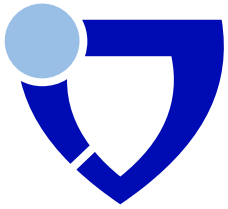


JACOBS
UNIVERSITY

MingJie Zhao Herbert Jaeger

Norm observable operator models

Technical Report No. 8
July, 2007

School of Engineering and Science

Norm observable operator models

MingJie Zhao Herbert Jaeger

School of Engineering and Science
Jacobs University Bremen gGmbH
Campus Ring 12
28759 Bremen
Germany

E-Mail: m.zhao, h.jaeger@jacobs-university.de
http://www.jacobs-university.de/

Summary

Observable operator models (OOMs), a recently developed matrix model class of stochastic processes [6], possesses several advantages over *hidden Markov models* (HMMs). Nevertheless, there is a critical issue, the *negative probability problem* (NPP), which remains unsolved in OOMs theory; and which has heavily prevented it from being an alternative to HMMs in practice. To avoid the NPP we introduce in this report a variation of OOM, the *norm observable operator models* (norm-OOMs).

Like OOMs, norm-OOMs describe stochastic processes also using *linear* observable operators. But norm-OOMs differ from OOMs in that they employ a nonlinear function acting on the state vectors, instead of the linear one used by OOMs, to compute probabilities. Under this nonlinear map, the family of all probability distributions can be embedded into a special inner product space. This provides novel insights into the relationship between the stochastic processes theory and linear algebra; and enables us to study stochastic processes by concepts and methods from linear algebra, a more convenient field of mathematics.

In this report the basic theory of norm-OOMs is set up; an iterative method for learning norm-OOMs is developed based upon the *maximum likelihood* (ML) principle; the advantages and limitations of norm-OOMs are discussed; and some problems for future investigation are outlined.

Contents

1 Introduction

Observable operator models (OOMs) are matrix models for describing stochastic processes developed recently [6]. Compared with *hidden Markov models* (HMMs), another model class of stochastic processes that has been widely and successfully used in many applications, OOMs have several attractive properties [7].

- OOMs are mathematically simpler than HMMs for they can be defined and manipulated in the framework of linear algebra.
- OOMs are more expressive in that any HMM has an equivalent OOM but not vice versa. Here the word "equivalent" means the two models describe the same stochastic process.
- The linear algebra nature of OOMs gives rise to a family of constructive, asymptotically consistent algorithms for estimating OOMs from empirical data [6, 8, 7]; whereas HMMs are typically trained by the (iterative) EM algorithm.

However, there is a crucial problem in OOMs theory remaining unsolved, namely the *negative probability problem* (NPP). More concretely, up to now no algebraic criterion is known for verifying whether an OOM-like system is indeed a *valid* OOM or not and, as the result, all the learning algorithms of OOMs usually obtain *invalid* models which might assign negative probabilities to some (rare) events, instead of small positive numbers. It is proven that the NPP can be reduced to verifying the existence of a common invariant convex cone under some given linear operators (the observable operators), which itself is also a very difficult problem, even from the viewpoint of mathematics.

In this report, we do not intend to address the NPP, but will set up another model class called *norm observable operator models* (norm-OOMs) in which the NPP is *avoided*. The idea of norm-OOMs is quite straightforward: the way that OOMs model stochastic processes can be extended for describing arbitrary numerical functions, so one can avoid the NPP by applying a nonnegative function on the state vectors of OOMs. In more detail, an OOM of some process can be seen as a dynamical system with state vectors $\boldsymbol{w}_t$ uniquely determined by the initial realization $a_1a_2\cdots a_t$ of the process up to time t and the probability that $a_1a_2\cdots a_t$ is observed computed via a linear function on $\boldsymbol{w}_t$: the sum of all elements of $\boldsymbol{w}_t$. So a natural way to get rid of "negative probabilities" is to use a nonnegative function $\sigma(\boldsymbol{w}_t)$ instead of the linear one used by OOMs. In particular, for norm-OOMs we use $\sigma(\boldsymbol{w}_t) = \|\boldsymbol{w}_t\|^2$, as indicated by its name.

While the NPP becomes a nonissue in norm-OOMs, new problems arise. First, as models of stochastic processes, norm-OOMs should satisfy all constraints from probability measures. So like the case of OOMs, here an important problem is

(P1) *whether we can characterize a valid norm-OOM in an algebraic way.*

Second, unlike an OOM living in the vector space spanned by the future conditional distributions of the underlying process (cf. Section 4 of [6]), which only has

the linear structure, a norm-OOM requires a normed vector space, involving both a linear structure and a topological structure. So the second important problem that should be discussed is

(P2) *on which normed vector space norm-OOMs can be defined.*

This report gives a general yet rigorous introduction to norm-OOMs, including the study of the above two theoretical problems and a *maximum likelihood* (ML) learning algorithm for estimating norm-OOMs from data. The report has the following organization. First we briefly review the construction of OOMs and the NPP in Section 2. Then a detailed introduction to the theory of norm-OOMs is presented in Section 3. In Section 4 we introduce the ML learning algorithm of norm-OOMs, whose performance is investigated in Section 5 through some numerical experiments. Finally, we make the conclusion and point out some future works in Section 6.

2 An Overview of OOM Theory

In this report we only consider discrete-time stochastic processes with values from a finite *alphabet*, say $O = \{1, 2, \cdots, \ell\}$. Such a stochastic process can be seen as a sequence of random variables $(Y_t)_{t\in\mathbb{N}}$ defined on some probability space $(\Omega, \mathcal{A}, \mu)$ and taking values on the same codomain O. To fully characterize the process (Y_t) one needs only to know the following family of finite joint distributions:

$$\left\{\Pr\left(Y_1 = a_1, \cdots, Y_n = a_n\right) := \mu\left(\textstyle\bigcap_{i=1}^{n} Y_i^{-1}(a_i)\right)\right\}_{n\in\mathbb{N}, a_i\in O} . \tag{2.1}$$

Other quantities can be computed from these joint probabilities, for example,

$$\begin{aligned} \Pr(Y_2 = b) &= \textstyle\sum_{a\in O} \Pr(Y_1 = a, Y_2 = b)\,, \\ \Pr(Y_2 = b | Y_1 = a) &= \Pr(Y_1 = a, Y_2 = b) / \Pr(Y_1 = a)\,. \end{aligned}$$

As (conditional) probabilities such as $\Pr(Y_1 = a_1, Y_2 = a_2, \cdots, Y_n = a_n)$ and $\Pr(Y_{n+1} = b_1, \cdots, Y_{n+k} = b_k | Y_1 = a_1, \cdots, Y_n = a_n)$ will be used very often in this report, it is convenient to introduce some shorthand notations for later use. Following the conventions of formal language theory, we denote by O^n the set of all sequence of length n of symbols from O; by O^* the set of all finite sequences of symbols in O, including the *empty sequence* ε, the sequences of length 0 which "consists" of no symbol at all. Thus $O^0 = \{\varepsilon\}$ and $O^* = \bigcup_{n\geqslant 0} O^n$. We shall use small letters with a bar (e.g., $\bar{a}, \bar{x}, \cdots$) to denote any element in O^*, i.e., any finite sequence $a_1 \cdots a_n$. For any two finite sequences $\bar{a} = a_1 \cdots a_n$ and $\bar{b} = b_1 \cdots b_k$, we write $P(\bar{a})$ for the joint probability $\Pr(Y_1 = a_1, \cdots, Y_n = a_n)$ and $P(\bar{b}|\bar{a})$ for the conditional probability $\Pr(Y_{n+1} = b_1, \cdots, Y_{n+k} = b_k | Y_1 = a_1, \cdots, Y_n = a_n)$. With these shorthands we can rewrite the family (2.1) as $\{P(\bar{a})\}_{\bar{a}\in O^*}$, which may be, and sometimes is, seen as a numerical function on the set O^*. Thus, the distribution of (Y_t) is uniquely characterized by the function $P : O^* \to \mathbb{R}$.

2.1 Linear dependent functions and processes

We denote by $\mathcal{F}$ the family of all real-valued functions defined on O^*. By the discussion above it is clear that $\mathcal{F}$ contains the family of all discrete processes in the sense that each process is uniquely characterized by a member of $\mathcal{F}$. Moreover, constructing an OOM from a given stochastic process is actually the procedure of deriving an OOM-like system from the function $h \in \mathcal{F}$ defined by $h(\bar{a}) = P(\bar{a})$, the probability distribution of the given process. In the following we will briefly review this procedure, see Section 4 of [6] for the detail.

The set $\mathcal{F}$ canonically becomes a real vector space in which the vector addition and scalar multiplication are defined pointwise. For each symbol $a \in O$ we define a *left appending operator* l_a on the space $\mathcal{F}$ by setting $(l_a f)(\bar{x}) = f(a\bar{x})$ for all $f \in \mathcal{F}$ and $\bar{x} \in O^*$. One can easily see that each l_a is a linear operator: for any $f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$ it holds that

$$(l_a(f+g))(\bar{x}) = (f+g)(a\bar{x}) = f(a\bar{x}) + g(a\bar{x}) = (l_a f + l_a g)(\bar{x})$$
$$\text{and} \quad (l_a(\alpha f))(\bar{x}) = (\alpha f)(a\bar{x}) = \alpha f(a\bar{x}) = \alpha(l_a f)(\bar{x}) = (\alpha l_a f)(\bar{x})$$

for all $\bar{x} \in O^*$. So $l_a(f+g) = l_a f + l_a g$ and $l_a(\alpha f) = \alpha l_a f$, which is what we want to prove. Iteratively applying left appending operators l_a on a fixed function $h \in \mathcal{F}$, we can evaluate the value of h on any sequence $\bar{a} = a_1 a_2 \cdots a_n \in O^*$, as follows:

$$\begin{aligned} h(a_1 a_2 \cdots a_n) &= (l_{a_1} h)(a_2 \cdots a_n) = (l_{a_2} l_{a_1} h)(a_3 \cdots a_n) \\ &= \cdots = (l_{a_n} \cdots l_{a_2} l_{a_1} h)(\varepsilon) := \sigma l_{\bar{a}} h \,, \end{aligned} \tag{2.2}$$

where $l_{\bar{a}}$ denotes the composition of $l_{a_1}, l_{a_2}, \cdots, l_{a_n}$ in *reverse* order; and σ is the linear functional on $\mathcal{F}$ which maps each $f \in \mathcal{F}$ to the real $f(\varepsilon)$. By (2.2) we see that the structure $(\mathcal{F}, \{l_a\}_{a \in O}, \sigma)$ provides another (algebraic) representation of the family $\mathcal{F}$, which allows us to calculate any $h \in \mathcal{F}$. This is natural and trivial since nothing is "thrown away" or "compressed" in the derivation of the structure $(\mathcal{F}, \{l_a\}_{a \in O}, \sigma)$. This structure, however, can be reduced to a smaller one if we are only interested in a single function $h \in \mathcal{F}$ with a special property, as shown below.

Let $h \in \mathcal{F}$ be a fixed function and $\mathcal{H}$ the subspace of $\mathcal{F}$ spanned by the vectors $\{l_{\bar{a}} h : \bar{a} \in O^*\}$. It is clear that $\mathcal{H}$ is invariant under the operation of each l_a, that is, $f \in \mathcal{H}$ implies $l_a f \in \mathcal{H}$. So we can restrict the domain of l_a's and σ to the vector space $\mathcal{H}$, getting a new set of linear operators and a new linear functional on $\mathcal{H}$ which we denote by the same symbols l_a and σ, respectively, trusting to the reader's good sense to avoid confusion. Therefore, we obtain a smaller system $(\mathcal{H}, \{l_a\}_{a \in O}, h, \sigma)$ specially for the function $h \in \mathcal{F}$ in which $h(\bar{a}) = \sigma l_{\bar{a}} h$ for any sequence $\bar{a} \in O^*$.

In practice, for computational reasons we are more interested in those functions h for which the linear space $\mathcal{H} = \text{span}\{l_{\bar{a}} h : \bar{a} \in O^*\}$ is of finite, say m, dimension. Such functions will be called *linearly dependent functions* (LDFs) in this report. For any LDF h of dimension m, we select a basis $\{g_1, g_2, \cdots, g_m\}$ of $\mathcal{H}$ such that

$g_i(\varepsilon) = 1$ for all $i = 1, 2, \cdots, m$. [1] Under this basis, the space $\mathcal{H}$ is isomorphic to $\mathbb{R}^m$, in which the linear functional σ is represented by $\mathbf{1}$, the row vector of units with its size determined by the context; each operator l_a is represented by a matrix $\tau_a \in \mathbb{R}^{m\times m}$ and the function h is represented by some $\boldsymbol{w}_0 \in \mathbb{R}^m$, called the *initial state*. All in all, the abstract system $(\mathcal{H}, \{l_a\}_{a\in O}, h, \sigma)$ is now represented by the concrete one $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$, where the "standard" functional $\mathbf{1}$ is omitted. It follows from (2.2) that, for any $\bar{a} = a_1 a_2 \cdots a_n \in O^*$,

$$h(a_1 a_2 \cdots a_n) = \mathbf{1}\tau_{a_n} \cdots \tau_{a_2}\tau_{a_1}\boldsymbol{w_0} := \mathbf{1}\tau_{\bar{a}}\boldsymbol{w}_0\,, \tag{2.3}$$

where, like the operator $l_{\bar{a}}$, $\tau_{\bar{a}}$ denotes the composition $\tau_{a_n} \cdots \tau_{a_2}\tau_{a_1}$, with the agreement that $\tau_\varepsilon = I_m$, the identity matrix of order m.

Definition 1 *A stochastic process (Y_t) is called a* linearly dependent process *(LDP)* [3, 5, 6] *if its distribution $P(\bar{a})$, when seen as a numerical function $h(\bar{a}) := P(\bar{a})$, is a LDF. A triple $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ is an* observable operator model *(OOM) of (Y_t), if* (2.3) *holds for all $\bar{a} \in O^*$. In an OOM $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$, the linear operators τ_a are called* observable operators *and $\boldsymbol{w}_0$ is the* initial state.

It follows from the above discussion that any LDP (Y_t) has an OOM. But a LDP (Y_t) can be described by different OOMs via (2.3). These OOMs are said to be *equivalent* to each other. An OOM $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ is *minimal* if it has the minimal dimension m in its equivalence class.

By the above construction of OOMs from the distribution $h(\bar{a}) := P(\bar{a})$ one easily sees that any minimal OOM of a LDP (Y_t) has the same dimension m as the abstract vector space $\mathcal{H}$. Furthermore, by mapping the function h defined by (2.3) back to the space $\mathcal{F}$ and then constructing its subspace $\mathcal{H} = \text{span}\{l_{\bar{a}}h : \bar{a} \in O^*\}$ one gets a systematic routine for minimizing a given OOM, as shown in Section 14.5 of [7]. This allows us to only consider minimal OOMs in the sequel.

As two equivalent minimal OOMs are just two different representations of the same abstract structure $(\mathcal{H}, \{l_a\}_{a\in O}, h, \sigma)$ under different bases, they are related to each other via a basis transition matrix ϱ which maps the standard functional $\mathbf{1}$ to itself. Thus,

Theorem 1 *Two minimal OOMs $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ and $(\mathbb{R}^m, \{\tau'_a\}_{a\in O}, \boldsymbol{w}'_0)$ are equivalent if and only if there is a linear bijection $\varrho : \mathbb{R}^m \to \mathbb{R}^m$, such that* (i) $\mathbf{1}\varrho = \mathbf{1}$*;* (ii) $\boldsymbol{w}'_0 = \varrho\boldsymbol{w}_0$ *and* (iii) $\varrho\tau_a\varrho^{-1} = \tau'_a$ *for all $a \in O$.*

— See Proposition 14.6 of [7] for the proof.

We now characterize the family of all m-dimensional LDFs, i.e., all functions $h \in \mathcal{F}$ with $\dim \mathcal{H} = \text{rank}\{l_{\bar{a}}h : \bar{a} \in O^*\} = m$. Since such a function is uniquely determined by some triple $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$, which has finitely many $(\ell m^2 + m$

[1] This can be done in three steps: first select an arbitrary basis $\{g_1, g_2, \cdots, g_m\}$, then there must be some k for which $g_k(\varepsilon) \neq 0$, pick one such k; next let $g_i \leftarrow g_i + g_k$ for any i satisfying $g_i(\varepsilon) = 0$, so that $g_i(\varepsilon) \neq 0$ for all i; finally set $g_i \leftarrow g_i/g_i(\varepsilon)$ for all i.

for the case here) free parameters, we conclude that an m-dimensional LDF h can be reconstructed from its values on a finite subset of O^*. To obtain an intuitive feeling of the situation we consider the number of constraints that the function h should satisfy. Assume that $\{l_{\bar{a}_1}h, l_{\bar{a}_2}h, \cdots, l_{\bar{a}_m}h\}$ is a basis of $\mathcal{H}$ and that $h = \sum_{j=1}^m \alpha_j (l_{\bar{a}_j} h)$. Then $h(\bar{x}) = \sum_{j=1}^m \alpha_j h(\bar{a}_j \bar{x})$ for all $\bar{x} \in O^*$. So if there exist $\bar{x} = \bar{b}_1, \bar{b}_2, \cdots, \bar{b}_m$ such that the resulting m equations (on the parameters $\alpha_1, \alpha_2, \cdots, \alpha_m$) are linearly independent, one can compute α_i's from the values $h(\bar{b}_i)$ and $h(\bar{a}_j \bar{b}_i)$ $(i, j = 1, 2, \cdots, m)$. Now it becomes clear that the function h should satisfy infinitely many constraints $h(\bar{x}) = \sum_{j=1}^m \alpha_j h(\bar{a}_j \bar{x})$. So the degree of freedom of h might be finite — actually it is, as we will see soon. Here the remaining problem is: *can we find $\bar{x} = \bar{b}_1, \bar{b}_2, \cdots, \bar{b}_m$ to get m independent equations on α_j's*? The answer is summarized in the following theorem.

Theorem 2 (1) *For any linearly independent functions $g_1, g_2, \cdots, g_m$ in $\mathcal{F}$ there exist $\bar{b}_1, \bar{b}_2, \cdots, \bar{b}_m \in O^*$ such that the matrix $[g_j(\bar{b}_i)]_{i,j=1,2,\cdots,m}$ is invertible.*

(2) *Let h be an m-dimensional LDF and σ the linear functional defined on $\mathcal{H} = \mathrm{span}\{l_{\bar{a}}h : \bar{a} \in O^*\}$ that maps each $f \in \mathcal{H}$ to the real $f(\varepsilon)$. For any two finite subsets $A = \{\bar{a}_j\}_{j=1,2,\cdots,M}$ and $B = \{\bar{b}_i\}_{i=1,2,\cdots,N}$ of O^*, we write $h(A, B)$ for the $N \times M$ matrix with $h(\bar{a}_j \bar{b}_i)$ as its (i, j)-th entry. Then* (a) $\dim \mathcal{G} = m$, *where $\mathcal{G} = \mathrm{span}\{\sigma l_{\bar{a}} : \bar{a} \in O^*\}$;* (b) $\mathrm{rank}\, h(A, B) \leqslant m$ *for any finite subsets $A, B \subseteq O^*$; and* (c) *there exist $A, B \subseteq O^{<m}$ such that* $\mathrm{rank}\, h(A, B) = m$, *where $O^{<m} := \bigcup_{k<m} O^k$ denotes the set of sequences of length $< m$.*

— See Appendix A.1 for the proof.

2.2 Learning algorithms of OOMs

The parts (2-b,c) of Theorem 2 represent an essential algebraic criterion should be fulfilled by any m-dimensional LDF $h(\bar{a})$. When used to describe some LDP (Y_t) via $P(\bar{a}) = h(\bar{a})$, the function $h(\bar{a})$ should satisfy three more conditions:

1. $h(\varepsilon) = 1$; (the probability of the whole space is 1)
2. $\sum_{a \in O} h(\bar{x}a) = h(\bar{x})$ for all $\bar{x} \in O^*$; (additivity of probability measure)
3. $h(\bar{a}) \geqslant 0$ for all $\bar{a} \in O^*$. (nonnegativity of probability measure)

In any *minimal* OOM $(\mathbb{R}^m, \{\tau_a\}_{a \in O}, \boldsymbol{w}_0)$ of (Y_t), these conditions can be equivalently restated as (i) $\mathbf{1}\boldsymbol{w}_0 = 1$; (ii) $\mathbf{1} \sum_{a \in O} \tau_a = \mathbf{1}$; and (iii) $\mathbf{1}\tau_{\bar{a}}\boldsymbol{w}_0 \geqslant 0$ for any $\bar{a} \in O^*$. Furthermore, by the definition of m-dimensional LDPs and Theorem 2-(2a), the minimality of $(\mathbb{R}^m, \{\tau_a\}_{a \in O}, \boldsymbol{w}_0)$ requires that (iv) the vector spaces $\mathrm{span}\{\tau_{\bar{a}}\boldsymbol{w}_0 : \bar{a} \in O^*\}$ and $\mathrm{span}\{\mathbf{1}\tau_{\bar{a}} : \bar{a} \in O^*\}$ both have dimension m. Conversely, the discussion in Subsection 2.1 shows that any m-dimensional LDP (Y_t) can be modelled by such a structure $(\mathbb{R}^m, \{\tau_a\}_{a \in O}, \boldsymbol{w}_0)$ via $P(\bar{a}) = \mathbf{1}\tau_{\bar{a}}\boldsymbol{w}_0$. This gives us the following theorem, which can be seen as an algebraic definition of OOMs.

