist another irreducible SFT Ω_F and factor maps $\widetilde{g}$ and $\widetilde{\pi}$ with $\mathrm{degree}(\widetilde{g}) \leq \mathrm{degree}(g)$ such that the following diagram commutes:

$$\begin{array}{ccc} \Omega_F & \xrightarrow{\widetilde{\pi}} & \Omega_C \\ {\scriptstyle \widetilde{g}}\big\downarrow & & \big\downarrow{\scriptstyle g} \\ X & \xrightarrow[\pi]{} & Y \end{array} \tag{6.4}$$

Apply Case I to $\widetilde{\pi}$ to get a degree 1 sofic measure ν^* on Ω_F which $\widetilde{\pi}$ sends to μ'. Then $\widetilde{g}(\nu^*)$ is a sofic measure of degree at most n which π sends to ν. $\square$

To complete the proof of Theorem 6.5 by proving Lemma 6.8, we must recall some background on magic words. Suppose $X = \Omega_A$ is SFT and $\pi : \Omega_A \to Y$ is a 1-block factor map. Any X-word v is mapped to a Y-word πv of equal length. Given a Y-word $w = w[1, n]$ and an integer i in $[1, n]$, set $d(w, i) = |\{w'_i : \pi w' = w\}|$. As in [17], the *resolving degree* $\delta(\pi)$ of π is defined as the minimum of $d(w, i)$ over all allowed w, i, and w is a *magic word* for π if for some i, $d(w, i) = \delta(\pi)$. (For finite-to-one maps, these are the standard magic words of symbolic dynamics [66, 58]; some of their properties are still useful in the infinite-to-one case. The junior author confesses an error: [17, Theorem 7.1] is wrong. The resolving degree is not in general invariant under topological conjugacy, in contrast to the finite-to-one case.)

If a magic word has length 1, then it is a *magic symbol*. As remarked in [17, Lemma 2.4], the argument of [58, Proposition 4.3.2] still works in the infinite-to-one case to show that π is topologically equivalent to a 1-block code from a one step irreducible SFT for which there is a magic symbol. (Factor maps π, ϕ are topologically equivalent if there exist topological conjugacies α, β such that $\alpha\phi\beta = \pi$.)

Proposition 6.6. *Suppose X is SFT; $\pi : X \to Y$ is a 1-block factor map; a is a magic symbol for π; aQa is a Y-word; and $a'Q'a''$ is an X-word such that $\pi(a'Q'a'') = aQa$. Then the image of the cylinder $\mathcal{C}_0[a'Q'a'']$ equals the cylinder $\mathcal{C}_0[aQa]$.*

Proof. Suppose $PaQaR$ is a Y-word, with preimage X-words $P^j a^j Q^j (a_*)^j R^j$, say $1 \le j \le J$, with the 1-block code acting by erasing $*$ and superscripts. Because a is a magic symbol, there must exist some j such that $a_j = a'$, and there must exist some k such that $(a_*)^k = a''$. Because X is a 1-step SFT, $P^j a'Q'a''R^k$ is an X-word, and it maps to $PaQaR$. This shows that the image of $\mathcal{C}_0[a'Q'a']$ is dense in $\mathcal{C}_0[aQa]$ and therefore, by compactness, equal to it. □

Corollary 6.7. *Suppose $\pi : X \to Y$ is a factor map from an irreducible SFT X to a sofic subshift Y. Then there is a residual set of points in Y which lift to doubly transitive points in X.*

Proof. Without loss of generality, we assume π is a 1-block factor map, X is a 1-step SFT, and there is a magic symbol a for π. Let $v_n = a'P_n a'$, $n \in \mathbb{N}$, be a set of X-words such that every X-word occurs as a subset of some P_n and a' is a symbol sent to a. The set E_n of points in X which see the words $v_1, v_2, \dots v_n$ both in the future and in the past is a dense open subset of X. It follows from Proposition 6.6 that each πE_n is open. For every n, E_n contains E_{n+1}, so $\pi(\cap_n E_n) = \cap_n \pi E_n$. Thus the set $\cap_n E_n$ of doubly transitive points in X maps to a residual subset of Y. □

We do not know whether in Corollary 6.7 every doubly transitive point of Y must lift to a doubly transitive point of X.

