Norm Observable Operator Models

Ming-Jie Zhao* mingjie.zhao@gmail.com
Herbert Jaeger h.jaeger@jacobs-university.de

October 25, 2009

Abstract

Hidden Markov models (HMMs) are one of the most popular and successful statistical models for time series. Observable operator models (OOMs) are generalizations of HMMs which exhibit several attractive advantages. In particular, a variety of highly efficient, constructive and asymptotically correct learning algorithms are available for OOMs. However, the OOM theory suffers from the *negative probability problem* (NPP): a given, learnt OOM may sometimes predict negative "probabilities" for certain events. It was recently shown that it is undecidable whether a given OOM will eventually produce such negative values.

We propose a novel variant of OOMs, called *norm observable operator models* (NOOMs), which avoid the NPP by design. Like OOMs, NOOMs use a set of linear operators to update system states. But differing from OOMs, they represent probabilities by the square of the norm of system states, thus precluding negative "probability" values. While being free of the NPP, NOOMs retain most advantages of OOMs. For example, NOOMs also capture (some) processes that cannot be modelled by HMMs. More importantly, in principle NOOMs can be learnt from data in a constructive way; and the learnt models are asymptotically correct. We also prove that NOOMs capture all Markov chain (MC) describable processes.

This contribution presents the mathematical foundations of NOOMs, discusses the expressiveness of the model class, and explains how a NOOM can be estimated from data constructively.

1 Introduction and Overview

Hidden Markov models (HMMs) (Bengio, 1999) have been intensely studied from many angles (Ephraim and Merhav, 2002) and successfully applied in a wide range of fields, including speech processing (Rabiner, 1989), control engineering (Elliott et al.,

*Corresponding author.

1995) and biosequence analysis (Durbin et al., 2000). However, the EM (Baum-Welch) algorithm (Dempster et al., 1977; Baum et al., 1970), as *the* predominant learning algorithm of HMMs, is not entirely satisfactory due to slow convergence and the presence of local minima.

Observable operator models (OOMs) (Jaeger, 2000) are a proper generalization of HMMs (in that they extend the range of processes that can be modelled) which leads to fast and asymptotically correct learning algorithms. The key element in OOMs is to identify a sequence of observations $(a_t)_{t=1,2,...}$ with a sequence of linear operators $(\tau_{a_t})_{t=1,2,...}$ acting on system states, called *observable operators*; and to evaluate probabilities of certain outcomes by a linear functional on the state space. See Jaeger (1999) for an introduction to the general theory of OOMs, including infinite-dimensional, continuous-time, and continuous-valued processes.

In a machine learning context, we are particularly interested in discrete-time, finite-valued OOMs of finite dimension. Such OOMs can be naturally represented, under a proper basis of the state space, in a matrix formalism that is structually similar to a standard matrix formalism of HMMs. The theory of finite-dimensional OOMs, as well as their matrix representations, is presented in Jaeger et al. (2005).

From an algebraic point of view, OOMs differ from HMMs in that the matrix entries in OOMs can be any real numbers, whereas the analog entries in HMMs, interpretable as probabilities, must be nonnegative. This range extension in model parameters endows OOMs with remarkable mathematical features:

- OOMs are more expressive than HMMs, that is, every stochastic process that can be described by HMMs can also be modelled by OOMs, but not vice versa. In this sense, HMMs are a subclass of OOMs. See Jaeger (2000) for a simple 3-dimensional OOM example (the *probability clock*) that cannot be modelled by any finite-dimensional HMM.

- OOMs can be conveniently analyzed with methods from linear algebra (where an analogous treatment of HMMs is encumbered by the nonnegativity constraint in model parameters), giving rise to a purely linear algebraic way to describe and analyze stochastic processes (Faigle and Schönhuth, 2006, 2007; Schönhuth, 2006).

- The linear algebra nature of OOMs leads to a basic constructive procedure for estimating OOMs from data (Jaeger, 2000), based upon which a variety of highly efficient, constructive and asymptotically correct learning algorithms have been developed (Jaeger et al., 2005; Zhao et al., 2009).

However, extending the range of model parameters from nonnegative numbers to real numbers also raises a painful *negative probability problem* (NPP) (Jaeger et al., 2005) in the OOM theory. The NPP surfaces in various ways:

- An OOM is, by and large, represented by a finite set of matrices. A simple algorithm employs these matrices to calculate probabilities of events in a stochastic

process. Now, given a set of matrices, will this algorithm ever yield negative numbers? This decision problem is particularly relevant in the context of learning OOMs, where the set of matrices is estimated from data, and one would naturally want to determine whether the obtained model is "valid" in the sense of never producing negative "probabilities". Much effort has been spent on this problem in our group during the past years. The quest came to an end by a result due to Wiewiora (2008): "*It is an undecidable problem to decide whether or not a given collection of matrices (using them as a OOM without considering the nonnegativity constraint, i.e., the third condition from Table 1) produces only nonnegative values.*"

- Thus far no nontrivial sufficient conditions for a collection of matrices to be a valid OOM has been found. Consequently, none of the existing OOM learning algorithms (Jaeger et al., 2005; Zhao et al., 2009) takes the nonnegative-output constraint into account. Practical experience shows that models learnt by these algorithms are *typically* invalid in that they yield "negative probabilities" for some (rare) sequences.

We hasten to add that well-working heuristic methods for circumventing the NPP exist, e.g., the one presented in Appendix J of Jaeger et al. (2005). Such methods simply replace negative "probabilities", when encountered, with a small positive value and renormalize the OOM state. Yet, both for theoretical reasons and for avoiding such heuristics, it is highly desirable to have an alternative to OOMs which is free of the NPP, while keeping as many of the good properties of OOMs as possible. To this end, we propose in this paper a novel variant of OOMs, *norm observable operator models* (NOOMs).

NOOMs, like OOMs, associate each possible outcome a with a linear observable operator φ_a, and use these operators to update system states. But they take the square of the norm of system states for probabilities of certain outcomes, rather than values of a special linear functional of system states as OOMs do. This design makes the above mentioned NPP a nonissue for NOOMs.

This paper is theory-oriented and is mainly devoted to the mathematical foundation of NOOMs (Section 3). In particular, we show a mathematical construction of (possibly infinite-dimensional) coordinate-free NOOMs from discrete-time, finite-valued stochastic processes. Besides this fundamental result, we also discuss some related topics that (we thought) are of theoretical or practical interest, as outlined below.

By using the notation of Kronecker product (Brewer, 1978), we show how any m-dimensional NOOM can be converted into an equivalent finite-dimensional OOM. Therefore, NOOMs are revealed to be a subclass of OOMs [1] . It is currently unknown whether this inclusion is proper. We show, however, that NOOMs are powerful

[1] Unless otherwise stated, we were talking about finite-size models (HMMs with finitely many hidden states, finite-dimensional OOMs, etc.). See Section 2 for the mathematical definition of the dimension of an OOM.

enough to capture all Markov chain (MC) describable processes and some processes that cannot be modelled by HMMs, e.g., the above-mentioned probability clock (see Section 5). Since to construct such non-HMM NOOMs and to convert them into equivalent OOMs are relatively easy, this actually provides an efficient way to construct non-HMM OOMs from scratch, which was not at all an easy task before.

For generic models such as OOMs and NOOMs, learnability is of critical importance from an application perspective. NOOMs, while making the NPP a nonissue, retain most of the favorable features of OOMs in this respect. Specifically, they also admit a constructive and asymptotically correct learning algorithm – a dual commodity of which neither part is available for HMMs. A basic version of such a learning algorithm of NOOMs is outlined in Section 6.

Both OOMs and NOOMs encode predictions of future events into system states. There have been some related models on the basis of the same principle developed in the literature, most recently and most conspicuously, *predictive state representations* (PSRs). PSRs (Littman et al., 2001) generalize partially observable Markov decision processes (POMDPs) just like OOMs generalize HMMs. PSRs have been proposed to describe discrete, stochastic input-output systems and are more closely related to input-output OOMs (IO-OOMs). A thorough comparison between the two model classes is given in Jaeger (1998). (IO-)OOMs and PSRs use a linear functional on system states to evaluate the probabilities of certain future events, and thereby fall prey to the NPP.

The rest of the paper has the following organization. Section 2 reviews the basic OOM theory in a more abstract and more general fashion than that in Jaeger (1998); Jaeger et al. (2005), clarifying the main problems to be solved in the paper. The general mathematical foundation of NOOMs is established in Section 3, where we prove that any discrete stochastic process can be modelled by some abstract NOOM (of possibly infinite dimension). In Section 4, we demonstrate how NOOMs can be utilized to generate and predict stochastic time series. We study the expressiveness of NOOMs in Section 5, in which a NOOM version of the probability clock is presented (to show that some, actually many, NOOMs are "beyond HMMs"); and a general way for constructing non-HMM OOMs is introduced. We then briefly explain in Section 6 how a NOOM estimate can be obtained, in principle, from empirical data in a constructive way. Finally, we summarize the paper in Section 7.

2 A Review of Basic OOM Theory

Let us consider the class of all discrete-time stochastic processes $(X_t)_{t=1,2,\ldots}$ with each X_t taking values from a common finite alphabet $O = \{a^1, a^2, \ldots, a^\ell\}$. It is well known that, to completely describe the distribution of such a process (X_t) one needs only to specify, for each finite sequence $a_1 \ldots a_n$ of symbols from O, the initial joint probability of the form

$$\{\Pr(X_1 = a_1, \ldots, X_n = a_n) =: P(a_1 \ldots a_n)\}_{a_1,\ldots,a_n \in O,\, n=0,1,2,\ldots} \,. \tag{1}$$

We hence identify the distribution of a process (X_t) with (the set of) its initial probabilities (of finite strings), which we denote for brevity by $P(a_1 \dots a_n)$.

To simplify notation we denote by small letters with a bar finite sequences over O, e.g., $\bar{a} = a_1 a_2 \dots a_n$; and by O^* the set of all such finite sequences, including the *empty sequence* ϵ. The probability distribution (1) can then be simply written as $\{P(\bar{a})\}_{\bar{a} \in O^*}$, with the obvious agreement that $P(\epsilon) = 1$. Moreover, in the sequel we will often use phrases like "a process $P(\bar{a})$", which, obviously, should be understood as "a stochastic process with the distribution specified by $\{P(\bar{a})\}_{\bar{a} \in O^*}$".

It was proven in Jaeger (1998, 2000) that any process $P(\bar{a})$ can be modelled by an abstract linear system of the form $(\mathcal{H}, \{T_a\}_{a \in O}, w_0, \sigma)$, where $\mathcal{H}$ is a real vector space (possibly of infinite dimension), the T_a's are linear operators on $\mathcal{H}$, $w_0 \in \mathcal{H}$ is the initial state of the system and σ is a linear functional on $\mathcal{H}$, through the formula

$$P(\bar{a}) = P(a_1 a_2 \dots a_n) = \sigma T_{a_n} T_{a_{n-1}} \cdots T_{a_1} w_0 =: \sigma T_{\bar{a}} w_0 \,,$$

where $T_{\bar{a}}$ denotes, for any sequence $\bar{a} = a_1 a_2 \dots a_n \in O^*$, the *reverse-ordered* product $T_{a_n} T_{a_{n-1}} \cdots T_{a_1}$, with the agreement that $T_\epsilon = \mathrm{id}_{\mathcal{H}}$ — the identity operator on $\mathcal{H}$.

We call the structure $(\mathcal{H}, \{T_a\}_{a \in O}, w_0, \sigma)$ an *abstract OOM* of the process $P(\bar{a})$. Its dimension is defined as the dimension of the space $\mathcal{H}$ [2] . Obviously $P(\bar{a})$ can be seen as a real-valued function on O^*, with some special properties that will be discussed in detail in Subsection 2.2. In other words, the class of distributions of stochastic processes can be embedded into the class $\mathcal{F}$ of functions from O^* into $\mathbb{R}$. We will later extend the concept of (abstract) OOMs to the larger class $\mathcal{F}$ (Subsection 2.1).

In machine learning applications, we are particularly interested in modeling stochastic processes by OOMs of finite dimension. The class of processes that admit finite-dimensional OOMs have been characterized independently several times and are now often termed *linearly dependent processes* (LDPs). They have a long history of mathematical investigations (Heller, 1965; Ito et al., 1992). The formal definition of LDPs (independent of OOMs) will be given in Subsection 2.1.

Now let $P(\bar{a})$ be a LDP specified by an OOM $(\mathcal{H}, \{T_a\}_{a \in O}, w_0, \sigma)$ of dimension $\dim \mathcal{H} = m$. By selecting a proper basis $\{w_1, w_2, \dots, w_m\}$ of the space $\mathcal{H}$ with the property $\sigma w_i = 1$ $(i = 1, 2, \dots, m)$ [3] , we obtain the following representation:

$$\mathcal{H} \to \mathbb{R}^m \,, \quad T_a \to \tau_a \in \mathbb{R}^{m \times m} \,, \quad w_0 \to \boldsymbol{w}_0 \in \mathbb{R}^m \,, \quad \sigma \to \mathbf{1}_m^{\mathsf{T}} \,, \tag{2}$$

where $\mathbf{1}_m$ denotes the m-dimensional vector of units. We hence get a matrix representation of the abstract system $(\mathcal{H}, \{T_a\}_{a \in O}, w_0, \sigma)$, namely $(\mathbb{R}^m, \{\tau_a\}_{a \in O}, \boldsymbol{w}_0)$,

[2] This definition of dimension of OOMs is different from that in Jaeger (2000).

[3] This can be done in three steps: first select an arbitrary basis $\{w_1, w_2, \dots, w_m\}$ of $\mathcal{H}$, then there must be some k such that $\sigma w_k \neq 0$ since σ is clearly not a zero functional, pick one such k; next let $w_i \leftarrow w_i + w_k$ for any i with $\sigma w_i = 0$, we therefore have $\sigma w_i \neq 0$ for all i; finally set $w_i \leftarrow (\sigma w_i)^{-1} w_i$ for all i.

in which the "default" element $\mathbf{1}_m^\mathsf{T}$ has been omitted. We will sometimes call such matrix representations of finite-dimensional OOMs *concrete* OOMs.