Theorem 3 *A triple $(\mathbb{R}^m, \{\tau_a\}_{a \in O}, \boldsymbol{w}_0)$ with $\tau_a \in \mathbb{R}^{m \times m}$ and $\boldsymbol{w}_0 \in \mathbb{R}^m$ is an OOM if, and only if,* (i) $\mathbf{1}\boldsymbol{w}_0 = 1$; (ii) $\mathbf{1} \sum_{a \in O} \tau_a = \mathbf{1}$ *and* (iii) $\mathbf{1}\tau_{\bar{a}}\boldsymbol{w}_0 \geqslant 0$ *for any $\bar{a} \in O^*$.*

It is a minimal *OOM iff further* (iv) *the vector spaces* $\mathrm{span}\{\tau_{\bar{a}}\boldsymbol{w}_0 : \bar{a} \in O^*\}$ *and* $\mathrm{span}\{\mathbf{1}\tau_{\bar{a}} : \bar{a} \in O^*\}$ *both have dimension* m.

The above three theorems play a fundamental role when deriving a general procedure for *reconstructing* minimal OOMs of a given m-dimensional LDP (Y_t) from its distribution $P(\bar{a})$ and a basic algorithm for *learning* OOMs from training data. We first consider the reconstructing procedure. Given an m-dimensional LDP (Y_t) and its distribution $h(\bar{a}) = P(\bar{a})$, by Theorem 2-(2c) we can select two subsets of O^*, say $A = \{\bar{a}_j\}_{j=1}^M$ and $B = \{\bar{b}_i\}_{i=1}^N$, such that the matrix $h(A, B)$ (cf. Theorem 2) has rank m. Besides this condition, we also require that

$$\mathbf{1}h(A, B) = h(A, \{\varepsilon\}) = \mathbf{1}\textstyle\sum_{a\in O} h(A, aB) \quad \text{and} \quad \mathbf{1}h(\{\varepsilon\}, B) = 1\,,$$

where aB denotes, for each $a \in O$, the subset $\{a\bar{b}_i\}_{i=1}^N$ of O^*. A possible choice of such a subset B is $B = O^r$ with r being sufficiently large.

Assume $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ is a minimal OOM of (Y_t), then by the definition of $h(A, B)$ and (2.3) we can compute

$$\begin{aligned}
h(A, B) &= \big[h(\bar{a}_j\bar{b}_i)\big]_{i\leqslant N, j\leqslant M} = \big[\mathbf{1}\tau_{\bar{b}_i} \cdot \tau_{\bar{a}_j}\boldsymbol{w}_0\big]_{i\leqslant N, j\leqslant M} =: \pi(B)\omega(A)\,,\\
h(A, \{\varepsilon\}) &= \big[h(\bar{a}_j)\big]_{j\leqslant M} = \big[\mathbf{1} \cdot \tau_{\bar{a}_j}\boldsymbol{w}_0\big]_{j\leqslant M} = \mathbf{1}\omega(A)\,,\\
h(\{\varepsilon\}, B) &= \big[h(\bar{b}_i)\big]_{i\leqslant N} = \big[\mathbf{1}\tau_{\bar{b}_i} \cdot \boldsymbol{w}_0\big]_{i\leqslant N} = \pi(B)\boldsymbol{w}_0\,,\\
h(A, aB) &= \big[h(\bar{a}_j a\bar{b}_i)\big]_{i\leqslant N, j\leqslant M} = \big[\mathbf{1}\tau_{\bar{b}_i}\tau_a\tau_{\bar{a}_j}\boldsymbol{w}_0\big]_{i\leqslant N, j\leqslant M} = \pi(B)\tau_a\omega(A)\,,
\end{aligned}$$

where $\pi(B)$ is the $N \times m$ matrix with i-th row $\mathbf{1}\tau_{\bar{b}_i}$, $\omega(A)$ is the $m \times M$ matrix with $\tau_{\bar{a}_j}\boldsymbol{w}_0$ as its j-th column. Then by $\mathbf{1}h(A, B) = h(A, \{\varepsilon\})$ we know $\mathbf{1}\pi(B)\omega(A) = \mathbf{1}\omega(A)$. As $h(A, B)$, and hence $\omega(A)$, has rank m, we conclude that $\mathbf{1}\pi(B) = \mathbf{1}$.

Since the matrix $h(A, B)$ has rank m, there exists $U \in \mathbb{R}^{m\times N}$ such that (1) $Uh(A, B)$ has rank m; and (2) $\mathbf{1}U = \mathbf{1}$. Let $Q := [Uh(A, B)]^\dagger$ be the pseudo-inverse of $Uh(A, b)$ and $\varrho := U\pi(B) \in \mathbb{R}^{m\times m}$. Then, as $Uh(A, B)$ is of full row rank, $Uh(A, B)Q = U\pi(B)\omega(A)Q = I_m$; and $\mathbf{1}\varrho = \mathbf{1}U\pi(B) = \mathbf{1}\pi(B) = \mathbf{1}$. Now it is clear that $\varrho^{-1} = \omega(A)Q$, and that

$$\begin{aligned}
\varrho\tau_a\varrho^{-1} &= U\pi(B)\tau_a\omega(A)Q = U \cdot h(A, aB) \cdot Q\,, &(2.4)\\
\varrho\boldsymbol{w}_0 &= U\pi(B)\boldsymbol{w}_0 = U \cdot h(\{\varepsilon\}, B)\,. &(2.5)
\end{aligned}$$

By Theorem 1, we actually have constructed, from the distribution $h(\bar{a}) = P(\bar{a})$ of the process (Y_t), an equivalent OOM $(\mathbb{R}^m, \{\varrho\tau_a\varrho^{-1}\}_{a\in O}, \varrho\boldsymbol{w}_0)$ of the "original" model $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$.

For the learning task, the distribution $h(\bar{a})$ is unknown and so one cannot reconstruct OOMs of (Y_t) via (2.4), (2.5). Instead, a sampling path $\bar{s} = s_1 s_2 \cdots$ of (Y_t) is provided, based on which one is required to estimate an OOM of (Y_t). A natural idea for this learning problem is to use the training data $\bar{s}$ to estimate the interesting probabilities $h(\bar{a}) = \hat{P}(\bar{a})$, by counting the occurrence number of $\bar{a}$ in $\bar{s}$; and then construct an OOM by (2.4) and (2.5). The overall procedure can be described as follows.

1. Select (depending on the training data $\bar{s}$) two subsets $A = \{\bar{a}_j\}_{j=1}^{M}$ and $B = \{\bar{b}_i\}_{i=1}^{N}$ of O^*; and estimate the probabilities $\hat{P}(\bar{a})$ by

$$\hat{P}(\bar{a}) = \frac{\text{occurrence number of } \bar{a} \text{ in } \bar{s}}{|\bar{s}| - |\bar{a}| + 1},$$

where $|\bar{s}|$ denotes the length of $\bar{s}$ and the sequence $\bar{a}$ runs over $\bar{b}_i$, $\bar{a}_j\bar{b}_i$ and $\bar{a}_j a\bar{b}_i$ ($a \in O$), so that we can construct the matrices $\hat{P}(A, B)$, $\hat{P}(A, aB)$ and the vector $\hat{P}(\{\varepsilon\}, B)$. Note that, here the selection of A, B should satisfy

$$\mathbf{1}\hat{P}(A, B) = \mathbf{1}\textstyle\sum_{a\in O}\hat{P}(A, aB)\,, \tag{2.6}$$

$$\mathbf{1}\hat{P}(\{\varepsilon\}, B) = 1\,; \tag{2.7}$$

and make $\hat{P}(A, B)$ have rank larger than m, the model dimension.

2. Design a matrix $U \in \mathbb{R}^{m\times N}$ such that $\text{rank}\{U\hat{P}(A, B)\} = m$ and $\mathbf{1}U = \mathbf{1}$.

3. Estimate the observable operators and the initial state respectively by

$$\hat{\tau}_a = U\hat{P}(A, aB)[U\hat{P}(A, B)]^{\dagger}\,, \tag{2.8}$$

$$\hat{\boldsymbol{w}}_0 = U\hat{P}(\{\varepsilon\}, B)\,. \tag{2.9}$$

This basic learning algorithm has some variations, for example,

- For stationary LDPs, the corresponding *minimal* OOMs have the property $(\sum_{a\in O}\tau_a)\boldsymbol{w}_0 = \boldsymbol{w}_0$. This condition, together with the condition (i) from Theorem 3, uniquely determines the initial state $\boldsymbol{w}_0$. So for stationary processes one needs only to estimate the observable operators τ_a.

- In practice, one can use the counting matrices $[\#(A, B)]$, $[\#(A, aB)]$ etc. instead of the probability matrices $\hat{P}(A, B)$, $\hat{P}(A, aB)$. But one should keep the same counting factor for $\hat{P}(A, B)$ and $\hat{P}(A, aB)$. That is, one gets $[\#(A, B)]$ from the sample except the last symbol but $[\#(A, aB)]$ from the whole sequence.

2.3 The negative probability problem of OOMs

In the preceding subsection we introduced the basic procedure for reconstructing or learning OOMs from the known distribution $h(\bar{a})$ or a sampling path $\bar{s}$ of an m-dimensional LDP (Y_t). If the accurate values of $h(\bar{a})$ are known, then by the fact that (Y_t) is a LDP of dimension m we know the structure evaluated by (2.4) and (2.5) is a *valid* OOM according to Theorem 3. In the learning task, however, only approximated values of $h(\bar{a})$ can be obtained from the training data $\bar{s}$. So a natural and crucial problem here is whether such approximation of $h(\bar{a})$ will violate the conditions from Theorem 3. In other words, we would ask whether the system computed by (2.8) and (2.9) is a valid OOM.

By Equations (2.6)–(2.9), one should have no difficulty to verify that $\mathbf{1}\hat{\boldsymbol{w}}_0 = 1$ and $\mathbf{1}\sum_{a\in O}\hat{\tau}_a = \mathbf{1}$. Unfortunately, the third condition of Theorem 3: $\mathbf{1}\tau_{\bar{a}}\boldsymbol{w}_0 \geqslant 0$ for any $\bar{a} \in O^*$, does not always hold for (actually is usually violated by) the structure $(\mathbb{R}^m, \{\hat{\tau}_a\}_{a\in O}, \hat{\boldsymbol{w}}_0)$. More precisely, the learnt structure may produce negative values as the probability of some rare paths $\bar{a}$ of the underlying process. We call this phenomenon the *negative probability problem* (NPP) of OOMs. As the learning algorithms of OOMs developed so far are all based the basic algorithm presented in Subsection 2.2 — roughly speaking, they only differ in the design of the auxiliary matrix U — it is fair to say up to now no theoretically satisfying algorithm for estimating OOMs from data has been developed.

When trying to repair the above NPP of OOMs, we immediately meet the problem of how to characterize the class of valid OOMs. In other words, we need a general method for checking whether a given a given structure $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ is a valid OOM. The frustrating thing is even this verification problem is proven very hard — note that, one cannot directly use Theorem 3 since the condition (iii) actually consists of infinitely many inequalities. In this direction, the following proposition gives us an equivalent statement of the condition (iii), in terms of (*invariant*) *convex cones*.

Theorem 4 *A triple* $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$*, in which* $\tau_a \in \mathbb{R}^{m\times m}$*,* $\boldsymbol{w}_0 \in \mathbb{R}^m$ *and the conditions* (i, ii, iv) *from Theorem 3 hold, forms an OOM if and only if there is a* proper convex cone $K \subseteq \mathbb{R}^m$ *(i.e., K is closed under vector addition and scalar multiplication by a nonnegative real; and satisfies $K \cap (-K) = \{\mathbf{0}\}$ and* $\operatorname{span} K = \mathbb{R}^m$*), such that* (i) $\boldsymbol{w}_0 \in K$*;* (ii) $\mathbf{1}\boldsymbol{v} \geqslant 0$ *for all* $\boldsymbol{v} \in K$ *and* (iii) K *is invariant under each operator* τ_a*, that is,* $\tau_a\boldsymbol{v} \in K$ *for any* $a \in O$ *and* $\boldsymbol{v} \in K$*.*

See Proposition 6 of [6] for a detailed discussion about this theorem. It should be noticed that, the theorem also provides no means to decide whether a given system $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ is a valid OOM, since it is non-constructive w.r.t. K. In fact, even in mathematics the problem of whether a common invariant convex cone K under a given family of linear operators τ_a exists or not is still an open problem, at least as the authors know.

As the conclusion, the NPP of OOMs consists of two parts:

- practically, the existing algorithms can only learn "almost OOMs" which sometimes produce "negative probabilities";
- theoretically, there is no general means known to decide whether a given system $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ is an OOM of some LDP or not.

The previous discussion indicates that neither of the above two problems can be *attacked* at the current stage. This difficult situation motivated us to study other variations of OOMs to *avoid* the NPP.

3 Norm Observable Operator Models

To avoid the NPP of OOMs, we introduce in this section another similar model class, *norm observable operator models* (norm-OOMs), which computes the probability of an initial sequence $\bar{a} \in O^*$ by $P(\bar{a}) = \|\tau_{\bar{a}}\boldsymbol{w}_0\|^2$, where, for vectors $\boldsymbol{x}$ in $\mathbb{R}^m$, $\|\boldsymbol{x}\|$ denotes the Euclidean norm, i.e., $\|\boldsymbol{x}\| := \sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{x}}$. To distinguish the two kinds of OOMs, we will use $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ to denote a norm-OOM throughout the report. It is obvious that a norm-OOM always produces nonnegative values and so does not suffer from the NPP. However, as stated in Section 1, the function $P : O^* \to \mathbb{R}$, at the same time as the probability measure of some process, should satisfy $P(\varepsilon) = 1$ and the condition $P(\bar{x}) = \sum_{a\in O} P(\bar{x}a)$ for all $\bar{x} \in O^*$. For norm-OOMs, these two conditions are equivalent to

Theorem 5 *The triple* $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ *forms a norm-OOM of some process iff* (i) $\|\boldsymbol{u}_0\| = 1$ *and* (ii) $\sum_{a\in O} \|\varphi_a\varphi_{\bar{x}}\boldsymbol{v}_0\|^2 = \|\varphi_{\bar{x}}\boldsymbol{v}_0\|^2$ *for all* $\bar{x} \in O^*$.

While the nonnegativity of probabilities is automatically satisfied by norm-OOMs, it seems that the condition (ii) of Theorem 5 is more difficult to cope with than the analogous part of OOMs which can be reduced to the constraint $\mathbf{1}(\sum_{a\in O} \tau_a) = \mathbf{1}$. However, the condition (ii) from Theorem 5 can be rewritten as

$$(\varphi_{\bar{x}}\boldsymbol{u}_0)^\mathsf{T}(\textstyle\sum_{a\in O} \varphi_a^\mathsf{T}\varphi_a)(\varphi_{\bar{x}}\boldsymbol{u}_0) = (\varphi_{\bar{x}}\boldsymbol{u}_0)^\mathsf{T}(\varphi_{\bar{x}}\boldsymbol{u}_0)\,, \quad (\forall \bar{x} \in O^*)$$

for which an obvious sufficient condition is $\sum_{a\in O} \varphi_a^\mathsf{T}\varphi_a = I_m$. Conversely, from any stochastic process with certain properties we can construct its norm-OOM $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ such that $\sum_{a\in O} \varphi_a^\mathsf{T}\varphi_a = I_m$, as presented in Subsection 3.2. These two facts lead to the following definition of *standard norm-OOMs*.

Definition 2 *A* standard norm-OOM *is any triple* $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ *with the properties* (i) $\|\boldsymbol{u}_0\| = 1$*; and* (ii) $\sum_{a\in O} \varphi_a^\mathsf{T}\varphi_a = I_m$.

Unlike the definition of OOMs, this definition of standard norm-OOMs is simple enough to serve as a practical algebraic criterium for verifying whether a given triple $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ is a standard norm-OOM. According to the discussion above, we will only consider standard norm-OOMs and simply call them norm-OOMs in the sequel.

In the following we will set up the basic theory of norm-OOMs which includes (1) defining an inner-product space $\mathcal{D}$ on which norm-OOMs can be constructed; (2) constructing norm-OOMs $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ on the space $\mathcal{D}$ from the distribution $P(\bar{a})$ of some process such that $P(\bar{a}) = \|\varphi_{\bar{a}}\boldsymbol{u}_0\|^2$ for all finite sequences $\bar{a}$ in O^*; (3) using norm-OOMs as generators and predictors; and (4) the expressiveness of norm-OOMs.

3.1 The inner-product space $\mathcal{D}$

In this subsection we construct the inner-product space $\mathcal{D}$ according to the follow procedure: (1) first the set $\mathcal{S}$ of all nonnegative functions $f : O^* \to \mathbb{R}^+$ such that

$P(\bar{a}) = f^2(\bar{a})$ is a probability distribution is considered; (2) as $\mathcal{S}$ is not a vector space, it is then embedded into a vector space $\mathcal{B}$ which serves as the underlying space; (3) next the family $\mathcal{S}$ is extended to a convex cone $\mathcal{D}_0^+$ in $\mathcal{B}$, on which a binary linear function Q is defined; (4) the function Q is extended to the subspace $\mathcal{D}_0$ spanned by $\mathcal{D}_0^+$, and a subspace $\mathcal{N}$ of $\mathcal{D}_0$ is defined via Q; (5) finally we define $\mathcal{D}$ to be the *quotient space* $\mathcal{D}_0/\mathcal{N}$ and induce an inner product on $\mathcal{D}$ from the function Q. Now we discuss these objects one by one.

The vector space $\mathcal{B}$. As mentioned above, the family $\mathcal{S}$ consists of all functions $f : O^* \to \mathbb{R}^+$ such that $P(\bar{a}) = f^2(\bar{a})$ is a probability distribution. In other words, here the family $\mathcal{S}$ is defined by

$$\mathcal{S} := \{f \in \mathcal{F} : f(\varepsilon) = 1, f(\bar{x}) \geqslant 0, f^2(\bar{x}) = \textstyle\sum_{a\in O} f^2(\bar{x}a) \text{ for all } \bar{x} \in O^*\}\,. \quad (3.1)$$

It is clear that each $f \in \mathcal{S}$ specifies a stochastic process with $P(\bar{a}) = f^2(\bar{a})$ as its distribution; and that each process (Y_t) determines an $f \in \mathcal{S}$ via $f(\bar{a}) = \sqrt{P(\bar{a})}$. In this sense, we can identify the family of stochastic processes with the set $\mathcal{S}$. However, the family $\mathcal{S}$, as a subset of $\mathcal{F}$, is not a convenient object to operate from the viewpoint of linear algebra, for it is neither a vector space nor invariant under the left appending operators l_a. So we need to construct another vector space $\mathcal{B}$ which contains $\mathcal{S}$ as a subset and is invariant under the operation of l_a's.

For each $n \geqslant 0$, define a (nonlinear) mapping S_n from $\mathcal{F}$ to $\mathbb{R}$ by setting

$$S_n(f) = \textstyle\sum_{\bar{a}\in O^n} f^2(\bar{a}) \quad \text{for all } f \in \mathcal{F}\,. \quad (3.2)$$

One can easily see that $S_n(f) = 1$ for all $f \in \mathcal{S}$ and nonnegative integers n. Let $\mathcal{B}$ be the family of those functions $f \in \mathcal{F}$ for which the set $\{S_n(f)\}_{n\geqslant 0}$ is upper bounded by some constant $C_f \in \mathbb{R}$, then $\mathcal{S}$ is clearly a subset of $\mathcal{B}$. Moreover, $\mathcal{B}$ forms a subspace of $\mathcal{F}$ that is invariant under the operation of l_a's, as stated by Theorem 6. This allows us to restrict the operators l_a on the space $\mathcal{B}$.

Theorem 6 *The set $\mathcal{B} = \{f \in \mathcal{F} : S_n(f) \leqslant C_f < \infty\}$ is a subspace of $\mathcal{F}$ that is invariant under the left appending operators l_a, i.e., $l_a f \in \mathcal{B}$ for all $f \in \mathcal{B}$.*

— See Appendix A.2 for the proof.

Now for each $n = 0, 1, \cdots$ we define a binary function on $\mathcal{F}$ by

$$Q_n(f, g) = \textstyle\sum_{\bar{a}\in O^n} f(\bar{a})g(\bar{a})\,. \quad (\forall f, g \in \mathcal{F}) \quad (3.3)$$

It follows from the proof of Theorem 6 that $Q_n(f, g) \leqslant \frac{1}{2}[S_n(f) + S_n(g)]$ for any $f, g \in \mathcal{B}$, which means there is a constant $C_{f,g} \in \mathbb{R}$ such that $Q_n(f, g) \leqslant C_{f,g}$ for all n. Conversely, by the definition of Q_n and S_n we have $S_n(f) = Q_n(f, f)$ for all $f \in \mathcal{F}$. So the family $\mathcal{B}$ is also characterized by the inequalities $Q_n(f, g) \leqslant C_{f,g}$ with $n = 0, 1, 2, \cdots$.

