Algebraic Reduction of Hidden Markov Models

Tommaso Grigoletto and Francesco Ticozzi

***Abstract*—The problem of reducing a Hidden Markov Model (HMM) to one of smaller dimension that exactly reproduces the same marginals is tackled by using a system-theoretic approach. Realization theory tools are extended to HMMs by leveraging suitable algebraic representations of probability spaces. We propose two algorithms that return coarse-grained equivalent HMMs obtained by stochastic projection operators: the first returns models that exactly reproduce the single-time distribution of a given output process, while in the second the full (multi-time) distribution is preserved. The reduction method exploits not only the structure of the observed output, but also its initial condition, whenever the latter is known or belongs to a given subclass. Optimal algorithms are derived for a class of HMM, namely observable ones.**

I. Introduction

Hidden Markov processes are an ubiquitous class of stochastic models that has extensive application in modeling and prediction for speech [1], [2], biological systems [3]–[6], information and communication systems [7]–[9]. Dedicated optimal control and estimation methods have been developed for this class of models, see e.g. [10]–[12].

In the development of the realization theory for HMMs, two related yet well distinct problems emerge: *constructing an HMM from data*, and *reducing an existing model*, when possible, to an equivalent one of smaller size. For an analysis and review of the first one, see for example [13], [14], and more recent results in [15]. In this paper, we shall focus on the reduction problem. Besides its theoretical interest, methods for model reduction are critical in effectively addressing problems in large-scale systems [16]–[18]. A characterization of equivalent HMMs, that is, models that produce the same output marginals of a given one, is proposed in [19]. Their treatment of equivalent HMMs is based on the definition of *effective* spaces, which specify equivalence classes of HMMs, representing the HMM analogue of minimal realizations spaces for linear systems. In the same paper, the authors pose the problem of finding a minimal equivalent HMM. As a reduction to the effective space does not guarantee to preserve the positivity of the model, the problem has so far remained unsolved.

In this paper, we show how effective spaces can be extended so that the reduced model remains an HMM. In fact, we propose a general approach to the model reduction problem that is based on an *algebraic description of probability spaces*. While this is done very frequently and almost implicitly, we take a deeper look into the algebraic structures and the associated representations. In particular, we shall need minimal algebraic models that represent a set of random variables (r.v) and conditional expectations. Such an algebraic approach has been developed to generalize the classical Kolmogorov description to the non-commutative case so that it suitably covers quantum mechanics [20]–[22], but it has proven useful in many other areas, from random matrix theory (see, e.g. the insightful introduction [23]) to algebraic statistics [24]. In our setting, the algebraic framework and the induced matrix representations allow us to leverage on observability and reachability ideas in the characterization of equivalent models, as well as linear-algebraic algorithms that compute reduced models. Our approach remains deeply rooted in the system-theoretic analysis of the dynamical model and can be seen as a way to construct *reduced stochastic realizations* for an HMM. Furthermore, the proofs of effectiveness for the proposed methods all hinge on a result of model reduction for switched linear systems, In order to maintain the focus on HMM, the latter is presented in Appendix A).

In what follows, we deal with reductions of a given HMM that *exactly* reproduce the marginals of the original systems. This allows us to clearly illustrate the working and theoretical foundation of the method: extension to approximate reduction will be the focus of upcoming work.

Similar problems have been studied from different perspectives: in particular, the concept of lumpability of Markov processes [25], which induces coarse-grained processes analogous to those presented here, has been employed to characterize a class of exactly reducible HMMs (2-lumpable systems), see [26] and references therein. Other works, as [27] and references therein, reframe the problem using cellular automata for hidden information sources and study reductions of Markov transition kernels within this abstract approach.

The differences between our approach and the existing results are manifold, both in the tools used and the nature of the results. In the proposed framework, we introduce and solve two types of reduction problems: preserving only the single-time marginal, or the full (multi-time) distribution of the outcomes. We show that the former, which is of interest in model reduction of master equations for statistical models or mixing processes and algorithms [28], can lead to further reduction and smaller final models, as one might expect. In addition, our reductions leverage not only the structure of the measured process, but also the particular initial distribution of the HMM. We show that the initial conditions are indeed critical for obtaining minimal reductions in many situations, in particular, when the original model is initialized in an equilibrium density. The method hinges on the use of conditional expectations as projections for obtaining a reduced representation of the dynamics. While the idea is certainly not new to the control community, see e.g. the derivation of Kalman filters [10], [29], in this work we develop it in an algebraic framework. After representing a conditional expectation as a linear operator, we construct stochastic, non-

T. Grigoletto and F. Ticozzi are with the Department of Information Engineering, University of Padova, Via Gradenigo 6, 35131 Padova, Italy. Emails: tommaso.grigoletto@phd.unipd.it, ticozzi@dei.unipd.it.

square factorization of its dual with respect to the inner product associated to the expectation: the factors are then used to obtain the reduced probabilistic description, preserving its stochastic character. Lastly, we make direct contact with system-theoretic ideas in a linear-algebraic framework, which allows for effective, practically implementable algorithms for the reduction process. In fact, while the whole analysis could be carried out in the infinite-dimensional case, we here restrict to the finite case: in order to derive computable algorithms a finite-dimensional approximation would be needed anyway.

The structure of the paper is as follows: In Section II we review the fundamentals of the algebraic probabilistic models needed for our aims. The approach is directly borrowed from non-commutative probability [22], [30] and its use in quantum theory, where the algebras used for embedding the probability space need not be commutative (and are typically infinite-dimensional [21]), and can then be used to model quantum systems [20]. As remarked above, in this work we only use commutative, finite-dimensional associative algebras, represented as $\mathbb{R}^n$ endowed with its element-wise product. Subsection II-B is focused on conditional expectations as linear maps on algebras, their duals, and their representations. These are some of the key tools in the development of our method.

Section III is devoted to introducing the notation and the problems of interest, namely obtaining reduced models that reproduce either the single-time marginals or the multi-time marginals of a given HMM, while Section IV presents some preliminary results that build upon [19] from an explicit system-theoretic perspective. The main results of the section are obtained specializing a switched-system result that we derive in Appendix A to maintain the focus on HMMs. The key ideas we leverage to obtain reduced HMMs are described in V, where a class of reduction algorithms for the single-time marginal problem is developed. Section VI then extends and adapts these ideas to the multi-time marginal problem. A key point in our analysis is that, in order to develop the algorithms, we must switch from the abstract quotient spaces of [19] to a representative effective subspace. We show that the choice of representative has a non-trivial effect on the reduction itself. How to select this and other parameters used in the algorithms is discussed in Section VII, where we provide optimal choices for a class of models that includes observable HMMs and Markov chains. The same choices prove to be optimal in all the tested examples, also in the presence of non-observable components of the reachable space. Some particularly instructive examples are given in Section VIII, and an outlook on future developments is provided with the concluding remarks in Section IX.

A. Basic Notation

In the following, we typically denote vectors $\boldsymbol{v} \in \mathbb{R}^n$ in boldface, and matrices in capitals $V \in \mathbb{R}^{n\times m}$. We denote $\mathbf{1}$, the vector of all ones, and $\mathbf{0}$ the vector of all zeros. The matrix transpose of V is V^T. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$ and the standard basis $\{\boldsymbol{e}_i\}$ for $\mathbb{R}^n$, we define its support as the vector space $\mathrm{supp}(\boldsymbol{x}) = \mathrm{span}\{\boldsymbol{e}_i | \boldsymbol{e}_i^T\boldsymbol{x} \neq 0\}$. Given a vector space $\mathcal{V} \subseteq \mathbb{R}^n$, its *support* is defined as the vector space $\mathrm{supp}(\mathcal{V}) = \mathrm{span}\{\boldsymbol{e}_i | \exists \boldsymbol{x} \in \mathcal{V} \text{ s.t } \boldsymbol{e}_i^T\boldsymbol{x} \neq 0\}$. $\mathrm{diag}(\cdot)$ is the operator that, given a vector $\boldsymbol{v}$, $\mathrm{diag}(\boldsymbol{v})$ returns a diagonal matrix with $[\mathrm{diag}(\boldsymbol{v})]_{i,i} = \boldsymbol{v}_i$.

II. Algebraic approach to probability theory

The central idea in algebraic probability models is to represent all the key ingredients of a classical probabilistic model as elements of a suitable algebra $\mathscr{A}$, endowed with a probability functional (or state) p. In the following sections, we start from a probability space $(\Omega, \Sigma, \mathbb{P})$ and briefly review how to construct an algebraic representation $(\mathscr{A}, p)$, with $\mathscr{A} \subseteq \mathbb{R}^n$. Correspondingly, we show that any pair $(\mathscr{A}, p)$ admits a classical representation. This allows for a natural probabilistic interpretation of the proposed reduction method.

A. Fundamentals of algebraic probabilistic models

1) Events and σ-Algebras: Throughout the rest of this article, we will consider finite-dimensional probability spaces $(\Omega, \Sigma, \mathbb{P})$. Without loss of generality, we can assume $\Omega = \{1, \dots, n\}$.

The first step in the construction entails the vector representation of events. The latter are in 1-to-1 correspondence to indicator functions: let $I_E(\omega)$ be the indicator function associated with the event E. Since the probability space is finite-dimensional, we can further associate indicator functions to vectors in $\mathbb{R}^n$. In particular, each indicator of an elementary event $\omega \in \Omega$ can be associated to its corresponding vector of the standard basis, i.e. $\boldsymbol{e}_\omega \in \mathbb{R}^n$. Similarly, we can define *indicator vectors* for any event $E \in \Sigma$ as $\boldsymbol{f}_E = \sum_{\omega \in E} \boldsymbol{e}_\omega$. For these vectors, $(\boldsymbol{f}_E)_\omega = 1$ if $\omega \in E$ and zero otherwise. Notice that $\boldsymbol{f}_\Omega = \mathbf{1}$, and $\boldsymbol{f}_\varnothing = \mathbf{0}$.

Let us denote with $\mathcal{F}_\Sigma$ the set of indicator vectors of the events of the σ-algebra Σ. Let $\wedge$ denote the element-wise product $(\boldsymbol{v} \wedge \boldsymbol{w})_i = \boldsymbol{v}_i\boldsymbol{w}_i$, $\vee$ denote the modified sum operation defined as $\boldsymbol{v} \vee \boldsymbol{w} = \boldsymbol{v} + \boldsymbol{w} - \boldsymbol{v} \wedge \boldsymbol{w}$ and $\neg$ denote the negation operation defined as $\neg\boldsymbol{v} = \mathbf{1} - \boldsymbol{v}$. By construction, the set $\mathcal{F}_\Sigma$ equipped with the operations $\wedge, \vee, \neg$ is isomorphic to the σ-algebra Σ with $\cap, \cup, \bar{\cdot}$. In the following, we refer to $\mathcal{F}_\Sigma$ as a *vector σ-algebra*, and we will drop the subscript when unnecessary.

A *vector partition of* Ω is a subset $\mathcal{P} \subseteq \mathcal{F}\backslash\{\mathbf{0}\}$ such that $\boldsymbol{f}_i \wedge \boldsymbol{f}_j = \mathbf{0}$, for all $\boldsymbol{f}_i, \boldsymbol{f}_j \in \mathcal{P}$, $i \neq j$ and $\mathbf{1} = \vee_{\boldsymbol{f}_j \in \mathcal{P}} \boldsymbol{f}_j$. The *finest resolution* in $\mathcal{F}$ is a partition $\mathrm{res}(\mathcal{F})$ such that $\boldsymbol{f} = \vee_{\boldsymbol{f}_j \in \mathrm{res}(\mathcal{F})} c_j \boldsymbol{f}_j$ with $c_j \in \{0, 1\}$, for all $\boldsymbol{f} \in \mathcal{F}$.

Note that $\mathrm{res}(\mathcal{F})$ is not necessarily equal to the standard basis of $\mathbb{R}^n$ since, in general, Σ is contained but not equal to the power set of Ω. We shall also denote $\mathrm{res}(\Sigma)$ to indicate the finest resolution of a classical σ-algebra.