One sees that a concrete OOM $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ has the same structure as a (transition-emission) HMM (Bourlard and Bengio, 2002; Jaeger, 2000), but with model parameters from a larger domain (real numbers) than HMMs (nonnegative numbers). It is a (nontrivial) consequence of this added algebraic freedom that OOMs can capture more processes than HMMs (Jaeger, 2000). Moreover, a concrete OOM $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ calculates a distribution $P(\bar{a})$ in a linear way:

$$P(\bar{a}) = P(a_1 a_2 \dots a_n) = \mathbf{1}_m^\mathsf{T} \tau_{a_n} \tau_{a_{n-1}} \cdots \tau_{a_1} \boldsymbol{w}_0 =: \mathbf{1}_m^\mathsf{T} \tau_{\bar{a}} \boldsymbol{w}_0 \,, \tag{3}$$

enabling us to study stochastic processes by methods from linear algebra encumbered by the nonnegative-paramater constraint of the analog HMM mechanism. However, as we pointed out before, this elegance has its price: the undecidable NPP. Using (3), the NPP can now be stated more precisely as follows: *To check whether a given matrix system* $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$ *yields a negative value* $P(\bar{a}) = \mathbf{1}_m^\mathsf{T} \tau_{\bar{a}} \boldsymbol{w}_0 < 0$ *on some* $\bar{a} \in O^*$ *is an undecidable problem* (Wiewiora, 2008).

A natural way one could think to back out of the NPP is to wrap the right hand side (r.h.s.) of (3) by some nonlinear function which only has nonnegative values. There are several natural choices for this, for example, one might require that $P(\bar{a}) = (\mathbf{1}_m^\mathsf{T} \tau_{\bar{a}} \boldsymbol{w}_0)^2$ (giving rise to *quadratic OOMs (QOOMs)*) or $P(\bar{a}) = \|\tau_{\bar{a}} \boldsymbol{w}_0\|^2$ (this gives *norm observable operator models (NOOMs)*). The first author investigated both alternatives. While the present contribution focusses solely on the latter, here is a short account of the status of QOOM research. QOOMs admit a straightforward re-use of the known learning algorithms for OOMs, and thus were the first of the two new model classes that were explored. It turned out however that it is not easy to check whether a given set of matrices, considered as a candidate QOOM, satisfies the fundamental summation constraint required from models of stochastic processes (see Table 1 below). This circumstance hindered swift progress in QOOMs, and made us turn toward NOOMs instead.

2.1 OOMs for linearly dependent functions

To pave the ground for the subsequent development of a theory of NOOMs, we will first generalize linearly dependent processes to *linearly dependent functions* (LDFs).

Throughout the paper we shall denote by $\mathcal{F}$ the set of all real-valued functions defined on O^*. With addition and scalar multiplication defined pointwise, $\mathcal{F}$ canonically becomes a real vector space. For each symbol $a \in O$, we define an operator L_a on the vector space $\mathcal{F}$ by setting $(L_a f)(\bar{x}) := f(a\bar{x})$ for any $f \in \mathcal{F}$ and $\bar{x} \in O^*$. These operators are called *left-appending operators*. Note that each left-appending operator L_a is a linear operator on $\mathcal{F}$. In fact, for any $f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$, we have

$$\begin{aligned}
&(L_a(f+g))(\bar{x}) = (f+g)(a\bar{x}) = f(a\bar{x}) + g(a\bar{x}) = (L_a f + L_a g)(\bar{x}) \,, \\
&(L_a(\alpha f))(\bar{x}) = (\alpha f)(a\bar{x}) = \alpha f(a\bar{x}) = \alpha (L_a f)(\bar{x}) = (\alpha L_a f)(\bar{x})
\end{aligned}$$

for all $\bar{x} \in O^*$. It then follows that $L_a(f+g) = L_a f + L_a g$ and $L_a(\alpha f) = \alpha L_a f$.

Iteratively applying left-appending operators on a function $h \in \mathcal{F}$, we get a linear formula to evaluate the value of h at any sequence $\bar{a} = a_1 a_2 \ldots a_n \in O^*$:

$$\begin{aligned} h(a_1 a_2 \ldots a_n) &= (L_{a_1} h)(a_2 \ldots a_n) = (L_{a_2} L_{a_1} h)(a_3 \ldots a_n) \\ &= \ldots = (L_{a_n} \cdots L_{a_2} L_{a_1} h)(\epsilon) := \sigma L_{\bar{a}} h \,, \end{aligned} \tag{4}$$

where, as before, $L_{\bar{a}}$ denotes the *reverse*-ordered composition of $L_{a_1}, \ldots, L_{a_n}$; and σ is the linear functional on $\mathcal{F}$ that maps each $f \in \mathcal{F}$ to the real number $f(\epsilon)$. By (4) we see that $(\mathcal{F}, \{L_a\}_{a \in O}, \sigma)$ provides an algebraic representation of the family $\mathcal{F}$ that allows us to calculate the value $h(\bar{a})$ of any $h \in \mathcal{F}$ on any $\bar{a} \in O^*$.

For any single $h \in \mathcal{F}$ we define $\mathcal{F}^h$ to be the subspace of $\mathcal{F}$ spanned by the functions $\{L_{\bar{a}} h : \bar{a} \in O^*\}$ and call it the space *induced* by h. It is clear that $\mathcal{F}^h$ is invariant under the operation of each L_a, that is, $f \in \mathcal{F}^h$ implies $L_a f \in \mathcal{F}^h$. So we can restrict the domain of L_a's and σ to the space $\mathcal{F}^h$, getting a new set of linear operators and a new linear functional on $\mathcal{F}^h$ which we will denote by the same symbols L_a and σ, respectively, trusting in the reader's good sense to avoid confusion. We thus obtain a smaller system $(\mathcal{F}^h, \{L_a\}_{a \in O}, h, \sigma)$ in which again $h(\bar{a}) = \sigma L_{\bar{a}} h$ holds for all $\bar{a} \in O^*$.

We call a function $h \in \mathcal{F}$ a *linearly dependent function* (LDF) if it induces a finite-dimensional space $\mathcal{F}^h$; and a process (X_t) a *linearly dependent process* (LDP) if its distribution $P(\bar{a})$, regarded as a member of $\mathcal{F}$, is a LDF. One notes that LDPs can be seen as a subclass of LDFs.

A remark on terminology: to the best of our knowledge, the generalization from LDPs to LDFs has not been done previously in the literature; and "LDF" (inherited from the term "LDP") is a new terminology with the special meaning explained above. Specifically, the word "LDFs" should not be understood in the standard sense as "(a set of) functions that are linearly dependent" [4] .

Assume now $h \in \mathcal{F}$ is a LDF with $\dim \mathcal{F}^h = m$, then, as in (2), we can select a proper basis of $\mathcal{F}^h$ and get a matrix model $(\mathbb{R}^m, \{\tau_a\}_{a \in O}, \boldsymbol{w}_0)$ which describes the function h just like an OOM describes a distribution (see equation (3)); we hence also call the model $(\mathbb{R}^m, \{\tau_a\}_{a \in O}, \boldsymbol{w}_0)$ an OOM of h. Obviously, the value of a LDF h at any $\bar{a} \in O^*$ can be computed from its OOM by the formula $h(\bar{a}) = \mathbf{1}_m^{\mathsf{T}} \tau_{\bar{a}} \boldsymbol{w}_0$, with the understanding that $\tau_\epsilon = I_m$, the identity matrix of order m.

2.2 Probability constraints on models of LDPs

In the previous subsection we have re-derived (and broadened) the concept of OOMs in an abstract-to-concrete, general-to-special manner; and have shown that any LDF, and hence any LDP, can be modelled by some (finite-dimensional) OOM. In this subsection we will focus our attention on LDPs and their OOMs.

As mentioned before, LDPs can be seen as a special subclass of LDFs characterized by three conditions which are essentially inherited from the well-known

[4] See, e.g., `http://mathworld.wolfram.com/LinearlyDependentFunctions.html`.

Kolmogorov axioms of probability theory. We will refer to these conditions as *probability constraints*. They are listed in Table 1 together with their reflections in OOMs (where $\bar{x}$ runs over O^* and $\bar{x}a$ denotes the concatenation of $\bar{x}$ and a; see Jaeger (2000) for a detailed explanation and proofs of these probability contraints).

Table 1: Probability constraints for LDPs and their OOMs.

probability constraints	in OOM $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0)$
unity constraint: $P(\epsilon) = 1$	$\mathbf{1}_m^\intercal \boldsymbol{w}_0 = 1$
summation constraint: $P(\bar{x}) = \sum_{a\in O} P(\bar{x}a)$	$\mathbf{1}_m^\intercal \sum_{a\in O} \tau_a = \mathbf{1}_m^\intercal$
nonnegativity constraint: $P(\bar{x}) \geqslant 0$	$\mathbf{1}_m^\intercal \tau_{\bar{x}} \boldsymbol{w}_0 \geqslant 0$

As one can see from the second column of the above table, for OOMs of LDPs, checking the first two probabilty constraints is straightforward, whereas the third one is (as we know today thanks to Wiewiora (2008)) undecidable. We therefore turn to variants of the OOM class for which the nonnegativity constraint is automatically fulfilled.

We have already mentioned two such variants right before Subsection 2.1, viz. *quadratic observable operator models* (QOOMs) and *norm observable operator models* (NOOMs). They compute a distribution $P(\bar{a})$ by $P(\bar{a}) = (\mathbf{1}_m^\intercal \varphi_{\bar{a}} \boldsymbol{u}_0)^2$ and $P(\bar{a}) = \|\varphi_{\bar{a}} \boldsymbol{u}_0\|^2$, respectively. Here $\|\cdot\|$ denotes Euclidean norm and different symbols have been used for operators and initial states to distinguish the two OOM variants from normal "linear" OOMs. This paper will only investigate NOOMs, for which the probability constraints read as in Table 2.

Table 2: Probability constraints for LDPs and their NOOMs.

probability constraints	in NOOM $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$
unity constraint: $P(\epsilon) = 1$	$\|\boldsymbol{u}_0\| = 1$
summation constraint: $P(\bar{x}) = \sum_{a\in O} P(\bar{x}a)$	$\|\varphi_{\bar{x}} \boldsymbol{u}_0\|^2 = \sum_{a\in O} \|\varphi_a \varphi_{\bar{x}} \boldsymbol{u}_0\|^2$
nonnegativity constraint: $P(\bar{x}) \geqslant 0$	$\|\varphi_{\bar{x}} \boldsymbol{u}_0\|^2 \geqslant 0$

For NOOMs, the unity constraint is easy to check and the nonnegativity constraint a nonissue. The summation constraint, however, contains infinitely many equality constraints and seems to be much more difficult than its linear OOM analog. To algebraically characterize the summation constraint for NOOMs, and giving a simple decision procedure, is one of the main results of this paper.

3 Norm Observable Operator Models

In the previous section we defined a NOOM as a triple $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ that satisfies the three conditions listed in Table 2. By the definition of the Euclidean norm $\|\boldsymbol{x}\| := \sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{x}}$, the summation constraint for NOOMs can be rewritten as

$$(\varphi_{\bar{x}}\boldsymbol{u}_0)^\mathsf{T}(\textstyle\sum_{a\in O}\varphi_a^\mathsf{T}\varphi_a)(\varphi_{\bar{x}}\boldsymbol{u}_0) = (\varphi_{\bar{x}}\boldsymbol{u}_0)^\mathsf{T}(\varphi_{\bar{x}}\boldsymbol{u}_0)\,. \quad (\forall \bar{x}\in O^*) \tag{5}$$

It is hence clear that $\sum_{a\in O}\varphi_a^\mathsf{T}\varphi_a = I_m$ is a sufficient condition for the summation constraint (5). This motivates us to introduce the concept of *standard NOOMs*.

Definition 1 *A (finite-dimensional)* standard NOOM *is a system* $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ *with* $\varphi_a \in \mathbb{R}^{m\times m}$ *and* $\boldsymbol{u}_0 \in \mathbb{R}^m$*, such that* $\sum_{a\in O}\varphi_a^\mathsf{T}\varphi_a = I_m$ *and* $\|\boldsymbol{u}_0\| = 1$*.*

A remark on terminology: in the above defintion the qualifier *standard* refers to the property $\sum_{a\in O}\varphi_a^\mathsf{T}\varphi_a = I_m$. For expository reasons we have introduced it here for finite-dimensional models in matrix representations. But it is straightforward to generalize this definition of standard NOOMs to possibliy infinite-dimensional models in a coordinate-free representation, as is revealed by Theorem 1.

Theorem 1 *For any stochastic process* $P(\bar{a})$ *there exist an inner product space* $\mathcal{E}$*, a set of linear operators* $\varphi_a : \mathcal{E}\to\mathcal{E}$ $(a\in O)$ *and an initial vector* $u_0\in\mathcal{E}$*, such that* $\|u_0\| = 1$*,* $\sum_{a\in O}\varphi_a^*\varphi_a = \mathrm{id}_\mathcal{E}$ *(the identity map on* $\mathcal{E}$*) and* $P(\bar{a}) = \|\varphi_{\bar{a}}u_0\|^2$ *for all* $\bar{a}\in O^*$*, where* φ_a^* *denotes the* adjoint *of* φ_a*, a linear operator on* $\mathcal{E}$ *defined by the property "*$\langle\varphi_a^* v, w\rangle = \langle v, \varphi_a w\rangle$ *for all* $v, w\in\mathcal{E}$*".*

Conversely, any triple $(\mathcal{E}, \{\varphi_a\}_{a\in O}, u_0)$ *with the properties* $\sum_{a\in O}\varphi_a^*\varphi_a = \mathrm{id}_\mathcal{E}$ *and* $\|u_0\| = 1$ *describes some stochastic process via* $P(\bar{a}) = \|\varphi_{\bar{a}}u_0\|^2$*.*

This theorem gives rise to, for each process $P(\bar{a})$, a system $(\mathcal{E}, \{\varphi_a\}_{a\in O}, u_0)$ of an inner product space $\mathcal{E}$ of possibly infinite dimension, a set $\{\varphi_a\}_{a\in O}$ of linear operators on $\mathcal{E}$ and an initial state $u_0\in\mathcal{E}$ with the properties $\sum_{a\in O}\varphi_a^*\varphi_a = \mathrm{id}_\mathcal{E}$ and $\|u_0\| = 1$, which describes $P(\bar{a})$ by $P(\bar{a}) = \|\varphi_{\bar{a}}u_0\|^2$. We call the triple $(\mathcal{E}, \{\varphi_a\}_{a\in O}, u_0)$ a (standard) abstract NOOM of the process $P(\bar{a})$, with the qualifier "standard" now referring to the condition $\sum_{a\in O}\varphi_a^*\varphi_a = \mathrm{id}_\mathcal{E}$.