Lemma 6.8. *Suppose $\alpha : X \to Z$ and $\beta : Y \to Z$ are factor maps of irreducible sofic subshifts. Then there is an irreducible SFT W with factor maps $\widetilde{\alpha}$ and $\widetilde{\beta}$ such that* degree$(\widetilde{\beta}) \le$ degree(β) *and the following diagram commutes.*

$$\begin{tikzcd} W \arrow[r, "\widetilde{\alpha}"] \arrow[d, "\widetilde{\beta}"'] & Y \arrow[d, "\beta"] \\ X \arrow[r, "\alpha"'] & Z \end{tikzcd} \tag{6.5}$$

Proof. First, suppose X and Y are SFT. The intersection of any two residual sets in Z is nonempty, so by Corollary 6.7 we may find x and y, doubly transitive in X and Y respectively, such that $\alpha x = \beta y$. Let Ω_F be the irreducible component of the fiber product $\{(u, v) \in X \times Y : \alpha x = \beta y\}$ built from α and β to which the point (x, y) is forward asymptotic, and let $\widetilde{\beta}, \widetilde{\alpha}$ be restrictions to Ω_F of the coordinate projections. These restrictions must be surjective. Note that degree$(\widetilde{\beta}) \leq$ degree(β).

If X and Y are not necessarily SFT, then there are degree 1 factor maps from irreducible SFT's, $\rho_1 : \Omega_A \to X$ and $\rho_2 : \Omega_B \to Y$, and we can apply the first case to find $\widetilde{\alpha\rho_1}$ and $\widetilde{\beta\rho_2}$ in the diagram with respect to the pair $\alpha\rho_1, \beta\rho_2$. Now for $\widetilde{\alpha}$ and $\widetilde{\beta}$ we use the maps $\rho_1\widetilde{\alpha\rho_1}$ and $\rho_2\widetilde{\beta\rho_2}$. □

Acknowledgment. This article arose from the October 2007 workshop "Entropy of Hidden Markov Processes and Connections to Dynamical Systems" at the Banff International Research Station, and we thank BIRS, PIMS, and MSRI for hospitality and support. We thank Jean-René Chazottes, Masakazu Nasu, Sujin Shin, Peter Walters and Yuki Yayama for very helpful comments. We are especially grateful to Uijin Jung and the two referees for extremely thorough comments and corrections. Both authors thank the Departamento de Ingeniería Matemática, Center for Mathematical Modeling, of the University of Chile and the CMM-Basal Project, and the second author also the Université Pierre et Marie Curie (University of Paris 6) and Queen Mary University of London, for hospitality and support during the preparation of this article. Much of Section 4 is drawn from lectures given by the second author in a graduate course at the University of North Carolina, and we thank the students who wrote up the notes: Rika Hagihara, Jessica Hubbs, Nathan Pennington, and Yuki Yayama.

References

1. Atef M. Abdel-Moneim and Frederick W. Leysieffer, *Weak lumpability in finite Markov chains*, J. Appl. Probab. **19** (1982), no. 3, 685–691. MR MR664854 (84a:60077)
2. ———, *Lumpability for nonirreducible finite Markov chains*, J. Appl. Probab. **21** (1984), no. 3, 567–574. MR MR752021 (85k:60089)
3. Rashid Ahmad, *An algebraic treatment of Markov processes*, Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the Eighth European Meeting of Statisticians (Tech. Univ. Prague, Prague, 1974), Vol. A, Reidel, Dordrecht, 1977, pp. 13–22. MR MR0488303 (58 #7854)
4. Michael Arbib, *Realization of stochastic systems*, Ann. Math. Statist. **38** (1967), 927–933. MR MR0225606 (37 #1199)
5. Jonathan Ashley, *Resolving factor maps for shifts of finite type with equal entropy*, Ergodic Theory Dynam. Systems **11** (1991), no. 2, 219–240. MR MR1116638 (92d:58056)
6. ———, *An extension theorem for closing maps of shifts of finite type*, Trans. Amer. Math. Soc. **336** (1993), no. 1, 389–420. MR MR1105064 (93e:58048)
7. François Bancilhon, *A geometric model for stochastic automata*, IEEE Trans. Computers **C-23** (1974), no. 12, 1290–1299. MR MR0406736 (53 #10522)
8. Leonard E. Baum and Ted Petrie, *Statistical inference for probabilistic functions of finite state Markov chains*, Ann. Math. Statist. **37** (1966), 1554–1563. MR MR0202264 (34 #2137)
9. J. Berstel and C. Reutenauer, *Rational Series and Their Languages*, Springer-Verlag, 1988.
10. Z. I. Bezhaeva and V. I. Oseledets, *Erdős measures, sofic measures, and Markov chains*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **326** (2005), no. Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 13, 28–47, 279–280. MR MR2183214 (2006h:60119)
11. P. Billingsley, *Probability and Measure*, John Wiley and Sons, Inc., 1995.
12. M. Binkowska and B. Kaminski, *Classification of ergodic finitary shifts*, Ann. Sci. Univ. Clermont-Ferrand II Probab. Appl. (1984), no. 2, 25–37.
13. D. Blackwell, *The entropy of functions of finite state Markov chains*, Trans. First Prague Conf. Inf. Theory, Stat. Dec. Functions, Random Proc., 1957, pp. 13–20.

