Approximation of stationary processes by hidden Markov models

Finesso, L.; Grassi, A.; Spreij, P.

DOI
10.1007/s00498-010-0050-7

Publication date
2010

Document Version
Final published version

Published in
Mathematics of control, signals, and systems

Math. Control Signals Syst. (2010) 22:1–22
DOI 10.1007/s00498-010-0050-7

ORIGINAL ARTICLE

Approximation of stationary processes by hidden Markov models

Lorenzo Finesso · Angela Grassi · Peter Spreij

Received: 24 June 2006 / Accepted: 26 June 2010 / Published online: 11 July 2010

Abstract Stochastic realization is still an open problem for the class of hidden Markov models (HMM): given the law Q of an HMM find a finite parametric description of it. Fifty years after the introduction of HMMs, no computationally effective realization algorithm has been proposed. In this paper we direct our attention to an approximate version of the stochastic realization problem for HMMs. We aim at the realization of an HMM of assigned complexity (number of states of the underlying Markov chain) which best approximates, in Kullback Leibler divergence rate, a given stationary law Q. In the special case of Q being the law of an HMM this corresponds to solving the approximate realization problem for HMMs. In general there is no closed form expression of the Kullback Leibler divergence rate, therefore we replace it, as approximation criterion, with the informational divergence between the Hankel matrices of the processes. This not only has the advantage of being easy to compute, while providing a good approximation of the divergence rate, but also makes the problem amenable to the use of nonnegative matrix factorization (NMF) techniques. We propose a three step algorithm, based on the NMF, which realizes an optimal HMM. The viability of the algorithm as a practical tool is tested on a few examples of HMM order reduction.

Keywords Hidden Markov model · Approximation · Kullback–Leibler divergence · Divergence rate · Nonnegative matrix factorization · Hankel matrix

A. Grassi was supported by a grant of Regione Veneto (Azione Biotech 3—DGR 2017/03-07-07) to CNR-ISIB.

L. Finesso · A. Grassi
ISIB-CNR, Corso Stati Uniti 4, 35127 Padova, Italy

P. Spreij (✉)
Korteweg-de Vries Institute for Mathematics, Universiteit van Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
e-mail: spreij@uva.nl

1 Introduction

Hidden Markov models (HMM) are a simple, yet very rich, class of stochastic processes which has become ubiquitous in several areas of control, signals and systems. Let $Y = (Y_t)_{t\in\mathbb{N}}$ be a stationary stochastic process taking values in a finite set. Following [20,21], we call Y an HMM, if it is equal in law to the *output* process η of a *finite stochastic system* $\xi = (X, \eta)$. A (finite) stochastic system, to be formally defined later in the paper, has the property that the bivariate process ξ, also taking values in a finite set, is jointly Markov. Note that, as a consequence, Y has the same law as $f(\xi)$, where f is the projection on the second component of ξ, which shows that an HMM can always be represented as a deterministic function of a Markov chain. In principle, since functions of Markov chains generally are not Markovian, the HMM can exhibit complex dynamic behaviors with possibly long dependence on the past.

The probabilistic characterization of HMMs was first given by Heller [12] who solved the following problem. Among all finitely valued stationary processes Y, characterize those whose laws are HMMs. To some extent his elegant result is not satisfactory. Even if Y is known to be an HMM [12], it does not provide a procedure to solve the *realization problem*, that is to deduce from the law of Y, say Q, a finite stochastic system whose output law is equal to Q. Such a stochastic system is then called a (*weak*) *realization* of Y.

A realization is fully determined by a finite set of parameters, typically the transition probability matrix of the process ξ. The *size* of a realization is the cardinality of the state space of the Markov chain X. The problem of constructing realizations of HMMs starting from their law Q has attracted the attention of workers in the area of Stochastic Realization Theory. An early reference is Picci and Van Schuppen [21], see also Anderson [1]. More recent references with related results are Vidyasagar [24] and Vanluyten et al. [23]. While some of the issues have been clarified, a constructive and computationally feasible realization algorithm, producing the parameters of a realization, is still missing.

In the present paper we focus on the simpler *approximate* realization problem which can be roughly formulated as follows. Given the law Q of an HMM, find a realization of assigned size, which best approximates Q in divergence rate, a most natural criterion from a statistical perspective. Unfortunately there exists no general, closed form, analytic expression of the divergence rate. To obviate this difficulty we reformulate the criterion in terms of the informational divergence between some positive matrices, representing the finite dimensional distributions of the processes. The approximate realization problem becomes then amenable to the use of nonnegative matrix factorization (NMF) techniques. Specifically we propose a three step, NMF based, optimization procedure to construct the parameters of the best approximate realization. The advantage of this approach is that it can also be used, in principle, to approximate any given stationary law Q by that of an HMM.

The remainder of the paper is organized as follows. Section 2 contains preliminaries on HMMs. In Sect. 3 the Hankel matrix of finite dimensional distributions of stationary processes is introduced. Section 4 establishes the existence of the divergence rate between a stationary process and an HMM. In Sect. 5 the realization problem is posed, as well as an approximate version of it, in terms of divergence rate. In Sect. 6 an

algorithm to find the best approximation is proposed. The concluding Sect. 7 contains numerical examples.

This paper develops and extends some preliminary ideas presented in [8].

2 Preliminaries on HMMs

Let $(Y_t)_{t\in\mathbb{N}}$ be a discrete time stationary stochastic process defined on a given probability space $\{\Omega, \mathcal{A}, P\}$ and with values in the finite set (alphabet) $\mathcal{Y}$. Denote by $\mathcal{Y}^*$ the set of all finite strings of symbols from the alphabet $\mathcal{Y}$, with the addition of the empty string ϕ. For any $v \in \mathcal{Y}^*$, let $|v|$ be the length of the string v. By convention $|\phi| = 0$. If $u, v \in \mathcal{Y}^*$, denote by uv the string obtained by concatenation of v to u.

For any $n \in \mathbb{N}$, let $\mathcal{Y}^n$ be the set of all strings of length n, with the obvious inclusion $\mathcal{Y}^n \subset \mathcal{Y}^*$. We denote by $Y_{t+1}^+ = (Y_{t+1}, Y_{t+2}, \ldots)$ the future of the process Y after t and by $Y_t^- = (\ldots, Y_{t-1}, Y_t)$ the past of the process Y up to t. The event $(Y_s, \ldots, Y_t) = v$ is represented by $Y_s^t = v$, for any $v \in \mathcal{Y}^*$ with $|v| = t - s + 1$. By convention $\{Y_t^+ = \phi\} = \{Y_t^- = \phi\} = \Omega$. For any $v \in \mathcal{Y}^*$ we use $Y_{t+1}^+ = v$ as a shorthand notation for the event $Y_{t+1}^{t+|v|} = v$. Since Y is stationary, the probability distribution of the sequence Y_t^+ is independent of t and it induces a map $p : \mathcal{Y}^* \to [0, 1]$ with the following properties

(a) $p(v) = P(Y_t^+ = v) \quad \forall v \in \mathcal{Y}^*$
(b) $p(\phi) = 1$
(c) $0 \le p(v) \le 1 \quad \forall v \in \mathcal{Y}^*$
(d) $\sum_{v\in\mathcal{Y}^n} p(uv) = p(u) \quad \forall u \in \mathcal{Y}^* \quad \forall n \in \mathbb{N}.$

The map p represents the finite dimensional probability distributions of the process Y, sometimes referred to as *pdf*.

All the hidden Markov models considered in this paper will be in discrete time and with values in a finite set. The basic definitions are taken from [20], to which the reader is referred for detailed derivations.