By Theorem 1, the class of standard NOOMs has the same expressiveness as the class of general NOOMs. In other words, any general NOOM allows an equivalent standard NOOM such that both models describe the same process. In this sense, the equality $\sum_{a\in O}\varphi_a^\mathsf{T}\varphi_a = I_m$ can be seen as the sufficient and necessary condition for the summation constraint (5). Note that, according to its definition, it is trivial to check whether a given system $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ is a standard NOOM or not, whereas we emphasize again that the analog problem is undecidable for OOMs. In the sequel we shall consider only standard NOOMs and simply call them NOOMs.

The rest of the section is devoted to the proof to Theorem 1. While the *conversely* part is easy to prove (by using the Kolmogorov Extension Theorem, see Appendix A for the detail), the proof to the first part is not trivial. It will be divided into two parts: we first (Subsection 3.1) define an inner product space $\mathcal{D}$ which plays a similar role for NOOMs as the vector space $\mathcal{F}$ plays for OOMs of LDFs; and then (Subsection 3.2) construct an NOOM of a given process $P(\bar{a})$ in the space $\mathcal{D}$.

3.1 The Inner Product Space $\mathcal{D}$

The construction of the inner product space $\mathcal{D}$ is further divided into three steps:

1. The set $\mathcal{P}$ of all nonnegative functions $p(\bar{a})$ with $p^2(\bar{a})$ satisfying the three probability constraints is considered. These functions will be referred to as *probability amplitudes* (a terminology borrowed from quantum machanics).

2. As $\mathcal{P}$ is not a linear subspace of $\mathcal{F}$, we embed it into a convex cone $\mathcal{B}^+$ (in $\mathcal{F}$), on which a bilinear function $(\!|\cdot,\cdot|\!)$ is defined. The domain of $(\!|\cdot,\cdot|\!)$ is then extended to the subspace $\mathcal{B}$ spanned by $\mathcal{B}^+$, yielding a *semidefinite inner product* space $(\mathcal{B}, (\!|\cdot,\cdot|\!))$ that contains $\mathcal{P}$ as a subset.

3. A *null* subspace $\mathcal{N}$ of $\mathcal{B}$ is defined through the semidefinite inner product $(\!|\cdot,\cdot|\!)$; and $\mathcal{D}$ is then defined to be the *quotient space* $\mathcal{B}/\mathcal{N}$, on which an inner product $\langle\cdot,\cdot\rangle$ is induced from $(\!|\cdot,\cdot|\!)$.

The family $\mathcal{P}$ of probability amplitudes. Let $\mathcal{P}$ be the family consisting of all nonnegative functions $p : O^* \to \mathbb{R}^+$ such that $p^2(\bar{a})$ satisfies all the probability constraints. In other words, here $\mathcal{P}$ is a subset of $\mathcal{F}$ defined by

$$\mathcal{P} := \left\{p \in \mathcal{F} \,:\, p(\epsilon) = 1,\ \forall\, \bar{x} \in O^*,\ p(\bar{x}) \geqslant 0,\ p^2(\bar{x}) = \textstyle\sum_{a\in O} p^2(\bar{x}a)\right\}. \qquad (6)$$

It is clear that each $p \in \mathcal{P}$ specifies a stochastic process distribution $P(\bar{a}) := p^2(\bar{a})$; and that each process $P(\bar{a})$ determines a probability amplitude $p \in \mathcal{P}$ via $p(\bar{a}) = \sqrt{P(\bar{a})}$. We may therefore identify each distribution $P(\bar{a})$ with its probability amplitude $p(\bar{a})$; and view $\mathcal{P}$ as the class of distributions of stochastic processes.

The convex cone $\mathcal{B}^+$ and the subspace $\mathcal{B}$ in $\mathcal{F}$. The family $\mathcal{P}$ of probability amplitudes, as a subset of $\mathcal{F}$, is neither a vector space nor invariant under the left-appending operators L_a. We thus need to find a larger subset of $\mathcal{F}$ that contains $\mathcal{P}$ as a subset and, at the same time, is an invariant subspace of $\mathcal{F}$ under the operation of L_a's.

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

$$\mathcal{B}^+ := \{f \in \mathcal{F} \,:\, \forall\, \bar{x} \in O^*,\ f(\bar{x}) \geqslant 0,\ f^2(\bar{x}) \geqslant \textstyle\sum_{a\in O} f^2(\bar{x}a)\}. \qquad (7)$$

Then it is clear that $\mathcal{P} \subseteq \mathcal{B}^+$. Moreover,

Theorem 2 *$\mathcal{B}^+$ is a* convex cone *in $\mathcal{F}$ pointed at* 0 *(the zero function), that is, for any $f, g \in \mathcal{B}^+$ it holds that* (i) *$-f \in \mathcal{B}^+$ implies $f = 0$;* (ii) *$\alpha f \in \mathcal{B}^+$ for any $\alpha \geqslant 0$; and* (iii) *$f + g \in \mathcal{B}^+$. Furthermore, $\mathcal{B}^+$ is invariant under the operators L_a, in other words, $L_a f \in \mathcal{B}^+$ whenever $f \in \mathcal{B}^+$.*

Proof: The assertions (i) and (ii) can be easily deduced from equation (7). To

prove (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 Cauchy's inequality and (7) tell us

$$\left[\sum_{a\in O} f(\bar{x}a)g(\bar{x}a)\right]^2 \leqslant \left[\sum_{a\in O} f^2(\bar{x}a)\right]\cdot\left[\sum_{a\in O} g^2(\bar{x}a)\right] \leqslant f^2(\bar{x})g^2(\bar{x})\,. \tag{8}$$

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

For each $n \in \mathbb{N} = \{0, 1, 2, \ldots\}$ we define a binary function Q_n on $\mathcal{F}$, by

$$Q_n(f, g) := \sum_{\bar{a}\in O^n} f(\bar{a})g(\bar{a})\,, \qquad (\forall\, f, g \in \mathcal{F})\,. \tag{9}$$

Then, for any $f, g \in \mathcal{B}^+$, the sum of the inequalities $f(\bar{x})g(\bar{x}) \geqslant \sum_{a\in O} f(\bar{x}a)g(\bar{x}a)$ (cf. equation (8)) over all $\bar{x} \in O^n$ reveals that $(Q_n(f, g))_{n=0,1,2,\ldots}$ is a decreasing sequence lower bounded by 0, so the limit

$$(\!|f, g|\!) := \lim_{n\to\infty} Q_n(f, g) = \lim_{n\to\infty} \sum_{\bar{a}\in O^n} f(\bar{a})g(\bar{a}) \tag{10}$$

exists and takes values in $[0, \infty)$.

Now let $\mathcal{B}$ be the subspace of $\mathcal{F}$ spanned by vectors in $\mathcal{B}^+$. As $\mathcal{B}^+$ is a convex cone, we know $\mathcal{B}$ consists exactly of those functions $h \in \mathcal{F}$ which can be written as the difference of two members from $\mathcal{B}^+$:

$$\mathcal{B} := \operatorname{span}\mathcal{B}^+ = \{f - g \,:\, f, g \in \mathcal{B}^+\}\,. \tag{11}$$

Since $\mathcal{B}^+$ is invariant under each L_a (see Theorem 2), by (11) we know $\mathcal{B}$ is an invariant subspace of the linear operators L_a. This allows us to restrict the operation of L_a's on the space $\mathcal{B}$ in the sequel. To extend the domain of the binary function $(\!|\cdot, \cdot|\!)$ defined by (10) from $\mathcal{B}^+ \times \mathcal{B}^+$ to $\mathcal{B} \times \mathcal{B}$, we need the following lemma to which the proof is trivial and omitted here.

Lemma 1 *Let $(a_n^i)_{n=0,1,2,\ldots}$, $(b_n^i)_{n=0,1,2,\ldots}$ $(i = 1, 2, \ldots, k)$ be $2k$ real sequences 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 $(\sum_{i=1}^k a_n^i)_{n=0,1,2,\ldots}$ and $(\sum_{i=1}^k b_n^i)_{n=0,1,2,\ldots}$ are two convergent sequences with the same limit $\sum_{i=1}^k c^i$.*

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

$$Q_n(f, g) = Q_n(f_1, g_1) + Q_n(f_2, g_2) - Q_n(f_1, g_2) - Q_n(f_2, g_1)\,. \tag{12}$$

Since $f_i, g_i \in \mathcal{B}^+$, each of the four items on the r.h.s. of the above equality converges to a nonnegative number when $n \to \infty$; and their sum $Q_n(f, g)$ is independent of the choice of f_1, f_2, g_1, g_2. Thus, by Lemma 1 the limit $(\!|f, g|\!) = \lim_{n\to\infty} Q_n(f, g)$ exists and, by (12), assumes values in $\mathbb{R}$.

By its definition (cf. equation (10)), one easily sees that the function $(\!|f, g|\!)$ is

(a) *nonnegative definite*: $(\!(f, f)\!) \geqslant 0$;
(b) *symmetric*: $(\!(f, g)\!) = (\!(g, f)\!)$;
(c) *linear* (in f and g): $(\!(\alpha f + \beta h, g)\!) = \alpha(\!(f, g)\!) + \beta(\!(h, g)\!)$,
$(\!(f, \alpha g + \beta h)\!) = \alpha(\!(f, g)\!) + \beta(\!(f, h)\!)$.

In the above list, f, g, h are arbitrary functions in $\mathcal{B}$ and $\alpha, \beta \in \mathbb{R}$. We proceed to construct an inner product space $\mathcal{D}$ from $\mathcal{B}$ and $(\!(\cdot, \cdot)\!)$, taking a route via semidefinite inner products.

Semidefinite inner product and seminorm To be self-contained this subsection introduces the construction of inner products (norms) from semidefinite inner products (seminorms) in an abstract vector space V. After giving the formal definition of semidefinite inner products and seminorms, we show that from any semidefinite inner product a seminorm can be defined. We then argue that points with zero seminorm form a linear subspace N of V, and so the quotient space V/N is well defined. Finally we define on this quotient space an inner product (norm). Readers familiar with this standard mathematical treatment may skip this subsection.

Let V be a vector space over $\mathbb{R}$. A *semidefinite inner product* on V is any binary function $(\!(\cdot, \cdot)\!) : V \times V \to \mathbb{R}$ that is nonnegative definite, symmetric and linear in both arguments. A *seminorm* on V is any nonnegative unary function $[\![\cdot]\!]$ that satisfies (i) $[\![\alpha x]\!] = |\alpha| \cdot [\![x]\!]$ and (ii) $[\![x+y]\!] \leqslant [\![x]\!] + [\![y]\!]$ for all $x, y \in V$ and $\alpha \in \mathbb{R}$. Note that, by the condition (i) we know $[\![0]\!] = 0$ for any seminorm $[\![\cdot]\!]$ on V.

Like each inner product $\langle \cdot, \cdot \rangle$ induces naturally a norm $\| \cdot \|$, each semidefinite inner product $(\!(\cdot, \cdot)\!)$ induces a seminorm $[\![\cdot]\!]$ by putting $[\![x]\!] := \sqrt{(\!(x, x)\!)}$ for each $x \in V$. To see this, we need an inequality, namely,

Lemma 2 (Cauchy-Schwarz inequality for semidefinite inner products) *Let V be a vector space and $(\!(\cdot, \cdot)\!)$ a semidefinite inner product on V. Let $[\![\cdot]\!] : V \to \mathbb{R}$ be defined by $[\![x]\!] = \sqrt{(\!(x, x)\!)}$. Then, for any $x, y \in V$, $|(\!(x, y)\!)| \leqslant [\![x]\!] \cdot [\![y]\!]$.*

Proof: By the definition of semidefinite inner product, we have

$$0 \leqslant (\!(x - \alpha y, x - \alpha y)\!) = [\![x]\!]^2 - 2\alpha(\!(x, y)\!) + \alpha^2 [\![y]\!]^2 \ . \tag{13}$$

If $[\![y]\!] = 0$, then (13) reads: $[\![x]\!]^2 - 2\alpha(\!(x, y)\!) \geqslant 0$ $(\forall \alpha \in \mathbb{R})$. Thus $(\!(x, y)\!) = 0$ and the desired inequality follows. If $[\![y]\!] \neq 0$, the quadratic $[\![x]\!]^2 - 2\alpha(\!(x, y)\!) + \alpha^2 [\![y]\!]^2$ (on α) assumes no negative value. So its discriminant is nonpositive, which implies the claim. □

We now show that $[\![x]\!] = \sqrt{(\!(x, x)\!)}$ is indeed a seminorm. The conditions $[\![x]\!] \geqslant 0$ and $[\![\alpha x]\!] = |\alpha| \cdot [\![x]\!]$ are easy to verify. To prove $[\![x+y]\!] \leqslant [\![x]\!] + [\![y]\!]$, we calculate $[\![x+y]\!]^2 = [\![x]\!]^2 + 2(\!(x, y)\!) + [\![y]\!]^2 \leqslant [\![x]\!]^2 + 2[\![x]\!] \cdot [\![y]\!] + [\![y]\!]^2 = ([\![x]\!] + [\![y]\!])^2$, where the second inequality follows from Lemma 2. We thus get $[\![x + y]\!] \leqslant [\![x]\!] + [\![y]\!]$.