The subspace $\mathcal{D}_0$ of $\mathcal{B}$. Let $\mathcal{D}_0^+$ be the subset of $\mathcal{F}$ consisting of those functions $f \in \mathcal{F}$ which are nonnegative and $f^2(\bar{x}) \geqslant \sum_{a\in O} f^2(\bar{x}a)$ for all $\bar{x} \in O^*$, i.e.,

$$\mathcal{D}_0^+ := \{f \in \mathcal{F} : f(\bar{x}) \geqslant 0, f^2(\bar{x}) \geqslant \textstyle\sum_{a\in O} f^2(\bar{x}a) \text{ for } \forall \bar{x} \in O^*\}\,. \tag{3.4}$$

By (3.1) (3.4) we know $\mathcal{S} \subseteq \mathcal{D}_0^+$. Furthermore, it follows from the definition of S_n: $S_n(f) = \sum_{\bar{a}\in O^n} f^2(\bar{a})$ that $\{S_n(f)\}_{n=0,1,2,\cdots}$ forms a decreasing sequence for any $f \in \mathcal{D}_0^+$; thus $S_n(f) \leqslant S_0(f)$ for all n and hence $\mathcal{D}_0^+$ is a subset of $\mathcal{B}$.

Theorem 7 *$\mathcal{D}_0^+$ is a* convex cone *in $\mathcal{B}$ pointed at* 0 *(the zero function). That is, for any $f, g \in \mathcal{D}_0^+$ and any $\alpha \geqslant 0$,* (i) *$-f \in \mathcal{D}_0^+$ implies $f = 0$;* (ii) *$\alpha f \in \mathcal{D}_0^+$ and* (iii) *$f + g \in \mathcal{D}_0^+$. Furthermore, $\mathcal{D}_0^+$ is invariant under the operators l_a, that is, $l_a f \in \mathcal{D}_0^+$ whenever $f \in \mathcal{D}_0^+$.* — See Appendix A.3 for the proof.

Now let $\mathcal{D}_0$ be the subspace of $\mathcal{B}$ spanned by $\mathcal{D}_0^+$. By Theorem 7, $\mathcal{D}_0$ is the set of those functions $h \in \mathcal{B}$ which can be written as $h = f - g$ with $f, g \in \mathcal{D}_0^+$:

$$\mathcal{D}_0 = \operatorname{span} \mathcal{D}_0^+ = \{f - g : f, g \in \mathcal{D}_0^+\}\,. \tag{3.5}$$

Since $\mathcal{D}_0^+$ is invariant under the linear operators l_a, by (3.5) we see that $\mathcal{D}_0$ is also invariant under l_a. So in the sequel we will restrict the operation of l_a's on the space $\mathcal{D}_0$. For any $f, g \in \mathcal{D}_0^+$, the summation of $f(\bar{x})g(\bar{x}) \geqslant \sum_{a\in O} f(\bar{x}a)g(\bar{x}a)$ (cf. eqn (A.2)) over all $\bar{x} \in O^n$ reveals that $\{Q_n(f, g)\}_{n=0,1,2,\cdots}$ forms a decreasing sequence lower bounded by 0, so the binary function

$$Q(f, g) := \lim_{n\to\infty} Q_n(f, g) \tag{3.6}$$

is well defined on the set $\mathcal{D}_0^+ \times \mathcal{D}_0^+$ and takes values from $[0, \infty)$. To extend the domain of the function $Q(f, g)$ to $\mathcal{D}_0 \times \mathcal{D}_0$, we need the following lemma.

Lemma 1 *Let $\{a_n^i\}_{n=0,1,2,\cdots}$, $\{b_n^i\}_{n=0,1,2,\cdots}$ $(i = 1, 2, \cdots, k)$ be $2k$ sequences of real numbers such that $\sum_{i=1}^k a_n^i = \sum_{i=1}^k b_n^i$ and $\lim_{n\to\infty} a_n^i = c^i$ for all $i \leqslant k$. Then $\{\sum_{i=1}^k a_n^i\}_{n=0,1,2,\cdots}$ and $\{\sum_{i=1}^k b_n^i\}_{n=0,1,2,\cdots}$ are two convergent sequences with the same limit $\sum_{i=1}^k c^i$.* — The proof is trivial and omitted here.

For any $f, g \in \mathcal{D}_0$, let $f = f_1 - f_2$ and $g = g_1 - g_2$ with $f_i, g_i \in \mathcal{D}_0^+$ $(i = 1, 2)$ be one of their decompositions, respectively. Then, by the linearity of Q_n,

$$Q_n(f, g) = Q_n(f_1, g_1) + Q_n(f_2, g_2) - Q_n(f_1, g_2) - Q_n(f_2, g_1)\,. \tag{3.7}$$

When n tends to infinity, the four items on the right hand side (r.h.s.) of equality (3.7) each converge to a nonnegative number since $f_i, g_i \in \mathcal{D}_0^+$; and their sum $Q_n(f, g)$ is independent of the choice of f_1, f_2, g_1, g_2. So by Lemma 1 the limit $Q(f, g) = \lim_{n\to\infty} Q_n(f, g)$ exists and, by (3.7), assumes values in $\mathbb{R}$.

So far we have defined the vector space $\mathcal{D}_0$ and the binary function $Q(\cdot, \cdot)$ on $\mathcal{D}_0 \times \mathcal{D}_0$ which clearly has the following three properties.

Theorem 8 *For any $f, g \in \mathcal{D}_0$,* (i) $Q(f,f) \geqslant 0$*;* (ii) $Q(f,g) = Q(g,f)$*; and* (iii) $Q(f,g)$ *is linear in* f*, i.e.,* $Q(\alpha f_1 + \beta f_2, g) = \alpha Q(f_1, g) + \beta Q(f_2, g)$*.*

Theorem 8 shows the function Q is nonnegative definite, symmetric and bilinear, only "one step" from being an *inner product* on the vector space $\mathcal{D}$. [2] We call any such function Q a *semi-definite* inner product on $\mathcal{D}$. More generally,

Definition 3 *Let V be a vector space over $\mathbb{R}$. A binary function $Q : V \times V \to \mathbb{R}$ is a* semi-definite inner product *(on V) if, for any $x, y, z \in V$ and $\alpha, \beta \in \mathbb{R}$,* (i) $Q(x,x) \geqslant 0$*;* (ii) $Q(x,y) = Q(y,x)$*; and* (iii) $Q(\alpha x + \beta y, z) = \alpha Q(x,z) + \beta Q(y,z)$*.*

Next we discuss some properties of the above "abstract" semi-definite inner product space (V, Q). Like inner products, semi-definite inner products are also interesting and useful from both theoretical and practical aspects. For instance, each semi-definite inner product Q on some vector space V induces a *pseudo norm* q on the same space V via $q(x) = \sqrt{Q(x,x)}$. Here, by "pseudo norm" we mean any function $q : V \to \mathbb{R}$ such that (i) $q(x) \geqslant 0$; (ii) $q(\alpha x) = |\alpha| \cdot q(x)$ and (iii) $q(x+y) \leqslant q(x) + q(y)$ for any $x, y \in V$ and $\alpha \in \mathbb{R}$. To see that $q(x) = \sqrt{Q(x,x)}$ is indeed a pseudo norm on V, we first prove an inequality, namely

Lemma 2 (Cauchy-Schwarz inequality for semi-definite inner products) *Let Q be any semi-definite inner product on the vector space V and q the unitary function on V defined by $q(x) = \sqrt{Q(x,x)}$. Then, for any $x, y \in V$, $|Q(x,y)| \leqslant q(x)q(y)$.*
— See Appendix A.4 for the proof.

Now we are ready to show that $q(x)$ is a pseudo norm. The conditions $q(x) \geqslant 0$ and $q(\alpha x) = |\alpha| \cdot q(x)$ are easy to verify. For the third condition, we calculate $q^2(x+y) = q^2(x) + 2Q(x,y) + q^2(y) \leqslant q^2(x) + 2q(x)q(y) + q^2(y) = [q(x) + q(y)]^2$, where the second inequality follows from Lemma 2. So $q(x+y) \leqslant q(x) + q(y)$.

The quotient space $\mathcal{D}$. Return to our concrete vector space $\mathcal{D}_0$, on which we have shown the function Q defined by (3.6) is a semi-definite inner product, and so induces the pseudo norm $q : \mathcal{D}_0 \to \mathbb{R}$ with

$$q(f) := \sqrt{Q(f,f)} = \lim_{n\to\infty} \left\{ \sum_{\bar{a} \in O^n} f^2(\bar{a}) \right\}^{\frac{1}{2}} = \lim_{n\to\infty} \{S_n(f)\}^{\frac{1}{2}} \ . \tag{3.8}$$

For any $f \in \mathcal{S}$, by (3.1) we know $S_n(f) = 1$ and hence $q(f) = 1$. As the pseudo norm $q(f)$ can be explained as the "length" of the vector f, we get a geometrical description of the family of stochastic processes: any process (Y_t), described by some function $f \in \mathcal{S}$ via $P(\bar{a}) = f^2(\bar{a})$ and seen to be identical to f, lies on the unit sphere of the semi-definite inner product space $(\mathcal{D}_0, Q)$. Similarly, the

[2] An inner product on a vector space V is any function $Q : V \times V \to \mathbb{R}$ that has all the properties from Definition 3 and is positive definite: $Q(x,x) = 0$ if and only if $x = 0$.

quantity $q(f-g)$ can be regarded as the "distance" between f and g. When f, g are members of $\mathcal{S}$, i.e., they describe two processes, we have

$$q^2(f-g) = q^2(f) - 2Q(f,g) + q^2(g) = 2\left[1 - Q(f,g)\right]. \tag{3.9}$$

So the quantity $Q(f,g)$ can be seen as the degree of similarity between f and g (note that $0 \leqslant Q(f,g) \leqslant 1$ for any $f, g \in \mathcal{S}$).

We call a function $f \in \mathcal{D}_0$ a *null* function if $q(f) = 0$; and denote by $\mathcal{N}$ the set of all null functions. By the inequality $q(f+g) \leqslant q(f) + q(g)$ and the identity $q(\alpha f) = |\alpha| \cdot q(f)$ one can easily see that $\mathcal{N}$ is a subspace of $\mathcal{D}_0$. So we can define the *quotient space*

$$\mathcal{D} := \mathcal{D}_0/\mathcal{N} = \{[f] : f \in \mathcal{D}_0\}, \tag{3.10}$$

where $[f]$ denotes, for any $f \in \mathcal{D}_0$, the *equivalence class* $\{g \in \mathcal{D}_0 : f - g \in \mathcal{N}\}$ of f. It is well known that, $\mathcal{D}$ canonically becomes a vector space over $\mathbb{R}$ under the addition and scalar multiplication

$$[f] + [g] := [f+g], \quad \alpha[f] := [\alpha f]; \quad (\forall f, g \in \mathcal{D}_0, \forall \alpha \in \mathbb{R})$$

whose definitions are independent of the choice of the "representative" elements f, g in their corresponding equivalence class $[f]$ or $[g]$.

On the vector space $\mathcal{D}$ define a binary function $\langle \cdot, \cdot \rangle : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ by setting

$$\langle [f], [g] \rangle := Q(f,g) \quad \text{for any} \quad f, g \in \mathcal{D}_0. \tag{3.11}$$

Here we should show the value of the r.h.s. of (3.11) is independent of the choice of f and g (in their equivalence class), i.e., $Q(f', g') = Q(f,g)$ for any $f' \in [f]$ and $g' \in [g]$. Noting that $f' \in [f]$ iff $f' - f \in \mathcal{N}$ iff $q(f'-f) = 0$, we know that

$$Q(f', h) = Q(f,h) + Q(f'-f, h) = Q(f,h)$$

for any $h \in \mathcal{D}_0$ since, by Lemma 2, $|Q(f'-f, h)| \leqslant q(f'-f)q(g) = 0$. Similarly, $Q(h,g) = Q(h,g')$ for any $h \in \mathcal{D}_0$. Thus, $Q(f',g') = Q(f,g') = Q(f,g)$. Since Q is a semi-definite inner product (on $\mathcal{D}_0$), by (3.11) one can easily verify that $\langle \cdot, \cdot \rangle$ is also a semi-definite inner product (on $\mathcal{D}$). Now assume $\langle [f], [f] \rangle = 0$, i.e., $Q(f,f) = q^2(f) = 0$, then $f \in \mathcal{N} = [0]$ and so $[f] = [0]$. This means $\langle \cdot, \cdot \rangle$ is actually an inner product on $\mathcal{D}$, which naturally induces the norm

$$\|[f]\| = \sqrt{\langle [f], [f] \rangle} = \sqrt{Q(f,f)} = q(f). \quad (\forall f \in \mathcal{D}_0) \tag{3.12}$$

Thus far we have finished the first step towards the construction of norm-OOMs: defining the inner product space $\mathcal{D}$.

Members of $\mathcal{D}$ that describes some stochastic process. We now consider the members $[f]$ ($f \in \mathcal{D}_0$) of $\mathcal{D}$ that actually represent a stochastic process. In other words, we want to characterize the subset $\mathcal{D}_{\mathrm{S}} := \{[f] : f \in \mathcal{S}\}$ of $\mathcal{D}$.

Firstly, if $f \in \mathcal{S}$, then by (3.12) and (3.8), we have $\|[f]\| = 1$. This means $\mathcal{D}_{\mathrm{S}}$ is located on the unit sphere of $\mathcal{D}$. To describe the subset $\mathcal{D}_{\mathrm{S}}$ in more detail, we define $\mathcal{D}^+$ to be the set of equivalence classes $[f]$ in $\mathcal{D}$ induced by members f from $\mathcal{D}_0^+$, i.e., $\mathcal{D}^+ := \{[f] : f \in \mathcal{D}_0^+\}$. As $\mathcal{S} \subseteq \mathcal{D}_0^+$, we know $\mathcal{D}_{\mathrm{S}} \subseteq \mathcal{D}^+$. Furthermore,

Theorem 9 *For any $f \in \mathcal{D}_0^+$ with $\|[f]\| = 1$, there exists a $g \in \mathcal{S}$ such that $[f] = [g]$, or equivalently, $q(f - g) = 0$.* — See Appendix A.5 for the proof.

So $\mathcal{D}_{\mathrm{S}}$ is the intersection of the unit sphere in $\mathcal{D}$ and the subset $\mathcal{D}^+$.

Secondly, by Theorem 7 we know $\mathcal{D}_0^+$ is a convex cone in the space $\mathcal{D}_0$. A natural problem is whether the subset $\mathcal{D}^+ = \{[f] : f \in \mathcal{D}_0^+\}$ is also a convex cone in the corresponding space $\mathcal{D} = \{[f] : f \in \mathcal{D}_0\}$. The answer is yes, as stated in the following theorem.

Theorem 10 (i) *Let $f, g \in \mathcal{D}_0^+$ be such that $q(f + g) = 0$, then $q(f) = q(g) = 0$;* (ii) *$\mathcal{D}^+$ is a convex cone in $\mathcal{D}$ pointed at $[0]$.* — See Appendix A.6 for the proof.

Finally, as pointed out earlier, each stochastic process can be characterized by a function $f \in \mathcal{S}$, which is now represented in the space $\mathcal{D}$ by the equivalence class $[f]$. From mathematics view, here the bracket $[\cdot]$ can be seen as a function from $\mathcal{D}_0$ *onto* $\mathcal{D}$; and the image of $\mathcal{S}$ under this function is $\mathcal{D}_{\mathrm{S}}$. Intuitively, this means the subset $\mathcal{D}_{\mathrm{S}}$ is "rich" enough to contain all stochastic processes and "pure" enough to exclude any unnecessary objects. Like the family $\mathcal{S}$, one may ask

"*can we also identify the family of stochastic processes with the subset $\mathcal{D}_S$?*".

An equivalent statement of the above problem is whether the function $[\cdot]$, when restricted on $\mathcal{S}$, is injective; or, whether it is possible for two different members f, g of $\mathcal{S}$ to have the same representation $[f] = [g]$ in $\mathcal{D}_{\mathrm{S}}$? This problem is indeed crucial because it actually ask whether $\mathcal{D}_{\mathrm{S}}$ is "fine" enough to distinguish any two different processes.

Theorem 11 *For any $f, g \in \mathcal{S}$, if $[f] = [g]$, i.e., if $q(f - g) = 0$, then $f = g$.*
— See Appendix A.7 for the proof.

The above three theorems give us a clear insight into the relationship between the family of stochastic processes and the families $\mathcal{S}$ and $\mathcal{D}_{\mathrm{S}}$; and the structure of the subsets $\mathcal{D}_{\mathrm{S}}$ and $\mathcal{D}^+$ in the space $\mathcal{D}$.

- The family $\mathcal{S}$ is isomorphic (i.e., one-to-one corresponding) to $\mathcal{D}_{\mathrm{S}}$ via the function $[\cdot]$; and both $\mathcal{S}$ and $\mathcal{D}_{\mathrm{S}}$ can be identified with the family of stochastic processes.
- $\mathcal{D}_{\mathrm{S}}$ is the intersection of the unit sphere and the convex cone $\mathcal{D}^+$ in the space $\mathcal{D}$. This means the family of stochastic processes can be embedded into the inner product space $\mathcal{D}$, with each process represented (uniquely) by a point on the unit sphere and in the "positive orthant" $\mathcal{D}^+$.

3.2 Constructing norm-OOMs in the space $\mathcal{D}$

In Theorem 7 we have illustrated that all left appending operators l_a leave the subspace $\mathcal{D}_0$ invariant, so we can restrict the operation of l_a's on the space $\mathcal{D}_0$. It is well known that each such restricted operator l_a induces naturally a linear operator $[l_a]$ on the quotient space $\mathcal{D}$ via

$$[l_a][f] := [l_a f] \quad \text{for all } f \in \mathcal{D}_0\,. \tag{3.13}$$

Now let $f \in \mathcal{S}$ and $\bar{a} \in O^*$ be fixed. By the definition of S_n one can easily verify that, for any $n = 0, 1, 2, \cdots$, $f^2(\bar{a}) = \sum_{\bar{x}\in O^n} f^2(\bar{a}\bar{x}) = S_n(l_{\bar{a}} f)$. Letting $n \to \infty$, we get $f(\bar{a}) = q(l_{\bar{a}} f) = \|[l_{\bar{a}} f]\|$. Assume $\bar{a} = a_1 a_2 \cdots a_n$, then by (3.13),

$$[l_{\bar{a}} f] = [l_{a_n} \cdots l_{a_1} f] = [l_{a_n}][l_{a_{n-1}} \cdots l_{a_1} f] = \cdots = [l_{a_n}] \cdots [l_{a_1}][f]\,. \tag{3.14}$$

So if we write $[l]_{\bar{a}}$ for the composition $[l_{a_n}] \cdots [l_{a_1}]$, then the identity (3.14) shows that $[l_{\bar{a}} f] = [l_{\bar{a}}][f] = [l]_{\bar{a}}[f]$ and therefore

Theorem 12 *For any $f \in \mathcal{S}$ and $\bar{a} \in O^*$, $f(\bar{a}) = \|[l_{\bar{a}}][f]\| = \|[l]_{\bar{a}}[f]\|$.*

Furthermore, the function Q defined by (3.6) has the following property.

Theorem 13 *for any $f, g \in \mathcal{D}_0$, it holds that $\sum_{a\in O} Q(l_a f, l_a g) = Q(f, g)$, i.e., $\sum_{a\in O} \langle [l_a][f], [l_a][g]\rangle = \langle [f], [g]\rangle$.* — See Appendix A.8 for the proof.

The above two theorems make the foundation for constructing norm-OOMs of a stochastic process (Y_t) from its distribution $P(\bar{a})$. Let h be the function on O^* defined by $h(\bar{a}) = \sqrt{P(\bar{a})}$, then $h \in \mathcal{S}$ and Theorem 12 tells us $h(\bar{a}) = \|[l]_{\bar{a}}[h]\|$ for all $\bar{a} \in O^*$. To get a matrix/vector representation of $[l]_{\bar{a}}$ and $[h]$, we consider the space $\mathcal{H} := \text{span}\{l_{\bar{a}} h : \bar{a} \in O^*\}$ and the subset $\mathcal{D}_h := \{[f] : f \in \mathcal{H}\}$ of $\mathcal{D}$. Since $\mathcal{D}_0$ is invariant under l_a's and since $h \in \mathcal{S} \subseteq \mathcal{D}_0$, by its definition we know $\mathcal{H}$ is a subspace of $\mathcal{D}_0$ which is also invariant under the operation of l_a's. It follows that $\mathcal{D}_h$ is a subspace of $\mathcal{D} = \{[f] : f \in \mathcal{D}_0\}$ and that $\mathcal{D}_h$ is invariant under $[l_a]$'s. Thus, we can restrict the linear operators $[l_a]$ on the space $\mathcal{D}_h$.