Definition 2.1 A pair $(X, Y) = (X_t, Y_t)_{t\in\mathbb{N}}$ of stochastic processes taking values in the finite set $\mathcal{X} \times \mathcal{Y}$ is said to be a *stationary finite stochastic system* (SFSS) if

(i) (X, Y) is jointly stationary,
(ii) for all $t \in \mathbb{N}, \sigma \in \mathcal{X}^*, v \in \mathcal{Y}^*$ it holds that

$$P\left(Y_{t+1}^+ = v, X_{t+1}^+ = \sigma | X_t^-, Y_t^-\right) = P\left(Y_{t+1}^+ = v, X_{t+1}^+ = \sigma | X_t\right). \quad (1)$$

The processes X and Y are called, respectively, the *state* and the *output* of the SFSS. The cardinalities of $\mathcal{X}$ and $\mathcal{Y}$ are denoted by N and m, respectively.

To the best of our knowledge Definition 2.1 goes back to [4], where it was given in an input/output context not needed here. From property (1) it follows immediately that

1. (X, Y) is a Markov chain.
2. X is a Markov chain.

3. The past and the future of Y at time t are conditionally independent given X_t, i.e. for all $t \in \mathbb{N}$ and $v \in \mathcal{Y}^*$

$$P\left(Y_{t+1}^{+} = v | X_t, Y_t^{-}\right) = P\left(Y_{t+1}^{+} = v | X_t\right). \tag{2}$$

Definition 2.2 A stochastic process $Y = (Y_t)_{t\in\mathbb{N}}$ with values in $\mathcal{Y}$ is a HMM, if it is equal in law to the output of a SFSS. Any such SFSS is called a representation of Y. The cardinality N of $\mathcal{X}$ is called *size* of the representation. The smallest N for which a representation exists is called *order* of the HMM.

Remark 2.3 If Y is an HMM, equal in law to the output of a SFSS, one can always replace the latter with Y, when probabilities in terms of Y are to be computed. This convention is followed in the remainder of the paper.

The probability distribution of a stationary HMM is specified by

- the m nonnegative matrices $\{M(y), y \in \mathcal{Y}\}$ of size $N \times N$ with elements

$$m_{ij}(y) = P(Y_{t+1} = y, X_{t+1} = j \mid X_t = i), \tag{3}$$

- a probability (row) vector π of size N, such that $\pi = \pi A$, where

$$A := \sum_y M(y).$$

The matrix A is the transition matrix of the Markov chain X and π is an invariant vector of A. Since the state space $\mathcal{X}$ is finite, A always admits an invariant vector, see [19], which is unique if A is irreducible.

Definition (3) extends to strings $v \in \mathcal{Y}^*$ as follows.

Definition 2.4 Let v be a string in $\mathcal{Y}^*$ of arbitrary length, k say. Then $M(v) \in \mathbb{R}_+^{N\times N}$ is defined by

$$m_{ij}(v) = P\left(Y_{t+1}^{t+k} = v, X_{t+k} = j | X_t = i\right).$$

An immediate consequence of (1) is that the following semigroup property holds

$$M(uv) = M(u)M(v) \quad \forall u, v \in \mathcal{Y}^*.$$

Let $w \in \mathcal{Y}^*$, then $p(w) = \pi M(w)e$, where $e = (1, \ldots, 1)^\top \in \mathbb{R}^N$. For any pair of strings u and v in $\mathcal{Y}^*$, one then has

$$p(uv) = \pi M(u)M(v)e. \tag{4}$$

and if $w = y_1 \cdots y_n$, then

$$p(w) = \pi M(y_1) \cdots M(y_n)e. \tag{5}$$

The *factorization hypothesis*

$$\begin{aligned} &P(Y_{t+1} = y, X_{t+1} = j \mid X_t = i) \\ &\quad = P(Y_{t+1} = y \mid X_{t+1} = j) P(X_{t+1} = j \mid X_t = i), \quad \forall t, y, i, j \end{aligned} \tag{6}$$

is widely used in the signal processing literature, see [22]. Under (6) it is possible to reparametrize the pdf. Define the *readout* matrix $B \in \mathbb{R}_+^{N \times m}$ with elements

$$b_{iy} := P(Y_t = y \mid X_t = i)$$

and the diagonal matrices

$$B_y := \operatorname{diag}\{b_{1y}, b_{2y}, \ldots, b_{Ny}\}. \tag{7}$$

The factorization hypothesis then reads

$$M(y) = AB_y, \tag{8}$$

from which (5) turns into the classical Baum formula, see [2],

$$p(w) = \pi A B_{y_1} \cdots A B_{y_n} e. \tag{9}$$

Note that if $Y = f(X)$, a deterministic function of X, then $b_{iy} \in \{0, 1\}$ with $b_{iy} = 1$ iff $f(i) = y$ and (6) holds. As it was recalled in the Introduction, enlarging the size it is always possible to represent an HMM as a deterministic function of an MC. The factorization hypothesis therefore is not restrictive, in principle, but it is not always desirable to work with HMMs of large size.

3 Hankel matrices

The Hankel matrix **H** of stationary process is a matricial representation of its finite dimensional distributions. In the special case of HMMs **H** has positive factorization properties which will be instrumental for the construction of the approximate realizations in Sect. 5.

3.1 Hankel matrix of a stationary process

Following [1], define for a given $n \in \mathbb{N}$ two ordered sets, listing the strings of $\mathcal{Y}^n$. The ordered set $\mathcal{Y}^n_{flo}$ lists the strings in *first lexical order* (*flo*), i.e., lexicographically reading from right to left. The ordered set $\mathcal{Y}^n_{llo}$ lists the strings in *last lexical order* (*llo*), i.e. lexicographically reading from left to right. For $\mathcal{Y} = \{0, 1\}$ and $n = 2$ the orders are $\mathcal{Y}^2_{flo} = (00, 10, 01, 11)$ and $\mathcal{Y}^2_{llo} = (00, 01, 10, 11)$.

The *flo* induces a complete enumeration of $\mathcal{Y}^*$, denoted by $\mathcal{Y}^*_{flo}$, which is obtained by first listing the empty string, followed by the strings of $\mathcal{Y}^1_{flo}$, followed by the

strings of $\mathcal{Y}_{flo}^2$ and so on. In a similar way one constructs $\mathcal{Y}_{llo}^*$. For $\mathcal{Y} = \{0, 1\}$ the two enumerations are

$$\mathcal{Y}_{flo}^* = (\phi, 0, 1, 00, 10, 01, 11, 000, 100, 010, 110, 001, 101, 011, 111, \ldots)$$

and

$$\mathcal{Y}_{llo}^* = (\phi, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, 110, 111, \ldots).$$

Definition 3.1 For a stationary process with pdf p the Hankel matrix $\mathbf{H}$ is the infinite matrix with elements $p(u_\alpha v_\beta)$, where u_α and v_β run through $\mathcal{Y}_{flo}^*$ and $\mathcal{Y}_{llo}^*$, respectively.

As an example the upper left corner of $\mathbf{H}$ is written below, for the case of a binary process. To improve readability the sequences u_α and v_β are displayed along the borders.

	ϕ	0	1	00	01	10	11	$\cdots$
ϕ	1	$p(0)$	$p(1)$	$p(00)$	$p(01)$	$p(10)$	$p(11)$	$\cdots$
0	$p(0)$	$p(00)$	$p(01)$	$p(000)$	$p(001)$	$p(010)$	$p(011)$	$\cdots$
1	$p(1)$	$p(10)$	$p(11)$	$p(100)$	$p(101)$	$p(110)$	$p(111)$	$\cdots$
00	$p(00)$	$p(000)$	$p(001)$	$p(0000)$	$p(0001)$	$p(0010)$	$p(0011)$	$\cdots$
10	$p(10)$	$p(100)$	$p(101)$	$p(1000)$	$p(1001)$	$p(1010)$	$p(1011)$	$\cdots$
01	$p(01)$	$p(010)$	$p(011)$	$p(0100)$	$p(0101)$	$p(0110)$	$p(0111)$	$\cdots$
11	$p(11)$	$p(110)$	$p(111)$	$p(1100)$	$p(1101)$	$p(1110)$	$p(1111)$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

The example clearly shows the natural block structure of the $\mathbf{H}$ matrix. For fixed integers $K \geq 0$ and $L \geq 0$ the (K, L) block of $\mathbf{H}$ is

$$\mathbf{H}_{KL} := \|p(u_i v_j)\|, \tag{10}$$

a matrix of size $m^K \times m^L$, where u_i, $i = 1, \ldots, \gamma = m^K$ and v_j, $j = 1, \ldots, \delta = m^L$ run through $\mathcal{Y}_{flo}^K$ and $\mathcal{Y}_{llo}^L$, respectively.