2) Random variables: Random variables (r.v.) are Σ-measurable functions $X(\omega) : \Omega \to \mathbb{A} \subset \mathbb{R}$, where $\mathbb{A} = \{x_i\}$ is the finite set of outcomes of X, called the alphabet. Let $E_i = X^{-1}(x_i)$. An r.v. X can also be represented as linear combination of indicator function $X(\omega) = \sum_{i=0}^{|\mathbb{A}|} x_i I_{E_i}(\omega)$.

Using the vector representation $\boldsymbol{f}_{E_i}$ of indicator functions I_{E_i} in the previous equation, each X can also be represented as a vector

$$\boldsymbol{x} = \sum_{i=1}^{|\mathbb{A}|} x_i \boldsymbol{f}_{E_i} \in \mathbb{R}^n$$

such that $\{\boldsymbol{f}_i\} \subset \mathcal{F}_\Sigma$ forms a partition of Ω. Notice that in the vector formalism, the notion of $\mathcal{F}_\Sigma$-measurability is equivalent to the condition $\boldsymbol{x} \in \operatorname{span}\{\mathcal{F}_\Sigma\}$. Here and elsewhere, the boldface font $\boldsymbol{x}$ is used for (vector representations of) r.v.s, while x denotes the corresponding outcome. As we show below, $\operatorname{span}\{\mathcal{F}_\Sigma\}$ has the property of being an *algebra*, namely a vector space (or subspace) that is closed under the element-wise product $\wedge$. An algebra is *unital* if it contains $\mathbf{1}$. The whole $\mathbb{R}^n$ is then an unital algebra, and we denote its subalgebras using the script font, e.g. $\mathscr{A}$. A non-unital algebra $\mathscr{A}$ still contains the vector $\mathbf{1}_{\mathscr{A}}$, which has entries 1 on the support of $\mathscr{A}$ and 0 otherwise and acts as the product identity in $\mathscr{A}$.

The following proposition collects some known facts which clarify the relation between $\mathcal{F}_\Sigma$ and $\mathscr{A} = \operatorname{span}\{\mathcal{F}_\Sigma\}$ and proves that it is indeed an algebra.

Proposition 1. *If $\mathcal{F} \subset \mathbb{R}^n$ is a vector σ-algebra, then $\mathscr{A} = \operatorname{span}\{\mathcal{F}\}$ is the smallest subalgebra in $\mathbb{R}^n$ containing $\mathcal{F}$, and it is unital. Conversely, let $\mathscr{A}$ be any unital subalgebra in $\mathbb{R}^n$ and $\operatorname{idem}(\mathscr{A}) := \{\boldsymbol{f} \in \mathscr{A} | \boldsymbol{f} \wedge \boldsymbol{f} = \boldsymbol{f}\} \subset \mathscr{A}$ be the set of idempotent vectors in $\mathscr{A}$. Then $\operatorname{idem}(\mathscr{A})$ is the smallest σ-algebra such that every element in $\mathscr{A}$ is $\mathcal{F}$-measurable and $\operatorname{res}(\operatorname{idem}(\mathscr{A}))$ forms an orthogonal basis for $\mathscr{A}$.*

A proof of this proposition is reported in Appendix B for completeness. This proposition shows that, not only does the space of $\mathcal{F}_\Sigma$-measurable random variables form an unital subalgebra, but, more importantly, given any unital subalgebra $\mathscr{A}$, it is possible to find the *minimal* (vector) σ-algebra that makes every random variable in $\mathscr{A}$ measurable. For convenience, in the following, we refer to $\operatorname{res}(\operatorname{idem}(\mathscr{A}))$ as $\operatorname{res}(\mathscr{A})$.

3) Probability and expectations: Let now consider a probability measure $\mathbb{P} : \Omega \to [0, 1]$. For any probability measure $\mathbb{P}[\cdot]$ on Σ we can define a vector as follows

$$\boldsymbol{p} := \sum_{\omega \in \Omega} \frac{\mathbb{P}[\omega]}{\langle \boldsymbol{f}_\omega, \boldsymbol{f}_\omega \rangle} \boldsymbol{f}_\omega.$$

Then, for any $\boldsymbol{f}_E \in \mathcal{F}_\Sigma$ it is immediate to verify that $\mathbb{P}[E] = \langle \boldsymbol{p}, \boldsymbol{f}_E \rangle$. In particular, notice that if we can write $\boldsymbol{p} := \sum_{\boldsymbol{f}_r \in \operatorname{res}(\mathscr{A})} p_r \boldsymbol{f}_r$, we find that $\boldsymbol{p}$ can be interpreted as a random variable in the same algebra, $\boldsymbol{p} \in \mathscr{A}$.

A vector $\boldsymbol{p}$ is said to be a probability vector if $\boldsymbol{p}_i \geqslant 0$ for all i and $\mathbf{1}^T \boldsymbol{p} = 1$. The set of probability vectors in $\mathscr{A}$ is defined as $\mathcal{D}(\mathscr{A}) := \{\boldsymbol{p} \in \mathscr{A} | \boldsymbol{p}_i \geqslant 0 \quad \forall i, \quad \mathbf{1}^T \boldsymbol{p} = 1\}$. Note that $\mathcal{D}(\mathscr{A}) = \mathcal{D}(\mathbb{R}^n) \cap \mathscr{A}$.

Consider a r.v. X and let us denote again with $\boldsymbol{f}_i$ the indicator function associated to the outcome x_i. It then holds that $\mathbb{P}[X = x_i] = \langle \boldsymbol{p}, \boldsymbol{f}_i \rangle$. Similarly, we can compute the expectation of a random variable as $\mathbb{E}[\boldsymbol{x}] = \sum_j x_j \mathbb{P}[E_j] = \sum_j x_j \langle \boldsymbol{p}, \boldsymbol{f}_j \rangle = \langle \boldsymbol{p}, \boldsymbol{x} \rangle$.

In summary, we have shown that an unital subalgebra $\mathscr{A}$ can subsume both the σ-algebra and the space of measurable random variables of a given probability space. Moreover, it is equivalent to a probability space when paired with a positive linear functional, associated to the inner product with a probability vector $\boldsymbol{p}$. Conversely, given a pair $(\mathscr{A}, \boldsymbol{p})$, we can always construct a (classical) probability space associated with the pair. This can be done by choosing $\Omega = \{1, \ldots, n\}$ and the underlying σ-algebra Σ associated to $\operatorname{idem}(\mathscr{A})$ as in Proposition 1. Lastly, $\boldsymbol{p}$ represents the probability distribution associated with the functional $\mathbb{P}[E] = \langle \boldsymbol{p}, \boldsymbol{f}_E \rangle$.

B. Stochastic maps and Conditional Expectations

Let us now focus on the maps between probability vectors.

Consider two unital subalgebras $\mathscr{F}$ of $\mathbb{R}^n$ and $\mathscr{G}$ of $\mathbb{R}^m$. A linear map between probability vectors $P[\cdot] : \mathcal{D}(\mathscr{F}) \to \mathcal{D}(\mathscr{G})$, $\boldsymbol{p} \mapsto \boldsymbol{q} = P[\boldsymbol{p}]$ is called a *stochastic map.* Such a map can be represented as a *(column)-stochastic matrix* $P \in \mathbb{R}^{m \times n}$, i.e. a matrix such that $(P)_{i,j} \geqslant 0 \; \forall i, j$ and $\mathbf{1}_m^T P = \mathbf{1}_n^T$.

In the following, the main task will be to find reduced descriptions of linear dynamics associated with stochastic maps. In doing this, we exploit the properties of a particular class of stochastic maps: the duals of conditional expectations.

Recall that the conditional expectation of an r.v. given a σ-algebra Σ with finest resolution $\operatorname{res}(\Sigma)$ can be written as follows:

$$\mathbb{E}[X|\Sigma] = \sum_{E \in \operatorname{res}(\Sigma)} \frac{\mathbb{E}[I_E X]}{\mathbb{E}[I_E]} I_E(\omega). \tag{1}$$

Let consider a vector r.v. $\boldsymbol{x} \in \mathscr{F} \subseteq \mathbb{R}^n$, a unital algebra $\mathscr{A} \subseteq \mathscr{F}$ with $\{\boldsymbol{a}_i\} = \operatorname{res}(\mathscr{A})$ and $d = \dim(\mathscr{A}) < n$, and the underlying probability measure $\boldsymbol{p}$. Following the previous definition, we can define the conditional expectation for the vector r.v. with respect to an algebra $\mathscr{A}$:

$$\mathbb{E}_{\boldsymbol{p}}[\boldsymbol{x}|\mathscr{A}] := \sum_{j=1}^{d} \frac{\langle \boldsymbol{p}, \boldsymbol{x} \wedge \boldsymbol{a}_j \rangle}{\langle \boldsymbol{p}, \boldsymbol{a}_j \rangle} \boldsymbol{a}_j.$$

Noticing that it is a linear operator acting on $\boldsymbol{x}$ we can represent it as a matrix $\mathbb{E}_{|\mathscr{A}, \boldsymbol{p}} \in \mathbb{R}^{n \times n}$, namely:

$$\mathbb{E}_{|\mathscr{A}, \boldsymbol{p}} = \sum_{j=1}^{d} \frac{\boldsymbol{a}_j (\boldsymbol{p} \wedge \boldsymbol{a}_j)^T}{\langle \boldsymbol{p}, \boldsymbol{a}_j \rangle}. \tag{2}$$

Consider the inner product of the conditional expectation of $\boldsymbol{x}$ with a probability distribution $\boldsymbol{q}$, which we have shown to correspond to its expectation. The dual of the conditional expectation is then a map on the probability distribution defined as:

$$\left\langle \boldsymbol{q}, \mathbb{E}_{|\mathscr{A}, \boldsymbol{p}} \boldsymbol{x} \right\rangle = \left\langle \mathbb{E}_{|\mathscr{A}, \boldsymbol{p}}^T \boldsymbol{q}, \boldsymbol{x} \right\rangle$$

which gives in

$$\mathbb{E}_{|\mathscr{A}, \boldsymbol{p}}^T = \sum_{j=1}^{d} \frac{(\boldsymbol{p} \wedge \boldsymbol{a}_j) \boldsymbol{a}_j^T}{\langle \boldsymbol{p}, \boldsymbol{a}_j \rangle}. \tag{3}$$

It is immediate to verify that $\mathbb{E}_{|\mathscr{A}, \boldsymbol{p}}^T$ is stochastic.

The conditional expectation and its adjoint are orthogonal projectors with respect to a modified inner product. Notice that $\boldsymbol{p} \wedge \mathscr{A} = \operatorname{span}\{\boldsymbol{p} \wedge \boldsymbol{a}_i\} = \operatorname{diag}(\boldsymbol{p}) \mathscr{A}$.

Lemma 1. *Let consider the modified inner product $\langle \boldsymbol{v}, \boldsymbol{w} \rangle_{\boldsymbol{p}} = \mathbb{E}_{\boldsymbol{p}}[\boldsymbol{v} \wedge \boldsymbol{w}]$, with $\boldsymbol{p} > \boldsymbol{0}$. Then $\mathbb{E}_{|\mathscr{A},\boldsymbol{p}}$ is the orthogonal projector onto $\mathscr{A}$ with respect to the inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{p}}$ and $\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$ is the orthogonal projector onto $\boldsymbol{p} \wedge \mathscr{A}$ with respect to the inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{p}^{-1}}$.*

The proof of this lemma is reported in Appendix B for completeness.

Remark 1. Note that the above Lemma also implies that $\mathbb{E}_{|\mathscr{A},\boldsymbol{p}}$ acts as the identity on $\mathscr{A}$ while $\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$ acts as the identity on $\boldsymbol{p} \wedge \mathscr{A}$. Furthermore, they are orthogonal projections for the standard inner product $\langle \cdot, \cdot \rangle$ if (and only if) $\boldsymbol{p} \in \mathcal{D}(\mathscr{A})$ and is positive, namely $\boldsymbol{p} = \sum_j \lambda_j \boldsymbol{a}_j \in \mathbb{R}^n$ with $\lambda_j > 0$, $\sum_j \lambda_j = 1$. In this case, we have $\mathbb{E}_{|\mathscr{A},\boldsymbol{p}} = \mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$.