Assume the vector space $\mathcal{D}_h$ is of finite, say m, dimension; and select an orthonormal basis of $\mathcal{D}_h$, i.e., a basis $\{[g_i] : g_i \in H, i = 1, 2, \cdots, m\}$ with the property $\langle [g_i], [g_j]\rangle = \delta_{ij}$, where δ_{ij} is the Kronecker symbol defined by $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. It is well known that each $[f] \in \mathcal{D}_h$ can be uniquely represented as a linear combination of $\{[g_1], [g_2], \cdots, [g_m]\}$:

$$[f] = \textstyle\sum_{i=1}^m \alpha_i(f)[g_i]\,, \quad \forall\, [f] \in \mathcal{D}_h$$

which actually defines a linear map $\pi : \mathcal{D}_h \to \mathbb{R}^m$ that sends each $[f]$ to the vector $\pi[f] = [\alpha_1(f), \alpha_2(f), \cdots, \alpha_m(f)]^{\mathsf{T}}$. Since the basis $\{[g_i]\}_{i=1}^m$ is orthonormal, by the linearity of the inner product $\langle \cdot, \cdot \rangle$ we have

$$\langle [f], [g]\rangle = \textstyle\sum_{i,j} \alpha_i(f)\alpha_j(g)\langle [g_i], [g_j]\rangle = \sum_{i,j} \alpha_i(f)\alpha_j(g)\delta_{ij} = \sum_{i=1}^m \alpha_i(f)\alpha_i(g)\,,$$

i.e., $\langle [f],[g]\rangle = \{\pi[f]\}^{\mathsf{T}}\{\pi[g]\} =: \langle \pi[f],\pi[g]\rangle$ for any $[f],[g] \in \mathcal{D}_h$. It follows from Theorem 13 that $\sum_{a\in O}\langle \pi[l_a][f],\pi[l_a][g]\rangle = \langle \pi[f],\pi[g]\rangle$ for any $[f],[g] \in \mathcal{D}_h$. Now let $\varphi_a \in \mathbb{R}^{m\times m}$ be the matrix representation of the linear operator $\pi \circ [l_a] \circ \pi^{-1} : \mathbb{R}^m \to \mathbb{R}^m$ under the standard basis of $\mathbb{R}^m$ and $\boldsymbol{u}_0 := \pi[h] \in \mathbb{R}^m$ the initial state. Then $\boldsymbol{u}_0^{\mathsf{T}}\boldsymbol{u}_0 = \{\pi[h]\}^{\mathsf{T}}\{\pi[h]\} = \langle [h],[h]\rangle = 1$ for $h \in \mathcal{S}$; and

$$\begin{aligned} \boldsymbol{e}_i^{\mathsf{T}}(\textstyle\sum_{a\in O}\varphi_a^{\mathsf{T}}\varphi_a)\boldsymbol{e}_j &= \textstyle\sum_{a\in O}\{\pi[l_a][g_i]\}^{\mathsf{T}}\{\pi[l_a][g_j]\} \\ &= \textstyle\sum_{a\in O}\langle [l_a][g_i],[l_a][g_j]\rangle \\ &= \langle [g_i],[g_j]\rangle \quad \text{(by Theorem 13)} \\ &= \delta_{ij}\,, \end{aligned}$$

where $\boldsymbol{e}_i$ denotes the i-th unit vector in $\mathbb{R}^m$. The above equality shows that $\sum_{a\in O}\varphi^{\mathsf{T}}\varphi = I_m$. Furthermore, for any $\bar{a} \in O^*$ it holds that $\pi[l]_{\bar{a}}[h] = \varphi_{\bar{a}}\boldsymbol{u}_0$, so by Theorem 12 we can compute

$$P(\bar{a}) = h^2(\bar{a}) = \langle [l]_{\bar{a}}[h],[l]_{\bar{a}}[h]\rangle = \{\pi[l]_{\bar{a}}[h]\}^{\mathsf{T}}\{\pi[l]_{\bar{a}}[h]\} = (\varphi_{\bar{a}}\boldsymbol{u}_0)^{\mathsf{T}}(\varphi_{\bar{a}}\boldsymbol{u}_0)\,.$$

In conclusion, from the distribution $P(\bar{a})$ of some process (Y_t), we constructed the structure $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ for which it holds that (1) $\|\boldsymbol{u}_0\| := \sqrt{\boldsymbol{u}_0^{\mathsf{T}}\boldsymbol{u}_0} = 1$; (2) $\sum_{a\in O}\varphi_a^{\mathsf{T}}\varphi_a = I_m$; and (3) $P(\bar{a}) = \|\varphi_{\bar{a}}\boldsymbol{u}_0\|^2$ for all $\bar{a} \in O^*$. Such structures, as defined earlier, are called norm-OOMs of (Y_t).

We now consider the problem of how to randomly construct a norm-OOM, which is important for both theoretical investigation and numerical simulation. The same problem exist for HMMs and OOMs. For HMMs, there is an easy and efficient way to create a random HMM of given dimension (i.e., number of hidden states) m over a given alphabet $O = \{1,2,\cdots,\ell\}$: one first randomly creates an $m \times m$ matrix A and an $m \times \ell$ matrix B, both consisting only of nonnegative entries; then normalize each row of A and B so that the two matrices both have row sums 1; the resulting A and B are the *transition* matrix and the *emission* matrix of the HMM, respectively. For OOMs, one can either create a HMM and then convert it to its equivalent OOM — easy but trivial; or, by Theorem 4, construct a proper convex cone K in $\mathbb{R}^m$ and then design τ_a's such that $\tau_a K \subseteq K$ and $\mathbf{1}\sum_{a\in O}\tau_a = \mathbf{1}$ — not easy at all: in fact, only a simple nontrivial example has been constructed so far, the "probability clock" [6]. So we actually have not found a nontrivial general way for constructing valid OOMs from scratch so far. Fortunately, a simple and efficient construction of random norm-OOMs is possible, as shown below.

Let φ be the $m\ell \times m$ matrix created by stacking the matrices φ_a below one another, i.e., $\varphi := [\varphi_1^{\mathsf{T}},\varphi_2^{\mathsf{T}},\cdots,\varphi_\ell^{\mathsf{T}}]^{\mathsf{T}}$. Then the condition (ii) from Definition 2 is equivalent to $\varphi^{\mathsf{T}}\varphi = I_m$, which means the columns of φ form an orthonormal set in $\mathbb{R}^{m\ell}$. So here we can first randomly create m $m\ell$-dimensional vectors; then, using the Gram-Schmidt process, make them an orthonormal set to get a $m\ell \times m$ matrix φ satisfying $\varphi^{\mathsf{T}}\varphi = I_m$; dividing φ into m blocks of equal size, we get the desired observable operators φ_a ($a \in O$). The initial state $\boldsymbol{u}_0$ can be any vector in $\mathbb{R}^m$ with norm 1.

Like the class of OOMs, we say that two norm-OOMs $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ and $(\mathbb{R}^n, \{\varphi'_a\}_{a\in O}, \boldsymbol{u}'_0)$ are equivalent if they describe the same process (Y_t). For the equivalence of two norm-OOMs, we have the following sufficient condition.

Theorem 14 *Let $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ and $(\mathbb{R}^n, \{\varphi'_a\}_{a\in O}, \boldsymbol{u}'_0)$ be two norm-OOMs such that $n \geqslant m$. Then they are equivalent to each other if there exists a matrix $\varrho \in \mathbb{R}^{n\times m}$ such that $\varrho^{\mathsf{T}}\varrho = I_m$, $\boldsymbol{u}'_0 = \varrho\boldsymbol{u}_0$ and $\varphi'_a = \varrho\varphi_a\varrho^{\mathsf{T}}$ for all $a \in O$.*

— The proof is obvious and omitted here.

3.3 Norm-OOMs as generators and predictors

We explain in this subsection how to generate and predict the paths of a process (Y_t) modelled by a norm-OOM $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ through $P(\bar{a}) = \|\varphi_{\bar{a}}\boldsymbol{u}_0\|^2$.

In the generation task we are required to randomly produce, at time steps $t = 1, 2, \cdots$, outcomes $a_1, a_2, \cdots \in O$, such that (i) at time $t = 1$, the probability of producing b is equal to $P(b) = \|\varphi_b\boldsymbol{u}_0\|^2$, and (ii) at each time $t = n+1$ ($n \in \mathbb{N}$), the probability of producing b (assume $\bar{a} := a_1a_2\cdots a_n$ have already been created) is equal to $P(b|\bar{a})$. From $P(\bar{a}) = \|\varphi_{\bar{a}}\boldsymbol{u}_0\|^2$, the conditional probability $P(b|\bar{a})$ can be expanded into

$$P(b|\bar{a}) = \frac{P(\bar{a}b)}{P(\bar{a})} = \frac{\|\varphi_b\varphi_{\bar{a}}\boldsymbol{u}_0\|^2}{\|\varphi_{\bar{a}}\boldsymbol{u}_0\|^2} =: \|\varphi_b\boldsymbol{u}_{\bar{a}}\|^2\,, \tag{3.15}$$

where $\boldsymbol{u}_{\bar{a}} = \varphi_{\bar{a}}\boldsymbol{u}_0/\|\varphi_{\bar{a}}\boldsymbol{u}_0\|$ is the *state vector* of the norm-OOM on $\bar{a}$. Note that all state vectors have norm 1 and can be incrementally calculated by

$$\boldsymbol{u}_\varepsilon = \boldsymbol{u}_0\,, \quad \boldsymbol{u}_{a_1a_2\cdots a_n} = \frac{\varphi_{a_n}\boldsymbol{u}_{a_1a_2\cdots a_{n-1}}}{\|\varphi_{a_n}\boldsymbol{u}_{a_1a_2\cdots a_{n-1}}\|}\,. \tag{3.16}$$

Based on the above discussion, we summarize the procedure for generating sample paths of a process from its norm-OOM $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$: (1) at time $t = 1$, generate a_1 according to the probability distribution $P(b) = \|\varphi_b\boldsymbol{u}_0\|^2$ and compute the state vector $\boldsymbol{u}_1 = \varphi_b\boldsymbol{u}_0/\|\varphi_b\boldsymbol{u}_0\|$; (2) assume we have got the $(n-1)$-th state vector $\boldsymbol{u}_{n-1}$, then generate a_n according to the distribution $P(b) = \|\varphi_b\boldsymbol{u}_{n-1}\|^2$ and update the state vector by $\boldsymbol{u}_n = \varphi_b\boldsymbol{u}_{n-1}/\|\varphi_b\boldsymbol{u}_{n-1}\|$.

Norm-OOMs can also be used as predictors: given an initial path $\bar{a} = a_1a_2\cdots a_n$ of the process up to time $t = n$, we want to calculate the probability that the next outcome is b. Again, this amounts to the computation of the conditional probabilities $P(b|\bar{a})$; and eqns (3.15)(3.16) can be employed here. But note that now the initial path $\bar{a}$ is not generated by the norm-OOM itself but is externally given.

In next subsection we will show that any norm-OOM can be converted to an equivalent OOM. So one can also first convert a given norm-OOM to its equivalent OOM; and then use this OOM as the generator/predictor to create/predict the outcome b that is going to occur at the next time step. See Section 3 of [6] for the detailed procedure.

Another usage of eqns (3.15)(3.16) is the evaluation of the probability $P(\bar{a})$ of an initial path $\bar{a} = a_1 a_2 \cdots a_n$. Notice that, here we cannot directly use the formular $P(\bar{a}) = \|\varphi_{\bar{a}} \boldsymbol{u}_0\|^2$, for the decrease of $P(\bar{a})$ with the increase of length n of $\bar{a}$ is so quick that numerically it is almost sure that $P(\bar{a}) = 0$, even for relatively small n. So instead of directly calculating $P(\bar{a})$, one should evaluate its logarithm $L(\bar{a}) := \log P(\bar{a})$. This can be done using eqns (3.15)(3.16), as follows: (1) for $k = 1, 2, \cdots, n$ compute $\boldsymbol{x}_k = \varphi_{a_k} \boldsymbol{u}_{k-1}$, $c_k = \|\boldsymbol{x}_k\|$ and $\boldsymbol{u}_k = c_k^{-1} \boldsymbol{x}_k$; (2) calculate $L(\bar{a}) = 2 \sum_{k=1}^{n} \log c_k$.

3.4 The expressiveness of norm-OOMs

In the last two subsections, we defined the class of norm-OOMs and presented a general method for constructing norm-OOMs from the distribution of the underlying process. In this subsection we will study the family of stochastic processes that can be described by (finite-dimensional) norm-OOMs.

We see an exemplary 2-dimensional norm-OOM over the alphabet $O = \{a, b\}$:

$$\varphi_a = \begin{bmatrix} .6c & -s \\ .6s & c \end{bmatrix}, \varphi_b = \begin{bmatrix} .8 & 0 \\ 0 & 0 \end{bmatrix}; \text{ and } \boldsymbol{u}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{3.17}$$

with $c = \cos(0.5)$ and $s = \sin(0.5)$. Intuitively, the operation of φ_a on the state vector $\boldsymbol{u} = [x, y]^\mathsf{T}$ is shrinking the value of x by factor 0.6 and then rotating the resulted vector by an angle $\theta = 0.5$; and that of φ_b is shrinking the value of x by 0.8 and discarding y, as depicted in Figure 1.

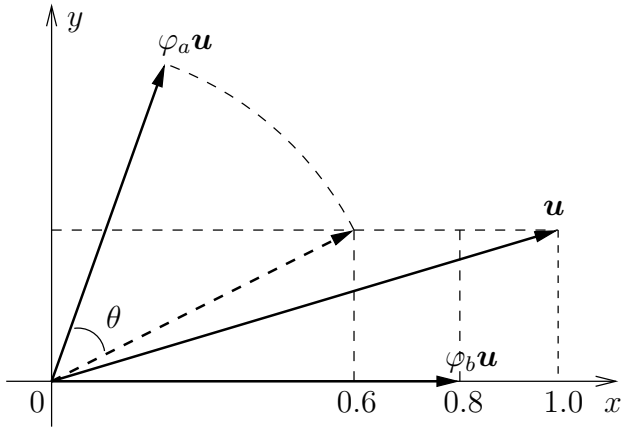


Figure 1: The operation of φ_a and φ_b on the vector $\boldsymbol{u}$.

From Figure 1 we see that, after the operation of φ_b, one always obtains a vector $\varphi_b \boldsymbol{u}$ lying on the x-axis. It then follows from (3.16) that

$$\boldsymbol{u}_{\bar{x}b} = \varphi_b \boldsymbol{u}_{\bar{x}} / \|\varphi_b \boldsymbol{u}_{\bar{x}}\| = [\pm 1, 0]^\mathsf{T};$$

and from (3.15) that

$$\begin{array}{rcccccl} P(a|\bar{x}b) & = & \|\varphi_a \boldsymbol{u}_{\bar{x}b}\|^2 & = & 0.36 & = & P(a|b)\,, \\ P(b|\bar{x}b) & = & \|\varphi_b \boldsymbol{u}_{\bar{x}b}\|^2 & = & 0.64 & = & P(b|b)\,; \end{array}$$

for any $\bar{x} \in O^*$. Thus, as $\boldsymbol{u}_0 = [1, 0]^\mathsf{T}$, the process described by the norm-OOM as in (3.17) is completely characterized by the family of the conditional probabilities $\{P(b|ba^n) = P(b|a^n) : n = 0, 1, 2, \cdots\}$, where a^n denotes the sequence consisting of n a's. Iteratively using eqns (3.15)(3.16), we are able to compute the values of $P(b|a^n)$. These conditional probabilities are plotted in Figure 2, from which it shows that the behavior of the norm-OOM specified by (3.17) is very similar to that of the so called probability clock [3] . As explained in [6], such *nonperiodic* oscillating behavior of $P(b|a^n)$ cannot be captured by (finite) HMMs.

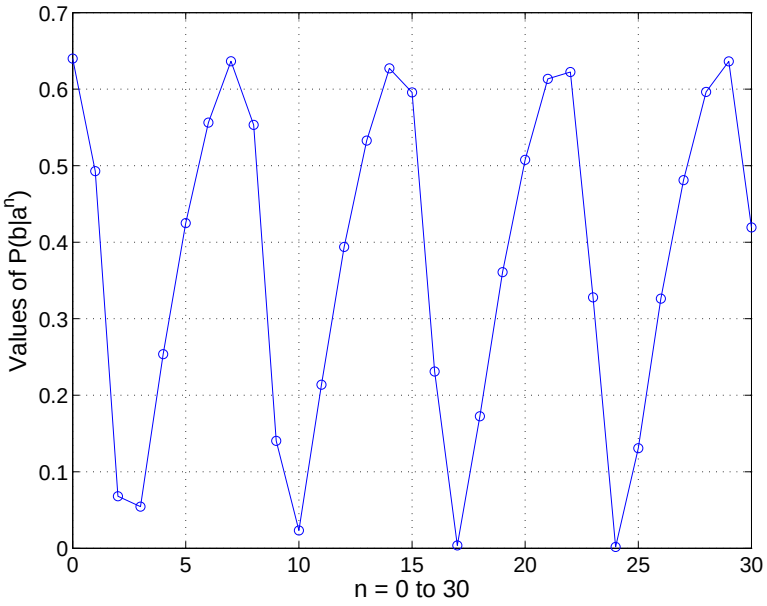


Figure 2: The curve of conditional probabilities $P(b|a^n)$.

This simple example shows that norm-OOMs, like OOMs, can describe some specific stochastic processes that cannot be modelled by HMMs. But norm-OOMs provide no more than that provided by OOMs. In other words, each norm-OOM can be equivalently converted to an OOM, as shown below.

Definition 4 *For two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, the* Kronecker product *of A and B, denoted $A \otimes B$, is the (blocked) matrix of size $mp \times nq$ with $a_{ij}B$ as its (i, j)-th block, where a_{ij} is the element of A at position (i, j). Furthermore, we write* vec(A) *for the mn-dimensional column vector formed by stacking the columns of A one below another.*

[3] A 3-dimensional OOM which represents a stochastic process that cannot be described by (finite) HMMs, see Section 6 of [6] for the detail.

For example, given

$$A = \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix},$$

their Kronecker product $A \otimes B$ and the vector $\mathrm{vec}(A)$ are

$$A \otimes B = \begin{bmatrix} 1 & 2 & -1 & -2 \\ 3 & 4 & -3 & -4 \\ 2 & 4 & 0 & 0 \\ 6 & 8 & 0 & 0 \end{bmatrix}, \qquad \mathrm{vec}(A) = \begin{bmatrix} 1 \\ 2 \\ -1 \\ 0 \end{bmatrix},$$

respectively. The reader is referred to [4] for a detailed introduction to the Kronecker product $A \otimes B$ and the "stack operator" $\mathrm{vec}(A)$, and to [1] for a quick reference. Here we only point out two facts that will be used when converting a norm-OOM to its equivalent OOM.

Theorem 15 *When dimensions are appropriate, (1)* $(A \otimes C)(B \otimes D) = AB \otimes CD$*; and (2)* $\boldsymbol{x}^{\mathsf{T}} A \boldsymbol{y} = [\mathrm{vec}(A)]^{\mathsf{T}}(\boldsymbol{x} \otimes \boldsymbol{y})$*. In particular, for* $\boldsymbol{x} \in \mathbb{R}^m$*, it holds that (3)* $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} = [\mathrm{vec}(I_m)]^{\mathsf{T}}(\boldsymbol{x} \otimes \boldsymbol{x})$.

Now let $(\mathbb{R}^m, \{\varphi_a\}_{a \in O}, \boldsymbol{u}_0)$ be a norm-OOM of some process (Y_t). Then by Theorem 15 the probabilities $P(\bar{a})$ for $\bar{a} = a_1 a_2 \cdots a_n$ can be evaluated by

$$\begin{aligned} P(\bar{a}) &= \|\varphi_{a_n} \cdots \varphi_{a_2} \varphi_{a_1} \boldsymbol{u}_0\|^2 \\ &= [\mathrm{vec}(I_m)]^{\mathsf{T}} (\varphi_{a_n} \cdots \varphi_{a_2} \varphi_{a_1} \boldsymbol{u}_0 \otimes \varphi_{a_n} \cdots \varphi_{a_2} \varphi_{a_1} \boldsymbol{u}_0) \\ &= [\mathrm{vec}(I_m)]^{\mathsf{T}} (\varphi_{a_n} \otimes \varphi_{a_n}) \cdots (\varphi_{a_2} \otimes \varphi_{a_2})(\varphi_{a_1} \otimes \varphi_{a_1})(\boldsymbol{u}_0 \otimes \boldsymbol{u}_0). \end{aligned}$$

Putting $\boldsymbol{\sigma} = [\mathrm{vec}(I_m)]^{\mathsf{T}}$, $\tau_a = \varphi_a \otimes \varphi_a$ and $\boldsymbol{w}_0 = \boldsymbol{u}_0 \otimes \boldsymbol{u}_0$, we get an "almost OOM" $(\mathbb{R}^{m^2}, \{\tau_a\}_{a \in O}, \boldsymbol{w}_0, \boldsymbol{\sigma})$, which computes the probabilities $P(\bar{a})$ by $P(\bar{a}) = \boldsymbol{\sigma} \tau_{\bar{a}} \boldsymbol{w}_0$, and is only "one step (from the functional $\boldsymbol{\sigma}$ to the standard one $\mathbf{1}$)" from "real OOMs". But it is easy to find a basis transition matrix ϱ in the space $\mathbb{R}^{m^2}$ such that $\mathbf{1}\varrho = \boldsymbol{\sigma}$, through which one gets an equivalent OOM $(\mathbb{R}^{m^2}, \{\varrho \tau_a \varrho^{-1}\}_{a \in O}, \varrho \boldsymbol{w}_0)$ of the original norm-OOM (and can then minimize it to get an equivalent minimal OOM).