The matrix $\mathbf{H}$ can be partitioned as

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{00} & \mathbf{H}_{01} & \cdots & \mathbf{H}_{0L} & \cdots \\ \mathbf{H}_{10} & \mathbf{H}_{11} & \cdots & \mathbf{H}_{1L} & \cdots \\ \vdots & \vdots & & \vdots & \\ \mathbf{H}_{K0} & \mathbf{H}_{K1} & \cdots & \mathbf{H}_{KL} & \cdots \\ \vdots & \vdots & & \vdots & \ddots \end{bmatrix}. \tag{11}$$

As the reader can readily see, the antidiagonal blocks $\mathbf{H}_{KL}$ (with $K + L$ constant) contain the same probabilities, although reshuffled. With abuse of language $\mathbf{H}$ is called a (block) Hankel matrix, although in a standard block Hankel matrix $\mathbf{H}_{KL}$ is *constant* along the antidiagonals.

Let $\mathcal{Y}^1_{flo} = \mathcal{Y}^1_{llo} = (y_1, \dots y_m)$. Because of the columns enumeration scheme, the block $\mathbf{H}_{K,L+1}$ of size $m^K \times m^{L+1}$ can be written as

$$\mathbf{H}_{K,L+1} = [\,\mathbf{H}_{KL}(y_1)\ \mathbf{H}_{KL}(y_2) \cdots \mathbf{H}_{KL}(y_m)\,], \tag{12}$$

where $\mathbf{H}_{KL}(y)$ is defined as

$$\mathbf{H}_{KL}(y) = \|p(u_i\ y\ v_j)\|_{i=1,\dots,\gamma,\, j=1,\dots,\delta}. \tag{13}$$

3.2 Hankel matrix of an HMM

The Hankel matrix $\mathbf{H}$ of a stationary HMM, and its blocks $\mathbf{H}_{KL}$, have special properties in two respects; they can be factored into smaller, positive matrices and there are recursive relations between neighboring blocks and block factors. We collect below, following [1] and introducing new ones as needed, the results that will be used later in the paper. All the properties stem from the basic formula (4)

$$p(u_i v_j) = \pi M(u_i) M(v_j) e.$$

Substituting (4) into (10) one gets the positive factorization of $\mathbf{H}_{KL}$,

$$\mathbf{H}_{KL} = \begin{bmatrix} \pi M(u_1) \\ \vdots \\ \pi M(u_\gamma) \end{bmatrix} \begin{bmatrix} M(v_1)e \cdots M(v_\delta)e \end{bmatrix} =: \mathbf{\Pi}_K \mathbf{\Gamma}_L, \tag{14}$$

where

$$\mathbf{\Pi}_K := \begin{bmatrix} \pi M(u_1) \\ \vdots \\ \pi M(u_\gamma) \end{bmatrix}, \quad \mathbf{\Gamma}_L := \begin{bmatrix} M(v_1)e \cdots M(v_\delta)e \end{bmatrix} \tag{15}$$

are matrices of sizes $m^K \times N$ and $N \times m^L$, respectively.

Comparing (11) with (14) one gets the positive factorization of $\mathbf{H}$,

$$\mathbf{H} = \begin{bmatrix} \mathbf{\Pi}_0 \\ \mathbf{\Pi}_1 \\ \vdots \\ \mathbf{\Pi}_K \\ \vdots \end{bmatrix} \begin{bmatrix} \mathbf{\Gamma}_0 \mathbf{\Gamma}_1 \cdots \mathbf{\Gamma}_L \cdots \end{bmatrix}.$$

Turning to the $\mathbf{H}_{KL}(y)$ matrices, note that their elements take the form

$$p\,(u_i\ y\ v_j) = \pi M(u_i) M(y\ v_j) e. \tag{16}$$

Substitution of (16) into (13) gives a positive factorizations of $\mathbf{H}_{KL}(y)$,

$$\mathbf{H}_{KL}(y) = \begin{bmatrix} \pi M(u_1) \\ \vdots \\ \pi M(u_\gamma) \end{bmatrix} \begin{bmatrix} M(y\, v_1)e \cdots M(y\, v_\delta)e \end{bmatrix} =: \mathbf{\Pi}_K\, \mathbf{\Gamma}_L(y),$$

where

$$\mathbf{\Gamma}_L(y) = \begin{bmatrix} M(y\, v_1)e \cdots M(y\, v_\delta)e \end{bmatrix} = M(y)\mathbf{\Gamma}_L. \tag{17}$$

There are several relations between neighboring blocks, and block factors. From definition (15), of $\mathbf{\Gamma}_L$, and the columns enumeration scheme one has

$$\mathbf{\Gamma}_{L+1} = \begin{bmatrix} \mathbf{\Gamma}_L(y_1) \cdots \mathbf{\Gamma}_L(y_m) \end{bmatrix}, \tag{18}$$

which, in view of (17), becomes

$$\mathbf{\Gamma}_{L+1} = \begin{bmatrix} M(y_1)\mathbf{\Gamma}_L \cdots M(y_m)\mathbf{\Gamma}_L \end{bmatrix}. \tag{19}$$

Defining the block matrices

$$\mathbf{M} := [M(y_1) \ldots M(y_m)], \quad \mathbf{\Gamma}_{(L)} := \begin{bmatrix} \mathbf{\Gamma}_L & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{\Gamma}_L \end{bmatrix}, \tag{20}$$

of sizes $N \times mN$ and $mN \times m^{L+1}$, respectively, we rewrite (19) compactly as

$$\mathbf{\Gamma}_{L+1} = \mathbf{M}\mathbf{\Gamma}_{(L)}. \tag{21}$$

4 Existence of the divergence rate

The divergence rate is the optimality criterion of choice for the approximation of stochastic processes. In this section, we review the definition of the divergence rate between processes, as previously given for instance in [13] for two HMMs, and show, under a technical condition, that the divergence rate between a stationary process and an HMM is well defined.

Consider a process Y with values in $\mathcal{Y}$ under two possibly different laws Q and P, probability measures on the path space $\mathcal{Y}^\infty$. Let $q(\cdot)$ and $p(\cdot)$ be the respective pdfs. For reasons of brevity, we write $q(Y_0^k)$ for the likelihood $q(Y_0, \ldots, Y_k)$ and likewise for $p(Y_0^k)$.

Definition 4.1 Let Q and P be probability measures on $\mathcal{Y}^\infty$. Define the (Kullback–Leibler) divergence rate of Q with respect to P as

$$D(Q\|P) := \lim_{n\to\infty} \frac{1}{n} E_Q \left[\log \frac{q\left(Y_0^{n-1}\right)}{p\left(Y_0^{n-1}\right)} \right] \tag{22}$$

if the limit exists and is finite.

The next theorem establishes, under some restrictions, that the divergence rate between a stationary process and a stationary HMM is well defined. The approach adopted for the proof is inspired by analogous results in [16] and [18], although the arguments given in [17], where the divergence rate between two HMMs is studied, could also be adapted. In the proof the following notation is used. If R is a set of real numbers, then $\min^+ R$ denotes the minimum of the strictly positive elements of R, if it exists, which is of course the case when R is finite and contains at least one positive number.