14. David Blackwell and Lambert Koopmans, *On the identifiability problem for functions of finite Markov chains.*, Ann. Math. Statist. **28** (1957), 1011–1015. MR MR0099081 (20 #5525)
15. Karl Bosch, *Notwendige und hinreichende Bedingungen dafür, daß eine Funktion einer homogenen Markoffschen Kette Markoffsch ist*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **31** (1974/75), 199–202. MR MR0383535 (52 #4416)
16. Paul E. Boudreau, *Functions of finite Markov chains and exponential type processes*, Ann. Math. Statist. **39** (1968), 1020–1029. MR MR0224161 (36 #7207)
17. Mike Boyle, *Putnam's resolving maps in dimension zero*, Ergodic Theory Dynam. Systems **25** (2005), no. 5, 1485–1502. MR MR2173429 (2006h:37013)
18. Mike Boyle and Selim Tuncel, *Infinite-to-one codes and Markov measures*, Trans. Amer. Math. Soc. **285** (1984), no. 2, 657–684. MR MR752497 (86b:28024)
19. C. Burke and M. Rosenblatt, *Consolidation of probability matrices*, Bull. Inst. Internat. Statist. **36** (1958), no. 3, 7–8. MR MR0120680 (22 #11429)
20. C. J. Burke and M. Rosenblatt, *A Markovian function of a Markov chain*, Ann. Math. Statist. **29** (1958), 1112–1122. MR MR0101557 (21 #367)
21. J.-R. Chazottes and E. Ugalde, *Projection of Markov measures may be Gibbsian*, J. Statist. Phys. **111** (2003), no. 5-6, 1245–1272. MR MR1975928 (2004d:37008)
22. ———, *On the preservation of Gibbsianness under symbol amalgamation*, preprint, 2009.
23. S. W. Dharmadhikari, *Functions of finite Markov chains*, Ann. Math. Statist. **34** (1963), 1022–1032. MR MR0152020 (27 #2001a)
24. ———, *Sufficient conditions for a stationary process to be a function of a finite Markov chain*, Ann. Math. Statist. **34** (1963), 1033–1041. MR MR0152021 (27 #2001b)
25. ———, *Exchangeable processes which are functions of stationary Markov chains*, Ann. Math. Statist. **35** (1964), 429–430. MR MR0161370 (28 #4577)
26. ———, *A characterisation of a class of functions of finite Markov chains*, Ann. Math. Statist. **36** (1965), 524–528. MR MR0172333 (30 #2552)
27. ———, *Splitting a single state of a stationary process into Markovian states*, Ann. Math. Statist. **39** (1968), 1069–1077. MR MR0224154 (36 #7200)
28. S. W. Dharmadhikari and M. G. Nadkarni, *Some regular and non-regular functions of finite Markov chains*, Ann. Math. Statist. **41** (1970), 207–213. MR MR0263161 (41 #7766)
29. T. Downarowicz and R.D. Mauldin, *Some remarks on output measures*, Topology Applications **152** (2005), 11–25.
30. Tomasz Downarowicz and Jacek Serafin, *Fiber entropy and conditional variational principles in compact non-metrizable spaces*, Fund. Math. **172** (2002), no. 3, 217–247. MR MR1898686 (2003b:37027)
31. Martin H. Ellis, *Lumping states of an irreducible stationary Markov chain*, unpublished manuscript.
32. ———, *The $\bar{d}$-distance between two Markov processes cannot always be attained by a Markov joining*, Israel J. Math. **24** (1976), no. 3-4, 269–273. MR MR0414820 (54 #2912)
33. Yariv Ephraim and Neri Merhav, *Hidden Markov processes*, IEEE Trans. Inform. Theory **48** (2002), no. 6, 1518–1569, Special issue on Shannon theory: perspective, trends, and applications. MR MR1909472 (2003f:94024)
34. R. V. Erickson, *Functions of Markov chains*, Ann. Math. Statist. **41** (1970), 843–850. MR MR0264769 (41 #9360)
35. M. Fannes, B. Nachtergaele, and L. Slegers, *Functions of Markov processes and algebraic measures*, Rev. Math. Phys. **4** (1992), no. 1, 39–64. MR MR1160137 (93g:82010)
36. Martin Fox and Herman Rubin, *Functions of processes with Markovian states*, Ann. Math. Statist. **39** (1968), 938–946. MR MR0232450 (38 #775)
37. ———, *Functions of processes with Markovian states. II*, Ann. Math. Statist. **40** (1969), 865–869. MR MR0243607 (39 #4928)
38. ———, *Functions of processes with Markovian states. III*, Ann. Math. Statist. **41** (1970), 472–479. MR MR0258099 (41 #2746)
39. H. Furstenberg, *Stationary Processes and Prediction Theory*, Annals of Mathematics Studies, vol. 44, Princeton University Press, 1960.
40. E. J. Gilbert, *On the identifiability problem for functions of finite Markov chains*, Ann. Math. Stat. **30** (1959), 688–697.