For example, the norm-OOM defined by (3.17) is equivalent to a 3-dimensional OOM with observable operators

$$\tau_a = \begin{bmatrix} -0.026 & -0.243 & 1.166 \\ 0.386 & 0.420 & -1.107 \\ 0 & 0.320 & 0.977 \end{bmatrix}, \quad \tau_b = \begin{bmatrix} 0.64 & 0.503 & -0.036 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix};$$

and initial state $\boldsymbol{w}_0 = [1, 0, 0]^{\mathsf{T}}$. As the eigenvalues of τ_a are 0.6 and $0.386 \pm 0.460i$ (complex numbers), the operation of τ_a includes a basis transition, vector rotation under the new basis, and the inverse basis transition; while the operation of τ_b is mapping any state vector $\boldsymbol{w} \in \mathbb{R}^3$ to a new one lying on the x-axis. This means

the norm-OOM (3.17) actually represents a "non-standard probability clock", as depicted in Figure 2.

The above discussion shows that norm-OOMs can be seen as a subclass of OOMs. Now we consider the problem in the reverse direction: which OOMs have equivalent norm-OOMs? So far there is little significant result obtained on this problem, except the following sufficient condition.

Theorem 16 *Any OOM* $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ *with positive parameters in which each row of each operator* τ_a *has at most one nonzero element has an equivalent norm-OOM* $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ *defined by* $\varphi_a = \sqrt{\tau_a}$ *and* $\boldsymbol{u}_0 = \sqrt{\boldsymbol{w}_0}$*, where the square root is defined entry-wise.* — See Appendix A.9 for the proof.

Notice that an m-states Markov chain (MC) can be represented as an m-dimensional OOM with each operator τ_a consisting of all but one zero columns. So by the above theorem we know any Markov chain has equivalent norm-OOMs.

Based upon the above discussion, the relationship between MCs, HMMs, norm-OOMs and OOMs can be expressed informally as

$$\text{norm-OOMs} \not\subseteq \text{HMMs}\,, \qquad \text{MCs} \subset \text{HMMs} \subset \text{OOMs}\,,$$
$$\text{MCs} \subset \text{norm-OOMs} \subseteq \text{OOMs}\,.$$

Thus far it is not clear whether or not "HMMs $\subseteq$ norm-OOMs" and "norm-OOMs = OOMs". We end this section with a picture depicting these relations.

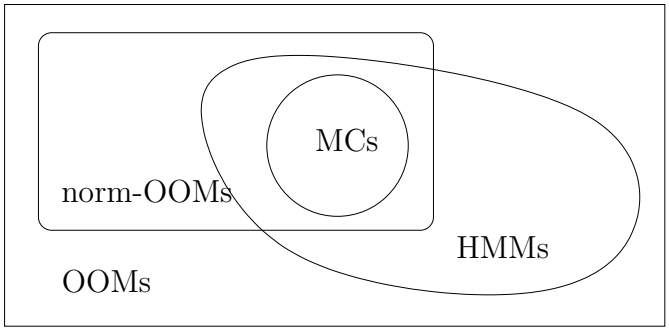


Figure 3: The relationships between MCs, HMMs, norm-OOMs and OOMs.

4 A Maximum-Likelihood Learning Algorithm

We introduce in this section an iterative algorithm for learning norm-OOMs from data based on the *maximum-likelihood* (ML) principle. The ML principle is quite simple: for our case it estimates a norm-OOM $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ from a given training sequence $\bar{s} = s_1 s_2 \cdots s_N$ by maximizing the likelihood $P(\bar{s}|\varphi_a, \boldsymbol{u}_0)$. In practice, for computational reasons we usually use the log-likelihood

$$L = \tfrac{1}{2}\log P(\bar{s}|\varphi_a, \boldsymbol{u}_0) = \tfrac{1}{2}\log(\boldsymbol{u}_0^{\mathsf{T}}\varphi_{s_1}^{\mathsf{T}}\varphi_{s_2}^{\mathsf{T}}\cdots\varphi_{s_N}^{\mathsf{T}}\varphi_{s_N}\cdots\varphi_{s_2}\varphi_{s_1}\boldsymbol{u}_0)\,, \tag{4.1}$$

instead of the likelihood $P(\bar{s}|\varphi_a, \boldsymbol{u}_0)$, as the target function. Note that in (4.1) a factor $\frac{1}{2}$ is introduced to simplify the computation hereafter. By Definition 2 the ML principle amounts to the following optimization problem

$$\begin{array}{ll} \text{maximize} & L = \frac{1}{2}\log(\boldsymbol{u}_0^{\mathsf{T}}\varphi_{s_1}^{\mathsf{T}}\varphi_{s_2}^{\mathsf{T}}\cdots\varphi_{s_N}^{\mathsf{T}}\varphi_{s_N}\cdots\varphi_{s_2}\varphi_{s_1}\boldsymbol{u}_0)\,, \\ \text{subject to} & \sum_{a\in O}\varphi_a^{\mathsf{T}}\varphi_a = I_m\,, \quad \boldsymbol{u}_0^{\mathsf{T}}\boldsymbol{u}_0 = 1\,; \end{array} \tag{4.2}$$

and the key point here is to develop an efficient method for this problem.

4.1 Two "local" operations on observable operators

As analytical solutions of (4.2) are unavailable, some (iterative) numerical method should be employed. This numerical method should keep the two constraints of (4.2) during the iterations, so that whenever the algorithm stops one immediately gets a valid norm-OOM. As before, we write φ for the matrix formed by stacking the matrices φ_a one below another. Then the identity $\sum_{a\in O}\varphi_a^{\mathsf{T}}\varphi_a = I_m$ can be rewritten as $\varphi^{\mathsf{T}}\varphi = I_m$. This means the columns of φ always form an orthonormal set in the space $\mathbb{R}^{m\ell}$. Now we see the log-likelihood L as a function of φ and $\boldsymbol{u}_0$ and assume $(\varphi^*, \boldsymbol{u}_0^*)$ is the optimal point of the problem (4.2). It is well known the two orthonormal sets φ and φ^* are related by some unitary matrix U of order $m\ell$, i.e., a matrix U with the property $UU^{\mathsf{T}} = I_{m\ell}$, through $\varphi^* = U\varphi$. Furthermore, concerning unitary matrices we have the following proposition.

Theorem 17 *A square matrix U of order $n \geqslant 2$ is a unitary matrix iff it is the product of some (simpler) matrices of the form $G(i,j,\theta)$ or $G'(i,j,\theta)$, where $i \neq j$ and $G(i,j,\theta)$ [4] , $G'(i,j,\theta)$ denote the matrices formed from the identity matrix I_n by changing its 2×2 submatrix at the cross of i-th and j-th rows/columns to*

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \qquad \begin{bmatrix} -\cos\theta & \sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$

respectively. Note that $G(i,j,\theta) = G(j,i,-\theta)$ and $G'(i,j,\theta) = G'(j,i,\pi-\theta)$, so in the above assertion we can require that $i < j$.

Furthermore, let $D(i)$ be the diagonal matrix with all diagonal entries being 1 except the i-th one which is -1. Then $G'(i,j,\theta) = G(i,j,\theta)D(i)$ and the theorem can be equivalently stated as: U is a unitary matrix iff it is the product of several $G(i,j,\theta)$'s and $D(i)$'s.

Although (we believe) this theorem should be a known result in matrix theory, a detailed proof (of our own) is provided in Appendix A.10 for completeness.

Let $\boldsymbol{\varphi}^i$ be the i-th row of the matrix φ. Then Theorem 17 tells us: starting from any initial matrix φ, we are able to reach the optimal one φ^* by (iteratively) using the following two "local" operations.

[4] Such matrices $G(i,j,\theta)$ are called *Givens matrices*.

- The operation $D(i)$ — select a row $\boldsymbol{\varphi}^j$ of φ and reverse its direction, i.e., $\boldsymbol{\varphi}^i_+ \leftarrow (-\boldsymbol{\varphi}^i)$, where the subscript + means the updated value.
- The operation $G(i,j,\theta)$ — select two rows $\boldsymbol{\varphi}^i$, $\boldsymbol{\varphi}^j$ of φ and do the linear combination $\boldsymbol{\varphi}^i_+ \leftarrow (\boldsymbol{\varphi}^i \cos\theta + \boldsymbol{\varphi}^j \sin\theta)$ and $\boldsymbol{\varphi}^j_+ \leftarrow (-\boldsymbol{\varphi}^i \sin\theta + \boldsymbol{\varphi}^j \cos\theta)$.

Furthermore, it is easy to verify that, after each of the above operations, the resulting matrix φ_+ still has the property $\varphi_+^{\mathsf{T}}\varphi_+ = I_m$, representing observable operators of a valid norm-OOM. From these observations we get a greedy scheme for adjusting the matrix φ iteratively so that the log-likelihood $L(\varphi, \boldsymbol{u}_0)$ increases for each iteration, as outlined below.

1. Select (cyclically or randomly) two rows $\boldsymbol{\varphi}^i$ and $\boldsymbol{\varphi}^j$ of φ.
2. Put $\boldsymbol{\varphi}^i_+ \leftarrow (-\boldsymbol{\varphi}^i)$ if this increases the value of $L(\varphi, \boldsymbol{u}_0)$.
3. Fix other parameters and consider the operations

$$\boldsymbol{\varphi}^i_+ \leftarrow (\boldsymbol{\varphi}^i \cos\theta + \boldsymbol{\varphi}^j \sin\theta) \quad \text{and} \quad \boldsymbol{\varphi}^j_+ \leftarrow (-\boldsymbol{\varphi}^i \sin\theta + \boldsymbol{\varphi}^j \cos\theta)\,. \tag{4.3}$$

   Then the log-likelihood L is a function of the single parameter $\theta \in [-\pi, \pi)$; and the gradient method can be used here. More precisely, we compute the derivative $L'(\theta) = \frac{\partial L}{\partial \theta}$ at $\theta = 0$ and set $\theta = \eta \cdot \operatorname{sgn}\{L'(0)\}$ or $\theta = \eta \cdot L'(0)$ in (4.3) to update the matrix φ, where $\eta > 0$ is the learning rate.
4. Repeat the above procedure until some termination criterion is satisfied.

4.2 The forward-backward algorithm for norm-OOMs

This subsection is devoted to the calculation of the derivative $\frac{\partial L}{\partial \theta}$ when (4.3) is applied to the matrix φ, as well as the derivative $\frac{\partial L}{\partial \boldsymbol{u}_0}$ for deriving the update rule of $\boldsymbol{u}_0$. To this end we need some short hand notations. For a numerical function f defined on some matrix X, we write $\frac{\partial f}{\partial X}$ for the matrix of the same size as X with (i,k)-th entry $\frac{\partial f}{\partial x_{ik}}$, where x_{ik} is the (i,k)-th element of X; and for a matrix function X defined on a single variable t, we use $\frac{\partial X}{\partial t}$ to denote the matrix of the same size as X with $\frac{\partial x_{ik}}{\partial t}$ as its (i,k)-th element. Then, when φ is changed by (4.3), we can compute

$$\begin{aligned}\left[\frac{\partial L}{\partial \theta}\right]_{\theta=0} &= \left[\left(\frac{\partial L}{\partial \boldsymbol{\varphi}^i_+}\right)\cdot\left(\frac{\partial \boldsymbol{\varphi}^i_+}{\partial \theta}\right)^{\mathsf{T}} + \left(\frac{\partial L}{\partial \boldsymbol{\varphi}^j_+}\right)\cdot\left(\frac{\partial \boldsymbol{\varphi}^j_+}{\partial \theta}\right)^{\mathsf{T}}\right]_{\theta=0} \\ &= \left(\frac{\partial L}{\partial \boldsymbol{\varphi}^i}\right)\cdot(\boldsymbol{\varphi}^j)^{\mathsf{T}} - \left(\frac{\partial L}{\partial \boldsymbol{\varphi}^j}\right)\cdot(\boldsymbol{\varphi}^i)^{\mathsf{T}}\,. \end{aligned} \tag{4.4}$$

Let $\psi := \frac{\partial L}{\partial \varphi}$ be the derivative of L w.r.t. φ and $\boldsymbol{\psi}^i$ the i-th row of ψ, then (4.4) can be rewritten as $\frac{\partial L}{\partial \theta}|_{\theta=0} = \boldsymbol{\psi}^i \cdot (\boldsymbol{\varphi}^j)^{\mathsf{T}} - \boldsymbol{\psi}^j \cdot (\boldsymbol{\varphi}^i)^{\mathsf{T}}$. But by its definition ψ is the $m\ell \times m$ matrix created by stacking $\psi_a := \frac{\partial L}{\partial \varphi_a}$'s one below another, so to get $\frac{\partial L}{\partial \theta}|_{\theta=0}$ we need to calculate $\frac{\partial L}{\partial \varphi_a}$'s.

From its definition $L = \frac{1}{2}\log(\boldsymbol{u}_0^\mathsf{T}\varphi_{s_1}^\mathsf{T}\varphi_{s_2}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\varphi_{s_N}\cdots\varphi_{s_2}\varphi_{s_1}\boldsymbol{u}_0)$ we know

$$\frac{\partial L}{\partial \varphi_{s_k}} = \frac{\varphi_{s_{k+1}}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\varphi_{s_N}\cdots\varphi_{s_1}\boldsymbol{u}_0\boldsymbol{u}_0^\mathsf{T}\varphi_{s_1}^\mathsf{T}\cdots\varphi_{s_{k-1}}^\mathsf{T}}{\boldsymbol{u}_0^\mathsf{T}\varphi_{s_1}^\mathsf{T}\varphi_{s_2}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\varphi_{s_N}\cdots\varphi_{s_2}\varphi_{s_1}\boldsymbol{u}_0}\,. \tag{4.5}$$

For $k = 1, 2, \cdots, N$ define

$$\boldsymbol{u}_k := \frac{\varphi_{s_k}\cdots\varphi_{s_1}\boldsymbol{u}_0}{\|\varphi_{s_k}\cdots\varphi_{s_1}\boldsymbol{u}_0\|}\,, \quad \boldsymbol{v}_k := \frac{\varphi_{s_{k+1}}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\boldsymbol{u}_N}{\boldsymbol{u}_{k-1}^\mathsf{T}\varphi_{s_k}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\boldsymbol{u}_N}\,.$$

Then (4.5) can be rewritten as

$$\frac{\partial L}{\partial \varphi_{s_k}} = \frac{\varphi_{s_{k+1}}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\boldsymbol{u}_N\boldsymbol{u}_{k-1}^\mathsf{T}}{\boldsymbol{u}_{k-1}^\mathsf{T}\varphi_{s_k}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\boldsymbol{u}_N} = \boldsymbol{v}_k\boldsymbol{u}_{k-1}^\mathsf{T}\,;$$

and it follows that $\psi_a = \frac{\partial L}{\partial \varphi_a} = \sum_{k:s_k=a}\boldsymbol{v}_k\boldsymbol{u}_{k-1}^\mathsf{T}$. By their definition the vectors $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ can be iteratively calculated as follows:

1. *forward procedure*: starting from the initial vector $\boldsymbol{u}_0$, for $k = 1, 2, \cdots, N$ compute $\boldsymbol{u}_k' = \varphi_{s_k}\boldsymbol{u}_{k-1}$, $c_k = \|\boldsymbol{u}_k'\|$ and $\boldsymbol{u}_k = c_k^{-1}\boldsymbol{u}_k'$.
2. *backward procedure*: starting from $\boldsymbol{v}_N = c_N^{-1}\boldsymbol{u}_N$, for $k = N-1, \cdots, 2, 1$ compute $\boldsymbol{v}_k = c_k^{-1}\varphi_{s_{k+1}}^\mathsf{T}\boldsymbol{v}_{k+1}$. — In fact, by the definition of c_k we have $\varphi_{s_k}\boldsymbol{u}_{k-1} = c_k\boldsymbol{u}_k$, thus

$$\boldsymbol{v}_k = \frac{\varphi_{s_{k+1}}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\boldsymbol{u}_N}{c_k\boldsymbol{u}_k^\mathsf{T}\varphi_{s_{k+1}}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\boldsymbol{u}_N} = c_k^{-1}\varphi_{s_{k+1}}^\mathsf{T}\boldsymbol{v}_{k+1}\,, \tag{4.6}$$

   with initial value $\boldsymbol{v}_N = \boldsymbol{u}_N/(c_N\boldsymbol{u}_N^\mathsf{T}\boldsymbol{u}_N) = c_N^{-1}\boldsymbol{u}_N$ since $\|\boldsymbol{u}_N\| = 1$.

In this way, we are able to compute the derivative matrix $\psi = \frac{\partial L}{\partial \varphi}$. Furthermore, using the auxiliary vectors $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$, we have

$$\frac{\partial L}{\partial \boldsymbol{u}_0} = \frac{\varphi_{s_1}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\varphi_{s_N}\cdots\varphi_{s_1}\boldsymbol{u}_0}{\boldsymbol{u}_0^\mathsf{T}\varphi_{s_1}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\varphi_{s_N}\cdots\varphi_{s_1}\boldsymbol{u}_0} = \frac{\varphi_{s_1}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\boldsymbol{u}_N}{\boldsymbol{u}_0^\mathsf{T}\varphi_{s_1}^\mathsf{T}\cdots\varphi_{s_N}^\mathsf{T}\boldsymbol{u}_N} = \varphi_{s_1}^\mathsf{T}\boldsymbol{v}_1\,. \tag{4.7}$$

So if we define $c_0 = 1$ in (4.6), then the above forward-backward procedure can be used also to calculate the derivative $\boldsymbol{v}_0 := \frac{\partial L}{\partial \boldsymbol{u}_0}$ of L w.r.t. the initial state $\boldsymbol{u}_0$. To get each vector $\boldsymbol{u}_k$ or $\boldsymbol{v}_k$ we should multiply a matrix to a vector, which costs m^2 flops; so the complexity of the forward-backward algorithm is $O(Nm^2)$.

4.3 Learning norm-OOMs from data

This subsection introduces in detail the learning algorithm of norm-OOMs. We first clarify the learning task as follows: given the training data $\bar{s} = s_1 s_2 \cdots s_N$ and the model dimension m, find a norm-OOM $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ so that the likelihood $P(\bar{s}|\varphi_a, \boldsymbol{u}_0)$ is maximal.

As was stated earlier, the learning algorithm modifies the observable operators φ_a iteratively: at each iteration only 1 or 2 rows selected from these operators, say $\boldsymbol{\varphi}^i$ and $\boldsymbol{\varphi}^j$, are updated via (possibly) $\boldsymbol{\varphi}^i_+ \leftarrow (-\boldsymbol{\varphi}^i)$ and (4.3), which should increase the likelihood of the estimated model on the given training sequence. For this we need to determine the parameter θ in (4.3). When seen as a function of θ, the log-likelihood $L(\theta)$ has the Taylor expansion $L(\theta) = L(0) + \theta \cdot \frac{\partial L}{\partial \theta}|_{\theta=0} + o(\theta)$. Substituting $\frac{\partial L}{\partial \theta}|_{\theta=0} = \boldsymbol{\psi}^i(\boldsymbol{\varphi}^j)^\mathsf{T} - \boldsymbol{\psi}^j(\boldsymbol{\varphi}^i)^\mathsf{T}$ into this expansion and omitting the infinitesimal (i.e., infinitely small) item $o(\theta)$, we get

$$L(\theta) \approx L(0) + \theta \cdot [\boldsymbol{\psi}^i(\boldsymbol{\varphi}^j)^\mathsf{T} - \boldsymbol{\psi}^j(\boldsymbol{\varphi}^i)^\mathsf{T}] \,.$$

So to increase the value of $L(\theta)$ one should set θ so that it has the same sign as $\boldsymbol{\psi}^i(\boldsymbol{\varphi}^j)^\mathsf{T} - \boldsymbol{\psi}^j(\boldsymbol{\varphi}^i)^\mathsf{T}$, e.g., $\theta = \eta \cdot \{\boldsymbol{\psi}^i(\boldsymbol{\varphi}^j)^\mathsf{T} - \boldsymbol{\psi}^j(\boldsymbol{\varphi}^i)^\mathsf{T}\}$ with $\eta > 0$. Furthermore, one may want to increase $L(\theta)$ as quickly as possible, which means we should select such indices i, j that make the absolute value of $\boldsymbol{\psi}^i(\boldsymbol{\varphi}^j)^\mathsf{T} - \boldsymbol{\psi}^j(\boldsymbol{\varphi}^i)^\mathsf{T}$ largest.