Theorem 4.2 *Let Y be a process with values in $\mathcal{Y}$. Let Q be an arbitrary stationary law of Y and P an HMM law. Assume that*

(i) *the distributions of all finite segments $(Y_0, \ldots, Y_{n-1})$ under Q are absolutely continuous with respect to those under P,*

(ii) *Q admits an invariant probability measure μ^* on $\mathcal{Y}$ i.e.*

$$\mu^*(y) = \sum_{y_0} Q(Y_1 = y|Y_0 = y_0)\mu^*(y_0),$$

(iii) *Y is geometrically ergodic under Q i.e., there exists $\rho \in (0, 1)$ such that*

$$|Q(Y_n = y|Y_0 = y_0) - Q(Y_n = y|Y_0 = y_0')| = O(\rho^n) \quad \forall y, y_0, y_0' \in \mathcal{Y}.$$

Then the limit in (22) exists and is finite.

The following technical lemma is needed for the proof of Theorem 4.2.

Lemma 4.3 *Under the assumptions of Theorem 4.2, there exists a constant $c \in (-\infty, 0)$ such that*

$$\lim_{n\to\infty} \frac{1}{n} \log p\left(Y_0^{n-1}\right) = c \quad Q\text{-a.s.} \tag{23}$$

Proof This is a special case of Proposition 4.3 of [18]. Assumption A of [18] is replaced with our assumptions (ii) and (iii). Assumption B of [18] plays no role in the present context. Assumption C of [18] can be dispensed with, since the alphabet is finite. □

Proof of Theorem 4.2 Rewrite (22) as the limit of

$$\frac{1}{n} E_Q \log q\left(Y_0^{n-1}\right) - \frac{1}{n} E_Q \log p\left(Y_0^{n-1}\right). \tag{24}$$

For the first term in (24) note that $-E_Q[\log q(Y_0^{n-1})]$ is the entropy of $q(Y_0^{n-1})$ and therefore $-\frac{1}{n}E_Q \log q(Y_0^{n-1})$ converges to $H(Q)$, the entropy rate of Q, which is finite, because of stationarity and the fact that $\mathcal{Y}$ is finite, see [10, Lemma 2.4.1]. Therefore it is sufficient to show that the second term in (24) has a finite limit. Let $y_0, \ldots, y_{n-1}$ be a string in $\mathcal{Y}^*$ with positive Q-probability. By absolute continuity, assumption (i), it also has positive P-probability. Since Y is an HMM under P, it follows from (5) that there are indices $i_0, \ldots, i_{n-1}$ such that

$$\pi_{i_0} m_{i_0 i_1}(y_0) \cdots m_{i_{n-1} i_n}(y_{n-1}) > 0.$$

Since the set R of all probabilities π_k and $m_{ij}(y)$ is finite, we have $\delta := \min^+ R > 0$. Hence, from the above displayed inequality, one concludes that $p(y_0^{n-1}) \geq \delta^{n+1}$. It follows that $p(Y_0^{n-1}) \geq \delta^{n+1}$ Q-a.s. and

$$\frac{n+1}{n} \log \delta \leq \frac{1}{n} \log p\left(Y_0^{n-1}\right) \leq 0 \quad Q\text{-a.s.}$$

Moreover, by Lemma 4.3

$$\lim_{n\to\infty} \frac{1}{n} \log p\left(Y_0^{n-1}\right) = c \quad Q\text{-a.s.}$$

Then the dominated convergence theorem can be applied to conclude that $\frac{1}{n}E_Q \log p(Y_0^{n-1})$ admits the finite limit c. □

Remark 4.4 It is possible to show a uniform version of Theorem 4.2, i.e., the uniform convergence of the divergence rate with respect to P, under more stringent conditions on the approximating model class. For details on a closely related problem we refer to [18], in particular Theorem 4.4.

A priori no extra information is available about the given stationary measure Q. Therefore it is useful to give simple conditions on the parameters $m_{ij}(y)$ of the HMM measure P to ensure the validity of the absolute continuity condition of Theorem 4.2 for any given stationary measure Q. One possibility is given in the following example.

Example 4.5 If Q is arbitrary, then in principle all probabilities $q(y_0^{n-1})$ can be strictly positive. The following sufficient condition entails that all $p(y_0^{n-1})$ are positive. Let Q and P be as in Theorem 4.2 with (i) replaced by

$$\sum_j m_{ij}(y) = P(Y_k = y | X_{k-1} = i) > 0, \quad \forall y \in \mathcal{Y}, \ \forall i \in \mathcal{X}. \tag{25}$$

We show that all finite strings have positive probability under P from which it follows that the limit in (22) exists. Let $\delta' = \min_{i,y} P(Y_k = y | X_{k-1} = i) > 0$, which is strictly

positive by (25). Then, for any $y \in \mathcal{Y}$

$$P\left(Y_k = y | Y_0^{k-1}\right) = \sum_i P\left(Y_k = y, X_{k-1} = i | Y_0^{k-1}\right)$$
$$= \sum_i P(Y_k = y | X_{k-1} = i) P\left(X_{k-1} = i | Y_0^{k-1}\right)$$
$$\geq \delta' \sum_i P\left(X_{k-1} = i | Y_0^{k-1}\right) = \delta'.$$

By iteration of this inequality applied to $p(y_0^{n-1}) = p(y_0) \prod_{k=1}^{n-1} p(y_k | y_0^{k-1})$, the result follows, since (25) also implies that $p(y_0) = \sum_{ij} \pi_i m_{ij}(y_0) > 0$.

Condition (25) may appear restrictive, but in absence of any additional knowledge about Q, one can not completely avoid it. To illustrate this, let us assume that Y is Markov under P. Since in principle all strings y_0^{n-1} may have positive Q-probability, the same must hold under P, which implies that all transition probabilities $A_{ij} > 0$, which is Condition (25). The existence of the divergence rate in this case is much easier to establish. Inspection of the proof of Theorem 4.2 reveals that the entropy term $-H(Q)$ remains, whereas one easily establishes by direct computation that

$$\frac{1}{n} E_Q \log p\left(Y_0^{n-1}\right) \to E_Q \log p(Y_1 | Y_0).$$

The condition $A_{ij} > 0$ guarantees finiteness of the divergence rate for arbitrary Q.

When additional information on Q is available, Condition (25) may be relaxed. For instance, if for some pair y_0, y_1 it is known that $q(y_0 y_1) = 0$, then $A_{y_0 y_1} = 0$ is allowed.

5 Approximate realization by HMMs

The weak stochastic realization problem for HMMs was formulated in [21] as follows. Let Y be an HMM whose law Q is known, e.g., via its pdf $q(\cdot)$. Find an SFSS (X, η) such that η has law Q. Any such SFSS is called a *(weak) realization* of Y. Note that the problem reduces to finding a set of parameters $M(y)$, such that $q(w) = \pi M(w) e$ for all $w \in \mathcal{Y}^*$, see Sect. 2.

A constructive approach to the solution of the realization problem has been proposed in [1], but an effective algorithm is still lacking. Replacing the hard constraint $q(w) = \pi M(w) e$ by an approximation criterion one can formulate a number of *approximate* realization problems. Specifically we are interested in finding an SFSS, of assigned size N, which best approximates, in divergence rate, the given HMM law Q. In view of the results of Sect. 4 it is apparent that the approximate realization of an HMM leads naturally to the more general problem of approximating a stationary law Q, which can be posed as follows.

Problem 5.1 Let Q, a stationary probability measure on $\mathcal{Y}^\infty$, and $N \in \mathbb{N}$ be given. Find a realization $\{M(y), y \in \mathcal{Y}\}$ of size N, of an HMM whose law P^* is closest to

Q in divergence rate, i.e., such that,

$$D(Q\|P^*) = \inf_P D(Q\|P), \tag{26}$$

where the infimum is taken over all HMM laws P corresponding to an SFSS of size N.

This problem is well defined under the conditions of Theorem 4.2, since the divergence rate is then guaranteed to exist.

Example 5.2 The minimization problem can be solved explicitly if P runs through the set of all stationary Markov laws, a subset of the HMM measures. Let P^* be such a law, defined by the transition probabilities

$$P^*(Y_{t+1} = j|Y_t = i) := Q(Y_{t+1} = j|Y_t = i). \tag{27}$$

A direct computation shows the *Pythagorean identity* [5]

$$D(Q\|P) - D(Q\|P^*) = D(P^*\|P),$$

which guarantees that P^* is the optimal approximating measure. A similar result holds for approximation by a k-step Markov chain. In [25] the Markov approximation problem has been analyzed in detail.