Geometrically, a vector space V equipped with a seminorm $[\![\cdot]\!]$ (or a semidefinite inner product $(\!(\cdot, \cdot)\!)$) can be seen as a "weak" distance space in which two distinct points $x, y \in V$ may have zero distance: $[\![x - y]\!] = 0$. However, if one identifies all

these zero-distanced points, he obtains a normed vector space (or an inner product space).

Let $[\![\cdot]\!]$ be a seminorm on the vector space V. We call a point $x \in V$ a *null element* (w.r.t. the seminorm $[\![\cdot]\!]$) if $[\![x]\!] = 0$. Let $N \subseteq V$ be the set of all null elements in V. By the conditions $[\![\alpha x]\!] = |\alpha| \cdot [\![x]\!]$ and $[\![x+y]\!] \leqslant [\![x]\!] + [\![y]\!]$ we know N is a subspace of V. We therefore can define the *quotient space* $V/N := \{[x] : x \in V\}$, where $[x]$ denotes, for each $x \in V$, the *equivalence class* $\{y \in V : x - y \in N\}$.

It is well known that the quotient space V/N forms a vector space under the addition and scalar multiplication operations

$$[x] + [y] := [x + y]\,, \quad \alpha[x] := [\alpha x]\,. \qquad (\forall\, x, y \in V\,,\ \ \forall\, \alpha \in \mathbb{R})$$

Note that the definitions above are independent of the choice of the "representative" elements x, y (in their equivalence classes $[x]$ or $[y]$). Finally, the vector space V/N becomes a normed space under the norm $\|[x]\| := [\![x]\!]$, $(\forall\, x \in V)$.

If the seminorm $[\![\cdot]\!]$ is induced from some semidefinite inner product $(\!(\cdot,\cdot)\!)$, we can correspondingly define an inner product $\langle\cdot,\cdot\rangle$ in the quotient space V/N:

$$\langle [x], [y] \rangle := (\!(x, y)\!)\,, \qquad (\forall\, x, y \in V)\,. \tag{14}$$

To see that the function $\langle\cdot,\cdot\rangle$ is well-defined, we need to prove $(\!(x, y)\!) = (\!(a, b)\!)$ for any $a \in [x]$ and $b \in [y]$. Noting that $a \in [x]$ iff $x - a \in N$ iff $[\![x - a]\!] = 0$, we know from Lemma 2 that, for any $z \in V$, $|(\!(x - a, z)\!)| \leqslant [\![x - a]\!] \cdot [\![z]\!] = 0$ and hence $(\!(x, z)\!) = (\!(a, z)\!) + (\!(x - a, z)\!) = (\!(a, z)\!)$. Similary, $(\!(z, y)\!) = (\!(z, b)\!)$ for any z. We thus get $(\!(x, y)\!) = (\!(a, y)\!) = (\!(a, b)\!)$. Moreover, since $(\!(\cdot,\cdot)\!)$ is a semidefinite inner product, by (14) one sees that $\langle\cdot,\cdot\rangle$ is also a semidefinite inner product. Now assume $\langle [x], [x] \rangle = 0$, then $(\!(x, x)\!) = [\![x]\!]^2 = 0$, thus $x \in N = [0]$ and so $[x] = [0]$. This proves that $\langle\cdot,\cdot\rangle$ is indeed an inner product on the space V/N. This inner product induces the norm $\|[x]\| = \sqrt{\langle [x], [x] \rangle} = \sqrt{(\!(x, x)\!)} = \sqrt{[\![x]\!]^2} = [\![x]\!]$, which is exactly what we have introduced above for the general case where only a seminorm is defined.

The quotient space $\mathcal{D}$ and its process-representing subset $\mathcal{D}_{\mathrm{P}}$. Now let us return to the vector space $\mathcal{B}$ defined by (11), on which we have constructed a semidefinite inner product (10). Therefore, as discussed above, we can define

1) the seminorm (on $\mathcal{B}$): $[\![f]\!] := \sqrt{(\!(f, f)\!)} = \lim_{n\to\infty} \sqrt{\sum_{\bar{a}\in O^n} f^2(\bar{a})}$;
2) the subspace (of $\mathcal{B}$): $\mathcal{N} := \{f \in \mathcal{B} : [\![f]\!] = 0\}$;
3) the quotient space: $\mathcal{D} := \mathcal{B}/\mathcal{N} = \{[f] : f \in \mathcal{B}\}$;
4) the inner product (on $\mathcal{D}$): $\langle [f], [g] \rangle := (\!(f, g)\!) = \lim_{n\to\infty} \sum_{\bar{a}\in O^n} f(\bar{a}) g(\bar{a})$.

In sum, so far we have obtained a semidefinite inner product space $\mathcal{B}$, which contains the family $\mathcal{P}$ of all probability amplitudes — we have $\mathcal{P} \subseteq \mathcal{B}^+ \subseteq \mathcal{B}$; and an inner product space $\mathcal{D}$ which can be seen as an "image" of $\mathcal{B}$. Next we will investigate those members $[f]$ ($f \in \mathcal{B}$) of $\mathcal{D}$ that actually represent a stochastic process, i.e., the subset $\mathcal{D}_{\mathrm{P}} := \{[f] : f \in \mathcal{P}\}$ of $\mathcal{D}$.

Firstly, by (6) and induction on n, we know $\sum_{\bar{a}\in O^n} p^2(\bar{a}) = 1$ for all $p \in \mathcal{P}$ and $n \in \mathbb{N}$. Thus, for any $[f] \in \mathcal{D}_{\rm P}$, $[f] = [p]$ for some $p \in \mathcal{P}$ and hence $\|[f]\| = \|[p]\| = [\![p]\!] = 1$. This means $\mathcal{D}_{\rm P}$ lies on the unit sphere of $\mathcal{D}$, i.e.,

$$\mathcal{D}_{\rm P} \subseteq \{[f] \in \mathcal{D} : \|[f]\| = 1\} =: \mathcal{D}_{\rm S}\,. \tag{15}$$

Furthermore, we define $\mathcal{D}^+$ to be the set of equivalence classes $[f]$ in $\mathcal{D}$ induced by members f from $\mathcal{B}^+$, i.e., $\mathcal{D}^+ := \{[f] : f \in \mathcal{B}^+\}$. As $\mathcal{P} \subseteq \mathcal{B}^+$, we know $\mathcal{D}_{\rm P} \subseteq \mathcal{D}^+$. This inclusion relation, together with (15) and Theorem 3 (with $\alpha = 1$), shows that $\mathcal{D}_{\rm P}$ is exactly the intersection of $\mathcal{D}_{\rm S}$ and $\mathcal{D}^+$: $\mathcal{D}_{\rm P} = \mathcal{D}_{\rm S} \cap \mathcal{D}^+$.

Theorem 3 *For each $f \in \mathcal{B}^+$ with $\|[f]\| = [\![f]\!] = \alpha$, there exists a $p \in \mathcal{P}$ such that $[f] = \alpha[p]$, or equivalently, $[\![f - \alpha p]\!] = 0$.*

Proof: If $\alpha = 0$, then any member of $\mathcal{P}$ can be taken as p. So we assume $\alpha > 0$.

For every $f \in \mathcal{B}^+$ define a sequence of functions $(f_n)_{n=0,1,\ldots}$ on O^* as $f_n(\bar{a}) := \sqrt{\sum_{\bar{x}\in O^n} f^2(\bar{a}\bar{x})}$. Then $f_0 = f$ and $f_n \geqslant 0$ for all n. By the definition of $\mathcal{B}^+$ (cf. equation (7)), we know $\sum_{\bar{x}\in O^n} f^2(\bar{a}\bar{x}) \geqslant \sum_{\bar{x}\in O^n, b\in O} f^2(\bar{a}\bar{x}b) = \sum_{\bar{x}\in O^{n+1}} f^2(\bar{a}\bar{x})$, which implies $f_n(\bar{a}) \geqslant f_{n+1}(\bar{a})$. It follows that, for each $\bar{a} \in O^*$, $(f_n(\bar{a}))_{n=0,1,2,\ldots}$ is a decreasing sequence with lower bound 0. Thus, the function

$$f_\infty(\bar{a}) := \lim_{n\to\infty} f_n(\bar{a}) = \lim_{n\to\infty} \sqrt{\textstyle\sum_{\bar{x}\in O^n} f^2(\bar{a}\bar{x})}\,, \qquad (\forall\, \bar{a} \in O^*) \tag{16}$$

is well defined and satisfies $f(\bar{a}) = f_0(\bar{a}) \geqslant f_1(\bar{a}) \geqslant \ldots \geqslant f_\infty(\bar{a}) \geqslant 0$. Moreover,

$$\sum_{b\in O} f_\infty^2(\bar{a}b) = \lim_{n\to\infty} \sum_{b\in O, \bar{x}\in O^n} f^2(\bar{a}b\bar{x}) = \lim_{n\to\infty} f_{n+1}^2(\bar{a}) = f_\infty^2(\bar{a})\,, \tag{17}$$

Let $p = \alpha^{-1} f_\infty$, then it follows from (17) that $\sum_{b\in O} p^2(\bar{a}b) = p^2(\bar{a})$ and from (16) that $p(\epsilon) = \alpha^{-1} \lim_{n\to\infty} \sqrt{\sum_{\bar{x}\in O^n} f^2(\bar{x})} = \alpha^{-1} [\![f]\!] = 1$. Therefore $p \in \mathcal{P}$.

As $f \geqslant f_\infty = \alpha p \geqslant 0$, by (9) we know $Q_n(f, \alpha p) \geqslant Q_n(\alpha p, \alpha p)$ for all n. These inequalities and the definition (10) imply that $(\!|f, \alpha p|\!) \geqslant (\!|\alpha p, \alpha p|\!) = [\![\alpha p]\!]^2 = \alpha^2$. Thus, $[\![f - \alpha p]\!]^2 = [\![f\,]\!]^2 - 2(\!|f, \alpha p|\!) + \alpha^2\, [\![\, p]\!]^2 = 2\alpha^2 - 2(\!|f, \alpha p|\!) \leqslant 0$. It then follows that $[\![f - \alpha p]\!] = 0$, i.e., $[f] = \alpha[p]$. □

Secondly, by Theorem 2 we know $\mathcal{B}^+$ is a convex cone in the space $\mathcal{B}$. A natural question arising here is whether the subset $\mathcal{D}^+ = \{[f] : f \in \mathcal{B}^+\}$ is also a convex cone in the space $\mathcal{D} = \{[f] : f \in \mathcal{B}\}$. The answer is yes, as stated in the following theorem.

Theorem 4 (i) *Let $f, g \in \mathcal{B}^+$ be such that $[\![f + g]\!] = 0$, then $[\![f]\!] = [\![g]\!] = 0$;* (ii) *$\mathcal{D}^+$ is a convex cone in $\mathcal{D}$ pointed at $[0]$.*

Proof: Since $f, g \in \mathcal{B}^+$, we know $f, g \geqslant 0$ and so $[\![f + g]\!] \geqslant [\![f]\!] \geqslant 0$. But $[\![f + g]\!] = 0$, thus $[\![f]\!] = 0$. By the same reason, $[\![g]\!] = 0$. This proves (i).

Let $[f], [g] \in \mathcal{D}^+$. By the definition of $\mathcal{D}^+$, there are $f', g' \in \mathcal{B}^+$ such that $[f] = [f']$ and $[g] = [g']$. As $\mathcal{B}^+$ is a convex cone, we have $f' + g' \in \mathcal{B}^+$ and $\alpha f' \in \mathcal{B}^+$ 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 $f, g \in \mathcal{B}^+$ satisfying $[f] = [h]$ and $[g] = [-h]$, i.e., $[\![f - h]\!] = [\![g + h]\!] = 0$. So $0 \leqslant [\![f + g]\!] \leqslant [\![f - h\,]\!] + [\![\,g + h]\!] = 0$. By (i) we know $[\![f]\!] = 0$, which means $[h] = [f] = [0]$. Therefore, $\mathcal{D}^+$ is pointed at $[0]$. $\square$

Finally, each probability amplitude $p(\bar{a}) \in \mathcal{P}$ (and hence each stochastic process $P(\bar{a}) = p^2(\bar{a})$) has a representation $[p]$ in the set $\mathcal{D}_{\rm P}$. This actually can be seen as a map that sends each process $P(\bar{a})$ to a member $[\sqrt{P}]$ of $\mathcal{D}_{\rm P}$. To establish the reverse map (from $\mathcal{D}_{\rm P}$ to stochastic processes), we need

Theorem 5 *For any $f, g \in \mathcal{P}$, if $[f] = [g]$, i.e., if $[\![f - g]\!] = 0$, then $f = g$.*

Proof: Since $f, g \in \mathcal{P}$, we have $[\![f]\!] = [\![g]\!] = 1$ and it follows from $0 = [\![f - g]\!]^2 = [\![f\,]\!]^2 + [\![\,g\,]\!]^2 - 2(\!|f, g|\!)$ that $(\!|f, g|\!) = 1$. For $f, g \in \mathcal{P} \subseteq \mathcal{B}^+$, $(Q_n(f, g))_{n=0,1,2,\ldots}$ (see (9) for the definition of Q_n) is a decreasing sequence with $Q_0(f, g) = f(\epsilon)g(\epsilon) = 1$ and $\lim_{n\to\infty} Q_n(f, g) = (\!|f, g|\!) = 1$. So $Q_n(f, g) = \sum_{\bar{a}\in O^n} f(\bar{a})g(\bar{a}) = 1$ for all n. But $f, g \in \mathcal{P}$ implies that $\sum_{\bar{a}\in O^n} f^2(\bar{a}) = \sum_{\bar{a}\in O^n} g^2(\bar{a}) = 1$. Thus,

$$\sum_{\bar{a}\in O^n} [f(\bar{a}) - g(\bar{a})]^2 = \sum_{\bar{a}\in O^n} f^2(\bar{a}) - 2\sum_{\bar{a}\in O^n} f(\bar{a})g(\bar{a}) + \sum_{\bar{a}\in O^n} g^2(\bar{a}) = 0\,,$$

which means $f(\bar{a}) = g(\bar{a})$ for all $\bar{a} \in O^n$ $(n = 0, 1, 2, \ldots)$ and therefore $f = g$. $\square$

Now for each $[f] \in \mathcal{D}_{\rm P}$, by the definition of $\mathcal{D}_{\rm P}$ and Theorem 5, there is a unique $p_f \in \mathcal{P}$ such that $[p_f] = [f]$. We therefore find a (unique) process $P(\bar{a}) = p_f^2(\bar{a})$ that is decribed by $[f]$.