Now we consider the modification of the initial state $\boldsymbol{u}_0$. Fix the observable operators φ_a and let $A = (\varphi_{s_N} \cdots \varphi_{s_1})^\mathsf{T} \varphi_{s_N} \cdots \varphi_{s_1}$, then problem (4.2) is equivalent to $\max\{\boldsymbol{u}_0^\mathsf{T} A \boldsymbol{u}_0 : \boldsymbol{u}_0^\mathsf{T} \boldsymbol{u}_0 = 1\}$. Since A is a symmetric, positive-definite matrix, it is well known that, when $\boldsymbol{u}_0$ is the *dominant eigenvector* of A, (i.e., the eigenvector w.r.t. the largest eigenvalue λ_1), $\boldsymbol{u}_0^\mathsf{T} A \boldsymbol{u}_0$ reaches its maximum which is exactly λ_1. However, the computation of $A = (\varphi_{s_N} \cdots \varphi_{s_1})^\mathsf{T} \varphi_{s_N} \cdots \varphi_{s_1}$ (which costs $O(Nm^3)$ flops) and the dominant eigenpair of A is too expensive to be used in our *iterative* learning algorithm; and an efficient approximated method is demanded. Here we employ the *power method* [5] and take the result of its first iteration as the updated initial state $(\boldsymbol{u}_0)_+$. More precisely, the initial state $\boldsymbol{u}_0$ is modified by

$$(\boldsymbol{u}_0)_+ = \|A\boldsymbol{u}_0\|^{-1} A\boldsymbol{u}_0 = \|\boldsymbol{v}_0\|^{-1} \boldsymbol{v}_0 \,,$$

where the second equality follows from (4.7) and the definition $\boldsymbol{v}_0 := \frac{\partial L}{\partial \boldsymbol{u}_0}$; and the vector $\boldsymbol{v}_0$ has been calculated by the forward-backward algorithm. So we need not to explicitly compute the matrix A and its eigenvectors.

Theorem 18 *Let $A \in \mathbb{R}^{m \times m}$ be a symmetric, positive-definite matrix. Let $\boldsymbol{x} \in \mathbb{R}^m$ have norm 1 and $\boldsymbol{y} = \|A\boldsymbol{x}\|^{-1} A\boldsymbol{x}$. Then $\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} \leqslant \boldsymbol{y}^\mathsf{T} A \boldsymbol{y}$, with the equality holds if and only if $\boldsymbol{y} = \pm \boldsymbol{x}$.* — See Appendix A.11 for the proof.

This theorem shows that, the updating law $(\boldsymbol{u}_0)_+ = \|\boldsymbol{v}_0\|^{-1} \boldsymbol{v}_0$ always increases the log-likelihood $L(\varphi, \boldsymbol{u}_0)$ and the increase is nonzero if $(\boldsymbol{u}_0)_+ \neq \pm \boldsymbol{u}_0$.

Summing the above discussion up, we propose a greedy learning algorithm which iteratively modifies the operators φ_a and the initial state $\boldsymbol{u}_0$, as below.

1. Randomly construct an m-dimensional norm-OOM $(\mathbb{R}^m, \{\varphi_a\}_{a \in O}, \boldsymbol{u}_0)$; and set the cyclic index $i = 1$.

[5] The power method is an iterative method for evaluating the dominant eigenpair of a matrix A with distinct eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ satisfying $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$.

2. Put $\boldsymbol{\varphi}^i_+ \leftarrow (-\boldsymbol{\varphi}^i)$ if this increases the log-likelihood L of the model.
3. Compute the vectors $\{\boldsymbol{u}_k\}_{k=0}^N$, $\{\boldsymbol{v}_k\}_{k=0}^N$ by the forward-backward algorithm; then construct the derivative matrix $\psi = \frac{\partial L}{\partial \varphi}$ from these vectors.
4. Find the index j that makes the absolute value of $\boldsymbol{\psi}^i(\boldsymbol{\varphi}^j)^\mathsf{T} - \boldsymbol{\psi}^j(\boldsymbol{\varphi}^i)^\mathsf{T}$ largest; and modify $\boldsymbol{\varphi}^i$, $\boldsymbol{\varphi}^j$ by (4.3) with $\theta = \eta \cdot \{\boldsymbol{\psi}^i(\boldsymbol{\varphi}^j)^\mathsf{T} - \boldsymbol{\psi}^j(\boldsymbol{\varphi}^i)^\mathsf{T}\}$, where $\eta > 0$ is small enough so that L is increased after the modification.
5. Re-compute the vector $\boldsymbol{v}_0$ and set $(\boldsymbol{u}_0)_+ = \|\boldsymbol{v}_0\|^{-1}\boldsymbol{v}_0$.
6. Move the index i to the next row of φ and repeat the above steps 2–5, until some termination condition is satisfied.

It is clear that each step of the above algorithm makes the target function $L(\varphi, \boldsymbol{u}_0)$ increase. It is also clear that if the algorithm reaches some local maximal point of $L(\varphi, \boldsymbol{u}_0)$, then $\frac{\partial L}{\partial \theta}|_{\theta=0} = \boldsymbol{\psi}^i(\boldsymbol{\varphi}^j)^\mathsf{T} - \boldsymbol{\psi}^j(\boldsymbol{\varphi}^i)^\mathsf{T} \neq 0$ for all indices i, j and

$$\frac{\partial}{\partial \boldsymbol{u}_0}[L - \lambda(\boldsymbol{u}_0^\mathsf{T}\boldsymbol{u}_0 - 1)] = \boldsymbol{v}_0 - 2\lambda\boldsymbol{u}_0 = 0\,,$$

where λ is the Lagrange multiplier whose value is determined by the condition $\|\boldsymbol{u}_0\| = 1$. It follows that $\boldsymbol{u}_0 = \pm\|\boldsymbol{v}_0\|^{-1}\boldsymbol{v}_0 = \pm(\boldsymbol{u}_0)_+$. Thus, if the algorithm does not reach a local maximal point of $L(\varphi, \boldsymbol{u}_0)$, then the above steps 4 or 5 will make L larger; and we conclude that *the algorithm converges to some local maximal point of* $L(\varphi, \boldsymbol{u}_0)$. Furthermore, note that the steps 2, 5 are not restricted to "local area" in the parameter space. This makes the algorithm differ from the standard gradient method; and (hopefully) enables it to jump out of some local optimum.

We end this section with an efficient variation of the above learning algorithm of norm-OOMs. In the above algorithm, the derivative matrix ψ computed in Step 3 is not fully used. In fact, only two rows of ψ is employed in Step 4 to modify the observable operators; and other rows are just ignored. So to make full use of the derivative matrix ψ, one may modify Step 4 as follows.

4'. Find $2n$ indices and divide them into n pairs such that for each pair (i, j), $|\boldsymbol{\psi}^i(\boldsymbol{\varphi}^j)^\mathsf{T} - \boldsymbol{\psi}^j(\boldsymbol{\varphi}^i)^\mathsf{T}|$ is large enough; and modify the corresponding $\boldsymbol{\varphi}^i$, $\boldsymbol{\varphi}^j$ by (4.3) with $\theta = \eta \cdot \{\boldsymbol{\psi}^i(\boldsymbol{\varphi}^j)^\mathsf{T} - \boldsymbol{\psi}^j(\boldsymbol{\varphi}^i)^\mathsf{T}\}$, where $\eta > 0$ is small enough so that L is increased after the modification.

One can easily see that the modified algorithm also converges to a local maximal point of $L(\varphi, \boldsymbol{u}_0)$.

5 Some Numerical Examples

In this section the performance of the presented learning algorithm of norm-OOMs are checked on two artificial datasets generated respectively by the norm-OOM specified by (3.17) and by a HMM with 5 hidden states and 3 outputs; and a real

world dataset, the novel "The One-Million Pound Bank-Note" written by Mark Twain. For comparison, we also trained HMMs from the same datasets using the EM algorithm [2]. There are two kinds of HMMs: the (usually used) *state-emission* HMMs (SE-HMMs), in which the outcomes are "emitted" by the hidden states; and the *transition-emission* HMMs (TE-HMMs), in which the symbol emitted at time t depends on the hidden states at times t and $t+1$.

The quality of learnt models is measured by the quantity *description accuracy* (DA), which is defined by

$$\mathrm{DA}(\mathfrak{A}, S) := f[1 + \mathrm{NLL}(\mathfrak{A}, S)] := f\left[1 + (S^{\#})^{-1} \log_{\ell} \Pr(S|\mathfrak{A})\right] ,$$

where $\mathfrak{A}$ is the model whose quality we want to measure; S is the dataset and $S^{\#}$ denotes the total number of symbols in S, i.e., the sum of length of all sequences in S; ℓ is the alphabet size; and f is a nonlinear function which maps the infinite interval $(-\infty, 1]$ to the finite one $(-1, 1]$ via

$$f(x) = \begin{cases} x & \text{if} \quad x \geqslant 0\,, \\ (1 - e^{-0.25x})/(1 + e^{-0.25x}) & \text{if} \quad x < 0\,. \end{cases}$$

Intuitively, $\mathrm{NLL}(\mathfrak{A}, S)$ is the *normalized log-likelihood* of $\mathfrak{A}$ on the dataset S per symbol and assumes values from $-\infty$ to 0. Therefore the range of $\mathrm{DA}(\mathfrak{A}, S)$ is the interval $(-1, 1]$: $\mathrm{DA} = 1$ means the model describes the data S perfectly well (it can predict the data with probability one); $\mathrm{DA} = 0$ means the model is irrelevant to the data for it provides no more information about the process than just randomly "guessing" such a data set; $\mathrm{DA} < 0$ means the model is even worse than a randomly created one, which, as one can imagine, rarely happens in practice. In short, the larger the DA-values are, the better the learnt model is. In the following numerical experiments the quality of all learnt models will be measured by their DA-values on the training sequence and testing sequences from the same source as that of the training data.

Modelling artificial systems. In the first experiment, 30 sequences of length 1000 are procured by the norm-OOM (3.17). The first 20 are used as training sequences from each of which several norm-OOMs (HMMs) with different dimensions $m \in \{2, 3, \cdots, 8\}$ are learnt; the other 10 sequences are used as the testing data on which the DA-values of each estimated model are computed and compared. We plotted in Figure 4-(**a**,**b**) the distribution of training and testing DA-values of the learnt models respectively; and in Figure 5-(**a**,**b**) the corresponding standard deviation. From these figures we see that norm-OOMs have higher description accuracy than that of HMMs. This is not surprising since the underlying system cannot be captured by HMMs, as shown before. Also observe that both HMMs and norm-OOMs suffer from the overfitting problem: when model dimension m becoming larger, the training DA-values increase but the test DA-values decrease.

In the second experiment, a HMM with transition matrix and emission matrix

$$A = \begin{bmatrix} .93 & .00 & .02 & .00 & .05 \\ .14 & .05 & .00 & .00 & .81 \\ .62 & .00 & .00 & .02 & .36 \\ .21 & .22 & .00 & .53 & .04 \\ .00 & .00 & .74 & .26 & .00 \end{bmatrix}, \quad B = \begin{bmatrix} .04 & .42 & .54 \\ .29 & .00 & .71 \\ .13 & .00 & .87 \\ .87 & .13 & .00 \\ .00 & .98 & .02 \end{bmatrix}$$

is used as the underlying system. Running this HMM we produced 20 training sequences of length 2000 and 10 testing sequences of length 1000. As before, several norm-OOMs and HMMs of dimensions m from 2 to 8 are trained and tested on the dataset. The results are presented in Figure 6 and 7. For this dataset, norm-OOMs show no significant advantages over HMMs, but the performance of norm-OOMs are still comparable to that of HMMs.

Modelling "The One-Million Pound Bank-Note". In this experiment norm-OOMs are taken to model a real-world stochastic source, a text source. We split the novel "The One-Million Pound Bank-Note" into two parts, the first half (of length 21042) was used as training sequence, the second half (of length 20569) as test sample. The text was simplified by putting all letters to lower case and reducing the set of punctuations to blanks, which left an alphabet of size 27. From this dataset, several norm-OOMs (HMMs) of dimension $m \in \{3, 6, \cdots, 21\}$ were trained and tested. The training and test DA-values of these learnt models are shown in Figure 8. From the figure we see that norm-OOMs and TE-HMMs have higher DA-values than SE-HMMs on both the training sequence and the test sample. This is because SE-HMMs have fewer free parameters than TE-HMMs and norm-OOMs of the same dimension. For high dimensional models, norm-OOMs are a little better than TE-HMMs. In particular, for the case of $m = 21$, TE-HMM has already overfitted the trainind data; whereas norm-OOM still works well. This example, as well as the above two experiments, suggests that norm-OOMs may be more expressive than HMMs.

6 Conclusion and Future Work

In this report we established the basic theory of norm-OOMs and proposed an iterative learning algorithm for estimating norm-OOMs from data. It is shown that any stochastic process can be represented by a point in an appropriate inner product space, the space $\mathcal{D}$. The inner product in $\mathcal{D}$ induces naturally a metric between stochastic processes. It is also shown that from any process an "abstract" norm-OOM can be constructed in the space $\mathcal{D}$. All these facts enable us to study stochastic processes in the space $\mathcal{D}$, by means from linear algebra and analysis.

Based upon the maximum-likelihood principle, an iterative algorithm is proposed for learning norm-OOMs from data. A significant feature of this algorithm

is that it only produces valid norm-OOMs (after each iteration); and so one can terminate the algorithm at any time. Two numerical experiments are implemented to illustrate the performance of the presented algorithm. It turns out that norm-OOMs are better than (at least comparable to) HMMs concerning the training and testing likelihood.

In Section 3 we studied the linear structures of the space $\mathcal{D}$ and its subsets $\mathcal{D}_{\rm S}$ and $\mathcal{D}^+$. It is also desirable to investigate their topological properties, which shed more light on the relation between the theory of stochastic processes and functional analysis. For instance, we have proven that *the metric space $\mathcal{D}^+$ with its distance induced by the inner product* (3.11) *is complete, i.e., each Cauchy sequence in $\mathcal{D}^+$ is convergent.* A more interesting and important theoretical problem is the distribution of processes that can be modelled by m-dimensional norm-OOMs in the space $\mathcal{D}$, for it actually asks to which accuracy we can approximate an arbitrary process by an m-dimensional norm-OOMs.

A Proofs

A.1 Proof of Theorem 2

(1) We induct on m. The case of $m = 1$ is trivial. Assume the assertion is true for $m = r$ and consider the case of $m = r + 1$. Let $g_1, g_2, \cdots, g_m \in \mathcal{F}$ be linearly independent, then $g_1, g_2, \cdots, g_r$ are also linearly independent. By the inductive hypothesis, there are $\bar{b}_1, \bar{b}_2, \cdots, \bar{b}_r \in O^*$ such that the matrix $A_r = [g_j(\bar{b}_i)]_{i,j=1,2,\cdots,r}$ is invertible. We claim that, for these sequences $\bar{b}_1, \bar{b}_2, \cdots, \bar{b}_r$, there is a $\bar{b}_m \in O^*$ which makes $A_m(\bar{b}_m) = [g_j(\bar{b}_i)]_{i,j=1,2,\cdots,m}$ a nonsingular matrix. If this is not true, then for any sequence $\bar{b}_m$ in O^*, we have $m > \operatorname{rank} A_m(\bar{b}_m) \geqslant \operatorname{rank} A_r = r$ and so $\operatorname{rank} A_m(\bar{b}_m) = r$. It follows that $A_m(\bar{b}_m)\boldsymbol{x}(\bar{b}_m) = 0$ for some nonzero vector

$$\boldsymbol{x}(\bar{b}_m) = [x_1(\bar{b}_m), x_2(\bar{b}_m), \cdots, x_m(\bar{b}_m)]^{\mathsf{T}} =: [\boldsymbol{y}^{\mathsf{T}}(\bar{b}_m), x_m(\bar{b}_m)]^{\mathsf{T}}. \tag{A.1}$$

As $\operatorname{rank} A_r = r$, we know $x_m(\bar{b}_m) \neq 0$. Without loss of generality, assume that $x_m(\bar{b}_m) = -1$ (otherwise we put $\boldsymbol{x} = -\boldsymbol{x}/x_m$). It follows from $A_m(\bar{b}_m)\boldsymbol{x}(\bar{b}_m) = 0$ and (A.1) that $A_r \cdot \boldsymbol{y}(\bar{b}_m) = [g_m(\bar{b}_1), g_m(\bar{b}_2), \cdots, g_m(\bar{b}_r)]^{\mathsf{T}}$ and so

$$\boldsymbol{x}(\bar{b}_m) = \big[[g_m(\bar{b}_1), g_m(\bar{b}_2), \cdots, g_m(\bar{b}_r)](A_r^{-1})^{\mathsf{T}}, -1\big]^{\mathsf{T}}$$

is a constant vector (not depending on $\bar{b}_m$). The last row of $A_m(\bar{b}_m)\boldsymbol{x} = 0$ shows $g_m(\bar{b}_m) = \sum_{j=1}^r x_j g_j(\bar{b}_m)$ for any $\bar{b}_m \in O^*$, contradicting the linear independence of $g_1, g_2, \cdots, g_m$. Therefore the assertion follows.

(2a) It is clear that $\mathcal{G} \subseteq \mathcal{H}^*$ and well known that $\dim \mathcal{H}^* = \dim \mathcal{H} = m$, where $\mathcal{H}^*$ is the dual space of $\mathcal{H}$ consisting of all linear functionals on $\mathcal{H}$. Thus $\dim \mathcal{G} \leqslant m$. Assume that $\dim \mathcal{G} = r < m$. Let $\{\sigma l_{\bar{b}_i}\}_{i=1,2,\cdots,r}$ be a basis of $\mathcal{G}$ and $\{l_{\bar{a}_j} h\}_{j=1,2,\cdots,m}$ a basis of $\mathcal{H}$. Then for any $j \leqslant m$ and $\bar{b} \in O^*$,

$$\begin{aligned}(l_{\bar{a}_j} h)(\bar{b}) &= (l_{\bar{b}} l_{\bar{a}_j})(\varepsilon) = \sigma l_{\bar{b}} l_{\bar{a}_j} h = \Big(\textstyle\sum_{i=1}^r \alpha_i(\bar{b})\sigma l_{\bar{b}_i}\Big)(l_{\bar{a}_j} h) \\ &= \textstyle\sum_{i=1}^r \alpha_i(\bar{b})\sigma l_{\bar{b}_i} l_{\bar{a}_j} h = \sum_{i=1}^r \alpha_i(\bar{b})(l_{\bar{a}_j} h)(\bar{b}_i).\end{aligned}$$