Consider the standard basis $\{\boldsymbol{e}_j\}$ for $\mathbb{R}^d$, where d is the dimension of $\mathscr{A}$. We can then construct a (full-rank) stochastic factorization of $\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$.

Proposition 2. *Define*

$$J = \sum_{j=1}^{d} \frac{(\boldsymbol{p} \wedge \boldsymbol{a}_j)\boldsymbol{e}_j^T}{\langle \boldsymbol{p}, \boldsymbol{a}_j \rangle} \in \mathbb{R}^{n \times d}, \quad R = \sum_{j=1}^{d} \boldsymbol{e}_j \boldsymbol{a}_j^T \in \mathbb{R}^{d \times n}. \tag{4}$$

Then J, R are stochastic matrices that satisfy $JR = \mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$, $RJ = I_d$, $\ker(R) = \mathscr{A}^\perp$ and $\ker(J^T) = (\boldsymbol{p} \wedge \mathscr{A})^\perp$.

Proof. J and R are clearly positive since both $\{\boldsymbol{a}_j\}$ and $\{\boldsymbol{e}_j\}$ are vectors of zeros and ones and $\boldsymbol{p}$ is positive. J is clearly stochastic, $\boldsymbol{1}_n^T J = \boldsymbol{1}_d^T$ since $\boldsymbol{1}_n^T(\boldsymbol{p} \wedge \boldsymbol{a}_j) = \langle \boldsymbol{p}, \boldsymbol{a}_j \rangle$. On the other hand, we have $\boldsymbol{1}_d^T R = \sum_{j=1}^d \boldsymbol{a}_j^T = \boldsymbol{1}_n$, since $\mathscr{A}$ is unital.

We can then observe that $\boldsymbol{a}_j^T(\boldsymbol{p} \wedge \boldsymbol{a}_k) = \langle \boldsymbol{p}, \boldsymbol{a}_j \wedge \boldsymbol{a}_k \rangle = \langle \boldsymbol{p}, \boldsymbol{a}_j \rangle \delta_{j-k}$ to conclude that $RJ = I_d$. Finally, if we consider $\boldsymbol{x} \in \mathscr{A}^\perp$, i.e. $\langle \boldsymbol{x}, \boldsymbol{a}_j \rangle = 0$ for all j we obtain $R\boldsymbol{x} = \boldsymbol{0}$ and, similarly, if $\boldsymbol{x} \in (\boldsymbol{p} \wedge \mathscr{A})^\perp$, i.e. $\langle \boldsymbol{x}, \boldsymbol{p} \wedge \boldsymbol{a}_j \rangle = 0$ for all j we obtain $J^T\boldsymbol{x} = \boldsymbol{0}$. □

This stochastic factorization induces a reduction in the probabilistic description. In fact, we have that for each distribution $\boldsymbol{q}$ and r.v. $\boldsymbol{p}$:

$$\begin{aligned} \left\langle \boldsymbol{q}, \mathbb{E}_{|\mathscr{A},\boldsymbol{p}}\boldsymbol{x} \right\rangle &= \left\langle \mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}\boldsymbol{q}, \boldsymbol{x} \right\rangle = \langle JR\boldsymbol{q}, \boldsymbol{x} \rangle \\ &= \left\langle R\boldsymbol{q}, J^T\boldsymbol{x} \right\rangle = \langle \check{\boldsymbol{q}}, \breve{\boldsymbol{x}} \rangle \end{aligned}$$

where we define the *reduced distribution* as $\check{\boldsymbol{q}} := R\boldsymbol{q} \in \mathcal{D}(\mathbb{R}^d)$ and reduced random variable $\breve{\boldsymbol{x}} := J^T\boldsymbol{x} \in \mathbb{R}^d$. This property shows that, given an unital algebra $\mathscr{A}$, it is possible to reduce the probabilistic description of the set of measurable events to the space $\mathbb{R}^d$ with $d = \dim(\mathscr{A})$. For this reason, we name R the stochastic reduction and J the stochastic injection.

In order to obtain smaller reduced models, it is useful to notice that even if $\mathscr{A}$ is a non-unital subalgebra of $\mathbb{R}^n$, namely the subalgebra has limited support, we can still use the reduction via factorization. In particular, we can use definitions (2), (3) and (4) to define orthogonal (for a modified product) projections on the algebra, their dual, and their factorization. We use the notation $\mathbb{E}_{|\mathscr{A},\boldsymbol{p}}$ for simplicity, even if these are not true conditional expectations. One relevant difference, in this case, is highlighted in the following.

Corollary 1. *Let $\mathscr{A}$ be a non-unital subalgebra and $\boldsymbol{p}$ be such that $\boldsymbol{p}_i > 0$ for all i, then $\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$ allows for a factorization $\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}} = JR$ with J and R as defined above. Moreover, J is stochastic while R is stochastic over the support of $\mathscr{A}$, i.e $\boldsymbol{1}_d^T R = \boldsymbol{1}^T_{\mathrm{supp}(\mathscr{A})}$ and $\boldsymbol{1}^T_{\mathrm{supp}(\mathscr{A})} J = \boldsymbol{1}_d^T$.*

Proof. The proof is the same as 2 with the only difference that $\sum_{j=1}^d \boldsymbol{a}_j^T = \boldsymbol{1}_{\mathrm{supp}(\mathscr{A})}$ and $\boldsymbol{1}^T_{\mathrm{supp}(\mathscr{A})}(\boldsymbol{p} \wedge \boldsymbol{a}_j) = \langle \boldsymbol{p}, \boldsymbol{a}_j \rangle$ holds, since $\mathscr{A}$ is not unital. □

III. HMM and Problem definition

Throughout the rest of this work, we consider stochastic processes that can be described as Markov processes or Hidden Markov processes (HMPs).

A stochastic process $\{\boldsymbol{x}_t\}$ is a collection of r.v.s taking values in the finite alphabet $\mathbb{A}_{\boldsymbol{x}}$, indexed by time t. Without loss of generality, we can assume $\mathbb{A}_{\boldsymbol{x}} = \{1, 2, \dots, n\}$. As the alphabet is independent of time, we can choose a fixed resolution of indicator vectors $\{\boldsymbol{f}_i\}$ with respect to which $\boldsymbol{x}_t$ is measurable at all times, the standard basis for $\mathbb{R}^n$ being the most compact one. With this choice, $\{\boldsymbol{x}_t\}$ is a sequence in $\mathbb{R}^n$. In the following we thus denote by $\boldsymbol{x}_{0:k}$ a stochastic process with $t = 0, \dots, k$, with $x_{0:k} \in \mathbb{A}_{\boldsymbol{x}}^{k+1}$ an ordered sequence of its outcomes, i.e. $x_{0:k} = x_0, x_1, \dots, x_k$, where $x_i \in \mathbb{A}_{\boldsymbol{x}}$ for all i and $|x_{0:k}| = k + 1$. Then, the joint probability of a sequence of outcomes can be written as $\mathbb{P}[\boldsymbol{x}_0 = x_0, \dots, \boldsymbol{x}_k = x_k] = \mathbb{P}[\boldsymbol{x}_{0:k} = x_{0:k}]$. A stochastic process $\{\boldsymbol{x}_t\}$ in $\mathbb{R}^n$ is an homogeneous Markov process if

$$\mathbb{P}[\boldsymbol{x}_{t+1} = x_{t+1} | \boldsymbol{x}_{0:t} = x_{0:t}] = \mathbb{P}[\boldsymbol{x}_{t+1} = x_{t+1} | \boldsymbol{x}_t = x_t]$$

and such probability is independent of t for all pairs x_{t+1}, x_t.

In this case, we have that there exists an initial probability vector $\boldsymbol{p}_0 \in \mathbb{R}^n$ and a stochastic matrix $P \in \mathbb{R}^{n \times n}$ called the *transition probability matrix* such that $\mathbb{P}[\boldsymbol{x}_0 = x_0] = \langle \boldsymbol{p}_0, \boldsymbol{f}_{x_0} \rangle$ and $\mathbb{P}[\boldsymbol{x}_{t+1} = x_{t+1} | \boldsymbol{x}_t = x_t] = \boldsymbol{f}^T_{x_{t+1}} P \boldsymbol{f}_{x_t}$ where $\boldsymbol{f}_{x_t}$ represents the elementary event associated with the outcome x_t.

The main focus of this work are partially-observed HMPs, better known as HMPs. The following definition adapts [13, Definitions 9.2 and 9.3] to our setting.

Definition 1 (Hidden Markov processes)**.** *A stochastic process $\{\boldsymbol{y}_t\}$ in $\mathbb{R}^m$ taking values in $\mathbb{A}_{\boldsymbol{y}}$ is an HMP if there exist a Markov process $\{\boldsymbol{x}_t\}$ in $\mathbb{R}^n$ taking values in $\mathbb{A}_{\boldsymbol{x}}$ such that $\{(\boldsymbol{y}_t, \boldsymbol{x}_t)\}$ is jointly Markov and $\mathbb{P}[\boldsymbol{y}_t = y_t, \boldsymbol{x}_t = x_t | \boldsymbol{y}_{t-1} = y_{t-1}, \boldsymbol{x}_{t-1} = x_{t-1}] = \mathbb{P}[\boldsymbol{y}_t = y_t, \boldsymbol{x}_t = x_t | \boldsymbol{x}_{t-1} = x_{t-1}]$ for all t.*

For HMPs, there exists an initial probability distribution $\boldsymbol{p}_0$ and transition probability matrix $P \in \mathbb{R}^{n \times n}$ defined as before, as well as a stochastic matrix $C \in \mathbb{R}^{m \times n}$, called *emission probability matrix*, such that $\mathbb{P}[\boldsymbol{y}_t = y_t | \boldsymbol{x}_t = x_t] = \boldsymbol{e}^T_{y_t} C \boldsymbol{f}_{x_t}$, where $\{\boldsymbol{e}_i\}$ is the standard basis for $\mathbb{R}^m$, and $\boldsymbol{e}_{y_t}$ represents the elementary event associated to y_t.

Definition 2 (Hidden Markov models)**.** *We define a* Hidden Markov Model (HMM) *as the couple $\theta = (P, C)$.*

The HMM θ and the initial distribution $\boldsymbol{p}_0$ completely characterize the evolution of the probability distributions, leaving n, m and the alphabets implicit. In fact, the marginal distribution evolution can be modeled by

$$\begin{cases} \boldsymbol{p}(t+1) = P\boldsymbol{p}(t) \\ \boldsymbol{q}(t) = C\boldsymbol{p}(t) \end{cases} \tag{5}$$

associated to θ and initial condition $\boldsymbol{p}(0) = \boldsymbol{p}_0$ and can then be computed as

$$\mathbb{P}_{\theta,\boldsymbol{p}_0}[\boldsymbol{y}_t = y_t] = \boldsymbol{e}_{y_t}^T C P^t \boldsymbol{p}_0.$$

Notice that we made the dependence on the HMM θ and initial distribution $\boldsymbol{p}_0$ explicit whenever necessary to distinguish distributions induced by different models.

We are ready to state the first of the problems we will address in the following sections.

Problem 1 (Single-time marginals)**.** *Given an HMM $\theta = (P, C)$ and a finite set of initial probability distributions $\mathcal{S} \subset \mathcal{D}(\mathbb{R}^n)$ find a reduced HMM $\check{\theta} = (\check{P}, \check{C})$ of dimension $d \leqslant n$ and a linear map $\Psi[\cdot] : \mathcal{S} \to \mathcal{D}(\mathbb{R}^d)$, $\boldsymbol{p}_0 \mapsto \check{\boldsymbol{p}}_0$ such that*

$$\mathbb{P}_{\theta,\boldsymbol{p}_0}[\boldsymbol{y}_t = y_t] = \mathbb{P}_{\check{\theta},\Psi[\boldsymbol{p}_0]}[\boldsymbol{y}_t = y_t]$$

for all $t \geqslant 0$ and for any initial conditions $\boldsymbol{p}_0 \in \mathcal{S}$.