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

- The family $\mathcal{P}$ is isomorphic (one-to-one corresponding) to $\mathcal{D}_{\rm P}$ via the map $[\cdot]$; and both $\mathcal{P}$ and $\mathcal{D}_{\rm P}$ can be identified with the family of stochastic processes.
- $\mathcal{D}_{\rm P}$ is the intersection of the unit sphere $\mathcal{D}_{\rm S}$ 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 $\mathcal{D}_{\rm S}$ and in the "positive orthant" $\mathcal{D}^+$.

3.2 Constructing NOOMs in the space $\mathcal{D}$

Since all left-appending operators L_a leave the subspace $\mathcal{B}$ invariant (cf. Theorem 2), we can restrict the operation of L_a's on the space $\mathcal{B}$. Moreover, from $[\![L_a f]\!] =$

$\lim_n \sqrt{\sum_{\bar{x}\in O^n} f^2(a\bar{x})} \leqslant \lim_n \sqrt{\sum_{\bar{x}\in O^{n+1}} f^2(\bar{x})} = [\![f]\!]$ we know L_a also leaves the subspace $\mathcal{N}$ invariant. Thus, as is well known, it induces naturally a linear operator $[L_a]$ on the quotient space $\mathcal{D} = \mathcal{B}/\mathcal{N}$, via

$$[L_a][f] := [L_a f]\,, \qquad (\forall\, f \in \mathcal{B})\,. \tag{18}$$

We see now an important property of the linear operators $[L_a]$, which ensures that a standard NOOM can be constructed for any stochastic process.

Theorem 6 *For any $f, g \in \mathcal{B}$, it holds that $\sum_{a\in O} (\!|L_a f, L_a g|\!) = (\!|f, g|\!)$; or, equivalently, $\sum_{a\in O} \langle [L_a][f], [L_a][g]\rangle = \langle [f], [g]\rangle$ for any $[f], [g] \in \mathcal{D}$.*

Proof: Direct computation shows that

$$\sum_{a\in O} (\!|L_a f, L_a g|\!) = \sum_{a\in O} \lim_{n\to\infty} \sum_{\bar{x}\in O^n} f(a\bar{x}) g(a\bar{x}) = \lim_{n\to\infty} \sum_{\bar{x}\in O^{n+1}} f(\bar{x}) g(\bar{x}) = (\!|f, g|\!)\,. \;\square$$

Iteratively using definition (18), we get, for any $\bar{a} = a_1 a_2 \ldots a_n$ and $f \in \mathcal{B}$,

$$[L_{\bar{a}} f] = [L_{a_n} \cdots L_{a_1} f] = [L_{a_n}][L_{a_{n-1}} \cdots L_{a_1} f] = \ldots = [L_{a_n}] \cdots [L_{a_1}][f]\,. \tag{19}$$

Writing $[L]_{\bar{a}}$ for the composition $[L_{a_n}] \cdots [L_{a_1}]$, then (19) can be shortly written as $[L_{\bar{a}} f] = [L]_{\bar{a}}[f]$. It should be noted that trivially $[L_a] = [L]_a$ for all $a \in O$, and furthermore $[L_{\bar{a}}] = [L]_{\bar{a}}$, since $[L_{\bar{a}}][f] = [L_{\bar{a}} f] = [L]_{\bar{a}}[f]$ for all $[f] \in \mathcal{D}$.

Theorem 7 *For any $p \in \mathcal{P}$ and $\bar{a} \in O^*$, $p(\bar{a}) = \|[L_{\bar{a}}][p]\| = \|[L]_{\bar{a}}[p]\|$.*

Proof: By (6) and induction on n, we obtain, for all $n \in \mathbb{N}$,

$$p^2(\bar{a}) = \textstyle\sum_{\bar{x}\in O^n} p^2(\bar{a}\bar{x}) = \sum_{\bar{x}\in O^n} [(L_{\bar{a}} p)(\bar{x})]^2\,.$$

Letting $n \to \infty$ in the above equality, we get $p^2(\bar{a}) = [\![L_{\bar{a}} p]\!]^2 = \|[L_{\bar{a}} p]\|^2$. It follows from (19) that $p(\bar{a}) = \|[L_{\bar{a}} p]\| = \|[L]_{\bar{a}}[p]\|$. $\square$

Theorem 7 gives rise to a "universal" system $(\mathcal{D}, \{[L_a]\}_{a\in O})$ for representing any probability amplitude $p(\bar{a})$, which plays the same role for NOOMs as the previously introduced system $(\mathcal{F}, \{L_a\}_{a\in O}, \sigma)$ (see page 7) plays for OOMs. Starting from $(\mathcal{D}, \{[L_a]\}_{a\in O})$ we can construct abstract NOOMs for any process $P(\bar{a}) = p^2(\bar{a})$, by a procedure similar to the one presented in Subsection 2.1 for deriving abstract OOMs for a given LDF from the system $(\mathcal{F}, \{L_a\}_{a\in O}, \sigma)$.

For any $p \in \mathcal{P}$, define $\mathcal{D}^p$ to be the subspace of $\mathcal{D}$ spanned by the vectors $[L]_{\bar{a}}[p]$ ($\bar{a} \in O^*$), i.e., $\mathcal{D}^p := \mathrm{span}\{[L]_{\bar{a}}[p] \,:\, \bar{a} \in O^*\}$. It is clear that the operators $[L_a]$ all leave the space $\mathcal{D}^p$ invariant and hence can be seen as operators on $\mathcal{D}^p$. We thus get an abstract NOOM $(\mathcal{D}^p, \{[L_a]\}_{a\in O}, [p])$ of the process $P(\bar{a}) = p^2(\bar{a})$, which, according to Theorem 7, computes the values of $P(\bar{a})$ via $P(\bar{a}) = \|[L]_{\bar{a}}[p]\|^2$.

Thus far we have almost finished the proof to the first part of Theorem 1: the system $(\mathcal{D}^p, \{[L]_a\}_{a\in O}, [p])$ constructed above fulfills all the conditions asserted in

the theorem, provided that we can define the adjoint operator $[L]_a^*$ of $[L]_a$ — so one gets $\sum_{a\in O}[L]_a^*[L]_a = \mathrm{id}_{\mathcal{D}^p}$ from Theorem 6. This amounts, in general, to proving that $\mathcal{D}^p$ is complete, which is a nonissue for the finite-dimensional case. For the general case where $\mathcal{D}^p$ is of infinite dimension, we currently do not know the proof or disproof of the completeness of $\mathcal{D}^p$. However, for the special case here, we can directly define an operator L_a^* on $\mathcal{F}$ (for each $a \in O$), which naturally induces the adjoint operator $[L]_a^*$ on the space $\mathcal{D}$, as follows:

$$(L_a^* f)(\epsilon) = f(\epsilon)\,; \quad (L_a^* f)(a_1 a_2 \dots a_n) = \begin{cases} 0 & \text{if } a_1 \neq a \\ f(a_2 \dots a_n) & \text{if } a_1 = a \end{cases} .$$

One should have no difficulty to see that (1) all L_a^* are linear operators on $\mathcal{F}$; (2) all L_a^* leave the set $\mathcal{B}^+$, and hence the subspace $\mathcal{B}$, invariant; (3) $[\![L_a^* f]\!] \leqslant [\![f]\!]$ for all $f \in \mathcal{B}$ and so L_a^* also leaves the space $\mathcal{N}$ invariant; (4) $Q_n(f, L_a g) = Q_{n+1}(L_a^* f, g)$ for all $n \in \mathbb{N}$ and $f, g \in \mathcal{B}$; so the induced operator $[L]_a^*$ on $\mathcal{D}$ satisfies

$$\langle [L]_a^*[f], [g]\rangle = (\!|L_a^* f, g|\!) = (\!|f, L_a g|\!) = \langle [f], [L]_a[g]\rangle\,, \qquad (\forall\, [f], [g] \in \mathcal{D})\,. \tag{20}$$

Equation (20) indicates that $[L]_a^*$ is the adjoint operator of $[L]_a$ in the vector space $\mathcal{D}$, whereas what we want is the adjoint of $[L]_a$ in the space $\mathcal{D}^p$ ($\mathcal{D}^p$ is not necessary invariant under the operation of $[L]_a^*$). To this end, we need further to (orthogonally) project the images of $[L]_a^*$ onto the space $\mathcal{D}^p$, yielding the composed operator $[L^{\mathrm{prj}}]_a^* := \mathrm{prj}_{\mathcal{D}^p} \circ [L]_a^*$, where $\mathrm{prj}_{\mathcal{D}^p}$ is the projection operator from $\mathcal{D}$ onto $\mathcal{D}^p$. It is clear that $[L^{\mathrm{prj}}]_a^*$ leaves the space $\mathcal{D}^p$ invariant and that for any $[f] \in \mathcal{D}$, $[L]_a^*[f] - [L^{\mathrm{prj}}]_a^*[f]$ is orthogonal to $\mathcal{D}^p$, i.e., $\langle [L]_a^*[f] - [L^{\mathrm{prj}}]_a^*[f], [g]\rangle = 0$ for all $[g] \in \mathcal{D}^p$. Thus,

$$\langle [L^{\mathrm{prj}}]_a^*[f], [g]\rangle = \langle [L]_a^*[f], [g]\rangle = \langle [f], [L]_a[g]\rangle\,, \qquad (\forall\, [f], [g] \in \mathcal{D}^p)$$

which means $[L^{\mathrm{prj}}]_a^*$ is the adjoint of $[L]_a$ in the space $\mathcal{D}^p$ and completes the proof of Theorem 1.

We now consider finite-dimensional NOOMs and their matrix representations. Assume the space $\mathcal{D}^p$ is of finite dimension m and select an orthonormal basis of $\mathcal{D}^p$, i.e., a basis $\{[g_1], [g_2], \dots, [g_m]\}$ with the property $\langle [g_i], [g_j]\rangle = \delta_{ij}$, where δ_{ij} is the Kronecker symbol defined as $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. As is well known, each $[f] \in \mathcal{D}^p$ can be uniquely represented as a linear combination of $\{[g_1], [g_2], \dots, [g_m]\}$ (with coefficients $\alpha_i(f) = \langle [f], [g_i]\rangle$):

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

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

$$\langle [f], [h]\rangle = \sum_{i,j} \alpha_i(f)\alpha_j(h)\langle [g_i], [g_j]\rangle = \sum_{i,j} \alpha_i(f)\alpha_j(h)\delta_{ij} = \sum_{i=1}^m \alpha_i(f)\alpha_i(h)\,, \tag{21}$$

i.e., $\langle [f],[h]\rangle = \{\pi[f]\}^{\mathsf{T}}\{\pi[h]\} =: \langle \pi[f],\pi[h]\rangle$. It follows from Theorem 6 that $\sum_{a\in O}\langle \pi[L_a][f],\pi[L_a][h]\rangle = \langle \pi[f],\pi[h]\rangle$ for any $[f],[h]\in\mathcal{D}^p$. 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 $\boldsymbol{u}_0 := \pi[p]\in\mathbb{R}^m$ the initial state. Then, since $p\in\mathcal{P}$, $\boldsymbol{u}_0^{\mathsf{T}}\boldsymbol{u}_0 = \{\pi[p]\}^{\mathsf{T}}\{\pi[p]\} = \langle [p],[p]\rangle = 1$. Furthermore,

$$\begin{aligned}\boldsymbol{e}_i^{\mathsf{T}}\big(\textstyle\sum_{a\in O}\varphi_a^{\mathsf{T}}\varphi_a\big)\boldsymbol{e}_j &= \textstyle\sum_{a\in O}\{\pi[L_a][g_i]\}^{\mathsf{T}}\{\pi[L_a][g_j]\} \quad \text{(by the definetion of } \varphi_a\text{)}\\ &= \textstyle\sum_{a\in O}\langle [L_a][g_i],[L_a][g_j]\rangle\\ &= \langle [g_i],[g_j]\rangle \quad \text{(by Theorem 6)}\\ &= \delta_{ij}\,,\end{aligned}$$

where $\boldsymbol{e}_i$ denotes the i-th unit vector in $\mathbb{R}^m$. We thus conclude $\sum_{a\in O}\varphi_a^{\mathsf{T}}\varphi_a = I_m$. Furthermore, for any $\bar{a} = a_1a_2\ldots a_n\in O^*$ it holds that

$$\pi[L]_{\bar{a}}[p] = (\pi[L_{a_n}]\pi^{-1})\cdots(\pi[L_{a_1}]\pi^{-1})(\pi[p]) = \varphi_{a_n}\cdots\varphi_{a_1}\boldsymbol{u}_0 = \varphi_{\bar{a}}\boldsymbol{u}_0\,.$$

This identity, together with Theorem 7, implies that

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

We thus have already constructed a NOOM $(\mathbb{R}^m,\{\varphi_a\}_{a\in O},\boldsymbol{u}_0)$ for the process $P(\bar{a})$ that is standard: it satisfies $\|\boldsymbol{u}_0\| = 1$ and $\sum_{a\in O}\varphi_a^{\mathsf{T}}\varphi_a = I_m$.