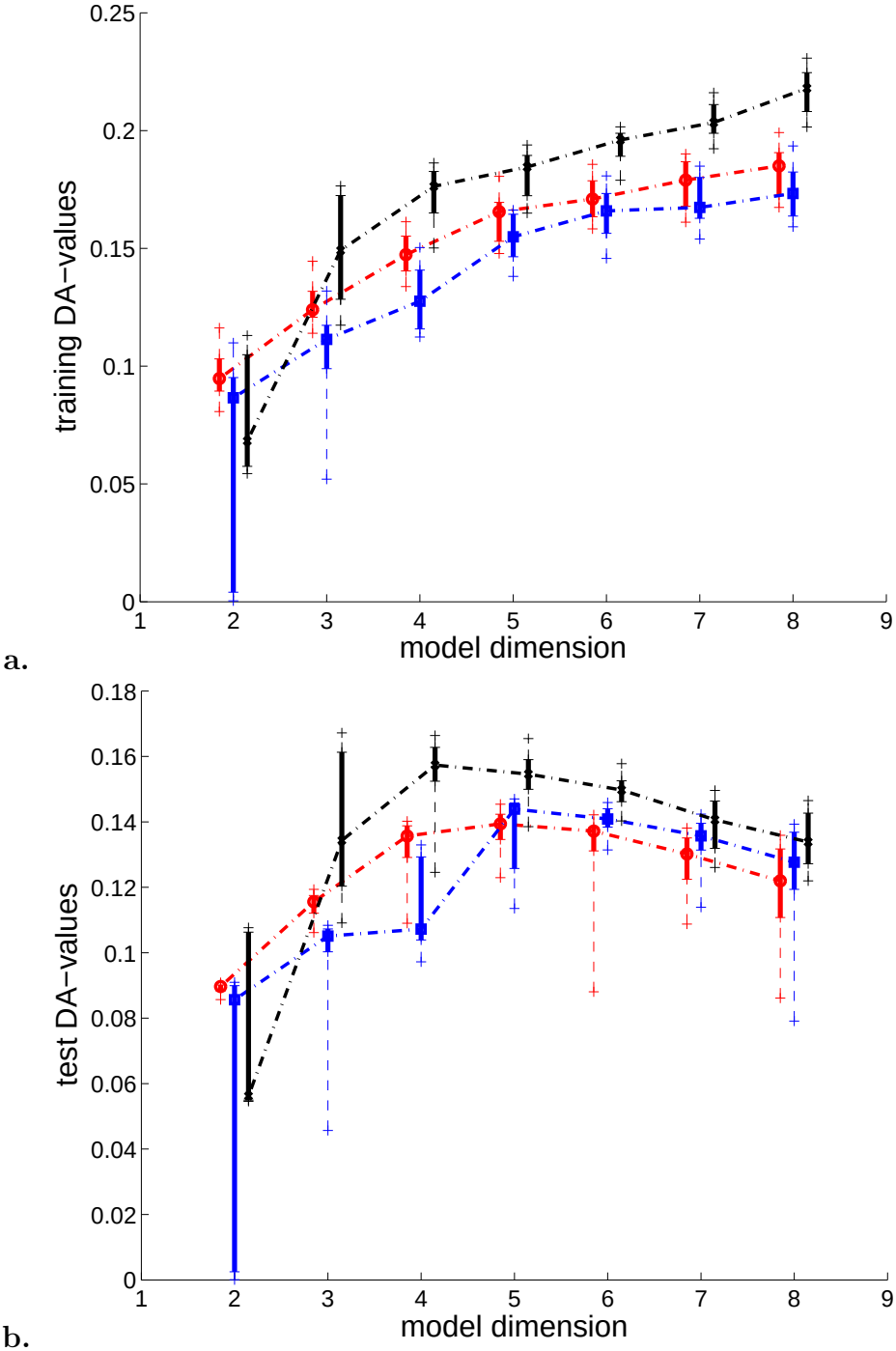
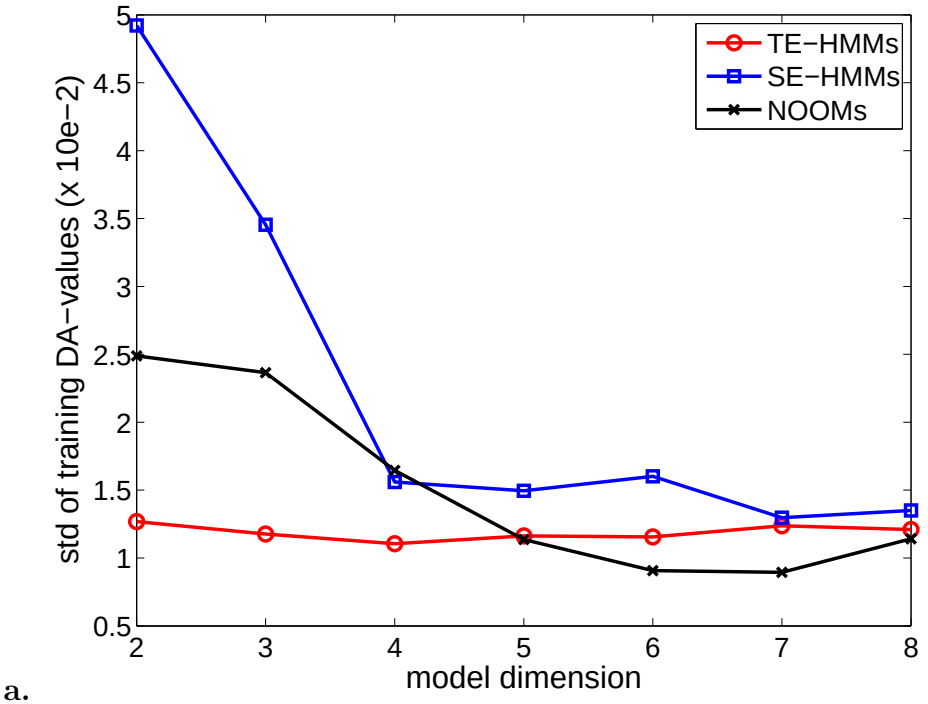
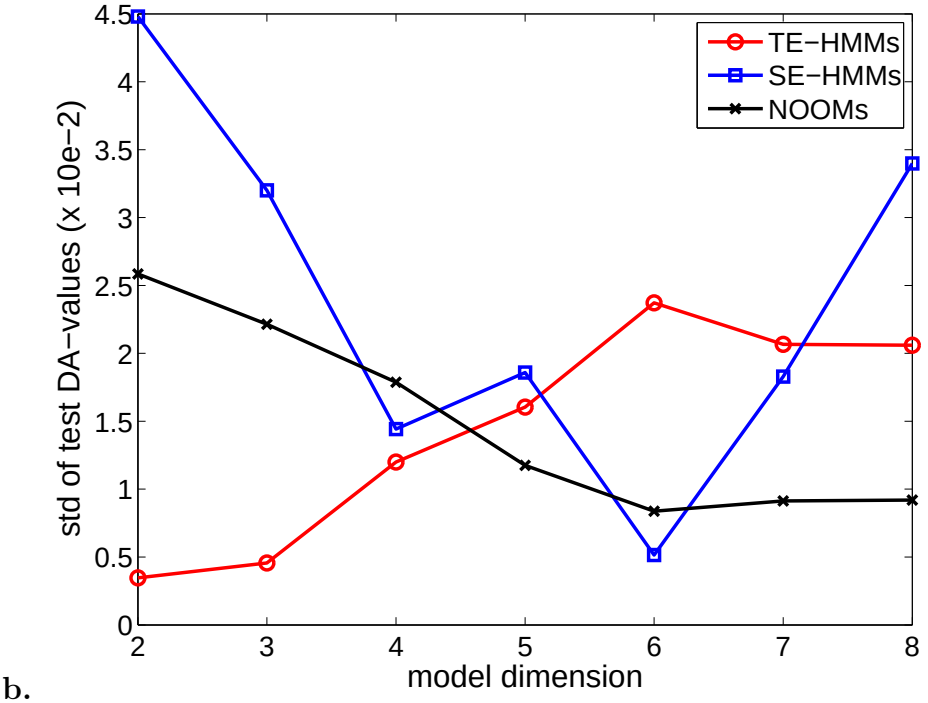


Figure 4: Training (**a**) and testing (**b**) DA-values of SE-HMMs (blue, mark: □), TE-HMMs (red, mark: ○) and norm-OOMs (black, mark: ×), in which the 10%-percentile, lower quantile, median, upper quantile and 90%-percentile positions are marked.

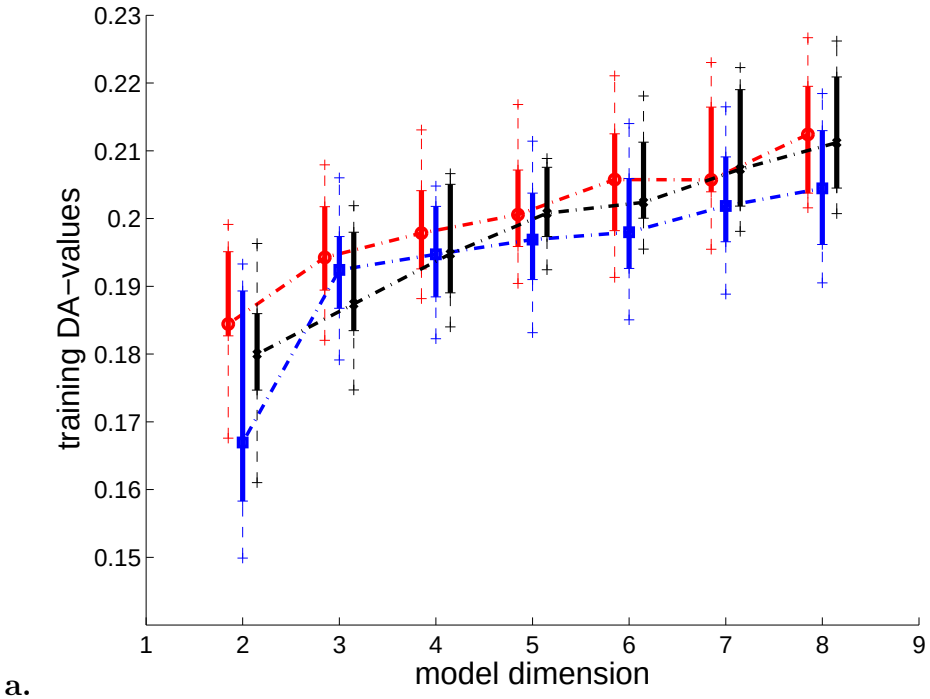


a.

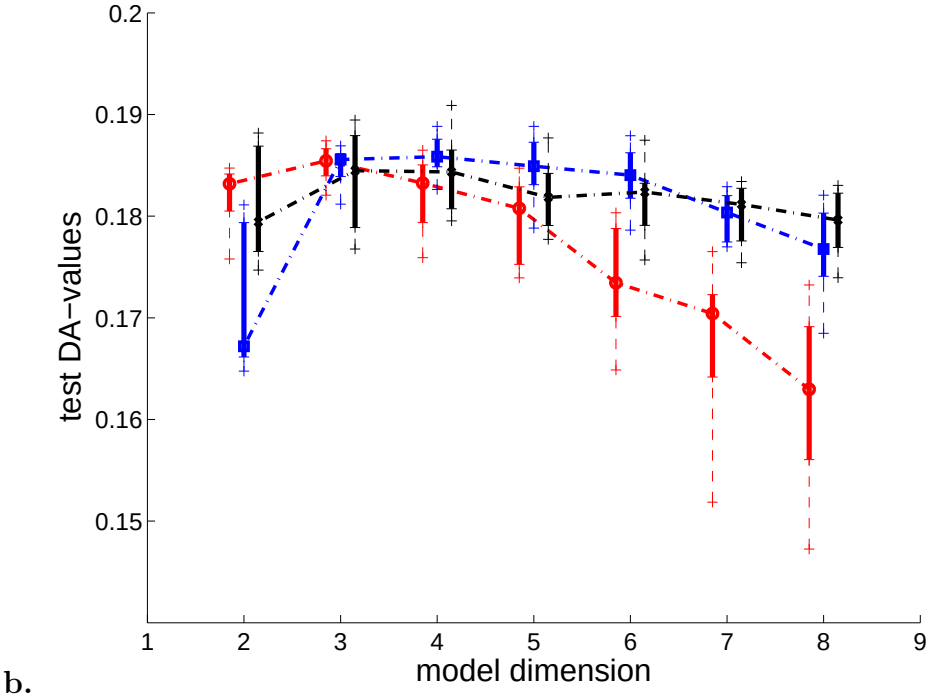


b.

Figure 5: Standard deviation of DA-values for SE-HMMs, TE-HMMs and norm-OOMs.

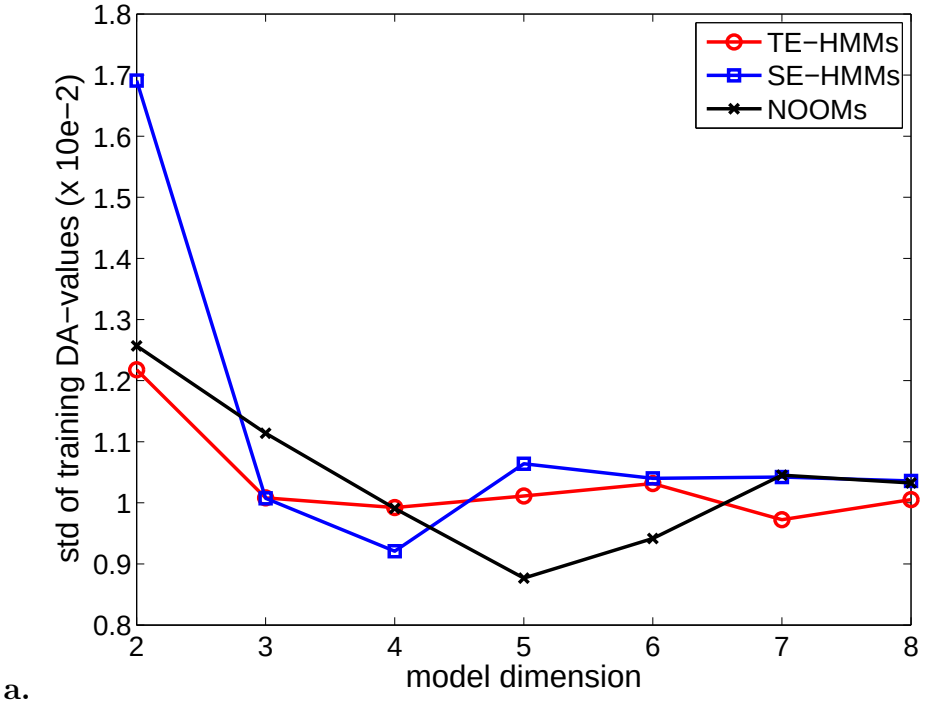


a.

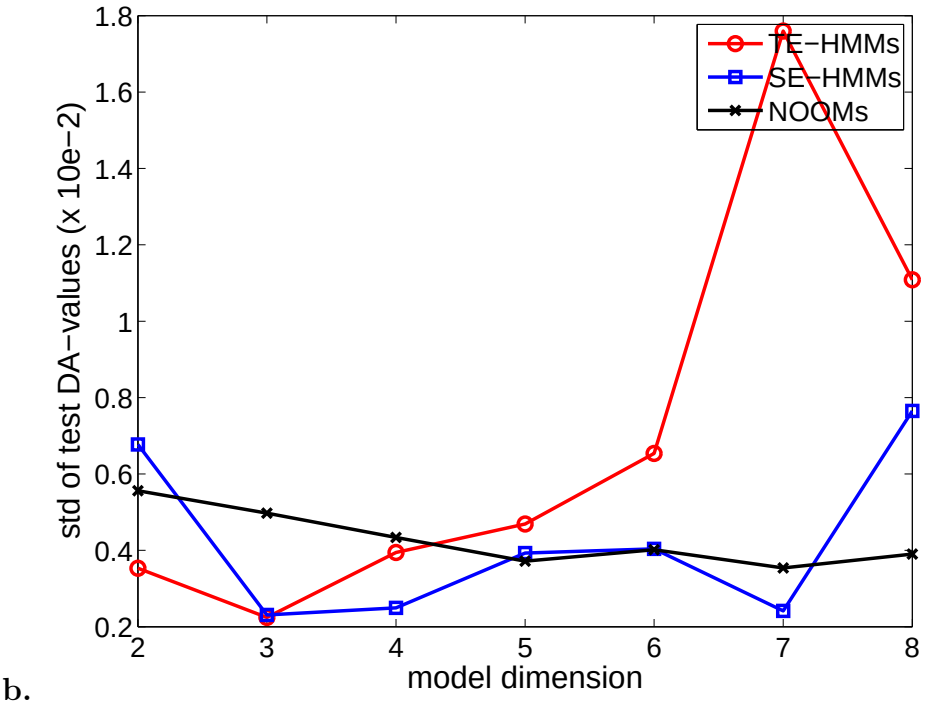


b.

Figure 6: Training (**a**) and testing (**b**) DA-values of SE-HMMs (blue, mark: □), TE-HMMs (red, mark: ○) and norm-OOMs (black, mark: ×) on the dataset generated by a 5-state-3-output HMM.

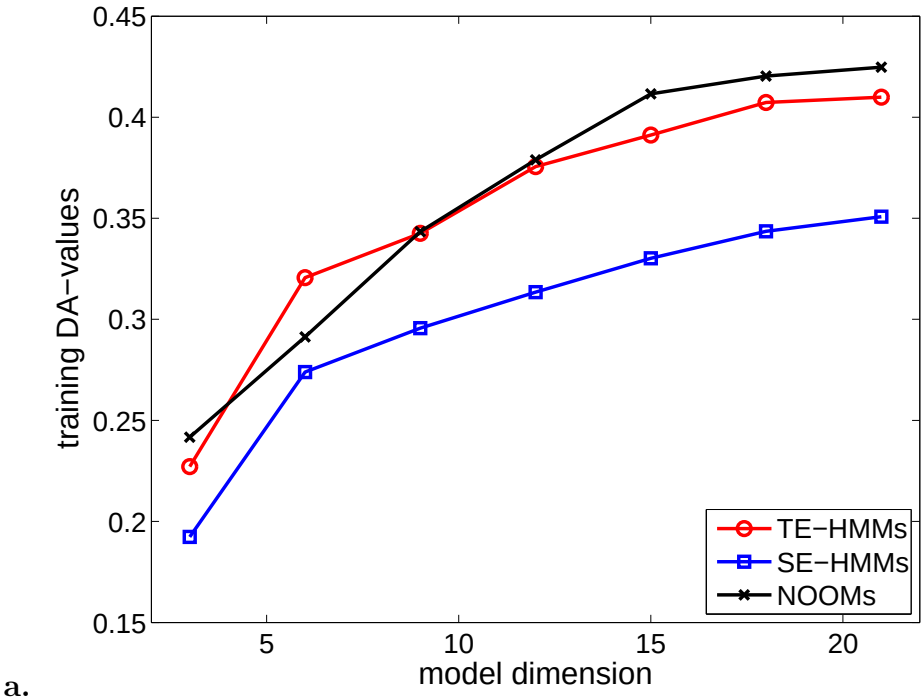


a.

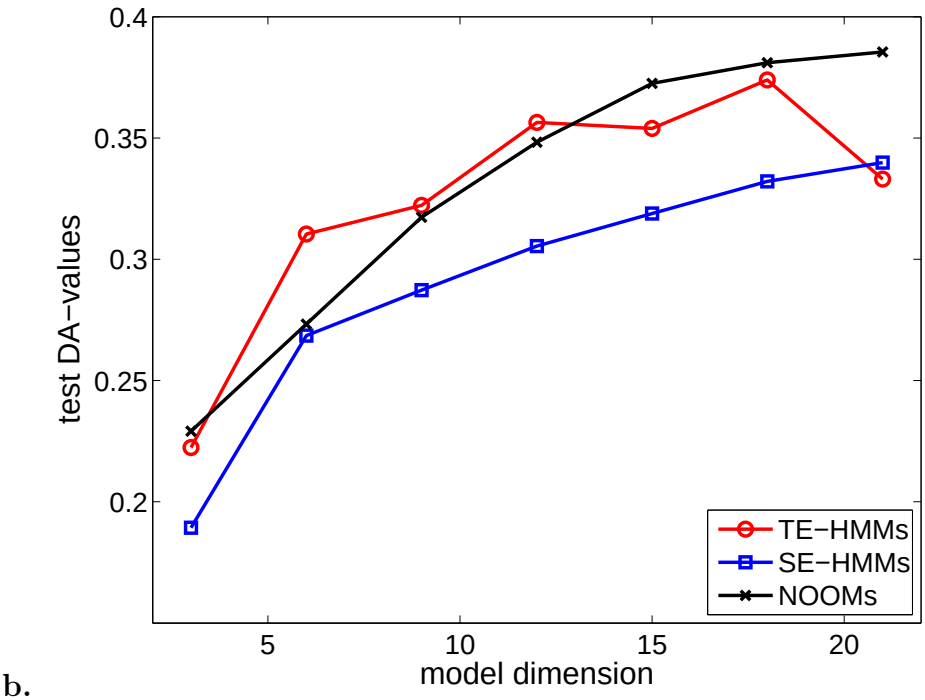


b.

Figure 7: Standard deviation of DA-values for SE-HMMs, TE-HMMs and norm-OOMs on the dataset generated by a 5-state-3-output HMM.



a.



b.

Figure 8: Training (**a**) and test (**b**) DA-values of SE-HMMs TE-HMMs and norm-OOMs on the text "The One-Million Pound Bank-Note".

This equality illustrates the vector $[(l_{\bar{a}_1}h)(\bar{b}), \cdots, (l_{\bar{a}_m}h)(\bar{b})]$ can be written as the linear combination $\sum_{i=1}^{r} \alpha_i(\bar{b})[(l_{\bar{a}_1}h)(\bar{b}_i), \cdots, (l_{\bar{a}_m}h)(\bar{b}_i)]$, which contradicts the fact that $\{l_{\bar{a}_j}h\}_{j=1,2,\cdots,m}$ is a basis of $\mathcal{H}$ and the statement (1). So $\dim \mathcal{G} = m$.

(2b) Since h is a LDF of dimension m, by (2.3) we know $h(\bar{a}) = \mathbf{1}\tau_{\bar{a}}\boldsymbol{w}_0$ for some structure $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$. It follows from the definition of $h(A, B)$ that

$$h(A, B) = \left[h(\bar{a}_j\bar{b}_i)\right]_{i\leqslant N, j\leqslant M} = \left[\mathbf{1}\tau_{\bar{b}_i} \cdot \tau_{\bar{a}_j}\boldsymbol{w}_0\right]_{i\leqslant N, j\leqslant M} =: \pi(B)\omega(A)\,,$$

where $\pi(B)$ is the $N\times m$ matrix with i-th row $\mathbf{1}\tau_{\bar{b}_i}$ and $\omega(A)$ the $m\times M$ matrix with $\tau_{\bar{a}_j}\boldsymbol{w}_0$ as its j-th column. So $\operatorname{rank} h(A, B) \leqslant \min\{\operatorname{rank}\pi(B), \operatorname{rank}\omega(A)\} \leqslant m$.