The second problem that we address targets multi-time probability distributions.

Problem 2 (Multi-time marginals)**.** *Given an HMM $\theta = (P, C)$ and a finite set of initial probability distributions $\mathcal{S} \subset \mathcal{D}(\mathbb{R}^n)$ find a reduced HMM $\check{\theta} = (\check{P}, \check{C})$ of dimension $d \leqslant n$ and a linear map $\Psi[\cdot] : \mathcal{S} \to \mathcal{D}(\mathbb{R}^d)$, $\boldsymbol{p}_0 \mapsto \check{\boldsymbol{p}}_0$ such that*

$$\mathbb{P}_{\theta,\boldsymbol{p}_0}[\boldsymbol{y}_{0:k} = y_{0:k}] = \mathbb{P}_{\check{\theta},\Psi[\boldsymbol{p}_0]}[\boldsymbol{y}_{0:k} = y_{0:k}]$$

for all sequences of the output process $y_{0:k}$ and for all initial conditions $\boldsymbol{p}_0 \in \mathcal{S}$.

Remark 2. Although Problem 2 is more natural than Problem 1 for the typical HMM setting, the latter is also interesting in particular cases, which include efficiently simulating an unmeasured stochastic evolution, and reproducing the mixing properties of lifted chains with more compact models. In fact, while we derive solutions of Problem 2 that are also solutions for Problem 1, the size of the effective multi-time reduced model is going to be in general significantly larger, as it must exactly reproduce all transition probabilities – see also Proposition 3 below.

Remark 3. As we pointed out before, in Problems 1 and 2 we have assumed that $\mathcal{S}$ is a finite set. This assumption can be relaxed since, as we show below, the proposed solution works for any initial condition contained in $\operatorname{span}\{\mathcal{S}\}$. For this reason, when dealing with linear spaces of initial conditions one can study the problem where $\mathcal{S}$ are the generators of the set.

IV. Preliminary results: A system theoretic viewpoint

Finding minimal realization of linear systems has been a central problem in control and system theory, for which well-established solutions are available. Nonetheless, when positivity is required on the reduced model, the minimal realization problem is, to the best of our knowledge, still open. In this section, we review some existing results, and extend and adapt them so that they can be used in our scenarios. In particular, we shall allow for non-minimal realizations in order to guarantee their positivity.

A. Single time-marginal problem

Let us start by considering model (5) with initial condition $\boldsymbol{p}_0 \in \mathcal{S}$. Let us define the *non-observable subspace* as:

$$\mathcal{N} := \ker \begin{bmatrix} C \\ CP \\ \vdots \\ CP^{n-1} \end{bmatrix}. \tag{6}$$

The subspace $\mathcal{N}$ can be characterized as the largest P-invariant subspace contained in $\ker C$ [31], [32]. In the case of HMM the non-observable subspace has another useful property.

Lemma 2. *For all $\boldsymbol{x} \in \mathcal{N}$ it holds $\mathbf{1}^T\boldsymbol{x} = 0$.*

Proof. From the definition of non-observable space, we have that $\boldsymbol{x} \in \mathcal{N}$ if and only if $CP^t\boldsymbol{x} = \mathbf{0}$ for all $t \geqslant 0$. If we then left-multiply by $\mathbf{1}^T$ on both sides we obtain $\mathbf{1}^T CP^t \boldsymbol{x} = \mathbf{1}^T P^t \boldsymbol{x} = \mathbf{1}^T\boldsymbol{x} = \mathbf{1}^T\mathbf{0} = 0$ for all $\boldsymbol{x} \in \mathcal{N}$. □

Next, define $\mathcal{R}$ as the smallest linear space that contains all probability distributions $\boldsymbol{p}(t)$ generated by the HMM for every $t \geqslant 0$ and any initial distribution $\boldsymbol{p}_0 \in \mathcal{S}$:

$$\mathcal{R} := \operatorname{span}\{P^t\boldsymbol{p}_0 | t \geqslant 0, \boldsymbol{p}_0 \in \mathcal{S}\}. \tag{7}$$

Remark 4. The space $\mathcal{R}$ is, in fact, the reachable subspace of a state-space model in the typical form:

$$\begin{cases} \tilde{\boldsymbol{p}}(t+1) = P\tilde{\boldsymbol{p}}(t) + B\boldsymbol{u}(t) \\ \boldsymbol{q}(t) = C\tilde{\boldsymbol{p}}(t) \end{cases} \tag{8}$$

where $B \in \mathbb{R}^{n\times|\mathcal{S}|}$ is a matrix whose columns are the initial conditions in $\mathcal{S}$. This model reproduces the trajectories of (5) for inputs corresponding to discrete impulses. The non-observable subspaces of (5) and (8) are the same, and the subspace $\mathcal{R}$ coincides with the reachable subspace of model (8), and thus shares the same properties: $\mathcal{R}$ is the smallest P-invariant subspace that contains $\operatorname{span}\{\mathcal{S}\}$. In light of this, we call the $\mathcal{R}$ defined above the *reachable subspace.*

Lastly, we call *effective subspace* $\mathcal{E}$ any subspace

$$\mathcal{E} \subseteq \mathbb{R}^n \text{ such that } (\mathcal{R} \cap \mathcal{N}) \oplus \mathcal{E} = \mathcal{R} \tag{9}$$

namely, a completion of the intersection $\mathcal{R} \cap \mathcal{N}$ to the reachable subspace $\mathcal{R}$. Notice that the choice of $\mathcal{E}$ is not unique, in fact, any representative of the quotient space $\mathcal{R}/(\mathcal{R} \cap \mathcal{N})$ is a suitable candidate for this choice. The most natural choice for the effective subspace is of course the orthogonal complement (with respect to the natural inner product) of $\mathcal{R} \cap \mathcal{N}$ in $\mathcal{R}$, which we shall denote with $\mathcal{E}_\perp$. Any other orthogonal complement, with respect to a modified inner product, would also be a suitable choice for $\mathcal{E}$.

Remark 5. The situation is reminiscent of the classical linear state-space analysis proposed by Rosenbrock [33], where all representatives of the quotient space $\mathcal{R}/(\mathcal{R}\cap\mathcal{N})$ are equivalent and associated to *minimal realizations*. In our case, however, $\mathcal{E}$ needs to be further extended to ensure positivity of the reduced dynamical matrix, a notion that depends on the chosen reference basis. For this reason, not all choices of the effective subspace are equivalent. While we will show how the algorithm we propose works with any choice of the effective subspace, in Section VII we will argue that the choice of the representative $\mathcal{E}$ of $\mathcal{R}/(\mathcal{R}\cap\mathcal{N})$ plays a key role in constructing an optimal reduction.

As we just recalled, the restriction of model (8) to (any) $\mathcal{E}$ corresponds to a minimal realization (yet not necessarily positive or stochastic). The next corollary shows that the same reduction method can be used for the HMM (5), while also allowing for extensions of the effective space. In this case, the minimality of the linear realization may be lost, but will later allow us to enforce positivity. The proof relies on a related result for general autonomous switching systems we present in detail in Appendix A.

Corollary 2. *Consider an effective subspace $\mathcal{E}$ for the HMM* (5) *and a subspace $\mathcal{V}$ such that $\mathcal{E}\subseteq\mathcal{V}$ with $d=\dim(\mathcal{V})$. Let $\Pi_{\mathcal{V}}$ be the orthogonal projection onto $\mathcal{V}$ with respect to an arbitrary inner product $\langle\cdot,\cdot\rangle$, such that $\Pi_{\mathcal{V}}(\mathcal{R}\cap\mathcal{N})\subseteq\mathcal{R}\cap\mathcal{N}$. Let $R:\mathbb{R}^n\to\mathbb{R}^d$ and $J:\mathbb{R}^d\to\mathcal{V}$ be two (non-square) factors of the orthogonal projection, $\Pi_{\mathcal{V}}=JR$.*

Define the reduced model $(\breve{P},\breve{C})=(RPJ,CJ)$ and the map $\breve{\boldsymbol{p}}_0=R\boldsymbol{p}_0$, for all $\boldsymbol{p}_0\in\mathcal{S}$. Then the linear systems associated with the pairs (P,C) and $(\breve{P},\breve{C})$ reproduce the same marginal distribution at a specific time instant, i.e.

$$CP^t\boldsymbol{p}_0=\breve{C}\breve{P}^t\breve{\boldsymbol{p}}_0$$

for all $t\geqslant 0$ and any initial condition $\boldsymbol{p}_0\in\operatorname{span}\{\mathcal{S}\}$.

Proof. This result follows from the application of Theorem 4 reported in Appendix A with only one $F_i=P$, $H=C$ and $\boldsymbol{x}(0)=\boldsymbol{p}_0$. □

In the following sections, we shall construct $\mathcal{V}$ so that the reduction is also an HMM.

B. Multi-time marginal problem

For the multi-time marginal problem, following on the seminal work [19], we will consider C for our initial model to have only zero or one entries, i.e. $C\in\{0,1\}^{m\times n}$. The assumption is not restrictive, as any Hidden Markov Process admits a realization with C of this type [13, Theorem 9.4].

The minimal reduction of the system producing the multi-time distribution can be obtained along the same lines. Calculating the probability of a sequence of events is however more involved: [19, Lemma 1] provides a closed form for such a computation. We report it here for completeness.

Lemma 3. *Given an HMM θ and an initial probability distribution $\boldsymbol{p}_0$, the probability of a sequence of outcomes is given by*

$$\mathbb{P}_{\theta,\boldsymbol{p}_0}[\boldsymbol{y}_{0:k}=y_{0:k}]=\mathbf{1}^TP_C^{y_{0:k}}\boldsymbol{p}_0$$

where

$$P_C^{y_{0:k}}=P_C^{y_{1:k}}\operatorname{diag}(\boldsymbol{e}_{y_0}^TC),$$

$$P_C^{y_{1:k}}=\prod_{i=k}^{1}P_C^{y_i},\quad P_C^{y_i}=\operatorname{diag}(\boldsymbol{e}_{y_i}^TC)P,\quad i>0.$$

In the above lemma, the multiplication by the diagonal matrices $\operatorname{diag}(\boldsymbol{e}_{y_t}^TC)$ accounts for the conditioning of $\boldsymbol{p}_t$ on the outcome $\boldsymbol{y}_t=y_t$. Without the latter we obtain the formulas for the single marginals.

In order to exploit system-theoretic tools, it is useful to write the probability of a sequence of outcomes as the output of a dynamical model. The dynamical model we are going to present next resembles the "observables representations of HMMs" described in [34]. Call $\psi(t)=\mathbb{P}(\boldsymbol{y}_{0:t}=y_{0:t})$. We can obtain its evolution as the output of a discrete-time, autonomous, switching, linear system described by

$$\begin{cases}\boldsymbol{\phi}(t+1)=P_C^{y_t}\boldsymbol{\phi}(t)\\ \boldsymbol{\psi}(t)=\mathbf{1}^T\boldsymbol{\phi}(t)\end{cases}\tag{10}$$

with initial condition $\boldsymbol{\phi}_{y_0}(1)=\operatorname{diag}(\boldsymbol{e}_{y_0}^TC)\boldsymbol{p}_0$, $P_C^{y_t}$ defined as in the previous lemma and where $\boldsymbol{\psi}(t)$ represents the probability associated to the sequence of events $y_{0:t}$. Clearly, the output $\boldsymbol{\psi}(t)$ depends on the sequence of $P_C^{y_t}$, which in turn depends on the outcomes of the sequence. The output at any time $k>0$ can be computed as $\boldsymbol{\psi}(y_{0:k})=\mathbf{1}^T\prod_{i=k}^{1}P_C^{y_i}\boldsymbol{\phi}_{y_0}(1)$, while for $l=0$ we have $\boldsymbol{\psi}(y_0)=\mathbf{1}^T\boldsymbol{\phi}_{y_0}(1)$, thus recovering the formulas of the lemma.