Clearly, the abstract system $(\mathcal{D}^p,\{[L_a]\}_{a\in O},[p])$ has (infinitely) many matrix representations (under different orhtonormal bases of $\mathcal{D}^p$). They are all standard concrete NOOMs that represent exactly the same process and are related to each other by a unitary matrix (the basis transition matrix). We shall utilize this fact to derive a constructive algorithm for learning NOOMs in Section 6.

4 NOOMs as Generators and Predictors

After having established the mathematical foundation for a theory of NOOMs, we explain in this section how to generate and predict the paths of a process (X_t) modelled by a standard, finite-dimensional NOOM $(\mathbb{R}^m,\{\varphi_a\}_{a\in O},\boldsymbol{u}_0)$ via $P(\bar{a}) = \|\varphi_{\bar{a}}\boldsymbol{u}_0\|^2$; and at the same time introduce some notations for later use.

The generation task requires one to randomly produce, at time steps $t = 1,2,\ldots$, outcomes $a_1,a_2,\ldots\in O$, such that (i) at time $t=1$, the probability that the symbol b is emitted is $P(b) = \|\varphi_b\boldsymbol{u}_0\|^2$, and (ii) at each time $t = n+1$ $(n = 1,2,\ldots)$, the probability of producing b (assuming that $\bar{a} := a_1a_2\ldots a_n$ has already been created) is

$$P(b|\bar{a}) := \frac{P(\bar{a}b)}{P(\bar{a})} = \frac{\|\varphi_b\varphi_{\bar{a}}\boldsymbol{u}_0\|^2}{\|\varphi_{\bar{a}}\boldsymbol{u}_0\|^2} =: \|\varphi_b\boldsymbol{u}_{\bar{a}}\|^2\,, \tag{22}$$

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

$$\boldsymbol{u}_\epsilon = \boldsymbol{u}_0\,,\quad \boldsymbol{u}_{a_1a_2\ldots a_n} = \frac{\varphi_{a_n}\boldsymbol{u}_{a_1a_2\ldots a_{n-1}}}{\|\varphi_{a_n}\boldsymbol{u}_{a_1a_2\ldots a_{n-1}}\|}\,. \tag{23}$$

Therefore, the procedure for generating sample paths of a process from its NOOM $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$ can be outlined as follows: at each time $t = 1, 2, \ldots$, generate a_t according to the probability distribution $P(b) = \|\varphi_b \boldsymbol{u}_{t-1}\|^2$ and compute the state vector $\boldsymbol{u}_t = \varphi_{a_t}\boldsymbol{u}_{t-1}/\|\varphi_{a_t}\boldsymbol{u}_{t-1}\|$.

NOOMs can also be used as predictors: given an initial path $\bar{a} = a_1 a_2 \ldots a_n$ of the process up to time $t = n$, we want to predict the probability that the next outcome is b. This amounts to the computation of the conditional probability $P(b|\bar{a})$; and equations (22)(23) can be employed here. But note that now the initial path $\bar{a}$ is not generated by the NOOM itself but is externally given.

In the next section we will show that any NOOM can be converted to an equivalent OOM. So one can also first convert a given NOOM to its equivalent OOM; and then use this OOM as the generator/predictor to create/predict the next outcome b. See Section 3 of Jaeger (2000) for a detailed explanation.

Equations (22)(23) also provide a way for evaluating the probabilities $P(\bar{a})$. Note 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 the length n of $\bar{a}$ is on the average exponentially fast, which would run us into numerical underflow problems. So instead of directly calculating $P(\bar{a})$, one should evaluate the log-likelihood $\mathrm{LL}(\bar{a}) := \log P(\bar{a})$. This can be done as follows: (1) for $t = 1, 2, \ldots, n$ compute $\boldsymbol{x}_t = \varphi_{a_t}\boldsymbol{u}_{t-1}$, $c_t = \|\boldsymbol{x}_t\|$ and $\boldsymbol{u}_t = c_t^{-1}\boldsymbol{x}_t$; (2) calculate $\mathrm{LL}(\bar{a}) = 2\sum_{t=1}^{n}\log c_t$.

5 On the Expressiveness of NOOMs

The previous section proved that any stochastic process admits a possibly infinite-dimensional NOOM. We now consider the class of processes that can be modelled by finite-dimensional NOOMs, which is of more practical interest in the machine learning context. We also present in this section a general procedure for creating (all) finite-dimensional NOOMs.

We first present a simple example (the NOOM version of the probability clock, see Section 6 of Jaeger (2000)) to illustrate that NOOMs, like OOMs, can describe some processes that cannot be modelled by HMMs. This is a 2-dimensional NOOM over the alphabet $O = \{a, b\}$, with the observable operators and initial state defined by

$$\varphi_a = \begin{bmatrix} .6c & -s \\ .6s & c \end{bmatrix}, \quad \varphi_b = \begin{bmatrix} .8 & 0 \\ 0 & 0 \end{bmatrix}; \quad \text{and} \quad \boldsymbol{u}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{24}$$

where $c = \cos(0.6)$ and $s = \sin(0.6)$. It follows from (23) and (24) that

$$\boldsymbol{u}_{\bar{x}b} = \varphi_b\boldsymbol{u}_{\bar{x}}/\|\varphi_b\boldsymbol{u}_{\bar{x}}\| = [\pm 1, 0]^{\mathsf{T}} = \pm\boldsymbol{u}_0\,, \qquad (\forall\,\bar{x} \in O^*)\,.$$

Note that, any $\bar{a} \in O^*$ is either $\bar{a} = aa\ldots a =: a^n$ (if b does not occur in $\bar{a}$) or $\bar{a} = \bar{x}ba^n$ (if b occurs at least once in $\bar{a}$). For the latter case, we compute

$$P(b|\bar{x}ba^n) = \frac{P(a^nb|\bar{x}b)}{P(a^n|\bar{x}b)} = \frac{\|\varphi_{a^nb}\boldsymbol{u}_{\bar{x}b}\|^2}{\|\varphi_{a^n}\boldsymbol{u}_{\bar{x}b}\|^2} = \frac{\|\varphi_{a^nb}\boldsymbol{u}_0\|^2}{\|\varphi_{a^n}\boldsymbol{u}_0\|^2} = \frac{P(a^nb)}{P(a^n)} = P(b|a^n)\,;$$

which is same as in the first case: $\bar{a} = a^n$. As $P(a|\bar{a}) + P(b|\bar{a}) = 1$ for all $\bar{a} \in O^*$, the process modelled by the NOOM (24) is completely described by the family of the conditional probabilities $\{P(b|a^n)\}_{n=0,1,\ldots}$. Iteratively using equations (22)(23), we computed the values of $P(b|a^n)$, and plotted them in Fig. 1 versus n. One observes that the behavior of the NOOM given by equation (24) is very similar to that of the probability clock. As explained in Jaeger (2000), such non-rational-periodic behavior of $P(b|a^n)$ cannot be captured by (finite) HMMs.

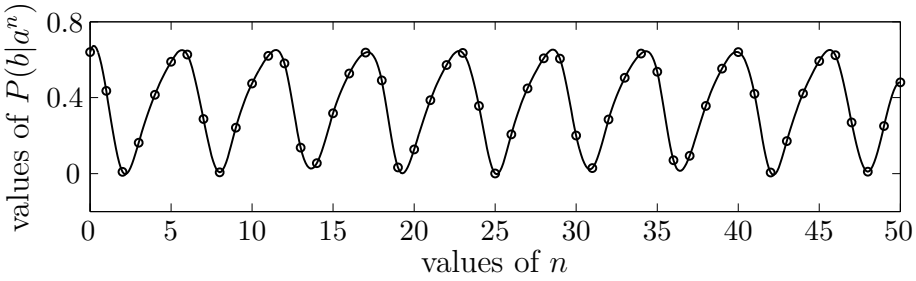

Figure 1: The conditional probabilities $P(b|a^n)$.

Although NOOMs are capable of capturing some non-HMM processes, they provide no more than OOMs, because each NOOM can be equivalently converted to an OOM, as shown next.

Definition 2 *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* $\mathrm{vec}(A)$ *for the mn-dimensional column vector formed by stacking the columns of A one below another.*

We mention two well-known properties of the Kronecker product $\otimes$ and the stacking operator $\mathrm{vec}(\cdot)$. See, e.g., Hom and Johnson (1989); Brewer (1978) for more detail.

Theorem 8 *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 NOOM of the process $P(\bar{a})$. Then by Theorem 8 the probabilities $P(\bar{a})$ for $\bar{a} = a_1 a_2 \ldots a_n$ can be calculated as

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

Writing $\boldsymbol{\sigma} = [\mathrm{vec}(I_m)]^\mathsf{T}$, $\tau_a = \varphi_a \otimes \varphi_a$ and $\boldsymbol{w}_0 = \boldsymbol{u}_0 \otimes \boldsymbol{u}_0$, we get a nonstandard OOM $(\mathbb{R}^{m^2}, \{\tau_a\}_{a\in O}, \boldsymbol{w}_0, \boldsymbol{\sigma})$ for the process $P(\bar{a})$, which can be easily converted to

an equivalent standard OOM $(\mathbb{R}^{m^2}, \{\varrho\tau_a\varrho^{-1}\}_{a\in O}, \varrho\boldsymbol{w}_0)$, via a basis transition matrix $\varrho \in \mathbb{R}^{m^2\times m^2}$ satisfying $\mathbf{1}_{m^2}^{\mathsf{T}}\varrho = \boldsymbol{\sigma}$, e.g., $\varrho = I_{m^2} + \frac{1}{m^2}\mathbf{1}_{m^2}(\boldsymbol{\sigma} - \mathbf{1}_{m^2}^{\mathsf{T}})$; which can be further converted to a minimal-dimensional OOM of the same process $P(\bar{a})$, see Jaeger et al. (2005) for the procedure in detail.

For instance, the NOOM defined by (24) is equivalent to a 3-dimensional OOM with initial state $\boldsymbol{w}_0 = [0.933, -0.170, 0.237]^{\mathsf{T}}$ and observable operators

$$\tau_a = 0.6\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\theta & -\sin\theta\\ 0 & \sin\theta & \cos\theta\end{bmatrix}, \quad \tau_b = 0.4\boldsymbol{w}_0\begin{bmatrix}1 & 0.483 & 3.157\end{bmatrix};$$

where $\theta = 1.1005$. One sees that the above OOM has the same structure as the probability clock given in Jaeger (2000) (cf. equation (6.1) therein). This proves the NOOM (24) indeed represents a probability clock and cannot be captured by any HMMs.

The fact that every NOOM has an equivalent OOM reveals the class of NOOMs as a subclass of OOMs. Now we consider the reverse problem: *which OOMs have equivalent NOOMs*? So far only little is known concerning this question, except the following sufficient condition.

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

Proof: Assume $[\tau_a]_{ik}$ is the only (possibly) nonzero element in the i-th row of τ_a. Then for any $\boldsymbol{x} = [x_1, x_2, \ldots, x_m]^{\mathsf{T}} \in \mathbb{R}^m$ 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 induction 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. □

Any Markov chain (MC) of m states can be equivalently represented as an m-dimensional OOM with each operator τ_a consisting of zero columns except the a-th column, which is equal to the corresponding column of the transition matrix of the MC. So by Theorem 9 we know any m-state Markov chain can be converted to an equivalent m-dimensional NOOM.

The currently known relationships between stochastic processes that can be captured by (finite) MCs, HMMs, NOOMs and OOMs can be summarized as follows:

$$\text{NOOMs} \not\subseteq \text{HMMs}, \quad \text{MCs} \subset \text{HMMs} \subset \text{OOMs}, \quad \text{MCs} \subset \text{NOOMs} \subseteq \text{OOMs}.$$

Thus far it remains unclear whether NOOMs also (like OOMs) contains HMMs as a subclass (i.e., HMMs $\subseteq$ NOOMs).

We now present a random construction of NOOMs matrices that is guaranteed to yield all NOOMs. According to Definition 1, to create a NOOM $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$, one needs only to make sure that $\sum_{a\in O} \varphi_a^{\mathsf{T}}\varphi_a = I_m$ and $\|\boldsymbol{u}_0\| = 1$. Let φ be the $m\ell$ by m matrix created by stacking the matrices φ_a below one another, i.e., $\varphi := [\varphi_1^{\mathsf{T}}, \varphi_2^{\mathsf{T}}, \ldots, \varphi_\ell^{\mathsf{T}}]^{\mathsf{T}}$. Then the condition $\sum_{a\in O} \varphi_a^{\mathsf{T}}\varphi_a = I_m$ can be rewritten as $\varphi^{\mathsf{T}}\varphi = I_m$, which means the columns of φ form an orthonormal set in $\mathbb{R}^{m\ell}$. In sum, we get the general procedure for constructing arbitrary NOOMs of dimension m as described in Algorithm 1.

Algorithm 1: A general procedure for creating random NOOMs

Given: the model dimension m
Want: a standard NOOM $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$
Procedure:

1. Randomly create m vectors in $\mathbb{R}^{m\ell}$.

2. Make them an orthonormal set by using the Gram-Schmidt procedure. This gives us an $m\ell \times m$ matrix φ with the property $\varphi^{\mathsf{T}}\varphi = I_m$.

3. Divide the φ into ℓ blocks φ_a $(a \in O)$ of equal size. These are the observable operators of the desired NOOM.

4. Finally, the initial state $\boldsymbol{u}_0$ can be any vector in $\mathbb{R}^m$ with norm 1.

One therefore obtains an efficient way to construct OOMs (which usually have no HMM equivalents): first create a random NOOM (by Algorithm 1) and then convert it to an equivalent OOM.

Thus far we have seen two methods to create concrete NOOMs:

Method-A: either from a process via the routine presented in Section 3;

Method-B: or from scratch by employing Algorithm 1.