(2c) Let $\mathcal{R}_n := \operatorname{span}\{\tau_{\bar{a}}\boldsymbol{w}_0 : \bar{a} \in O^{\leqslant n}\}$. Then (i) $\dim \mathcal{R}_0 = 1$; (ii) $\mathcal{R}_n \subseteq \mathcal{R}_{n+1}$ and so $\dim \mathcal{R}_n \leqslant \dim \mathcal{R}_{n+1}$; (iii) $\dim \mathcal{R}_n = \dim \mathcal{R}_{n+1}$ implies that $\mathcal{R}_n$ is invariant under each τ_a and therefore $\mathcal{R}_n = \mathcal{R}_{n+1} = \cdots = \mathcal{R} := \operatorname{span}\{\tau_{\bar{a}}\boldsymbol{w}_0 : \bar{a} \in O^*\}$; and (iv) $\dim \mathcal{R} = m$, since $\mathcal{R}$ is a representation of the abstract vector space $\mathcal{H} = \operatorname{span}\{l_{\bar{a}}h : \bar{a} \in O^*\}$, which, by the definition of m-dimensional LDFs, has dimension m. These facts reveal that $\dim \mathcal{R}_n = m$ for some $n < m$ and hence the matrix $\omega(O^{\leqslant n})$ is of full (row) rank. Similary, $\pi(O^{\leqslant r})$ is of full (column) rank for some $r < m$. Thus, $h(O^{\leqslant n}, O^{\leqslant r}) = \pi(O^{\leqslant r})\omega(O^{\leqslant n})$ has rank m.

A.2 Proof of Theorem 6

It is obvious that $f \in \mathcal{B}$ implies $\alpha f \in \mathcal{B}$ for any real number α. Now let $f, g \in \mathcal{B}$, then $S_n(f + g) = S_n(f) + S_n(g) + 2\sum_{\bar{a}\in O^n} f(\bar{a})g(\bar{a})$ and so, since $2ab \leqslant a^2 + b^2$ for any $a, b \in \mathbb{R}$, $S_n(f + g) \leqslant 2[S_n(f) + S_n(g)]$. This proves $\mathcal{B}$ is a vector space. Furthermore, the space $\mathcal{B}$ is invariant under l_a because $S_n(l_a f) \leqslant S_{n+1}(f)$.

A.3 Proof of Theorem 7

The assertions (i, ii) are clear. For (iii) we need only to show that $f(\bar{x})g(\bar{x}) \geqslant \sum_{a\in O} f(\bar{x}a)g(\bar{x}a)$. But by Cauchy's inequality we have

$$\left[\sum_{a\in O} f(\bar{x}a)g(\bar{x}a)\right]^2 \leqslant \left[\sum_{a\in O} f^2(\bar{x}a)\right]\left[\sum_{a\in O} g^2(\bar{x}a)\right] \leqslant f^2(\bar{x})g^2(\bar{x})\,; \qquad \text{(A.2)}$$

and the desired inequality follows. To see that $\mathcal{D}_0^+$ is invariant under l_a, it suffices to show $(l_a f)^2(\bar{x}) \geqslant \sum_{b\in O}(l_a f)^2(\bar{x}b)$, i.e., $f^2(a\bar{x}) \geqslant \sum_{b\in O} f^2(a\bar{x}b)$ for any $f \in \mathcal{D}_0^+$ and any $\bar{x} \in O^*$, which is clear by the definition of $\mathcal{D}_0^+$.

A.4 Proof of Lemma 2

By Definition 3, we see that for any $\alpha \in \mathbb{R}$ and $x, y \in V$,

$$0 \leqslant Q(x - \alpha y, x - \alpha y) = q^2(x) - 2\alpha Q(x, y) + \alpha^2 q^2(y)\,. \qquad \text{(A.3)}$$

If $q(y) = 0$, it follows from $q^2(x) - 2\alpha Q(x, y) \geqslant 0$ $(\forall \alpha \in \mathbb{R})$ that $Q(x, y) = 0$. Interchanging the roles of x and y, we see that $q(x) = 0$ implies $Q(x, y) = 0$. So

$|Q(x,y)| \leqslant q(x)q(y)$ if one of $q(x)$ and $q(y)$ is zero. If $q(x), q(y)$ both are nonzeros, letting $\alpha = \pm q(x)/q(y)$ in (A.3) we get $\pm Q(x,y)q(x)/q(y) \leqslant q^2(x)$, which implies $|Q(x,y)| \leqslant q(x)q(y)$.

A.5 Proof of Theorem 9

Let $f \in \mathcal{D}_0^+$ be fixed with $\|[f]\| = q(f) = 1$. For each $n = 0, 1, 2, \cdots$ we define a function $g_n \in \mathcal{F}$ by

$$g_n(\bar{a}) := \sqrt{\textstyle\sum_{\bar{x}\in O^n} f^2(\bar{a}\bar{x})}\,. \quad (\forall \bar{a} \in O^*)$$

It follows that $g_0 = f$ and $g_n \geqslant 0$ for all n. Moreover, by the definition of $\mathcal{D}_0^+$ (see (3.4)), $\sum_{\bar{x}\in O^n} f^2(\bar{a}\bar{x}) \geqslant \sum_{\bar{x}\in O^n, b\in O} f^2(\bar{a}\bar{x}b) = \sum_{\bar{x}\in O^{n+1}} f^2(\bar{a}\bar{x})$, which implies $g_n(\bar{a}) \geqslant g_{n+1}(\bar{a})$. Thus, for each $\bar{a} \in O^*$, $\{g_n(\bar{a})\}_{n=0,1,2,\cdots}$ is a decreasing number sequence lower bounded by 0; and the function

$$g(\bar{a}) := \lim_{n\to\infty} g_n(\bar{a}) = \lim_{n\to\infty} \sqrt{\textstyle\sum_{\bar{x}\in O^n} f^2(\bar{a}\bar{x})} \quad (\forall \bar{a} \in O^*) \tag{A.4}$$

is well defined and satisfies $0 \leqslant g(\bar{a}) \leqslant f(\bar{a})$ since $g_0 = f$. Now we compute

$$\sum_{b\in O} g^2(\bar{a}b) = \lim_{n\to\infty} \sum_{b\in O} g_n^2(\bar{a}b) = \lim_{n\to\infty} \sum_{b\in O, \bar{x}\in O^n} f^2(\bar{a}b\bar{x}) = \lim_{n\to\infty} g_{n+1}^2(\bar{a}) = g^2(\bar{a})\,,$$

and $g^2(\varepsilon) = \lim_{n\to\infty} g_n^2(\varepsilon) = \lim_{n\to\infty} \sum_{\bar{x}\in O^n} f^2(\bar{x}) = \lim_{n\to\infty} S_n(f) = q^2(f) = 1$ (see (3.2) and (3.8)). Therefore, the function g defined by (A.4) is a member of $\mathcal{S}$ and, as illustrated earlier, has the property $\|[g]\| = q(g) = 1$.

As $f \geqslant g \geqslant 0$, by (3.3) we have $Q_n(f,g) \geqslant Q_n(g,g)$ for all n, which, together with (3.6), implies $Q(f,g) \geqslant Q(g,g) = q^2(g) = 1$. Thus,

$$q^2(f-g) = q^2(f) - 2Q(f,g) + q^2(g) = 2[1 - Q(f,g)] \leqslant 0$$

and so $q(f-g) = 0$, i.e., $[f] = [g]$.

A.6 Proof of Theorem 10

Since $f, g \in \mathcal{D}_0^+$, we know $f, g \geqslant 0$ and so $q(f+g) \geqslant q(f) \geqslant 0$. But $q(f+g) = 0$, thus $q(f) = 0$. Similarly, $q(g) = 0$. This proves (i).

Let $[f], [g] \in \mathcal{D}^+$. By the definition of $\mathcal{D}^+$, there are $f', g' \in \mathcal{D}_0^+$ such that $[f] = [f']$ and $[g] = [g']$. As $\mathcal{D}_0^+$ is a convex cone, we have $f' + g' \in \mathcal{D}_0^+$ and $\alpha f' \in \mathcal{D}_0^+$ for any $\alpha \geqslant 0$. It follows that $[f] + [g] = [f'] + [g'] = [f' + g']$ and $\alpha[f] = \alpha[f'] = [\alpha f']$ both belong to $\mathcal{D}^+$. So $\mathcal{D}^+$ is a convex cone.

Now let $[h] \in \mathcal{D}^+$ be such that $-[h] = [-h]$ is also a member of $\mathcal{D}^+$. Then there exist $f, g \in \mathcal{D}_0^+$ satisfying $[f] = [h]$ and $[g] = [-h]$, i.e., $q(f-h) = q(g+h) = 0$. So $0 \leqslant q(f+g) \leqslant q(f-h) + q(g+h) = 0$. By (i) we know $q(f) = 0$, which means $[h] = [f] = [0]$. Therefore, $\mathcal{D}^+$ is pointed at $[0]$.

A.7 Proof of Theorem 11

By (3.9) we see that $Q(f,g)=1$. As $f,g\in\mathcal{D}_0^+$, $\{Q_n(f,g)\}_{n=0,1,2,\cdots}$ (see (3.3) for the definition of Q_n) forms a decreasing sequence with $Q_0(f,g)=f(\varepsilon)g(\varepsilon)=1$ and $\lim_{n\to\infty}Q_n(f,g)=Q(f,g)=1$. So $Q_n(f,g)=1$ for all n. By (3.2) we have $S_n(f-g)=S_n(f)-2Q_n(f,g)+S_n(g)=0$, which means $f(\bar{a})=g(\bar{a})$ for all $\bar{a}\in O^n$ $(n=0,1,2,\cdots)$ and so $f=g$.

A.8 Proof of Theorem 13

Derect computation shows that

$$\begin{aligned}\sum_{a\in O}Q(l_af,l_ag) &= \sum_{a\in O}\lim_{n\to\infty}\sum_{\bar{x}\in O^n}f(a\bar{x})g(a\bar{a})\\ &= \lim_{n\to\infty}\sum_{\bar{x}\in O^{n+1}}f(\bar{x})g(\bar{x})\\ &= Q(f,g)\,.\end{aligned}$$

A.9 Proof of Theorem 16

Assume $[\tau_a]_{ik}$ is the only (possibly) nonzero element in the i-th row of τ_a. Then for any $\boldsymbol{x}=[x_1,x_2,\cdots,x_m]^{\mathsf{T}}\in\mathbb{R}^m$ with all elements x_j being nonnegative, the i-th element of $\tau_a\boldsymbol{x}$ is $[\tau_a]_{ik}\cdot x_k$; and its square root $\sqrt{[\tau_a]_{ik}}\sqrt{x_k}$ is just the i-th element of $\sqrt{\tau_a}\sqrt{\boldsymbol{x}}=\varphi_a\sqrt{\boldsymbol{x}}$. So $\sqrt{\tau_a\boldsymbol{x}}=\varphi_a\sqrt{\boldsymbol{x}}$ for any $a\in O$ and any nonnegative $\boldsymbol{x}\in\mathbb{R}^m$. By inducting on the length of $\bar{a}\in O^*$, we can prove $\sqrt{\tau_{\bar{a}}\boldsymbol{w}_0}=\varphi_{\bar{a}}\sqrt{\boldsymbol{w}_0}=\varphi_{\bar{a}}\boldsymbol{u}_0$. Now it is clear that $\|\varphi_{\bar{a}}\boldsymbol{u}_0\|^2=\|\sqrt{\tau_{\bar{a}}\boldsymbol{w}_0}\|^2=\mathbf{1}\tau_{\bar{a}}\boldsymbol{w}_0$ and the assertion follows.

A.10 Proof of Theorem 17

The *if* part is clear since both $G(i,j,\theta)$ and $G'(i,j,\theta)$ are unitary and since the product of two unitary matrices is also unitary. To prove the *only if* part, we need two preparing propositions.

1. *Any unitary matrix A of order 2 is either $G(1,2,\alpha)$ or $G'(1,2,\alpha)$.* — Assume that $A=[a,b;c,d]$ in Matlab's notation, then by $AA^{\mathsf{T}}=I_2$ we know

$$a^2+b^2=c^2+d^2=1\,,\quad ac+bd=0\,.$$

Thus $a=\cos\alpha$, $b=\sin\alpha$, $c=\sin\beta$ and $d=\cos\beta$ for some $\alpha,\beta\in(-\pi,\pi]$. It follows that $ac+bd=\sin(\alpha+\beta)=0$ and so $\alpha+\beta=k\pi$ with $k=-1,0,1,2$. If $k=0$ or 2, then $A=G(1,2,\alpha)$; if $k=\pm1$, then $A=G'(1,2,\beta)$.

2. *For any vector $\boldsymbol{x}\in\mathbb{R}^n$ $(n\geqslant 2)$ there exists a matrix A which is the product of $G(i,j,\theta)$'s, such that $A\boldsymbol{x}=\|\boldsymbol{x}\|\boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the first unit vector in $\mathbb{R}^n$.* — Write $\boldsymbol{x}=[x_1,x_2,\cdots,x_n]^{\mathsf{T}}$. Let $y_2=\sqrt{x_1^2+x_2^2}$ and $\alpha_2\in(-\pi,\pi]$ be such that $x_1=y_2\cos\alpha_2$ and $x_2=y_2\sin\alpha_2$. Then $G(1,2,\alpha_2)\boldsymbol{x}=[y_2,0,x_3,\cdots,x_n]^{\mathsf{T}}$. Next let $y_3=\sqrt{y_2^2+x_3^2}$ and $\alpha_3\in(-\pi,\pi]$ be such that $y_2=y_3\cos\alpha_3$ and $x_3=y_3\sin\alpha_3$. Then $G(1,3,\alpha_3)G(1,2,\alpha_2)\boldsymbol{x}=[y_3,0,0,x_4,\cdots,x_n]^{\mathsf{T}}$. Repeating this operation,

we get $G(1,n,\alpha_n)\cdots G(1,3,\alpha_3)G(1,2,\alpha_2)\boldsymbol{x} = [y_n, 0, \cdots, 0]^{\mathsf{T}}$ with $y_n = \|\boldsymbol{x}\|$. This proves the assertion.

Now we prove the theorem by inducting on the order n of the matrix U. For the case of $n = 2$, the theorem follows from the above assertion 1. Assume the theorem is true for $n = k$. For $n = k+1$, we write $U = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_k, \boldsymbol{u}_n]$. By the assertion 2 we know there exist $\alpha_2, \cdots, \alpha_n \in (-\pi, \pi]$ such that

$$G(1,n,\alpha_n)\cdots G(1,3,\alpha_3)G(1,2,\alpha_2)\boldsymbol{u}_1 = \|\boldsymbol{u}_1\|\boldsymbol{e}_1 = \boldsymbol{e}_1\,,$$

for U is a unitary matrix and so $\|\boldsymbol{u}_1\| = 1$. It follows that

$$U' := G(1,n,\alpha_n)\cdots G(1,2,\alpha_2)U = \begin{bmatrix} 1 & \boldsymbol{c}^{\mathsf{T}} \\ \boldsymbol{0} & V \end{bmatrix}, \tag{A.5}$$

where $V \in \mathbb{R}^{k\times k}$ and $\boldsymbol{c} \in \mathbb{R}^k$. The matrices $G(1,j,\alpha_j)$'s and U in (A.5) are all unitary matrices, so is U'. By $U'(U')^{\mathsf{T}} = I_n$ we conclude $VV^{\mathsf{T}} = I_k$ and $\boldsymbol{c} = \boldsymbol{0}$. By the induction hypothesis, the matrix V, and hence U', can be written as the product of $G(i,j,\theta)$'s and $G'(i,j,\theta)$'s. Since $[G(i,j,\alpha)]^{-1} = G(i,j,-\alpha)$, by (A.5) we know U is also the product of $G(i,j,\theta)$'s and $G'(i,j,\theta)$'s.

A.11 Proof of Theorem 18

Since the matrix $A \in \mathbb{R}^{m\times m}$ is symmetric and positive-definite, it has the singular value decomposition $A = UDU^{\mathsf{T}}$, where U is a unitary matrix of order m, i.e., $U^{\mathsf{T}}U = I_m$; and $D = \mathrm{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_m\}$ with $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_m > 0$. For any vector $\boldsymbol{x}$ in $\mathbb{R}^m$ and $\boldsymbol{y} = \|A\boldsymbol{x}\|^{-1}A\boldsymbol{x}$, let $\boldsymbol{x}_0 = U^{\mathsf{T}}\boldsymbol{x}$ and $\boldsymbol{y}_0 = U^{\mathsf{T}}\boldsymbol{y}$. It follows that $\boldsymbol{y}_0 = \|D\boldsymbol{x}_0\|^{-1}D\boldsymbol{x}_0$, $\boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x} = \boldsymbol{x}_0^{\mathsf{T}}D\boldsymbol{x}_0$ and $\boldsymbol{y}^{\mathsf{T}}A\boldsymbol{y} = \boldsymbol{y}_0^{\mathsf{T}}D\boldsymbol{y}_0$. Furthermore, if $\boldsymbol{x}$ has norm 1, so does $\boldsymbol{x}_0$. Thus, we can assume $A = \mathrm{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_m\}$ without loss of generality.

As $\boldsymbol{y} = \|A\boldsymbol{x}\|^{-1}A\boldsymbol{x}$ and A is diagonal, the inequality $\boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x} \leqslant \boldsymbol{y}^{\mathsf{T}}A\boldsymbol{y}$ can be rewritten as $(\boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x})(\boldsymbol{x}^{\mathsf{T}}A^2\boldsymbol{x}) \leqslant \boldsymbol{x}^{\mathsf{T}}A^3\boldsymbol{x}$. Assume that $\boldsymbol{x} = [x_1, x_2, \cdots, x_m]^{\mathsf{T}}$ and let $f(s) := s^{-1}\log(\boldsymbol{x}^{\mathsf{T}}A^s\boldsymbol{x}) = s^{-1}\log(\sum_{i=1}^m x_i^2\sigma_i^s)$, where $s > 0$. If we can prove $f(s)$ is an increasing function, then $f(2) \leqslant f(3)$ and $f(1) \leqslant f(3)$, i.e.,

$$(\boldsymbol{x}^{\mathsf{T}}A^2\boldsymbol{x})^3 \leqslant (\boldsymbol{x}^{\mathsf{T}}A^3\boldsymbol{x})^2\,, \qquad (\boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x})^3 \leqslant \boldsymbol{x}^{\mathsf{T}}A^3\boldsymbol{x}\,.$$

The product of the above two inequalities implies $(\boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x})(\boldsymbol{x}^{\mathsf{T}}A^2\boldsymbol{x}) \leqslant \boldsymbol{x}^{\mathsf{T}}A^3\boldsymbol{x}$. To show that $f(s)$ is increase, we compute its derivative $f'(s) = s^{-2}\sum_{i=1}^m x_i^2 z_i \log(z_i)$, where $z_i = \sigma_i^s/(\sum_{k=1}^m x_k^2\sigma_k^s)$ for $i = 1, 2, \cdots, m$. To prove that $f'(s) \geqslant 0$, we need the following finite form of Jensen's inequality.

For a real convex function h, numbers z_i in its domain, and nonnegative weights w_i such that $\sum_{i=1}^m w_i = 1$, we have $h(\sum_{i=1}^m w_i z_i) \leqslant \sum_{i=1}^m w_i h(z_i)$, with the equality holds iff there exists a constant c such that, for each i, either $w_i = 0$ or $z_i = c$.

Set $w_i = x_i^2$ (note that $\boldsymbol{x}$ has norm 1) and $h(z) = z\log(z)$, then Jensen's inequality implies $f'(s) = s^{-2}\sum_{i=1}^m x_i^2 h(z_i) \geqslant s^{-2}h(\sum_{i=1}^m x_i^2 z_i) = s^{-2}h(1) = 0$. This proves the desired inequality.

Furthermore, $(\boldsymbol{x}^\mathsf{T}A\boldsymbol{x})(\boldsymbol{x}^\mathsf{T}A^2\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}A^3\boldsymbol{x}$ iff $f'(s) = 0$ for $s \in [1,3]$, which happens iff there is a constant c_s such that, for each i, either $x_i = 0$ or $z_i = \sigma_i^s/(\sum_{k=1}^m x_k^2\sigma_k^s) = c_s$, i.e., $\sigma_i = c$ (constant). It follows that $\boldsymbol{y} = \|A\boldsymbol{x}\|^{-1}A\boldsymbol{x} = \pm\boldsymbol{x}$.

References

[1] J. W. Brewer. Kronecker product and matrix calculus in system theory. *IEEE Transactions on Circuits and Systems*, CAS-25(9):772–781, September 1978.

[2] A. P. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[3] A. Heller. On stochastic processes derived from Markov chains. *Annals of Mathematical Statistics*, 36:1286–1291, 1965.

[4] R. A. Hom and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, 1989.

[5] H. Ito. *An algebraic study of discrete stochastic systems*. PhD thesis, The University of Tokyo, 1992.

[6] H. Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

[7] H. Jaeger, M. Zhao, K. Kretzschmar, T. Oberstein, D. Popovici, and A. Kolling. Learning observable operator models via the ES algorithm. In S. Haykin, J. Principe, T. Sejnowski, and J. McWhirter, editors, *New Directions in Statistical Signal Processing: from Systems to Brains*, chapter 20. MIT Press, 2005.

[8] K. Kretzschmar. Learning symbol sequences with observable operator models. GMD Report 161, Fraunhofer Institute AIS, 2003.