Given a finite set $\mathcal{S}$ of initial distributions of interest, the corresponding set of initial conditions for this model is $\Phi=\bigcup_{y_0}\operatorname{diag}(\boldsymbol{e}_{y_0}^TC)\mathcal{S}$.

Following the approach of [19] in a system-theoretic setting, we can define the reachable, non-observable, and effective subspaces for the multi-time problem. To avoid confusion with the previous definitions, we call these the *conditioned* subspaces and denote them with a $\mathcal{C}$ subscript. Given an HMM (P,C) and a set of initial conditions $\mathcal{S}$ we define the *conditioned non-observable subspace* as:

$$\mathcal{N}_{\mathcal{C}}:=\{\boldsymbol{v}\in\mathbb{R}^n|\mathbf{1}^TP_C^{y_{0:l}}\boldsymbol{v}=0,\quad\forall y_{0:l}\},\tag{11}$$

and the *conditioned reachable subspace* as

$$\mathcal{R}_{\mathcal{C}}:=\operatorname{span}\{P_C^{y_{0:l}}\boldsymbol{p}_0,\quad\forall y_{0:l},\quad\forall\boldsymbol{p}_0\in\mathcal{S}\}.\tag{12}$$

We can then define the *conditioned effective subspace* $\mathcal{E}_{\mathcal{C}}$ as a completion of the intersection $\mathcal{R}_{\mathcal{C}}\cap\mathcal{N}_{\mathcal{C}}$ to the conditioned reachable subspace $\mathcal{R}_{\mathcal{C}}$, i.e. $\mathcal{E}_{\mathcal{C}}\bigoplus(\mathcal{R}_{\mathcal{C}}\cap\mathcal{N}_{\mathcal{C}})=\mathcal{R}_{\mathcal{C}}$. As before, the choice of $\mathcal{E}_{\mathcal{C}}$ is not unique, as any representative of the quotient space $\mathcal{R}_{\mathcal{C}}/(\mathcal{R}_{\mathcal{C}}\cap\mathcal{N}_{\mathcal{C}})$ is a suitable choice.

The properties of these spaces have been described in [19, Lemma 3, Section 3]. We recap them in the following Lemma for the reader's convenience.

Lemma 4. *$\mathcal{N}_{\mathcal{C}}$ and $\mathcal{R}_{\mathcal{C}}$ are P-invariant, $\operatorname{diag}(\boldsymbol{e}_i^TC)$-invariant for all i and thus, $P_C^{y_{0:l}}$-invariant for all sequences $y_{0:l}$.*

A result similar to Cayley-Hamilton Theorem holds and lets us compute the spaces by using a finite number of generators:

$$\mathcal{N}_{\mathcal{C}}=\{\boldsymbol{v}\in\mathbb{R}^n|\mathbf{1}^TP_C^{y_{0:l}}\boldsymbol{v}=0,\quad\forall y_{0:l}\ \textit{s.t.}\ l<n\},\tag{13}$$

$$\mathcal{R}_{\mathcal{C}}=\operatorname{span}\{P_C^{y_{0:l}}\boldsymbol{p}_0,\quad\forall\boldsymbol{p}_0\in\mathcal{S},\quad\forall y_{0:l}\ \textit{s.t.}\ l<n\}.\tag{14}$$

We can then notice that $\mathcal{N}_{\mathcal{C}}$ is the non-observable subspace of model (10) see e.g. [35], $\mathcal{R}_{\mathcal{C}}$ is its reachable subspace and $\mathcal{E}_{\mathcal{C}}$ is its effective subspace. The second statement holds trivially, while the first holds because $\mathcal{N}_{\mathcal{C}}$ is $\operatorname{diag}(\boldsymbol{e}_i^T C)$-invariant for all i. The third follows by combining the first two.

An useful property of the propagator $P_C^{y_{0:k}}$ is proved in the following Lemma.

Lemma 5. *The sum over all sequences $y_{0:k}$ of the same length k of $P_C^{y_{0:k}}$ is equal to the k-th power of P, i.e.*

$$\sum_{y_{0:k}} P_C^{y_{0:k}} = P^k$$

Proof. The statement is simply proved by observing that $\sum_{y_i} \operatorname{diag}(\boldsymbol{e}_{y_i} C) = I$ for all i and summing over all the possible strings $y_{0:k}$, starting from the first character. □

The next Proposition shows that, in general, solving the multi-time marginal case requires a larger model than the single-time case defined before.

Proposition 3. *It holds that*

$$\ker C \supseteq \mathcal{N} \supseteq \mathcal{N}_{\mathcal{C}},$$

$$\mathcal{S} \subseteq \mathcal{R} \subseteq \mathcal{R}_{\mathcal{C}},$$

and also

$$\mathcal{E} \subseteq \mathcal{E}_{\mathcal{C}}.$$

The proof of this Lemma can be found in Appendix B.

Remark 6. This result clarifies the relation as well as the distinction between problems 2 and 1. In fact, this Proposition shows that, at least in principle, there could be a larger reduction if we are only interested in describing only the evolution of the marginal distribution at a specific time. Moreover, the conditioned effective subspace contains the effective subspace, thus showing, due to Corollary 2, that a solution for Problem 2 is also a solution for Problem 1.

We now propose a class of effective model reductions for the multi-time marginal problem.

Corollary 3. *Consider any conditioned effective subspace $\mathcal{E}_{\mathcal{C}}$ and subspace $\mathcal{V}$ such that $\mathcal{E}_{\mathcal{C}} \subseteq \mathcal{V}$ with $d = \dim(\mathcal{V})$, and let $\Pi_{\mathcal{V}}$ be the orthogonal projection onto $\mathcal{V}$ with respect to an inner product $\langle \cdot, \cdot \rangle$, such that $\Pi_{\mathcal{V}}(\mathcal{R}_{\mathcal{C}} \cap \mathcal{N}_{\mathcal{C}}) \subseteq \mathcal{R}_{\mathcal{C}} \cap \mathcal{N}_{\mathcal{C}}$. Let $R : \mathbb{R}^n \to \mathbb{R}^d$ and $J : \mathbb{R}^d \to \mathcal{V}$ be two (non-square) factors of the orthogonal projection, $\Pi_{\mathcal{V}} = JR$.*

Let then consider the reduced model $(\{\check{P}_C^{y_i}\}, \mathbf{1}_m^T) = (\{RP_C^{y_i}J\}, \mathbf{1}_n^T J)$ and the map $\check{\boldsymbol{\phi}}(1) = R\boldsymbol{\phi}(1)$ for all $\boldsymbol{\phi}(1) \in \Phi$. Then the two models described by equations (10) *and denoted by the couples $(\{P_C^{y_i}\}, \mathbf{1}_n^T)$ and $(\{\check{P}_C^{y_i}\}, \mathbf{1}_m^T)$ reproduce the same probability of a sequence of outcomes, i.e.*

$$\mathbf{1}_n^T \prod_{j=k}^{0} P_C^{y_j} \boldsymbol{\phi}(1) = \mathbf{1}_m^T \prod_{j=k}^{0} \check{P}_C^{y_j} \check{\boldsymbol{\phi}}(1)$$

for any sequence $y_{0:k}$ and any initial condition $\boldsymbol{\phi}(1) \in \operatorname{span}\{\Phi\}$.

Proof. This result follows from the application of Theorem 4 reported in Appendix A with $F_i = \operatorname{diag}(\boldsymbol{e}_{y_i}^T C)P$, $H = \mathbf{1}^T$, $\boldsymbol{x}(0) = \operatorname{diag}(\boldsymbol{e}_{y_0}^T C)\boldsymbol{p}_0$. □

Remark 7. At this point one may notice that Corollary 3 provides a reduction for model (10) which includes the conditioning as part of the dynamics and in general may not translate directly into a reduction of (5) in the HMM form $(\check{P}, \check{C}, \check{\mathcal{S}})$. Nevertheless, we anticipate here that the algorithm we propose in Section VI for the multi-time case provides a model in HMM form, thanks to Lemma 4. Thanks to Proposition 3 and Corollary 2 the obtained model also reproduces the single-time marginals.

Remark 8. The two main results in this section, Corollary 2 and 3, as well as the underlying Theorem 4 shown in Appendix A, have been stated for time-invariant dynamics for sake of simplicity. While it is possible to generalize the analysis to time-dependent systems, in that case, Cayley-Hamilton-type results do not apply and consequently, the computation of reachable and non-observable spaces may become impractical.

V. Single-time solution

In this section, we illustrate how to obtain solutions to Problem 1 appropriately choosing $\mathcal{V}$ in Corollary 2. We first discuss the intuition behind the method, next we present the proposed solution in form of a parametric algorithm, and prove that, under appropriate constraints, the algorithm indeed provides a solution. Finally, in Section VII, propose a way to choose the relevant parameters.

A. Intuition

The core idea behind the method stems from the fact that in order to define an HMM we need an underlying probability space and, as we have seen in Section II, any probability space is associated to an algebra. This directly suggests that, in order to preserve the (stochastic) HMM structure in the reduction it is natural to restrict the model to an algebra whose dual contains the effective subspace, and then use the dual of the conditional expectation to obtain a stochastic reduction.

More in detail, consider the two stochastic reduction matrices R and J obtained in Section II-B as factors of the dual of a conditional expectation $\mathbb{E}_{|\mathscr{A},\boldsymbol{p}}^T$, which is an orthogonal projection onto $\boldsymbol{p} \wedge \mathscr{A}$ with respect to the inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{p}^{-1}}$. Then, according to Corollary 2 we know that as long as $\mathcal{E} \subseteq \mathcal{V} = \boldsymbol{p} \wedge \mathscr{A}$, and $\mathbb{E}_{|\mathscr{A},\boldsymbol{p}}^T$ leaves $\mathcal{R} \cap \mathcal{N}$ invariant, then the reduced model reproduces the same marginal distribution as the original one.

In order to choose $\mathscr{A}$ such that $\mathcal{E} \subseteq \boldsymbol{p} \wedge \mathscr{A}$ we can $\wedge$-multiply left and right by $\boldsymbol{p}^{-1}$ obtaining $\boldsymbol{p}^{-1} \wedge \mathcal{E} \subseteq \mathscr{A}$. Let $\operatorname{alg}(\mathcal{X})$ denote the minimal sub-algebra of $\mathbb{R}^n$ containing the set $\mathcal{X}$. Then, if we define $\mathscr{A} := \operatorname{alg}(\boldsymbol{p}^{-1} \wedge \mathcal{E})$, we ensure that $\mathcal{E} \subseteq \boldsymbol{p} \wedge \mathscr{A}$ is satisfied and that the reduced model reproduces the same marginal at a single time.