We would however point out that the two methods do not procure the same set of NOOMs: Method-B is much more "productive" than Method-A. In fact, any concrete NOOM $(\mathbb{R}^m, \{\tau_a\}_{a\in O}, \boldsymbol{u}_0)$ created by Method-A is a matrix representation of some abstract NOOM $(\mathcal{D}^p, \{[L_a]\}_{a\in O}, [p])$ under some orthonormal basis of the space $\mathcal{D}^p$, in which it holds that $\langle [f], [h]\rangle = \{\pi[f]\}^{\mathsf{T}}\{\pi[h]\}$ for any $[f], [h] \in \mathcal{D}^p$ (see equation (21)). In particular, for $\bar{a}, \bar{b} \in O^*$, we have

$$(\varphi_{\bar{a}}\boldsymbol{u}_0)^{\mathsf{T}}(\varphi_{\bar{b}}\boldsymbol{u}_0) = \langle [L]_{\bar{a}}[p], [L]_{\bar{b}}[p]\rangle = (\!|L_{\bar{a}}p, L_{\bar{b}}p|\!) \geqslant 0\,, \tag{25}$$

since $L_{\bar{a}}p(\bar{x}) = p(\bar{a}\bar{x}) \geqslant 0$ and $L_{\bar{b}}p(\bar{x}) = p(\bar{b}\bar{x}) \geqslant 0$ for all $\bar{x} \in O^*$. Note that, equation (25) by no means indicates that the NOOM theory also suffers from the NPP; it just says that any NOOM created by Method-A has this special property, which is not necessarily true for NOOMs generated by Algorithm 1, as one can arbitrarily and independently change the sign of the observable operators and the initial state of a NOOM (so (25) is violated) but still get the same process.

We conclude this section with three comments:

- What we have presented in Section 3 is not the only way to construct NOOMs from stochastic processes. There are many other possibilities to consider, e.g., one may define a different semidefinite inner product on the space $\mathcal{B}$; or embed the family $\mathcal{P}$ of probability amplitudes into another subspace of $\mathcal{F}$ rather than the space $\mathcal{B}$ as we do in this paper.

- Consequently, $\dim \mathcal{D}^p = \infty$ does not imply that the associated process $P(\bar{a})$ admits no finite-dimensional NOOM.

- The randomized construction of NOOMs fills a gap in OOM research. In the past, it was difficult to created non-HMM OOMs. In fact, the only (and widely used) example appears to be the "probability clock" (Jaeger, 2000). According to our experience, if a NOOM is randomly created and then transformed to an OOM, the resulting OOM will typically be non-HMM. In this way, OOM research will benefit from a richer zoo of examples, always a booster for research.

6 Toward a Constructive Learning Algorithm for NOOMs

The first author has developed an iterative algorithm for learning NOOMs, on which a separate paper is being prepared. While revising the current paper, the principles of a constructive learning algorithm of NOOMs were found, which will be outlined in this section.

This constructive algorithm is based on (25). In more detail, as the probability $\hat{P}(\bar{a})$ of a given string $\bar{a}$ can be asymptotically correctly estimated by simply counting the number of its occurrences in the training dataset, in principle we are able to estimate the value of the semidefinite inner product $(\!|L_{\bar{a}}p, L_{\bar{b}}p|\!)$ from its definition (see equation (10)):

$$(\!|L_{\bar{a}}p, L_{\bar{b}}p|\!) \approx \sum_{\bar{x}\in O^N} \sqrt{\hat{P}(\bar{a}\bar{x})} \cdot \sqrt{\hat{P}(\bar{b}\bar{x})}\,. \quad (N \text{ sufficiently large}) \tag{26}$$

Equation (25) then allows us to estimate the system states of the form $\varphi_{\bar{a}}\boldsymbol{u}_0$ (up to a unitary transform, as we will prove soon), from which we compute the observable operator estimations $\hat{\varphi}_a$ by solving a set of linear equations.

We now describe the basic procedure of the NOOM learning algorithm. For simplicity we would assume the training dataset is large enough so that one can estimate the probability of any string $\bar{a} \in O^*$ with sufficient accuracy; or, there is an oracle who would, whenever we asked, tell us the true probability $P(\bar{a})$. We also assume the model dimension m is already known.

We first select a subset $\{\bar{c}_1, \ldots, \bar{c}_m\}$ of m finite strings in O^*; and estimate the value of $(\!|L_{\bar{a}}p, L_{\bar{b}}p|\!)$ by (26), where $\bar{a}$ and $\bar{b}$ run over the strings $\bar{c}_i$ and $\bar{c}_i a$ with $i = 1, 2, \ldots, m$ and $a \in O$. These estimated quantities are then collected in a

symmetric square matrix Q of order $(1+\ell)m$ (recall that $O = \{a^1, a^2, \cdots, a^\ell\}$), which can be naturally divided into $(1+\ell)\times(1+\ell)$ blocks of equal size $m \times m$:

$$Q = [Q_{ab}]_{a,b\in\{\epsilon\}\cup O}\,, \quad [Q_{ab}]_{ij} = (\!|L_{\bar{c}_i a}p, L_{\bar{c}_j b}p|\!)\,, \quad (i,j = 1,2,\ldots,m)\,. \tag{27}$$

As the matrix Q defined as above plays a central role in the learning algorithm, we will give it a special name: the *kernel matrix*.

In the following we would fix $\bar{c}_1 = \epsilon$ (to simplify the estimating of the initial state, see below) and assume that the choice of $\{\bar{c}_1, \ldots, \bar{c}_m\}$ makes the matrix $Q_{\epsilon\epsilon}$ nonsingular.

By (25), it is easy to see that $Q_{ab} = U_a^{\mathsf{T}} U_b$ $(a, b \in \{\epsilon\}\cup O)$, where the U_a's are $m \times m$ matrices defined by

$$U_a = [\varphi_{\bar{c}_1 a}\boldsymbol{u}_0, \ldots, \varphi_{\bar{c}_m a}\boldsymbol{u}_0] = \varphi_a \cdot [\varphi_{\bar{c}_1}\boldsymbol{u}_0, \ldots, \varphi_{\bar{c}_m}\boldsymbol{u}_0]\,. \tag{28}$$

It then follows that $U_a = \varphi_a U_\epsilon$ for $a \in O$. Moreover, the kernel matrix Q has the decompostion $Q = U^{\mathsf{T}}U$ with $U := [U_\epsilon, U_{a^1}, \ldots, U_{a^\ell}]$.

Lemma 3 *Let $A_1, A_2 \in \mathbb{R}^{m\times N}$ $(m \leqslant N)$ be two full rank matrices and assume that $A_1^{\mathsf{T}}A_1 = A_2^{\mathsf{T}}A_2$. Then there is a unitary matrix $\mu \in \mathbb{R}^{m\times m}$ such that $A_1 = \mu A_2$.*

Proof: For $i = 1, 2$, let $A_i = L_i D_i R_i^{\mathsf{T}}$ be the compact SVD of A_i, where $L_i \in \mathbb{R}^{m\times m}$ are unitary, $D_i \in \mathbb{R}^{m\times m}$ diagonal and $R_i \in \mathbb{R}^{N\times m}$. From $A_1^{\mathsf{T}}A_1 = A_2^{\mathsf{T}}A_2$, we see the symmetric matrix $A_1^{\mathsf{T}}A_1$ has eigenvalue decomposition $A_1^{\mathsf{T}}A_1 = R_1 D_1^2 R_1^{\mathsf{T}} = R_2 D_2^2 R_2^{\mathsf{T}}$ (here the eigenvalue 0 and its associated eigenvectors are not presented). It follows from the uniqueness of eigenvalues that $D_1^2 = D_2^2 = \mathrm{diag}\{\sigma_1^2 I_{m_1}, \ldots, \sigma_k^2 I_{m_k}\} =: D^2$, where $\sigma_1 > \ldots > \sigma_k > 0$ and m_j is the multiplicity of the eigenvalue σ_j^2 of $A_1^{\mathsf{T}}A_1$ $(j = 1, \ldots, k)$. Now, for $i = 1, 2$, we have $R_i = [R_i^{(1)}, \ldots, R_i^{(k)}]$, where $R_i^{(j)} \in \mathbb{R}^{N\times m_j}$ has orthonormal columns that form a basis of the eigenspace of $A_1^{\mathsf{T}}A_1$ with respect to the eigenvalue σ_j^2. Since the eigenspace is unique, the two matrices $R_1^{(j)}$ and $R_2^{(j)}$ are connected by some unitary matrix U_j of size $m_j \times m_j$: $R_1^{(j)} = R_2^{(j)}U_j$. Therefore, $R_1 = R_2 \cdot \mathrm{diag}\{U_1, \ldots, U_k\} =: R_2 U$, where $U \in \mathbb{R}^{m\times m}$ is a unitary matrix which, obviously, satisfies $UD = \mathrm{diag}\{\sigma_1 U_1, \ldots, \sigma_k U_k\} = DU$ and hence $DU^{\mathsf{T}} = U^{\mathsf{T}}D$. We thus get $A_1 = L_1 D R_1^{\mathsf{T}} = L_1 D U^{\mathsf{T}} R_2^{\mathsf{T}} = L_1 U^{\mathsf{T}} D R_2^{\mathsf{T}} = L_1 U^{\mathsf{T}} L_2^{\mathsf{T}} L_2 D R_2^{\mathsf{T}} = \mu A_2$, where $\mu := L_1 U^{\mathsf{T}} L_2^{\mathsf{T}} \in \mathbb{R}^{m\times m}$ is a unitary matrix. □

Now let $Q = RDR^{\mathsf{T}}$ be the (compact) SVD of the kernel matrix Q and define $W := D^{\frac{1}{2}}R^{\mathsf{T}}$, then $Q = W^{\mathsf{T}}W$. Were everything in its place, we would have $W \in \mathbb{R}^{m\times(1+\ell)m}$ and $U^{\mathsf{T}}U = W^{\mathsf{T}}W$ — in practice, the matrix W usually contains more than m rows and one should use the first m rows of W to replace the matrix W here. Therefore, by Lemma 3, there exists a unitary matrix $\mu \in \mathbb{R}^{m\times m}$ such that $U = \mu W$. Writing $W = [W_\epsilon, W_{a^1}, \ldots, W_{a^\ell}]$ with each $W_a \in \mathbb{R}^{m\times m}$ $(a \in \{\epsilon\}\cup O)$, we know from $U_a = \varphi_a \cdot U_\epsilon$ (see equation (28)) that $\mu W_a = \varphi_a \mu W_\epsilon$. Therefore,

$$\mu^{\mathsf{T}}\varphi_a\mu = W_a \cdot W_\epsilon^{-1}\,, \qquad (a \in O)\,. \tag{29}$$

Note that, since the matrix $Q_{\epsilon\epsilon}$ is nonsingular (assured by the selection of $\{\bar{c}_1, \ldots, \bar{c}_m\}$), one should have no difficulty to see that U_ϵ, and hence W_ϵ, is invertible. So the above equation is well defined. Moreover, as we have set $\bar{c}_1 = \epsilon$, the first column the matrix U is $U(:,1) = \boldsymbol{u}_0$ (in MatLab's notation). So by the equality $U = \mu W$ we know

$$\mu^{\mathsf{T}} \boldsymbol{u}_0 = W(:,1)\,. \tag{30}$$

The above two equations give rise to a NOOM estimation $(\mathbb{R}^m, \{\mu^{\mathsf{T}}\varphi_a\mu\}_{a\in O}, \mu^{\mathsf{T}}\boldsymbol{u}_0)$ that is obviously equivalent to the "true" model $(\mathbb{R}^m, \{\varphi_a\}_{a\in O}, \boldsymbol{u}_0)$.

In practice, only an estimate $\hat{Q}$ of the kernel matrix is available. Consequently, the model estimated by $\hat{\varphi}_a = \hat{W}_a \hat{W}_\epsilon^{-1}$ and $\hat{\boldsymbol{u}}_0 = \hat{W}(:,1)$ (here the symbols $\hat{W}_a$, $\hat{W}_\epsilon$ and $\hat{W}$ have obvious meaning) is usually not a valid NOOM — the two conditions from Definition 1 might be violated. One possible way to overcome this problem is to calculate a valid NOOM that is closest to the learnt model under some distance measure. For instance, we can compute a new NOOM $(\mathbb{R}^m, \{\varphi_a^*\}_{a\in O}, \boldsymbol{u}_0^*)$ from the estimated model by solving the following optimization problem:

$$\begin{array}{ll} \underset{\varphi_a^*, \boldsymbol{u}_0^*}{\text{minimze}} & \sum_{a\in O} \|\varphi_a^* - \hat{\varphi}_a\|_{\mathrm{F}}^2 + \|\boldsymbol{u}_0^* - \hat{\boldsymbol{u}}_0\|^2\,, \\ \text{subject to} & \sum_{a\in O} (\varphi_a^*)^{\mathsf{T}}(\varphi_a^*) = I_m\,, \quad \|\boldsymbol{u}_0^*\| = 1\,; \end{array} \tag{31}$$

where $\|X\|_{\mathrm{F}} := \sqrt{\mathrm{tr}(X^{\mathsf{T}}X)}$ denotes Frobenius norm of matrices. This problem can be analytically solved as follows. As before, write $\hat{\varphi}$ (resp. φ^*) for the $m\ell$ by m matrix created by stacking the matrices $\hat{\varphi}_a$ (resp. φ_a^*) below one another. Let $\hat{\varphi} = LDR^{\mathsf{T}}$ be the SVD of $\hat{\varphi}_a$, then $\varphi^* = LR^{\mathsf{T}}$ and $\boldsymbol{u}_0^* = \hat{\boldsymbol{u}}_0/\|\hat{\boldsymbol{u}}_0\|$ is the (unique) minimizer of (31). See Appendix B for the proof.

We now briefly illustrate the asymptotical correctness of the presented algorithm. When the size of training dataset and the value of N in (26) both go to infinity, the probability estimation $\{\hat{P}(\bar{a})\}_{\bar{a}\in O^*}$, and hence the kernel matrix $\hat{Q}$ estimated by (26) and (27), tend to their true value. This implies that the matrix $\hat{W}$ defined above is asymptotically correct, so is the NOOM estimation calculated by (29)(30).