41. Leonid Gurvits and James Ledoux, *Markov property for a function of a Markov chain: a linear algebra approach*, Linear Algebra Appl. **404** (2005), 85–117. MR MR2149655 (2006g:60108)
42. J. Hachigian and M. Rosenblatt, *Functions of reversible Markov processes that are Markovian*, J. Math. Mech. **11** (1962), 951–960. MR MR0145588 (26 #3118)
43. Olle Häggström, *Is the fuzzy Potts model Gibbsian?*, Ann. Inst. H. Poincaré Probab. Statist. **39** (2003), no. 5, 891–917. MR MR1997217 (2005f:82049)
44. Guangyue Han and Brian Marcus, *Analyticity of entropy rate of hidden Markov chains*, IEEE Trans. Inform. Theory **52** (2006), no. 12, 5251–5266. MR MR2300690 (2007m:62008)
45. ______, *Derivatives of entropy rate in special families of hidden Markov chains*, IEEE Trans. Inform. Theory **53** (2007), no. 7, 2642–2652. MR MR2319402 (2008b:94023)
46. G. Hansel and D. Perrin, *Rational probability measures*, Theor. Comp. Sci. **65** (1989), 171–188.
47. T. E. Harris, *On chains of infinite order*, Pacific J. Math. **5** (1955), 707–724. MR MR0075482 (17,755b)
48. A. Heller, *On stochastic processes derived from Markov chains*, Ann. Math. Stat. **36** (1965), 1286–1291.
49. Alex Heller, *Probabilistic automata and stochastic transformations*, Math. Systems Theory **1** (1967), 197–208. MR MR0235926 (38 #4227)
50. Paul W. Holland, *Some properties of an algebraic representation of stochastic processes*, Ann. Math. Statist. **39** (1968), 164–170. MR MR0221574 (36 #4626)
51. Yasuyoshi Inagaki, Teruo Fukumura, and Hiroyuki Matuura, *Some aspects of linear space automata*, Information and Control **20** (1972), 439–479. MR MR0332402 (48 #10729)
52. Robert B. Israel, *Convexity in the Theory of Lattice Gases*, Princeton University Press, Princeton, N.J., 1979, Princeton Series in Physics, With an introduction by Arthur S. Wightman. MR MR517873 (80i:82002)
53. F. P. Kelly, *Markovian functions of a Markov chain*, Sankhyā Ser. A **44** (1982), no. 3, 372–379. MR MR705461 (85d:60129)
54. John G. Kemeny and J. Laurie Snell, *Finite Markov Chains*, Springer-Verlag, New York, 1976, Reprinting of the 1960 original, Undergraduate Texts in Mathematics. MR MR0410929 (53 #14670)
55. Bruce Kitchens, *An invariant for continuous factors of markov shifts*, Proc.Amer.Math.Soc. **83** (1981), 825–828. MR MR0630029 (82k:28021)
56. ______, *Linear algebra and subshifts of finite type*, Conference in modern analysis and probability (New Haven, Conn., 1982), Contemp. Math., vol. 26, Amer. Math. Soc., Providence, RI, 1984, pp. 231–248. MR MR737405 (85m:28022)
57. Bruce Kitchens and Selim Tuncel, *Finitary measures for subshifts of finite type and sofic systems*, Mem. Amer. Math. Soc. **58** (1985), no. 338, iv+68. MR MR818917 (87h:58110)
58. Bruce P. Kitchens, *Symbolic Dynamics*, Universitext, Springer-Verlag, Berlin, 1998, One-sided, two-sided and countable state Markov shifts. MR MR1484730 (98k:58079)
59. S. C. Kleene, *Representation of events in nerve nets and finite automata*, Automata Studies (C. E. Shannon and J. McCarthy, eds.), Princeton University Press, 1956, pp. 3–42.
60. Yasuo Komota and Masayuki Kimura, *A characterization of the class of structurally stable probabilistic automata. I. Discrete-time case*, Internat. J. Systems Sci. **9** (1978), no. 4, 369–394. MR MR0490599 (58 #9937a)
61. ______, *A characterization of the class of structurally stable probabilistic automata. II. Continuous-time case*, Internat. J. Systems Sci. **9** (1978), no. 4, 395–424. MR MR0490600 (58 #9937b)
62. ______, *On Markov chains generated by Markovian controlled Markov systems: structural stability*, Internat. J. Systems Sci. **12** (1981), no. 7, 835–854. MR MR626281 (83c:93060)
63. Hans Künsch, Stuart Geman, and Athanasios Kehagias, *Hidden Markov random fields*, Ann. Appl. Probab. **5** (1995), no. 3, 577–602. MR MR1359820 (97a:60070)
64. François Ledrappier and Peter Walters, *A relativised variational principle for continuous transformations*, J. London Math. Soc. (2) **16** (1977), no. 3, 568–576. MR MR0476995 (57 #16540)
65. Frederick W. Leysieffer, *Functions of finite Markov chains*, Ann. Math. Statist. **38** (1967), 206–212. MR MR0207043 (34 #6859)

