Introduction

This volume is a collection of papers on hidden Markov processes (HMPs) involving connections with symbolic dynamics and statistical mechanics. The subject was the focus of a five-day workshop held at the Banff International Research Station (BIRS) in October 2007, which brought together thirty mathematicians, computer scientists, and electrical engineers from institutions throughout the world. Most of the papers in this volume are based either on work presented at the workshop or on problems posed at the workshop.

From one point of view, an HMP is a stochastic process obtained as the noisy observation process of a finite-state Markov chain; a simple example is a binary Markov chain observed in binary symmetric noise, i.e., each symbol (0 or 1) in a binary state sequence generated by a two-state Markov chain may be flipped with some small probability, independently from time instant to time instant. In another (essentially equivalent) viewpoint, an HMP is a process obtained from a finite-state Markov chain by partitioning its state set into groups and completely "hiding" the distinction among states within each group; more precisely, there is a deterministic function on the states of the Markov chain, and the HMP is the process obtained by observing the sequences of function values rather than sequences of states (and hence such a process is sometimes called a "function of a Markov chain").

HMPs are encountered in an enormous variety of applications involving phenomena observed in the presence of noise. These range from speech and optical character recognition, through target tracking, to biomolecular sequence analysis. HMPs are also important objects of study in their own right in many areas of pure and applied mathematics, including information theory, probability theory, and dynamical systems. An excellent survey of HMPs can be found in [3].

A central problem in the subject is computation of the entropy rate (sometimes known simply as entropy) of an HMP. The entropy rate of a process can

be regarded as the asymptotic exponential growth rate of the number of different sequences that can be generated by the process (after one discards certain sequences of abnormally small probability). The entropy rate is a measure of randomness of a process. It is one of the most fundamental concepts in information theory, because it also measures the incompressibility of a process. It is also closely related to important quantities in statistical mechanics [4].

There is a very simple closed-form formula for the entropy rate of a finite-state Markov chain, but there is no such simple formula for entropy rate of HMPs except in very special cases. However, more than fifty years ago Blackwell [2] discovered an expression for the entropy rate of an HMP as the integral of a very simple integrand with respect to a typically very complicated (and usually singular) measure on a simplex; his measure is the stationary measure for a continuous-valued Markov chain on the simplex. In some sense, Blackwell's formula demonstrates that computation of entropy rate of HMPs is intrinsically complicated. While his formula has been used to help estimate the entropy rate, it is primarily of theoretical interest. Shortly after publication of Blackwell's paper, Birch [1] discovered excellent general upper and lower bounds on the entropy rate. However, until recently, there had been very little progress, with only a few papers on the subject.

Closely related is the problem of computing the capacity of an information channel, especially a channel with memory. Roughly speaking, the capacity of a channel is defined as the maximum of a quantity known as mutual information rate between the input and corresponding output processes, over all possible input processes. For some channels, this amounts to maximizing the entropy rate of the output process, which would be an HMP if the input process were Markov.

Recently, the entropy rate problem and related problems have received a good deal of attention from people working in many different areas, primarily information theory, dynamical systems, statistical mechanics, and probability theory. In particular, there is considerable interest in symbolic dynamics on problems regarding the properties of images and pre-images of Markov and hidden Markov processes via factor maps between symbolic dynamical systems. These issues have also been studied in the wider context of Gibbs measures.

The papers in this volume address these and related themes.

Computation of entropy rate is the explicit focus of the papers by Ordentlich and Weissman, Peres and Quas, and Pollicott. Ordentlich and Weissman develop an alternative to Blackwell's continuous-valued Markov chain and use it to obtain improved bounds in various noise regimes over the binary symmetric channel; they also compare various approximation schemes via an analysis of complexity versus precision. Peres and Quas obtain explicit asymptotics

for the entropy rate of HMPs obtained as the noisy observation processes of Markov chains observed in a certain noise regime ("rare transitions") over the binary symmetric channel, thereby solving an open problem posed in the workshop. Pollicott develops a new numerical technique for approximating entropy rate, using ideas from dynamical systems and statistical mechanics to obtain approximations which are provably superexponentially convergent in several cases.

The paper by Han, Marcus, and Peres develops a complex version of the Hilbert metric on the real simplex in order to obtain estimates of the domain of analyticity of entropy rate as a function of the underlying Markov chain transition probabilities.

Pfister focuses on computation of capacity for certain finite-state channels. He develops a formula for the derivative of entropy rate as a function of the underlying Markov chain and applies this to obtain estimates on mutual information rates for the channels.

Boyle and Petersen give an in-depth survey of results on hidden Markov processes relating to symbolic dynamics and connections with probability, automata theory, and thermodynamics. The survey contains many results and open problems regarding hidden Markov processes and factor maps. Chazottes and Ugalde show that under certain factor maps the image of every Gibbs measure, defined by a certain type of potential function Φ, is also a Gibbs measure, with a potential function of a type determined by regularity properties of Φ and the factor map. Pollicott and Kempton obtain similar results for Gibbs measures defined by a related class of potential functions. Verbitskiy explores the relationship between HMPs and the thermodynamic formalism. He surveys work on the problem of computing the decay rate of the conditional probability of the present given the past, computation of entropy rate, identification of a potential for a given measure known to be Gibbs, and relations to Markov random field models.