We should however point out that, while being mathematically clear and simple, the above basic algorithm for learning NOOMs is still far from a practical method. There are technical problems that have to be solved before it can be utilized in practice. Especially, the estimation of the kernel matrix Q by equations (26) and (27) is a challenge: the summation in (26) contains ℓ^N items and would become computationally prohibitive even for modest values of N. To work out an efficient way to estimate the kernel matrix Q is therefore a main goal for future work, for which two (heuristic) methods are currently being investigated by the first author.

- To use variable-length strings $\bar{x}$ in the summation (26): the idea behind is that if for some $\bar{x}$ the product $\sqrt{\hat{P}(\bar{a}\bar{x})} \cdot \sqrt{\hat{P}(\bar{b}\bar{x})} =: \Pi_{\bar{a},\bar{b}}(\bar{x})$ is already very small, we need not to compute further the values of $\Pi_{\bar{a},\bar{b}}(\bar{x}a)$ $(a \in O)$ since in this case it holds that $\Pi_{\bar{a},\bar{b}}(\bar{x}) \approx \sum_{a\in O} \Pi_{\bar{a},\bar{b}}(\bar{x}a)$. We therefore get a greedy method to estimate the quantity $(\!|L_{\bar{a}}p, L_{\bar{b}}p|\!)$, as outlined below.

1. let $k = 0$, $A_k := \{\epsilon\}$ and compute $v_k = \sum_{\bar{x} \in A_k} \Pi_{\bar{a},\bar{b}}(\bar{x})$;
2. set $A_{k+1} := (A_k \setminus \{y\}) \cup \{\bar{y}a \,:\, a \in O\}$, where $\bar{y} = \arg\max_{\bar{x} \in A_k} \Pi_{\bar{a},\bar{b}}(\bar{x})$;
3. let $k \leftarrow k + 1$ and compute v_k as in step 1;
4. stop (and output $(\!|L_{\bar{a}}p, L_{\bar{b}}p|\!) \approx v_k$) if $v_{k-1} - v_k$ is less than some threshold θ, otherwise goto step 2.

- To use Monte Carlo sampling. In fact, by its definition we know that

$$(\!|L_{\bar{a}}p, L_{\bar{b}}p|\!) = \lim_{N \to \infty} \sum_{\bar{x} \in O^N} P(\bar{a}\bar{x}) \cdot \sqrt{P(\bar{b}\bar{x})/P(\bar{a}\bar{x})}\,.$$

  Therefore, letting $\{\bar{x}_1, \ldots, \bar{x}_K\}$ be a aubset of O^* sampled from the distribution $P(\cdot|\bar{a})$, we get the following approximation:

$$(\!|L_{\bar{a}}p, L_{\bar{b}}p|\!) \approx \frac{P(\bar{a})}{K} \sum_{k=1}^{K} \sqrt{P(\bar{b}\bar{x}_k)/P(\bar{a}\bar{x}_k)}\,.$$

The efficiency and accuracy of these two methods remains to be investigated in detail.

7 Conclusion

In this paper we first briefly reviewed the basic OOM theory, with a special emphasis on the negative probability problem (NPP). To avoid this NPP of OOMs, we proposed a novel variant of OOMs, called norm observable operator models (NOOMs). Although NOOMs look structurally similar to OOMs, the mathematical ground of NOOMs is actually quite different from that of OOMs. Specifically, the system states and operators of the two models are defined in entirely different spaces.

NOOMs avoid the NPP by design, while still being powerful enough to capture all MC-describable processes and some stochastic processes that cannot be modelled by HMMs. Furthermore, it is rather easy to construct NOOMs from scratch and convert an arbitrary NOOM to the equivalent OOM, which gives rise to an efficient way for creating non-trivial OOMs. Studying such non-HMM OOMs is helpful for further exploring the standard OOM theory, especially for finding nontrivial sufficient conditions that allow us to check whether a given collection of matrices form a valid OOM.

We outlined a constructive algorithm for learning NOOMs from data, proved its asymptotical correctness, and proposed two possible ways to estimate the kernel matrix Q without the need of an exponential size sample. Although a practical algorithm has not been completely worked out, it is a noteworthy initial result that NOOMs can in principle be learnt from data constructively and asymptotically correctly.

Although research on NOOMs is still in its initial stage, we perceive NOOMs as a model class of significant theoretical and practical interest. From the theoretical aspect, the NOOM approach embeds the family of stochastic processes into an inner product space, enabling us to describe and analyze stochastic processes with methods from not only linear algebra but also real analysis. Furthermore, NOOMs exhibit striking analogies to the formalism of quantum mechanics, a topic that the first author pursues in a separate line of investigations. From a practical perspective, NOOMs have the potential to develop into a viable alternative to HMMs and OOMs, possibly capturing the same class of processes as OOMs do, while avoiding the non-negativity problem: the original motif that started NOOM research.

We end the paper by pointing out some open theoretical problems which we deem to be of particular relevance for further progress.

- As we mentioned in the discussion at the end of Section 5, there may be other methods besides the one presented in Section 3 to construct standard NOOMs from a given process. Finding these NOOM constructions will shed more light on the relationship between NOOMs and the associated processes; and furthermore help to develop new learning algorithms (note the relation between the learning algorithm introduced in Section 6 and the construction of NOOMs in Section 3).

- Due to the fact that there are various methods to construct NOOMs from a process, currently we do not know how to check whether a given NOOM has minimal dimension in its equivalence class, which is an easy task for OOMs. But to check whether two NOOMs are equivalent to each other is relatively easy: one first converts them into equivalent OOMs, then check the equivalence of these with the methods known for OOMs.

- There are processes that can be described by NOOMs more compactly than by OOMs, such as the probability clock presented in this paper. To characterize these processes would also be an interesting reseach topic. — In general, if one creates a NOOM by, e.g., Algorithm 1, and converts it into the equivalent minimal-dimensional OOM, one typically obtains a model with higher dimension. We thus conjecture that, if a process can be modelled by NOOMs, its NOOM representation is usually more compact than its OOM representation.

Acknowledgements

We are very grateful to two anonymous reviewers for unusually careful reviews and very constructive suggestions. The research on NOOMs was supported by the Deutsche Forschungsgemeinschaft (DFG) project JA 1210/1-1&2.

References

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1):164–171.

Bengio, Y. (1999). Markovian models for sequential data. *Neural Computing Surveys*, 2:129–162.

Bourlard, H. and Bengio, S. (2002). Hidden Markov models and other finite state automata for sequence processing. In Arbin, M. A., editor, *The handbook of brain theory and neural networks*. MIT Press.

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

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

Durbin, R., Eddy, S., Krogh, A., and Mitchinson, G. (2000). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Elliott, R. J., Aggoun, L., and Moore, J. B. (1995). *Hidden Markov models: estimation and control*, volume 29. Springer Verlag, New York.

Ephraim, Y. and Merhav, N. (2002). Hidden Markov processes. *IEEE Transactions on Information Theory*, 48(6):1518–1569.

Faigle, U. and Schönhuth, A. (2006). Quantum predictor models. *Electronic Notes in Discrete Mathematics*, 25:149–155.

Faigle, U. and Schönhuth, A. (2007). Asymptotic mean stationarity of sources with finite evolution dimension. *IEEE Transactions on Information Theory*, 53:2342–2348.

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

Hom, R. A. and Johnson, C. R. (1989). *Topics in matrix analysis*. Cambridge University Press.

Ito, H., Amari, S.-I., and Kobayashi, K. (1992). Identifiability of hidden markov information sources and their minimum degrees of freedom. *IEEE Trans. Inform. Theory*, 38(2):324–333.

Jaeger, H. (1998). Discrete-time, discrete-valued observable operator models: a tutorial. GMD Report 42, GMD, Sankt Augustin.

Jaeger, H. (1999). Characterizing distributions of stochastic processes by linear operators. GMD Report 62, German National Research Center for Information Technology.

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

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

Littman, M. L., Sutton, R. S., and Satinder, S. (2001). Predictive representation of state. In *Advances in Neural Information Processing Systems*, volume 14, pages 1555–1561.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Schönhuth, A. (2006). *Diskretwertige stochastische Vektorräume (in German)*. PhD thesis, Faculty of Mathematics and Natural Sciences, University Köln.

Wiewiora, E. W. (2008). *Modeling probability distributions with predictive state representation*. PhD thesis, University of California, San Diego.

Zhao, M.-J., Jaeger, H., and Thon, M. (2009). A bound on modeling error in observable operator models and an associated learning algorithm. *Neural Computation*, 20(9):2687–2712.

A Kolmogorov Extension Theorem

We first give a mathematically rigorous statement of the KET for the general case, then use it to prove the "converse" part of Theorem 1.

Theorem 10 (the Kolmogorov Extension Theorem)

Let I be an arbitrary (possibly continuous) index set and $(B, \mathcal{B})$ a measurable space. For each finite subset J of I, let $P_J : \mathcal{B}^J \to [0,1]$ be a probability measure on the product measurable space $(B^J, \sigma(\mathcal{B}^J))$. For $J \subseteq K \subseteq I$, let π_{JK} be the natural projection from B^K onto B^J. For example, if $J = \{1,2\}$ and $K = \{1,2,3\}$, then π_{JK} maps each point $b_K = (b_1, b_2, b_3)$ in B^K to the point $b_J = (b_1, b_2) \in B^J$.

Assume now that the following consistency condition:

$$P_J(A) = P_K(\pi_{JK}^{-1}(A)), \quad (\forall\, A \in \sigma(\mathcal{B}^J),\ J \subseteq K \text{ are finite subsets of } I) \tag{32}$$

is met, then there is a unique measure P on the measurable space $(\Omega, \mathcal{E}) = (B, \mathcal{B})^I$ — Ω is the collection of all functions from I to B, and $\mathcal{H}$ is the σ-algebra generated by the finite-dimensional measurable rectangles, such that

$$P_J(A) = P(\pi_{JI}^{-1}(A)), \quad (\forall\, A \in \sigma(\mathcal{B}^J),\ J \subseteq I \text{ finite}).$$

In particular, for the special case of the paper, $B = O = \{a^1, a^2, \ldots, a^\ell\}$, $\mathcal{B}$ is the power set of O, and $I = \{0, 1, 2, \ldots\}$. The consistency condition (32) is then reduced to the summation constraint $P(\bar{a}) = \sum_{b \in O} P(\bar{a}b)$, which is warranted by the property $\sum_{a \in O} \varphi_a^* \varphi_a = \mathrm{id}_{\mathcal{E}}$ in standard NOOM representations. We thus proved the "converse" part of Theorem 1.

B The Analytical Solution to the Problem (31)

As mentioned before, let $\hat{\varphi}$ (resp. φ^*) be the $m\ell$ by m matrix created by stacking the matrices $\hat{\varphi}_a$ (resp. φ_a^*) below one another. Then the optimization problem (31) can be equivalently written as (after some simple computation)

$$\begin{array}{ll} \text{maximize} & \mathrm{tr}\{\hat{\varphi}^{\mathsf{T}}\varphi^*\} + \hat{\boldsymbol{u}}_0^{\mathsf{T}}\boldsymbol{u}_0^*\,, \\ \text{subject to} & (\varphi^*)^{\mathsf{T}}(\varphi^*) = I_m\,, \quad \|\boldsymbol{u}_0^*\| = 1\,. \end{array}$$

It is obvious that we can compute the optimal φ^* and $\boldsymbol{u}_0^*$ separately, as in

$$\begin{array}{rcl} (\varphi^*)_{\mathrm{opt}} & = & \arg\max_{\varphi^*}\{\mathrm{tr}\{\hat{\varphi}^{\mathsf{T}}\varphi^*\} \,:\, (\varphi^*)^{\mathsf{T}}(\varphi^*) = I_m\}\,, \\ (\boldsymbol{u}_0^*)_{\mathrm{opt}} & = & \arg\max_{\boldsymbol{u}_0^*}\{\hat{\boldsymbol{u}}_0^{\mathsf{T}}\boldsymbol{u}_0^* \,:\, \|\boldsymbol{u}_0^*\| = 1\}\,, \end{array}$$

from which we immediately see $(\boldsymbol{u}_0^*)_{\mathrm{opt}} = \hat{\boldsymbol{u}}_0/\|\hat{\boldsymbol{u}}_0\|$. To obtain $(\varphi^*)_{\mathrm{opt}}$, we take the SVD $\hat{\varphi} = LDR^{\mathsf{T}}$ of $\hat{\varphi}$, where $L \in \mathbb{R}^{m\ell\times m}$ and $R \in \mathbb{R}^{m\times m}$. The target function $\mathrm{tr}\{\hat{\varphi}^{\mathsf{T}}\varphi^*\}$ can then be written as $\mathrm{tr}\{\hat{\varphi}^{\mathsf{T}}\varphi^*\} = \mathrm{tr}\{RDL^{\mathsf{T}}\varphi^*\} = \mathrm{tr}\{DL^{\mathsf{T}}\varphi^*R\} = \mathrm{tr}\{DL^{\mathsf{T}}V\}$, where $V := \varphi^*R \in \mathbb{R}^{m\ell\times m}$ satisfies $V^{\mathsf{T}}V = I_m$ and so has orthonormal columns. Writing

$$D = \mathrm{diag}\{d_1, d_2, \ldots, d_m\}\,, \quad L = [\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_m]\,, \quad V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]\,,$$

we get $\mathrm{tr}\{\hat{\varphi}^{\mathsf{T}}\varphi^*\} = \sum_{i=1}^m d_i \cdot (\boldsymbol{l}_i^{\mathsf{T}}\boldsymbol{v}_i)$, which, obviously, reaches its maximum $\sum_{i=1}^m d_i$ when $\boldsymbol{v}_i = \boldsymbol{l}_i$ for all $i = 1, 2, \ldots, m$, that is, $L = V = \varphi^*R$. We thus get $(\varphi^*)_{\mathrm{opt}} = LR^{\mathsf{T}}$.