66. Douglas Lind and Brian Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, Cambridge, 1995. MR MR1369092 (97a:58050)
67. Richard W. Madsen, *Decidability of $\alpha(P^k) > 0$ for some k*, J. Appl. Probability **12** (1975), 333–340. MR MR0373011 (51 #9213)
68. Christian Maes and Koen Vande Velde, *The fuzzy Potts model*, J. Phys. A **28** (1995), no. 15, 4261–4270. MR MR1351929 (96i:82022)
69. B. Marcus, K. Petersen, and S. Williams, *Transmission rates and factors of Markov chains*, Contemp. Math. **26** (1984), 279–293.
70. Masakazu Nasu, *An invariant for bounded-to-one factor maps between transitive sofic subshifts*, Ergodic Theory Dynam. Systems **5** (1985), no. 1, 89–105. MR MR782790 (86i:28030)
71. David L. Neuhoff and Paul C. Shields, *Indecomposable finite state channels and primitive approximation*, IEEE Trans. Inform. Theory **28** (1982), no. 1, 11–18. MR MR651096 (83k:94025)
72. William Parry, *Intrinsic Markov chains*, Trans. Amer. Math. Soc. **112** (1964), 55–66. MR MR0161372 (28 #4579)
73. William Parry and Selim Tuncel, *Classification Problems in Ergodic Theory*, London Mathematical Society Lecture Note Series, vol. 67, Cambridge University Press, Cambridge, 1982, Statistics: Textbooks and Monographs, 41. MR MR666871 (84g:28024)
74. ______, *On the stochastic and topological structure of Markov chains*, Bull. London Math. Soc. **14** (1982), no. 1, 16–27. MR MR642417 (84i:28024)
75. A. Paz, *Word functions of pseudo-Markov chains*, Linear Algebra and Appl. **10** (1975), 1–5. MR MR0388543 (52 #9379)
76. Azaria Paz, *Introduction to probabilistic automata*, Academic Press, New York-London, 1971. MR MR0289222
77. K. Petersen, *Symbolic Dynamics*, http://www.math.unc.edu/Faculty/petersen/m261s98.pdf, 1998.
78. Karl Petersen, *Ergodic Theory*, Cambridge Studies in Advanced Mathematics, vol. 2, Cambridge University Press, Cambridge, 1989, Corrected reprint of the 1983 original. MR MR1073173 (92c:28010)
79. Karl Petersen, Anthony Quas, and Sujin Shin, *Measures of maximal relative entropy*, Ergodic Theory Dynam. Systems **23** (2003), no. 1, 207–223. MR MR1971203 (2004b:37009)
80. Karl Petersen and Sujin Shin, *On the definition of relative pressure for factor maps on shifts of finite type*, Bull. London Math. Soc. **37** (2005), no. 4, 601–612. MR MR2143740 (2005m:37066)
81. Robert R. Phelps, *Unique equilibrium states*, Dynamics and Randomness (Santiago, 2000), Nonlinear Phenom. Complex Systems, vol. 7, Kluwer Acad. Publ., Dordrecht, 2002, pp. 219–225. MR MR1975579 (2004c:28028)
82. Ian F. Putnam, *Lifting factor maps to resolving maps*, Israel J. Math. **146** (2005), 253–280. MR MR2151603 (2007i:37020)
83. M.O Rabin, *Probabilistic automata*, Information and Control **6** (1963), 230–245.
84. J. B. Robertson, *The mixing propoerties of certain processes related to Markov chains*, Math. Sys. Th **7** (1973), 39–43.
85. ______, *A spectral representation of the states of a measure preserving transformation*, Z. Wahr. verw. Geb. **27** (1973), 185–194.
86. Gerardo Rubino and Bruno Sericola, *On weak lumpability in Markov chains*, J. Appl. Probab. **26** (1989), no. 3, 446–457. MR MR1010934 (90j:60069)
87. ______, *A finite characterization of weak lumpable Markov processes. I. The discrete time case*, Stochastic Process. Appl. **38** (1991), no. 2, 195–204. MR MR1119981 (92g:60092)
88. A. Schönhuth, *Equations for hidden Markov models*, preprint, 2009.
89. M. P. Schützenberger, *On the definition of a family of automata*, Inf. and Control **4** (1961), 245–270.
90. Claude E. Shannon and Warren Weaver, *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, Ill., 1949. MR MR0032134 (11,258e)
91. Sujin Shin, *An example of a factor map without a saturated compensation function*, Ergodic Theory Dynam. Systems **21** (2001), no. 6, 1855–1866. MR MR1869074 (2002h:37020)
92. ______, *Measures that maximize weighted entropy for factor maps between subshifts of finite type*, Ergodic Theory Dynam. Systems **21** (2001), no. 4, 1249–1272. MR MR1849609 (2002i:37009)