To make this idea more concrete, we provide a simple illustrative example, which also highlights the importance of choosing the distribution $\boldsymbol{p}$ to be used in $\mathbb{E}_{|\mathscr{A},\boldsymbol{p}}^T$.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2023.3279209

Example 1. Let us consider the following HMM:

$$P = \begin{bmatrix} 2/5 & 0 & 1/5 \\ 0 & 2/5 & 1/5 \\ 3/5 & 3/5 & 3/5 \end{bmatrix}, \quad \mathcal{S} = \left\{ \begin{bmatrix} 1/5 \\ 1/5 \\ 3/5 \end{bmatrix} \right\}$$

$$C = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Notice that $\boldsymbol{p}_0$ is an equilibrium, $P\boldsymbol{p}_0 = \boldsymbol{p}_0$ thus the output distribution is equal to $\boldsymbol{q}(t) = \begin{bmatrix} 2/5 & 3/5 \end{bmatrix}^T$, $\forall t \geqslant 0$. We can then compute the following.

$$\mathcal{R} = \operatorname{span}\left\{ \begin{bmatrix} 1/5 \\ 1/5 \\ 3/5 \end{bmatrix} \right\}, \quad \mathcal{N} = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \right\}$$

and $\mathcal{R} \cap \mathcal{N} = \operatorname{span}\{\mathbf{0}\}$ and we can thus choose $\mathcal{E} = \mathcal{R}$. If we then choose $\boldsymbol{p} = \mathbf{1}$ we obtain

$$\mathscr{A} = \operatorname{alg}(\mathcal{R}) = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

and thus the relative factors of the dual of the conditional expectation are

$$R = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad J = \begin{bmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 0 & 1 \end{bmatrix}$$

and the associated reduced HMM is

$$\check{P} = \begin{bmatrix} 2/5 & 2/5 \\ 3/5 & 3/5 \end{bmatrix}, \quad \check{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \check{\boldsymbol{p}}_0 = \begin{bmatrix} 2/5 \\ 3/5 \end{bmatrix}$$

which correctly reproduces the output marginal distribution $\boldsymbol{q}(t) = \begin{bmatrix} 2/5 & 3/5 \end{bmatrix}^T$, $\forall t \geqslant 0$.

On the other hand, if we were to choose $\boldsymbol{p} = \boldsymbol{p}_0$ we would obtain a different result. In fact, in that case, we have

$$\mathscr{A} = \operatorname{alg}(\boldsymbol{p}^{-1} \wedge \mathcal{R}) = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

and thus the relative factors of the dual of the conditional expectation are $R = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$, and $J = \begin{bmatrix} 1/5 & 1/5 & 3/5 \end{bmatrix}^T$ and the associated reduced HMM is $\check{P} = 1$, $\check{C} = \begin{bmatrix} 2/5 & 3/5 \end{bmatrix}$ and $\boldsymbol{p}_0 = 1$ which also reproduces the output marginal distribution and is clearly minimal (optimal reduction). This shows that the choice of $\boldsymbol{p}$ is important if we are interested in minimizing the dimension of the reduced model.

B. Proposed solution

We now formalize the proposed method to solve Problem 1 in the following Algorithm. Let $\Gamma(\mathcal{R}, \mathcal{N})$ be a map that selects an effective space $\mathcal{E}$ given some $\mathcal{R}, \mathcal{N}$.

Notice that this algorithm depends, in addition to its inputs, on two parameters: the first one, $\boldsymbol{p}$, is a positive vector; the second one, is the map Γ that selects the effective subspace. We will discuss more in detail the choice of the effective subspace in Section VII.

We are finally ready to prove that Algorithm 1 solves the single-time marginal problem.

Algorithm 1: HMM reduction for problem 1

```
Input      : $(P, C)$, $\mathcal{S}$.
Parameters: $\boldsymbol{p}$, $\Gamma$.
Compute $\mathcal{R}$ and $\mathcal{N}$ using equations (7) and (6);
Compute $\mathcal{E} = \Gamma(\mathcal{R}, \mathcal{N})$;
Compute $\mathscr{A} := \operatorname{alg}(\boldsymbol{p}^{-1} \wedge \mathcal{E})$;
Compute $\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$ using equation (3) ;
If $\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}(\mathcal{R} \cap \mathcal{N}) \nsubseteq \mathcal{R} \cap \mathcal{N}$: redefine
  $\mathscr{A} := \operatorname{alg}(\boldsymbol{p}^{-1} \wedge \mathcal{R})$ and recompute $\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$ ;
Compute the factors $R$ and $J$ of $\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$ with the
  definition given in equation (4);
Output     : $(\check{P}, \check{C}) = (RPJ, CJ)$ and $R$.
```

Theorem 1. *For any choice of $\mathcal{E}$ and $\boldsymbol{p}$ positive i.e. $\boldsymbol{p}_i > 0$ $\forall i$, Algorithm 1 provides a solution to Problem 1.*

Proof. To prove the statement we have to prove that: i) The reduced model $\check{\theta} = (\check{P}, \check{C})$ and the linear map R provide the same marginal distribution at any time as the original model; ii) the reduced model $\check{\theta}$ is an HMM, and $R\boldsymbol{p}_0$ is a probability vector.

We shall start by proving the first point. We do so leveraging Corollary 2. First of all, we have that, for any vector $\boldsymbol{p}$ such that $\boldsymbol{p}_i > 0$ for all i the inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{p}}$ is positive-definite and thus well defined. Moreover, by definition of the algebra $\mathscr{A}$, we have that, for any choice of the effective subspace $\mathcal{E}$ it holds $\mathcal{E} \subseteq \boldsymbol{p} \wedge \mathscr{A}$ so, by choosing $\mathcal{V} = \boldsymbol{p} \wedge \mathscr{A}$, and using the restriction and injection map defined in equation (4), i) follows from Corollary 2 if case $\mathbb{E}^T_{\mathscr{A},\boldsymbol{p}}(\mathcal{R} \cap \mathcal{N}) \subseteq \mathcal{R} \cap \mathcal{N}$.

If $\mathbb{E}^T_{\mathscr{A},\boldsymbol{p}}(\mathcal{R} \cap \mathcal{N}) \nsubseteq \mathcal{R} \cap \mathcal{N}$, pick $\tilde{\mathcal{N}} = \{\mathbf{0}\}$ so that $\mathcal{R} \cap \tilde{\mathcal{N}} = \{0\}$ and Theorem 4 applies with $\mathcal{V} = \operatorname{alg}(\boldsymbol{p} \wedge \mathcal{R})$.

Regarding ii) we have that, if $\mathscr{A}$ is unital, then Proposition 2 ensures that J and R are stochastic and thus RPJ and CJ are stochastic and $R\boldsymbol{p}_0$ is a probability vector for any $\boldsymbol{p}_0$ probability vector. If $\mathscr{A}$ is not unital, because of Corollary 1 we have, that J is stochastic (and thus CJ is stochastic) but R is only stochastic over $\operatorname{supp}(\mathscr{A})$, i.e. $\mathbf{1}_d^T R = \mathbf{1}^T_{\operatorname{supp}(\mathscr{A})}$. We next show that this condition is sufficient to show that the reduced model is stochastic.

We shall first notice that $\operatorname{supp}(\mathcal{E}) = \operatorname{supp}(\mathscr{A}) \subsetneq \mathbb{R}^n$. Let assume that $\dim(\operatorname{supp}(\mathcal{E})) = k$. Then we can consider a permutation (that is a double-stochastic change of basis) T such that $T\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}' \mid \mathbf{0}^T_{n-k} \end{bmatrix}^T$ for all $\boldsymbol{x} \in \mathcal{E}$, with $\boldsymbol{x}' \in \mathbb{R}^k$. Then, since $\mathcal{E}$ is P-invariant

$$\left[\begin{array}{c} \boldsymbol{x}' \\ \hline \mathbf{0}_{n-k} \end{array}\right] = \underbrace{\left[\begin{array}{c|c} P_{11} & P_{12} \\ \hline P_{21} & P_{22} \end{array}\right]}_{TPT^T} \left[\begin{array}{c} \boldsymbol{x}' \\ \hline \mathbf{0}_{n-k} \end{array}\right] \in \mathcal{E}$$

and thus $P_{21} = 0$. This shows that $\operatorname{supp}(\mathcal{E})$ is P-invariant. Since P, T and T^T are stochastic, TPT^T is also stochastic. This implies that $\mathbf{1}_k^T P_{11} = \mathbf{1}_k^T$. Then, it holds that $\mathbf{1}^T_{\operatorname{supp}(\mathscr{A})} T^T T P T^T =$

$$\left[\begin{array}{c|c} \mathbf{1}_k^T & \mathbf{0}^T_{n-k} \end{array}\right] \left[\begin{array}{c|c} P_{11} & P_{12} \\ \hline 0 & P_{22} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{1}_k^T & \mathbf{0}^T_{n-k} \end{array}\right]$$

or, in other words $\mathbf{1}^T_{\mathrm{supp}(\mathscr{A})}P = \mathbf{1}^T_{\mathrm{supp}(\mathscr{A})}$. We can also verify that $\breve{P}$ is stochastic by verifying the following chain of equivalences: $\mathbf{1}_d^T RPJ = \mathbf{1}^T_{\mathrm{supp}(\mathscr{A})}PJ = \mathbf{1}^T_{\mathrm{supp}(\mathscr{A})}J = \mathbf{1}_d^T$ where the last equality comes from Corollary 1. Finally, to prove that $R\boldsymbol{p}_0$ is a probability vector we can observe that $\mathbf{1}_n^T\boldsymbol{p}_0 = \mathbf{1}^T\Pi_{\mathcal{E}}\boldsymbol{p}_0 + \underbrace{\mathbf{1}^T\Pi_{\mathcal{R}\cap\mathcal{N}}\boldsymbol{p}_0}_{=\mathbf{0}} = 1$ and then re-use the reasoning above. □

Remark 9. In the proof of Theorem 1 we stated that a positive vector $\boldsymbol{p}$ is necessary to have a well-defined inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{p}}$. This assumption, however, can be relaxed to the following: $\boldsymbol{p}$ is positive over $\mathrm{supp}(\mathcal{E}) = \mathrm{supp}(\mathscr{A})$, i.e. $\boldsymbol{p}_i > 0$ for all i such that $\boldsymbol{e}_i^T\boldsymbol{x} \neq 0$ for some $\boldsymbol{x} \in \mathcal{E}$. This is due to the fact that the values of $\boldsymbol{p}$ where $\mathcal{E}$ has no support has no role in the projection.

Although such $\boldsymbol{p}$ defines a positive semi-definite inner product over $\mathbb{R}^n$, it provides a positive definite inner product over $\mathrm{supp}(\mathcal{S})$ and this is sufficient to define the orthogonal projection onto $\mathscr{A}$. Consider, for example, the following case: assume $\mathrm{supp}(\mathscr{A}) \subsetneq \mathbb{R}^n$ then let $\boldsymbol{p}_s$ be a positive vector over the $\mathrm{supp}(\mathscr{A})$, $\boldsymbol{p}_n$ be a positive vector over the remaining support, i.e. s.t. $\boldsymbol{p} := \boldsymbol{p}_s + \boldsymbol{p}_n$, $\mathrm{supp}(\boldsymbol{p}) = \mathbb{R}^n$. We can then notice that $\boldsymbol{p} \wedge \boldsymbol{x} = \boldsymbol{p}_s \wedge \boldsymbol{x}$ and $\langle\boldsymbol{y},\boldsymbol{x}\rangle_{\boldsymbol{p}} = \langle\boldsymbol{y},\boldsymbol{x}\rangle_{\boldsymbol{p}_s}$ for all $\boldsymbol{x} \in \mathrm{supp}(\mathscr{A})$ and $\boldsymbol{y} \in \mathbb{R}^n$. This implies that

$$\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}} = \sum_{j=1}^{d}\frac{(\boldsymbol{p}\wedge\boldsymbol{a}_j)\boldsymbol{a}_j^T}{\langle\boldsymbol{a}_j,\boldsymbol{a}_j\rangle_{\boldsymbol{p}}} = \sum_{j=1}^{d}\frac{(\boldsymbol{p}_s\wedge\boldsymbol{a}_j)\boldsymbol{a}_j^T}{\langle\boldsymbol{a}_j,\boldsymbol{a}_j\rangle_{\boldsymbol{p}_s}} = \mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}_s}.$$

The role of the positivity of $\boldsymbol{p}$ will be further discussed in Section VII.

VI. Multi-time solution

The solution of Problem 2 follows the same ideas presented in the previous section. In fact, the algorithm we propose to solve Problem 2 is identical to the previous algorithm but for the involved subspaces. We now present our proposed method to solve Problem 2. This method takes the form of the following Algorithm, where Γ is defined as in the previous section.