Unfortunately, such appealing closed form solutions do not exist if the minimization is carried out over stationary HMM measures. No analytic expression is known for the divergence rate, when Q is arbitrary and P a genuine HMM measure. The situation does not improve when Q is an HMM measure, see [14] for an interesting discussion and [11] for recent results. The simplest non trivial example of an information quantity computed for HMMs was given in [3], where the entropy rate is expressed as an infinite series. To obviate this difficulty we approximate the criterion of Problem 5.1 with one which, in principle, is amenable to numerical computation. Specifically we replace the divergence rate $D(Q\|P)$ between the processes with the *informational divergence* between the corresponding Hankel matrices.

For two nonnegative numbers q and p their *informational divergence* is defined as $D(q\|p) = q\log\frac{q}{p} - q + p$ with the conventions $0/0 = 0, 0\log 0 = 0$ and $q/0 = \infty$ for $q > 0$. From the inequality $x\log x \geq x - 1$ it follows that $D(q\|p) \geq 0$ with equality iff $q = p$.

Definition 5.3 Let $\mathbf{M}, \mathbf{N} \in \mathbb{R}_+^{m\times n}$. The informational divergence of $\mathbf{M}$ relative to $\mathbf{N}$ is

$$D(\mathbf{M}\|\mathbf{N}) = \sum_{ij} D(M_{ij}\|N_{ij}) = \sum_{ij}\left(M_{ij}\log\frac{M_{ij}}{N_{ij}} - M_{ij} + N_{ij}\right). \tag{28}$$

It follows that $D(\mathbf{M}\|\mathbf{N}) \geq 0$ with equality iff $M = N$. If $\sum_{ij} M_{ij} = \sum_{ij} N_{ij} = 1$, the informational divergence reduces to the usual Kullback–Leibler divergence between

probability distributions

$$D(\mathbf{M}\|\mathbf{N}) = \sum_{ij} M_{ij} \log \frac{M_{ij}}{N_{ij}}. \tag{29}$$

Let Q and P be measures as in Theorem 4.2 and denote by $\mathbf{H}_{nn}^{Q}$ and $\mathbf{H}_{nn}$ the (n, n) blocks of their respective Hankel matrices. For notational convenience here, and later in the paper, the Hankel blocks of the given measure Q will always carry the superscript Q, those of the variable measure P will have no superscript. For all u_i in $\mathcal{Y}_{flo}^{n}$ and v_j in $\mathcal{Y}_{llo}^{n}$ the corresponding element of $\mathbf{H}_{nn}^{Q}$ is

$$q^{(2n)}(u_i v_j) := Q\left(Y_0^{2n-1} = u_i v_j\right).$$

Analogously, a typical element of $\mathbf{H}_{nn}$ is

$$p^{(2n)}(u_i v_j) := P\left(Y_0^{2n-1} = u_i v_j\right).$$

The informational divergence between the Hankel blocks is

$$D\left(\mathbf{H}_{nn}^{Q}\|\mathbf{H}_{nn}\right) = \sum_{u_i, v_j \in \mathcal{Y}^n} q^{(2n)}(u_i v_j) \log \frac{q^{(2n)}(u_i v_j)}{p^{(2n)}(u_i v_j)} \tag{30}$$

$$= E_Q\left[\log \frac{q\left(Y_0^{2n-1}\right)}{p\left(Y_0^{2n-1}\right)}\right] \tag{31}$$

which, when compared to the definition of divergence rate, provides the following

Theorem 5.4 *Assume that Q and P are as in Theorem 4.2. Then*

$$\lim_{n\to\infty} \frac{1}{2n} D\left(\mathbf{H}_{nn}^{Q}\|\mathbf{H}_{nn}\right) = D(Q\|P). \tag{32}$$

Theorem 5.4 motivates the use of $\frac{1}{2n} D(\mathbf{H}_{nn}^{Q}\|\mathbf{H}_{nn})$, for n properly chosen, as an approximation of the theoretical divergence rate $D(Q\|P)$. The choice of n is critical. Intuition suggests that a good approximation might require n to be large, while computational efficiency does require n to be small as the size of the Hankel blocks increases exponentially with n. Comments on the choice of the size parameter n are deferred to Sect. 7.1.

6 Algorithm for the approximate realization

In view of Theorem 5.4 the approximation Problem 5.1 will now be reformulated in order to make it more amenable to numerical computations.

Problem 6.1 Let Q, a stationary probability measure on $\mathcal{Y}^\infty$, and $N \in \mathbb{N}$ be given. For given n, let $\mathbf{H}_{nn}^Q$ be the Hankel block of Q. Find a realization $\{M(y), y \in \mathcal{Y}\}$ of size N, of an HMM whose Hankel block $\mathbf{H}_{nn}^*$ is closest to $\mathbf{H}_{nn}^Q$ in informational divergence, i.e.,

$$D\left(\mathbf{H}_{nn}^Q \| \mathbf{H}_{nn}^*\right) = \min_{\mathbf{H}_{nn}} D\left(\mathbf{H}_{nn}^Q \| \mathbf{H}_{nn}\right), \tag{33}$$

where the minimum is taken over all $\mathbf{H}_{nn}$, Hankel blocks of HMM laws corresponding to an SFSS of size N.

By the positive factorization property (14) of the Hankel blocks of HMMs, the minimization (33) reduces to the following approximate *Nonnegative Matrix Factorization* (NMF) problem

$$\min_{\mathbf{\Pi}_n, \mathbf{\Gamma}_n} D\left(\mathbf{H}_{nn}^Q \| \mathbf{\Pi}_n \mathbf{\Gamma}_n\right), \tag{34}$$

under the constraints $\mathbf{\Pi}_n \geq 0, \mathbf{\Gamma}_n \geq 0, e^\top \mathbf{\Pi}_n e = 1$ and $\mathbf{\Gamma}_n e = e$. The necessity of these constraints follows from the definitions of the factors $\mathbf{\Pi}_n$ and $\mathbf{\Gamma}_n$

A minimizing nonnegative factorization $(\mathbf{\Pi}_n^*, \mathbf{\Gamma}_n^*)$ always exists, see [9], Proposition 2.1, but Problem 6.1 also calls for the construction of the corresponding parameters $M^*(y)$. The analysis of the ideal case will serve as a guide. If Q were an HMM law, the following *exact* NMFs would hold by the results of Sect. 3.2

$$\mathbf{H}_{nn}^Q = \mathbf{\Pi}_n^Q \mathbf{\Gamma}_n^Q \tag{35}$$

$$\mathbf{H}_{n,n+1}^Q = \mathbf{\Pi}_n^Q \mathbf{\Gamma}_{n+1}^Q \tag{36}$$

$$\mathbf{\Gamma}_{n+1}^Q = \mathbf{M}^Q \mathbf{\Gamma}_{(n)}^Q \tag{37}$$

This can be considered as an ideal algorithm. Feeding into the system (35), (36), and (37) the inputs $(\mathbf{H}_{nn}^Q, \mathbf{H}_{n,n+1}^Q)$, which are known since Q is given, produces the output $\mathbf{M}^Q$, whose blocks contain the parameters $M^Q(y)$ sought for. In real situations the exact NMFs are not valid since Q might not be an HMM, or might be an HMM of order larger than N. This suggests constructing a three step algorithm where (35), (36), and (37) are substituted with approximate NMFs. The inputs of the algorithm will still be the given N and $(\mathbf{H}_{nn}^Q, \mathbf{H}_{n,n+1}^Q)$. The scheme below illustrates the three steps.