93. ———, *Relative entropy functions for factor maps between subshifts*, Trans. Amer. Math. Soc. **358** (2006), no. 5, 2205–2216 (electronic). MR MR2197440 (2006i:37026)
94. Charles B. Silio, Jr., *An efficient simplex coverability algorithm in E^2 with application to stochastic sequential machines*, IEEE Trans. Comput. **28** (1979), no. 2, 109–120. MR MR519218 (80d:68069)
95. Selim Tuncel, *Conditional pressure and coding*, Israel J. Math. **39** (1981), no. 1-2, 101–112. MR MR617293 (82j:28012)
96. Peter Walters, *An Introduction to Ergodic Theory*, Graduate Texts in Mathematics, vol. 79, Springer-Verlag, New York, 1982. MR MR648108 (84e:28017)
97. ———, *Relative pressure, relative equilibrium states, compensation functions and many-to-one codes between subshifts*, Trans. Amer. Math. Soc. **296** (1986), no. 1, 1–31. MR MR837796 (87j:28028)
98. Benjamin Weiss, *Subshifts of finite type and sofic systems*, Monatsh. Math. **77** (1973), 462–474. MR MR0340556 (49 #5308)
99. Y. Yayama, *Existence of a measurable saturated compensation function between subshifts and its applications*, preprint, 2009.
100. Jisang Yoo, *Measures of maximal relative entropy with full support*, preprint (2009).

Department of Mathematics, University of Maryland, College Park, MD 20742-4015 USA

E-mail address: mmb@math.umd.edu

Department of Mathematics, CB 3250, Phillips Hall, University of North Carolina, Chapel Hill, NC 27599 USA

E-mail address: petersen@math.unc.edu