Algorithm 2: HMM reduction for problem 2

```
Input      : $(P, C)$, $\mathcal{S}$.
Parameters: $\boldsymbol{p}$, $\Gamma$.
Compute $\mathcal{R}_{\mathcal{C}}$ and $\mathcal{N}_{\mathcal{C}}$ using equations (14) and (13);
Compute $\mathcal{E}_{\mathcal{C}} = \Gamma(\mathcal{R}_{\mathcal{C}}, \mathcal{N}_{\mathcal{C}})$;
Compute $\mathscr{A}_{\mathcal{C}} = \mathrm{alg}(\boldsymbol{p}^{-1} \wedge \mathcal{E}_{\mathcal{C}})$;
Compute $\mathbb{E}^T_{|\mathscr{A}_{\mathcal{C}},\boldsymbol{p}}$ using equation (3);
If $\mathbb{E}^T_{|\mathscr{A}_{\mathcal{C}},\boldsymbol{p}}(\mathcal{R}_{\mathcal{C}} \cap \mathcal{N}_{\mathcal{C}}) \nsubseteq \mathcal{R}_{\mathcal{C}} \cap \mathcal{N}_{\mathcal{C}}$: redefine
  $\mathscr{A}_{\mathcal{C}} := \mathrm{alg}(\boldsymbol{p}^{-1} \wedge \mathcal{R}_{\mathcal{C}})$ and recompute $\mathbb{E}^T_{|\mathscr{A}_{\mathcal{C}},\boldsymbol{p}}$;
Compute the factors $R$ and $J$ of $\mathbb{E}^T_{|\mathscr{A}_{\mathcal{C}},\boldsymbol{p}}$ with the
  definition given in equation (4);
Output     : $(\breve{P}, \breve{C}) = (RPJ, CJ)$ and $R$
```

We are finally ready to prove that Algorithm 2 solves the multi-time marginal problem.

Theorem 2. *For any choice of $\mathcal{E}_{\mathcal{C}}$ and $\boldsymbol{p}$ positive, i.e. $\boldsymbol{p}_i > 0$ $\forall i$, Algorithm 2 provides a solution to Problem 2.*

Proof. The proof of this theorem follows the lines of the proof of Theorem 1. In fact, the proof of the fact that the reduced HMM $\breve{\theta}$ is stochastic and $R\boldsymbol{p}_0$ is a probability vector is identical to the one given in 1. The only difference in the two proofs regards proof of the fact that the reduced model $\breve{\theta}$ with initial condition $R\boldsymbol{p}_0$ provides the same probability of a sequence of events as the model θ with initial condition $\boldsymbol{p}_0$.

From Corollary 3 we have that $(R\mathrm{diag}(\boldsymbol{e}_i^T C)PJ, \mathbf{1}^T J)$ with initial condition $R\mathrm{diag}(\boldsymbol{e}_i^T C)\boldsymbol{p}_0$ generates the same probability as the original model. Since $\mathcal{R}_{\mathcal{C}}$ and $\mathcal{N}_{\mathcal{C}}$ are both P and $\mathrm{diag}(\boldsymbol{e}_i^T C)$-invariant, Corollary 5 applies, thus leading to the reduced HMM $\breve{\theta} = (RPJ, CJ)$ and initial conditions $R\boldsymbol{p}_0$. □

VII. Choosing the algorithm's parameters

In this section, we discuss what is the best choice of the parameters for Algorithms 1 and 2. Being the structure of the two algorithms identical, we only discuss the optimal choice of $\mathcal{E}$ and $\boldsymbol{p}$: the results can be extended directly to $\mathcal{E}_{\mathcal{C}}$. The notion of optimality is related to the dimension of the reduced system, meaning: we want to find a choice of $\mathcal{E}$ and $\boldsymbol{p}$ positive such that the reduced model returned by Algorithm 1 has minimal dimension. This is equivalent to finding $\mathcal{E}$ and $\boldsymbol{p}$ such that $\mathrm{alg}(\boldsymbol{p}^{-1} \wedge \mathcal{E})$ has minimal dimension.

A. Optimal distributions for observable HMMs

We shall start the discussion by finding the optimal choice of $\boldsymbol{p}$ assuming that an effective subspace $\mathcal{E}$ is given. Before we prove the main result of this section, we shall first state the following useful result.

Lemma 6. *Given a vector space $\mathcal{W} \subseteq \mathbb{R}^n$ with generators $\{\boldsymbol{w}_i\}$, $\mathcal{W} = \mathrm{span}\{\boldsymbol{w}_i\}$ there exists a vector $\overline{\boldsymbol{w}} := \sum_i \lambda_i \boldsymbol{w}_i$, with $\lambda_i \neq 0$ for all i and such that $\mathrm{supp}(\overline{\boldsymbol{w}}) = \mathrm{supp}(\mathcal{W})$.*

The proof of this Lemma can be found in Appendix B.

Theorem 3. *Let consider a vector space $\mathcal{W} \subseteq \mathbb{R}^n$ and a vector $\overline{\boldsymbol{w}}$ as in Lemma 6. Then there exists a unique algebra $\mathscr{A}^*$ of minimal dimension such that $\mathcal{W} \subseteq \boldsymbol{x} \wedge \mathscr{A}^*$ for some $\boldsymbol{x} \in \mathbb{R}^n$. Moreover, $\mathscr{A}^* = \mathrm{alg}(\overline{\boldsymbol{w}}^{-1} \wedge \mathcal{W})$ and it is unital over the support of $\mathcal{W}$, i.e. $\mathbf{1}_{\mathrm{supp}(\mathcal{W})} \in \mathscr{A}^*$.*

Proof. The existence of such a $\overline{\boldsymbol{w}}$ is proved in Lemma 6.

Since $\mathscr{A} = \mathbb{R}^n$ satisfies $\mathcal{W} \subseteq \boldsymbol{x} \wedge \mathscr{A}$, for all $\boldsymbol{x} \in \mathbb{R}^n$ and its possible sub-algebras are finite (corresponding to the partition of n), $\mathscr{A}^*$ exists. To prove that it is an unique solution we proceed by contradiction. Let assume that there exist two different algebras $\mathscr{A}, \mathscr{B} \subseteq \mathbb{R}^n$ with minimal dimension $\dim(\mathscr{A}) = \dim(\mathscr{B})$ and two vectors $\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}} \in \mathbb{R}^n$ such that $\mathcal{W} \subseteq \overline{\boldsymbol{a}} \wedge \mathscr{A}$ and $\mathcal{W} \subseteq \overline{\boldsymbol{b}} \wedge \mathscr{B}$. From Proposition 1 we know that $\mathscr{A} = \mathrm{span}\{\boldsymbol{a}_j\}$ and $\mathscr{B} = \mathrm{span}\{\boldsymbol{b}_j\}$ where $\{\boldsymbol{a}_i\}$ and $\{\boldsymbol{b}_i\}$ are the finest resolutions in $\mathrm{idem}(\mathscr{A})$ and $\mathrm{idem}(\mathscr{B})$ respectively. Clearly, if $\boldsymbol{a}_i = \boldsymbol{b}_i$ for all i then $\mathscr{A} = \mathscr{B}$ which yields a contradiction. Therefore, we assume that there exists an index j such that $\boldsymbol{a}_j \neq \boldsymbol{b}_i$ for all i. We can then notice that for all $\boldsymbol{v} \in \mathcal{W}$, we can write $\boldsymbol{v} = \sum_i \mu_i \overline{\boldsymbol{a}} \wedge \boldsymbol{a}_i = \sum_i \nu_i \overline{\boldsymbol{b}} \wedge \boldsymbol{b}_i$.

For j such that $\boldsymbol{a}_j \neq \boldsymbol{b}_i$ or all i we can then write

$$\boldsymbol{a}_j \wedge \boldsymbol{v} = \mu_j \overline{\boldsymbol{a}} \wedge \boldsymbol{a}_j = \sum_i \nu_i \boldsymbol{a}_j \wedge \overline{\boldsymbol{b}} \wedge \boldsymbol{b}_i.$$

The first equality implies that over the support of each $\boldsymbol{a}_j$ every $\boldsymbol{v}$ must be proportional to $\overline{\boldsymbol{a}} \wedge \boldsymbol{a}_j$. The second equality, on the other hand, due to the fact $\boldsymbol{a}_j \neq \boldsymbol{b}_j$ implies at least two of the products $\boldsymbol{a}_j \wedge \overline{\boldsymbol{b}} \wedge \boldsymbol{b}_i$ must be non-zero. In order for the nontrivial sum to be always proportional to $\overline{\boldsymbol{a}} \wedge \boldsymbol{a}_j$ it must be that the coefficients ν_i appear always in a fixed ratio. Hence, the corresponding $\boldsymbol{b}_i$ can be substituted by their sum, and still, generate the full $\mathcal{W}$ when multiplied by a suitable vector $\overline{\boldsymbol{b}}$. This shows that $\mathscr{B}$ could not be a minimal algebra unless $\boldsymbol{a}_i = \boldsymbol{b}_i$ for all i, up to a reordering.

Let then $\mathscr{A}^*$ be the unique algebra of minimal dimension such that $\mathcal{W} \subseteq \boldsymbol{x} \wedge \mathscr{A}^*$ for some $\boldsymbol{w}$. From Proposition 1 we know that $\mathscr{A}^* = \mathrm{span}\{\boldsymbol{a}_j\}$ where $\{\boldsymbol{a}_i\}$ is the finest resolution in $\mathrm{idem}(\mathscr{A}^*)$. In particular $\{\boldsymbol{a}_j\}$ forms an orthogonal basis for $\mathscr{A}^*$ and its elements have completing mutually-orthogonal supports, i.e. $\mathrm{supp}(\boldsymbol{a}_k) \perp \mathrm{supp}(\boldsymbol{a}_j)$ for $k \neq j$ and $\sum_j \boldsymbol{a}_j = \mathbf{1}_{\mathrm{supp}(\mathcal{W})}$. We can then observe that $\boldsymbol{x} \wedge \mathscr{A}^* = \mathrm{span}\{\boldsymbol{x} \wedge \boldsymbol{a}_j\}$ and that the vectors $\boldsymbol{x} \wedge \boldsymbol{a}_j$ have complementary mutually-orthogonal supports. Then for $\mathcal{W} \subseteq \boldsymbol{x} \wedge \mathscr{A}^*$ to hold it must be that $\boldsymbol{w} = \sum_j \mu_j \boldsymbol{x} \wedge \boldsymbol{a}_j$ for all $\boldsymbol{w} \in \mathcal{W}$.