Algorithm

1. Law approximation step

Given: $\mathbf{H}_{nn}^Q$

Problem: $\min_{\mathbf{\Pi}_n, \mathbf{\Gamma}_n} D(\mathbf{H}_{nn}^Q \| \mathbf{\Pi}_n \mathbf{\Gamma}_n)$ constraints $e^\top \mathbf{\Pi}_n e = 1, \ \mathbf{\Gamma}_n e = e$

Solution: $(\mathbf{\Pi}_n^*, \mathbf{\Gamma}_n^*)$ of respective sizes $(m^n \times N)$ and $(N \times m^n)$.

2. Approximate realization step

Given: $\mathbf{H}^{Q}_{n,n+1}$ and $\mathbf{\Pi}^{*}_{n}$ from step 1

Problem: $\min_{\mathbf{\Gamma}_{n+1}} D(\mathbf{H}^{Q}_{n,n+1} \| \mathbf{\Pi}^{*}_{n}\mathbf{\Gamma}_{n+1})$ constraint $\mathbf{\Gamma}_{n+1}e = e$

Solution: $\mathbf{\Gamma}^{*}_{n+1}$ of size $(N \times m^{n+1})$.

3. Parametrization step

Given: $\mathbf{\Gamma}^{*}_{n}$ from step 1, $\mathbf{\Gamma}^{*}_{n+1}$ from step 2, $\mathbf{\Gamma}^{*}_{(n)}$ defined as in (20)

Problem: $\min_{\mathbf{M}} D\left(\mathbf{\Gamma}^{*}_{n+1} \| \mathbf{M}\mathbf{\Gamma}^{*}_{(n)}\right)$ constraint $\mathbf{M}e = e$

Solution: $\mathbf{M}^{*} = \left[M^{*}(y_1) \dots M^{*}(y_m)\right]$.

Note that the constraint $\mathbf{M}e = e$, imposed at step 3, corresponds to the requirement that the transition matrix of the underlying Markov chain be stochastic. The resulting $A^{*} = \sum_{y_i} M^{*}(y_i)$ is used as the transition matrix of the Markov chain in approximate model.

We discuss the behavior of the algorithm in two special cases. Issues concerning the numerical implementation are deferred to Sect. 7.

The algorithm when the true distribution is an HMM

It is desirable that under ideal conditions the algorithm behaves as expected. Consider first the case where the "true" law Q is actually that of a stationary HMM of order N. Equations (35), (36), and (37) then hold and the exact NMF's are full rank factorizations. The matrices $\mathbf{\Pi}^{*}_{n}, \mathbf{\Gamma}^{*}_{n}$, resulting from step 1, satisfy $\mathbf{\Pi}^{Q}_{n}\mathbf{\Gamma}^{Q}_{n} = \mathbf{\Pi}^{*}_{n}\mathbf{\Gamma}^{*}_{n}$ as this gives value zero to the informational divergence. It then holds that $\mathbf{\Pi}^{*}_{\mathbf{n}} = \mathbf{\Pi}_{\mathbf{n}}S$ and $S\mathbf{\Gamma}^{*}_{n} = \mathbf{\Gamma}_{n}$, for some invertible matrix S, with the property that $Se = e$. It also follows that $S\mathbf{\Gamma}^{*}_{n+1} = \mathbf{\Gamma}_{n+1}$ and one easily verifies that the matrices $M^{*}(y_i)$ from step 3 satisfy $SM^{*}(y_i) = M(y_i)S$. Consequently $SA^{*} = AS$ and $\pi^{*} = \pi S$ is an invariant vector of A^{*}. The probabilities $p^{*}(w) = \pi^{*}M^{*}(w)e$ induced by the output of the algorithm are therefore equal to the original probabilities $p(w) = \pi M(w)e$. As expected the output of the algorithm reproduces the original Q.

The algorithm under Markov approximation

Here we analyze the behavior of the algorithm when the given Q is any stationary law and P varies in the set of Markov laws, a subset of the HMMs. In Example 5.2 it was proved that, in this case, the optimal P^{*} is the Markov measure with transition probabilities $P^{*}(Y_{t+1} = j|Y_t = i) = q(j|i) := Q(Y_{t+1} = j|Y_t = i)$, see (27). As it will be proved below, also in this case the output of the algorithm is in agreement with the theoretical solution.

As Markov measures are special cases of HMMs, one can construct the corresponding $M(y)$ parametrization. Let $\mathcal{Y} = \{1, \dots, N\}$ be the space state of the Markov chain

with transition matrix A, then the matrices $M(y)$ assume the special structure

$$m_{ij}(y) = A_{ij}\delta_{jy}, \tag{38}$$

where δ_{jy} is the Kronecker delta. The corresponding matrix $\mathbf{\Pi}_n$ consists of the row vectors $\pi M(u)$, with $u = y_1 \dots y_n \in \mathcal{Y}_{flo}^n$. The generic row takes the form of an N-vector consisting of zeros and on the jth place $P(Y_{t+1}^{t+n} = u)$ iff $y_n = j$. Write $u = \tilde{u}y_n$, where $\tilde{u}$ runs through all strings of length $n-1$. It follows that $\mathbf{\Pi}_n$ has the following block-diagonal structure,

$$\mathbf{\Pi}_n = \begin{bmatrix} \mathbf{\Pi}_n^1 & 0 & \cdots & \cdots & 0 \\ 0 & \mathbf{\Pi}_n^2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \mathbf{\Pi}_n^N \end{bmatrix}, \tag{39}$$

where each block $\mathbf{\Pi}_n^j$ is a column vector consisting of the probabilities $P(Y_{t+1}^{t+n} = \tilde{u}j)$. In step 1 of the algorithm we therefore impose that the matrix $\mathbf{\Pi}_n$ has the block-diagonal structure (39). The Markov assumption does not impose any special structure on the matrices $\mathbf{\Gamma}_n$. Write

$$\mathbf{\Gamma}_n = \begin{bmatrix} \mathbf{\Gamma}_n^1 \\ \vdots \\ \mathbf{\Gamma}_n^N \end{bmatrix},$$

where the $\mathbf{\Gamma}_n^j$ are row vectors. Likewise decompose the Hankel matrix $\mathbf{H}_{nn}^Q$ of the given law Q as

$$\mathbf{H}_{nn}^Q = \begin{bmatrix} \mathbf{H}_{nn}^1 \\ \vdots \\ \mathbf{H}_{nn}^N \end{bmatrix}.$$

The minimization $D(\mathbf{H}_{nn}^Q \| \mathbf{\Pi}_n \mathbf{\Gamma}_n)$ in step 1, under the constraint $\mathbf{\Gamma}_n e = e$, reduces to the N (decoupled) minimization problems $D(\mathbf{H}_{nn}^j \| \mathbf{\Pi}_n^j \mathbf{\Gamma}_n^j)$ with constraints $\mathbf{\Gamma}_n^j e = e$. These problems can be solved *explicitly*, since the inner size of the factorization is equal to one. The solutions are

$$\mathbf{\Pi}_n^{*^j} = \mathbf{H}_{nn}^j e,$$

and

$$\mathbf{\Gamma}_n^{*^j} = \frac{1}{e^\top \mathbf{H}_{nn}^j e} e^\top \mathbf{H}_{nn}^j.$$

Stated in other terms, $\mathbf{\Pi}_n^{*^j}$ has typical elements $q(\tilde{u}j)$ and $\mathbf{\Gamma}_n^{*^j}$ has typical elements $\frac{q(jv)}{q(j)}$ (v a string of length n). In step 2 of the algorithm something similar takes place. The solution $\mathbf{\Gamma}_{n+1}^{*^j}$ has typical elements $\frac{q(jw)}{q(j)}$, where w is a string of length $n+1$. In step 3 of the algorithm, the matrix $\mathbf{M}$ takes the form

$$\mathbf{M} = [M^1, \cdots, M^N],$$

where, by virtue of (38), $M^j = [0, \ldots, 0, m^j, 0, \ldots, 0]$, with the column vector m^j on the jth place. It turns out that also this step of the algorithm has an *explicit* solution, given by $m_i^{*^j} = \frac{q(ij)}{q(i)}$. Hence the corresponding matrix of transition probabilities A^* has elements $A_{ij}^* = \frac{q(ij)}{q(i)}$, in agreement with the theoretical result of Example 5.2.