We thank the Banff International Research Station and its constituent institutions for running and supporting the workshop; the anonymous referees for their careful reading and reviewing of the papers; and the staff of Cambridge University Press for their expert handling of the publication.

References

[1] J. J. Birch. *Approximations for the entropy for functions of Markov chains*. Ann. Math. Statist. **33** (1962) 930–938

[2] D. Blackwell. *The entropy of functions of finite-state Markov chains*. In Trans. First Prague Conf. Information Theory, Statistical Decision Functions, Random Processes, 1957, pp. 13–20

[3] Y. Ephraim and N. Merhav. *Hidden Markov processes*. IEEE Trans. Inf. Theory **48** (2002) 1518–1569

[4] D. Ruelle. *Thermodynamic Formalism*: *The Mathematical Structures of Classical Equilibrium Statistical Mechanics*. Advanced Book Program, Addison-Wesley, Reading, MA, 1978

1

Hidden Markov processes in the context of symbolic dynamics

MIKE BOYLE

Department of Mathematics, University of Maryland, College Park, MD 20742-4015, USA
E-mail address: mmb@math.umd.edu

KARL PETERSEN

Department of Mathematics, CB 3250, Phillips Hall, University of North Carolina, Chapel Hill, NC 27599, USA
E-mail address: petersen@math.unc.edu

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.

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 [47, 56, 55, 71].

Entropy of Hidden Markov Processes and Connections to Dynamical Systems: Papers from the Banff International Research Station Workshop, ed. B. Marcus, K. Petersen, and T. Weissman. Published by Cambridge University Press.

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.

$$\mathcal{A}^* = \bigcup_{k=0}^{\infty} \mathcal{A}^k \tag{1}$$

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 sets 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 $x \neq y$

$$k(x,y) = \min\{|j| : x_j \neq y_j\} \quad \text{and} \quad d(x,y) = e^{-k(x,y)}. \tag{2}$$

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

$$\mathcal{C}_0(w) = \{y \in \Omega(\mathcal{A}) : y[0, n-1] = \omega\} \tag{3}$$

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

$$\mathcal{C}_i(\omega) = \{y \in \Omega(\mathcal{A}) : y[i, i+n-1] = \omega\}. \tag{4}$$

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 article, 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 article, 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

$$x[i, i+k-1] = x_i x_{i+1} \cdots x_{i+k-1} \tag{5}$$

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

$$\mathcal{L}(X) = \{w \in \mathcal{A}^* : \text{there are } n \in \mathbb{Z}, y \in X \text{ such that } w = y_n \cdots y_{n+|w|-1}\}. \tag{6}$$

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

$$X = \{x \in \Omega_d : \text{for each } i \le j, x_i x_{i+1} \cdots x_j \notin \mathcal{F}(X)\}. \tag{7}$$

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 the 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 $\mathcal{A} = \{0, 1, \ldots, d-1\}$ and entries from $\{0, 1\}$. Define

$$\Omega_M = \{\omega \in \mathcal{A}^{\mathbb{Z}} : \text{for all } n \in \mathbb{Z}, M(\omega_n, \omega_{n+1}) = 1\}. \tag{8}$$

These were called *topological Markov chains* by Parry [51]. 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 article.) 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

$$(\pi x)_i = \pi(x_{i-m} \cdots x_i \cdots x_{i+n}). \tag{9}$$

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) [33] *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

$$(\pi x)_i = x[i + r, i + r + 1, \ldots, i + r + k - 1] \quad \text{for all } x \in X. \tag{10}$$

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 one-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 one-step stationary (σ-invariant) Markov measure μ on the shift of finite type Ω_M by

$$\begin{aligned} \mu(\mathcal{C}_i(\omega[i,j])) &= \mu\{y \in \Omega_M : y[i,j] = \omega_i\omega_{i+1}\cdots\omega_j\} \\ &= p(\omega_i)P(\omega_i,\omega_{i+1})\cdots P(\omega_{j-1},\omega_j) \end{aligned} \tag{11}$$

(by the Kolmogorov extension theorem [6, p. 3ff.]).

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,

$$\mu(\mathcal{C}_0(x[0,i])|\mathcal{C}_0(x[-j,-1])) = \mu(\mathcal{C}_0(x[0,i])|\mathcal{C}_0(x[-k,-1])). \tag{12}$$

A measure is one-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 one-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 one-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

$$(\pi\mu)(E) = \mu(\pi^{-1}E) \quad \text{for each measurable } E \subset Y. \tag{13}$$

(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 [23]), 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 [3]).

The definition of hidden Markov measure raises several questions.