7 Numerical examples

In this section we present some numerical examples to illustrate the behavior of the proposed approximation algorithm. We first show, on a known example, that the approximation of the divergence rate proposed in Theorem 5.4 yields good results even for small values of n. The three step algorithm is then tested on a set of HMM model reduction problems. The computer code for all the examples is written in the R programming language and is available [27].

7.1 Hankel approximation of the divergence rate

The first issue that needs to be addressed is the use of the informational divergence between the finite Hankel matrices $\mathbf{H}_{nn}^Q$ and $\mathbf{H}_{nn}$ to approximate the divergence rate between the processes (Theorem 5.4). A priori n should be large enough to ensure that the asymptotics have set in, on other hand it should be small enough to avoid the curse of dimensionality, since the size of the Hankel matrices grows exponentially in n. We will see below, when Q and P are both HMM measures, that even small values of n are sufficient for a good approximation to hold. The Q and P HMM measures for this example are taken from [13]. Both are such that the factorization hypothesis (6) holds. Specifically, under Q, the matrix of transition probabilities is

$$A^Q = \begin{pmatrix} 0.80 & 0.15 & 0.05 & 0 \\ 0.07 & 0.75 & 0.12 & 0.06 \\ 0.05 & 0.14 & 0.80 & 0.01 \\ 0.001 & 0.089 & 0.11 & 0.80 \end{pmatrix}$$

and the readout matrix is

$$B^Q = \begin{pmatrix} 0.3 & 0.4 & 0.2 & 0.1 \\ 0.5 & 0.3 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.4 & 0.3 \\ 0.4 & 0.3 & 0.1 & 0.2 \end{pmatrix}.$$

Similarly P is the HMM with matrix of transition probabilities

$$A^P = \begin{pmatrix} 0.40 & 0.25 & 0.15 & 0.20 \\ 0.27 & 0.45 & 0.22 & 0.06 \\ 0.35 & 0.14 & 0.40 & 0.11 \\ 0.111 & 0.119 & 0.23 & 0.54 \end{pmatrix}$$

and readout matrix

$$B^P = \begin{pmatrix} 0.1 & 0.15 & 0.65 & 0.1 \\ 0.2 & 0.3 & 0.4 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.3 \\ 0.15 & 0.25 & 0.4 & 0.2 \end{pmatrix}.$$

It follows from (7), (8) that $M^Q(y) = A^Q B_y^Q$ and $M^P(y) = A^P B_y^P$.

No general, closed form, expression of the divergence rate $D(Q\|P)$ in terms of the parameters $M^Q(y)$ and $M^P(y)$ is available in the literature. A computationally efficient device for the numerical evaluation of $D(Q\|P)$, proposed in [13], is based on the Shannon–McMillan–Breiman (SMB) theorem:

$$D(Q\|P) = \lim_{T\to\infty} \frac{1}{T} \log \frac{q(Y_0, \ldots, Y_{T-1})}{p(Y_0, \ldots, Y_{T-1})} \qquad Q\text{-a.s.}$$

For a proof of the SMB Theorem in the context of HMMs see e.g. Theorem 2 of [17]. The procedure proposed in [13], based on the almost sure convergence in the SMB Theorem, is as follows. Simulate, according to Q, a chunk of a trajectory $y_0, \ldots, y_{T-1}$ of the process Y, and define

$$\hat{D}_T = \frac{1}{T} \log \frac{q(y_0, \ldots, y_{T-1})}{p(y_0, \ldots, y_{T-1})}. \tag{40}$$

where the numerator and the denominator are computed via the Baum formula (9). For T large enough $\hat{D}_T$ will be close to $D(Q\|P)$ for Q-almost all trajectories. In the table below we collect the results for three different simulation runs (*i.e.* three different trajectories of Y).

T	100	220	500	10^3	10^4	10^5	10^6
run 1 $\hat{D}_T$	0.1998	0.1817	0.1186	0.1357	0.1207	0.1022	0.1016
run 2 $\hat{D}_T$	0.1111	0.1432	0.1255	0.1140	0.0961	0.1018	0.1026
run 3 $\hat{D}_T$	0.2059	0.1436	0.1061	0.0942	0.1072	0.1049	0.1023

Note that, after about $T = 10^5$ simulation samples, the different runs stabilize around the value 0.10, in agreement with the trajectory independent behavior predicted by the SMB theorem. In [13] the approximate value 0.14 was obtained, with $T = 220$ simulation samples of only one trajectory of Y. Our runs, on a 25 years newer PC, are compatible with [13] for $T = 220$ but also indicate that it takes rather long ($T = 10^5$)

before convergence sets in. We conclude that 0.10 is a good approximation of the *theoretical* divergence rate $D(Q\|P)$.

The computation of $D_n := \frac{1}{2n} D(\mathbf{H}_{nn}^{Q} \| \mathbf{H}_{nn})$, assuming the above HMM specifications, has been carried out for $n = 2, 3, 4, 5$. The resulting values of D_n are listed below.

n	2	3	4	5
D_n	0.0976	0.0991	0.0998	0.1003

Comparing the two tables it clearly appears that D_n is close to the theoretical value $D(Q\|P)$, already for $n = 4$ and $n = 5$. This phenomenon is not completely surprising. In the case of an HMM Q, with a representation of size N, its pdf $q(\cdot)$ is completely determined by the values $q(w)$, for all strings w of length $2N$, see [7] for an easy proof, or [4] for more involved arguments leading to a proof that in fact length $2N-1$ suffices. It follows that the Hankel matrices with $n = N$ completely determine the laws of the corresponding HMM processes.

7.2 The algorithm in action

The implementation of the three step algorithm requires the numerical computation of three approximate NMFs. Lee and Seung [15] have proposed an iterative algorithm for the approximate NMF $\min_{W,H} D(V\|WH)$ under the constraint $He = e$. Its convergence properties, close to those of the EM [26], have been analyzed in [9]. By its very nature the iterative algorithm stops at the local minima of the objective function. In practice it is essential to start from many initial conditions (W_0, H_0) and select the best run. As a final remark note that our three NMFs satisfy different constraints. Moreover, in the second and third steps, one of the matrices W or H is fixed. In this case convergence takes place to a global minimum, which follows from application of results by Csiszár and Tusnády [6]. These differences have been taken into account in the implementation [27].

We have tested the algorithm in the context of *model order reduction*. In all the examples the given (true) law Q, is a binary valued HMM of size 4, to be approximated with the law P of a binary HMM of size 3.

In the first example, the true law Q of the 4 state HMM has matrix of transition probabilities

$$A = \begin{pmatrix} 0.325 & 0.325 & 0.025 & 0.325 \\ 0.300 & 0.375 & 0.025 & 0.300 \\ 0.33\overline{3} & 0.33\overline{3} & 0 & 0.33\overline{3} \\ 0.375 & 0.275 & 0.050 & 0.300 \end{pmatrix}$$

and readout matrix

$$B = \begin{pmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \\ 0.9 & 0.1 \\ 0.7 & 0.3 \end{pmatrix}.$$

Note that the third column of A has been chosen much smaller than the others. The third state of the hidden state process therefore appears only rarely and it is conceivable that, in this case, the 4 state HMM Q can be well approximated with a 3 state HMM P. We executed 10 runs of the algorithm generating randomly, for each run, the initial conditions required for the three steps. The number of iterations has been set empirically to 2,000 for the first step, 1,000 for the second, and 500 for the third. The table below gives, for each run, the divergence rate of Q with respect to the HMM induced by the initialization (second column) and with respect to the best approximation (third column).

Run	Initial value	Final value
1	8.96×10^{-3}	6.52×10^{-9}
2	9.02×10^{-3}	7.04×10^{-9}
3	9.41×10^{-3}	2.88×10^{-8}
4	9.11×10^{-3}	1.75×10^{-8}
5	9.02×10^{-3}	8.46×10^{-9}
6	8.75×10^{-3}	3.78×10^{-9}
7	9.89×10^{-3}	6.88×10^{-9}
8	9.58×10^{-3}	5.33×10^{-9}
9	9.16×10^{-3}	6.33×10^{-9}
10	9.13×10^{-3}	7.13×10^{-9}

Notice that for each of the 10 runs, the divergence rate drops by a factor of order 10^{-6}. The sixth run gives the lowest value of the divergence rate, which we consider as the best numerical approximation in this case.

In the second example, the 'true' law Q of the 4 state HMM has matrix of transition probabilities

$$A_{\text{random}} = \begin{pmatrix} 0.567 & 0.131 & 0.258 & 0.043 \\ 0.325 & 0.190 & 0.411 & 0.074 \\ 0.259 & 0.364 & 0.111 & 0.266 \\ 0.758 & 0.104 & 0.008 & 0.130 \end{pmatrix},$$

which was randomly generated, while the readout matrix B is as before. Following the procedure outlined above, and maintaining the same initial conditions, the 10 runs produced the following results.

Run	Initial value	Final value
1	1.33×10^{-2}	4.92×10^{-5}
2	1.26×10^{-2}	9.61×10^{-6}
3	1.34×10^{-2}	1.29×10^{-5}
4	1.25×10^{-2}	1.06×10^{-5}
5	1.19×10^{-2}	1.59×10^{-5}
6	1.28×10^{-2}	1.78×10^{-5}
7	1.36×10^{-2}	2.60×10^{-4}
8	1.24×10^{-2}	9.27×10^{-6}
9	1.34×10^{-2}	3.38×10^{-5}
10	1.22×10^{-2}	5.09×10^{-5}

Note that, in this case, the divergence rate roughly drops by a factor of order 10^{-3}. For this example the 8th run produces the best result.

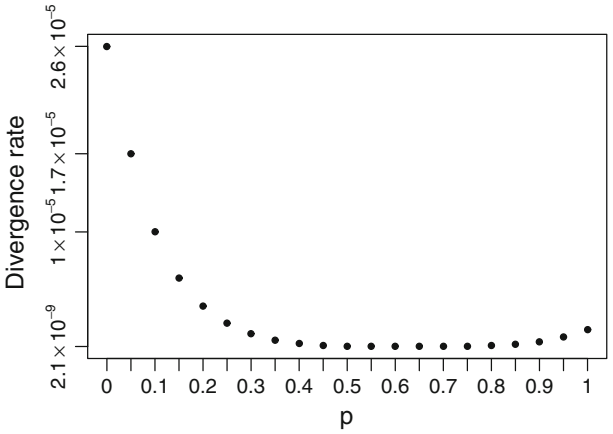

Fig. 1 Minimum divergence rate for $p = k/20$ ($k = 0, 1, \ldots, 20$)

Comparing the first two examples observe that, in agreement with what was expected, the almost degenerate HMM corresponding to the A of the first example can be approximated much better than the the HMM corresponding to A_{random}.

The final example provides a preliminary analysis of the sensitivity of the best approximation P to variations of the readout matrix of Q. We constructed a set of HMMs Q with common matrix of transition probabilities A, as in the first example, and generated a family of 21 readout matrices B_p, $p = k/20, k = 0, 1, \ldots, 20$ by changing the fourth row of the B used in the first example.

$$B_p = \begin{pmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \\ 0.9 & 0.1 \\ p & 1-p \end{pmatrix}.$$

To make comparisons meaningful, the initializations required by the three step algorithm were kept fixed for all 21 cases. Figure 1 shows, for each B_p, the corresponding minimal divergence rate. The minimum divergence for p in the range $[0.45, 0.75]$ is practically constant. This shows that the final outcome is hardly sensitive to values of p within that range. On the other hand, there is a strong dependence on p for $p < 0.45$.

References

1. Anderson BDO (1999) The realization problem for hidden Markov models. Math Control Signals Syst 12:80–120
2. Baum LE, Petrie T (1966) Statistical inference for probabilistic functions of finite Markov chains. Ann Math Stat 37:1554–1563

3. Blackwell D (1957) The entropy of functions of finite-state Markov chains Trans. of the first Prague conference on information theory, statistical decision functions, Random Processes, pp 13–20
4. Carlyle JW (1969) Stochastic finite-state system theory. In: Zadeh L, Polak L (eds) Systems theory, Chapter 10, McGraw-Hill, New York
5. Csiszár I (1975) I-divergence geometry of probability distributions and minimization problems. Ann Probab 3:146–158
6. Csiszár I, Tusnády G (1984) Information geometry and alternating minimization procedures. Stat Decis supplement issue 1:205–237
7. Finesso L (1990) Consistent estimation of the order for Markov and hidden Markov chains, PhD Thesis Report 91-1, Institute of Systems Research, University of Maryland College Park
8. Finesso L, Spreij PJC (2002) Approximate realization of finite hidden Markov chains. In: Proceedings of the 2002 IEEE information theory workshop, Bangalore, India, pp 90–93
9. Finesso L, Spreij PJC (2006) Nonnegative matrix factorization and I-divergence alternating minimization. Linear Algebra Appl 416:270–287
10. Gray RM (1990) Entropy and information theory. Springer, New York
11. Han G, Marcus B (2006) Analyticity of entropy rate of hidden Markov chains. IEEE Trans Inf Theory 52(12):5251–5266
12. Heller A (1965) On stochastic processes derived from Markov chains. Ann Math Stat 36:1286–1291
13. Juang BH, Rabiner LR (1985) A probabilistic distance measure for hidden Markov models. AT&T Tech J 64(20):391–408
14. Karan M, Anderson BDO, Williamson RC (1993) A note on the calculation of a probabilistic distance between hidden Markov models. In: Proc. ISPACS 93, 2nd international workshop on intelligent signal Proc. Comm. Systems, Sendai, 93–98
15. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
16. LeGland F, Mevel L (2000) Exponential forgetting and geometric ergodicity in HMMs. Math Control Signals Syst 13(1):63–93
17. Leroux BG (1992) Maximum-likelihood estimation for hidden Markov models. Stoch Process Appl 40:127–143
18. Mevel L, Finesso L (2004) Asymptotical statistics of misspecified hidden Markov models. IEEE Trans Autom Control 49(7):1123–1132
19. Norris JR (1998) Markov chains. Cambridge University Press, Cambridge
20. Picci G (1978) On the internal structure of finite state stochastic processes. In: Mohler RR, Ruberti A (eds) Recent developments in variable structure systems, economics and biology, Lecture notes in Economics and Mathematical Systems, vol 162, Springer, Berlin, pp 288–304
21. Picci G, van Schuppen JH (1984) On the weak finite stochastic realization problem. In: Korezlioglu H, Mazziotto G, Szpirglas J (eds) Filtering and control of random processes, Lecture Notes in Control and Information Sciences, vol 61, Springer, New York, pp 237–242
22. Rabiner LR, Juang BH (1986) An introduction to hidden Markov models. IEEE ASSP Mag 3(1):4–16
23. Vanluyten B, Willems JC, De Moor B (2006) Matrix factorization and stochastic state representations. In: Proceedings of the 45th IEEE conference on decision and control, San Diego, pp 4188–4193
24. Vidyasagar M (2005) The realization problem for hidden Markov models: the complete realization problem. In: Proceedings of the 44th conference on decision and control, Seville, pp 6632–6637
25. Vidyasagar M (2007) Stochastic modelling over a finite alphabet: approximation using the Kullback-Leibler divergence rate. In: Proceedings of the European control conference 2007, Kos, Paper ThA06.1
26. Wu CFJ (1983) On the convergence properties of the EM algorithm. Ann Stat 11:95–103
27. Software code in R for the numerical implementation of the three step algorithm, http://www.isib.cnr.it/~grassi/HMMs