By the above discussion we can write $\boldsymbol{w}_i = \sum_j \mu_j^i \boldsymbol{x} \wedge \boldsymbol{a}_j$ for each generator of $\mathcal{W}$. Notice that, for all j, $\mu_j^i \neq 0$ for at least one i. Let then use the definition of $\overline{\boldsymbol{w}}$ given in the statement and, substituting the form of the $\boldsymbol{w}_i$ we just reported we obtain $\overline{\boldsymbol{w}} = \sum_j \sigma_j \boldsymbol{x} \wedge \boldsymbol{a}_j$ with $\sigma_j = \sum_i \lambda_i \mu_j^i$. From the argument above, from the fact that $\lambda_i \neq 0$ for all i and from the fact that, by hypothesis, $\overline{\boldsymbol{w}}$ has maximal support, we have that $\sigma_j \neq 0$ for all j. Because of the structure of $\{\boldsymbol{x} \wedge \boldsymbol{a}_j\}$ we have that

$$(\boldsymbol{a}_j \wedge \overline{\boldsymbol{w}})^{-1} = \boldsymbol{a}_j \wedge \overline{\boldsymbol{w}}^{-1} = \sigma_j^{-1}(\boldsymbol{x} \wedge \boldsymbol{a}_j)^{-1} = \sigma_j^{-1} \boldsymbol{x}^{-1} \wedge \boldsymbol{a}_j,$$

and thus $\overline{\boldsymbol{w}}^{-1} = \sum_j \sigma_j^{-1} \boldsymbol{x}^{-1} \wedge \boldsymbol{a}_j$. From this we have that the vector space $\overline{\boldsymbol{w}}^{-1} \wedge \mathcal{W}$ is generated by vectors of the type

$$\overline{\boldsymbol{w}}^{-1} \wedge \boldsymbol{w}_i = \sum_{j,k} \sigma_j^{-1} \mu_k^i \boldsymbol{x}^{-1} \wedge \boldsymbol{a}_j \wedge \boldsymbol{x} \wedge \boldsymbol{a}_k = \sum_j \sigma_j^{-1} \mu_j^i \boldsymbol{a}_j.$$

This proves that $\overline{\boldsymbol{w}}^{-1} \wedge \mathcal{W} \subseteq \mathscr{A}^*$ and that any vector $\boldsymbol{v} \in \overline{\boldsymbol{w}}^{-1} \wedge \mathcal{W}$ can be written as $\boldsymbol{v} = \sum_i v_i \overline{\boldsymbol{w}}^{-1} \wedge \boldsymbol{w}_i = \sum_j \xi_j \boldsymbol{a}_j$ with $\xi_j := \sum_i v_i \sigma_j^{-1} \mu_j^i$. Let then consider any two vectors $\boldsymbol{v}, \boldsymbol{u} \in \overline{\boldsymbol{w}}^{-1} \wedge \mathcal{W}$ and compute their $\wedge$-product,

$$\boldsymbol{v} \wedge \boldsymbol{u} = \left(\sum_j \xi_j \boldsymbol{a}_j\right) \wedge \left(\sum_j \hat{\xi}_j \boldsymbol{a}_j\right) = \sum_j \xi_j \hat{\xi}_j \boldsymbol{a}_j.$$

This implies that $\mathrm{alg}(\overline{\boldsymbol{w}}^{-1} \wedge \mathcal{W}) \subseteq \mathscr{A}^*$. On the other hand, it trivially holds that $\mathcal{W} \subseteq \overline{\boldsymbol{w}} \wedge \mathrm{alg}(\overline{\boldsymbol{w}}^{-1} \wedge \mathcal{W})$. But then, since we assumed that $\mathscr{A}^*$ was the unique algebra of minimal dimension such that $\mathcal{W} \subseteq \boldsymbol{x} \wedge \mathscr{A}^*$ for some $\boldsymbol{x}$ it must hold that $\mathrm{alg}(\overline{\boldsymbol{w}}^{-1} \wedge \mathcal{W}) = \mathscr{A}^*$.

Finally, since $\overline{\boldsymbol{w}} \in \mathcal{W}$, then $\overline{\boldsymbol{w}}^{-1} \wedge \overline{\boldsymbol{w}} = \mathbf{1}_{\mathrm{supp}(\mathcal{W})} \in (\overline{\boldsymbol{w}}^{-1} \wedge \mathcal{W}) \subseteq \mathscr{A}^*$. □

Remark 10. Theorem 3 shows that, given any choice of the effective subspace, we can construct a vector $\overline{\boldsymbol{w}}$ such that the algebra $\mathrm{alg}(\overline{\boldsymbol{w}}^{-1} \wedge \mathcal{E})$ has minimal dimension. However, not all such $\overline{\boldsymbol{w}}$ are positive over the support of $\mathcal{E}$. As a matter of fact, it could happen that some choices of $\mathcal{E}$ do not contain any non-negative vector, while $\overline{\boldsymbol{w}} = \boldsymbol{p}$ being non-negative is fundamental to construct a stochastic reduction.

Theorem 3 is nonetheless sufficient to determine the optimal reduction for a class of HMMs, namely those for which $\mathcal{R}$ is "observable", i.e. $\mathcal{R} \cap \mathcal{N} = \varnothing$.

Proposition 4. *Let $\{\boldsymbol{r}_i\}$ be an N-dimensional set of positive generators of $\mathcal{R}$ and let $\overline{\boldsymbol{p}} := \sum \boldsymbol{r}_i/N$. Then, if $\mathcal{R} \cap \mathcal{N} = \{\mathbf{0}\}$, $\mathscr{A} := \mathrm{alg}(\overline{\boldsymbol{p}}^{-1} \wedge \mathcal{R})$ provides the optimal reduction.*

Proof. By hypothesis we have $\mathcal{R} \cap \mathcal{N} = \varnothing$. This implies that $\mathcal{E} = \mathcal{R}$. Then, using Theorem 3 we have that $\boldsymbol{p} = \sum_i \boldsymbol{r}_i/N$, provides the minimal dimension for $\mathrm{alg}(\boldsymbol{p}^{-1} \wedge \mathcal{R})$ and thus the optimal reduction. □

Notice that this result applies in particular fully observable HMMs, i.e. when the pair (P, C) is observable, and thus to finite-state Markov chains. In fact, the latter can be seen as HMMs with $C = I$. The corresponding optimal reduction is then a maximally-lumped version of the original process [25].

B. Effective subspace for the general case

In order to address the general case, in addition to a distribution $\mathbf{p}$ we also need to choose an effective subspace. Example 2 below illustrates that not all effective spaces are equivalent and lead to different dimensions for the reduced model, making this choice critical towards the optimality of the reduction. A natural candidate effective subspace is $\mathcal{E}_\perp$, the orthogonal complement (with respect to the natural inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$) of $\mathcal{R} \cap \mathcal{N}$ in $\mathcal{R}$. Let then $\{\boldsymbol{\varepsilon}_i\}_{i=1,\dots,d}$ be the set of generators of $\mathcal{E}_\perp$. Then any choice of the effective subspace can be described as $\mathcal{E} = \mathrm{span}\{\boldsymbol{\varepsilon}_i + \boldsymbol{n}_i\}_{i=1,\dots,d}$, where $\{\boldsymbol{n}_i\}_{i=1,\dots,d}$ is a set of vectors in $\mathcal{N}$.

We next show that the choice of the orthogonal complement $\mathcal{E}_\perp$ always allows for finding a *positive vector* $\overline{\boldsymbol{w}} = \overline{\boldsymbol{p}}$ as in the statement of Theorem 3, and hence a valid stochastic reduction. The following proposition is instrumental to this aim.

Proposition 5. *Let $\boldsymbol{p} \in \mathbb{R}^n$ be a probability vector, and let $\mathcal{V}$ be a vector space such that $\mathbf{1}^T \boldsymbol{v} = 0$ for all $\boldsymbol{v} \in \mathcal{V}$. Let then $\Pi_{\mathcal{V}}$ be the orthogonal projector on $\mathcal{V}$ with respect to the standard inner product $\langle \cdot, \cdot \rangle$. Then $\boldsymbol{q} := \boldsymbol{p} - \Pi_{\mathcal{V}} \boldsymbol{p}$ is a probability vector.*

Proof. Let us start by defining $\boldsymbol{w} := \mathbf{1}/2 - \boldsymbol{p}$. We can then write $\boldsymbol{p} = \mathbf{1}/2 - \boldsymbol{w}$ to notice that $\boldsymbol{p}_i \in [0, 1]$ if and only if $-1/2 \leqslant \boldsymbol{w}_i \leqslant 1/2$, that is if and only if $||\boldsymbol{w}||_\infty \leqslant 1/2$. Moreover, we have that $\mathbf{1}^T \boldsymbol{p} = 1$ if and only if $\mathbf{1}^T \boldsymbol{w} = (n-2)/2$. We can then compute $\boldsymbol{q}$:

$$\begin{aligned} \boldsymbol{q} &= \mathbf{1}/2 - \boldsymbol{w} - \Pi_{\mathcal{V}} \mathbf{1}/2 + \Pi_{\mathcal{V}} \boldsymbol{w} \\ &= \mathbf{1}/2 - \underbrace{(I - \Pi_{\mathcal{V}})}_{=:\Pi_{\mathcal{V}^\perp}} \boldsymbol{w} = \mathbf{1}/2 - \Pi_{\mathcal{V}^\perp} \boldsymbol{w} \end{aligned}$$

where we used the the hypothesis $\mathbf{1}^T \boldsymbol{v} = 0$ for all $\boldsymbol{v} \in \mathcal{V}$ to say that $\Pi_{\mathcal{V}} \mathbf{1} = \mathbf{0}$. Then, since $\Pi_{\mathcal{V}^\perp}$ is an orthogonal projection, and thus a contraction in norm, we have that $||\Pi_{\mathcal{V}^\perp}\boldsymbol{w}||_\infty \leqslant ||\boldsymbol{w}||_\infty$. Then, using the argument above, we have that $\boldsymbol{q}$ is a non-negative vector with $\boldsymbol{q}_i \in [0,1]$. Lastly we have that $\mathbf{1}^T\boldsymbol{q} = \mathbf{1}^T\mathbf{1}/2 - \mathbf{1}^T\boldsymbol{w} = n/2-(n-2)/2 = 1$. □

The result we are after is then obtained as a corollary of the previous one.

Corollary 4. *Let $\mathcal{E}_\perp$ be the orthogonal complement of $\mathcal{R}\cap\mathcal{N}$ to $\mathcal{R}$. Let $\{\boldsymbol{r}_i\}$ be an N dimensional set of probability vectors such that $\mathcal{R} = \operatorname{span}\{\boldsymbol{r}_i\}$. Then $\boldsymbol{\varepsilon}_i := \boldsymbol{r}_i - \Pi_{\mathcal{R}\cap\mathcal{N}}\boldsymbol{r}_i$ are such that $\mathcal{E}_\perp = \operatorname{span}\{\boldsymbol{\varepsilon}_i\}$. Moreover, $\overline{\boldsymbol{\varepsilon}} = \sum_i \boldsymbol{\varepsilon}_i/N$ satisfies $\operatorname{supp}(\overline{\boldsymbol{\varepsilon}}) = \operatorname{supp}(\mathcal{E})$ and $\overline{\boldsymbol{\varepsilon}}_i \geqslant 0$ for all i.*

Proof. From Lemma 2 we have that $\mathbf{1}^T\boldsymbol{x} = 0$ for all $\boldsymbol{x}\in\mathcal{N}$ and thus, by applying the proposition above on every generator of $\mathcal{R}$ we have that the set $\{\boldsymbol{\varepsilon}_i\}$ is a set of probability vectors. Being $\overline{\boldsymbol{\varepsilon}}$ a convex combination of probability vectors it is itself a probability vector and it shares the same support as $\mathcal{E}$. □

Other choices are possible, and the choice of the effective subspace can influence the dimension of the reduced model, as illustrated in the following example.

Example 2. Consider the following spaces:

$$\mathcal{R} = \operatorname{span}\left\{\begin{bmatrix}1/2\\1/2\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\1\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\1\end{bmatrix}\right\},\quad \mathcal{N} = \operatorname{span}\left\{\begin{bmatrix}0\\0\\1\\-1\end{bmatrix}\right\}$$

then we clearly have that $\mathcal{R}\cap\mathcal{N} = \mathcal{N}$. Let us denote with $\mathcal{E}_\perp$ the orthogonal complement of $\mathcal{R}\cap\mathcal{N}$ to $\mathcal{R}$, i.e.

$$\mathcal{E}_\perp = \operatorname{span}\left\{\begin{bmatrix}1/2\\1/2\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\1/2\\1/2\end{bmatrix}\right\}.$$

We can easily notice that $\mathcal{E}_\perp$ is an unital algebra. Let us now consider another completion $\mathcal{E}$ of $\mathcal{R}\cap\mathcal{N}$ to $\mathcal{R}$. In general, we can write

$$\mathcal{E} = \operatorname{span}\left\{\begin{bmatrix}1\\1\\a\\-a\end{bmatrix},\begin{bmatrix}0\\0\\1+b\\1-b\end{bmatrix}\right\}$$

for some values $a,b\in\mathbb{R}$. We can then consider two cases. First, if $a=0$ and $b\neq 0$ we can choose $\overline{\boldsymbol{v}} = \begin{bmatrix}1 & 1 & 1+b & 1-b\end{bmatrix}^T$ thus obtaining $\operatorname{alg}(\overline{\boldsymbol{v}}^{-1}\wedge\mathcal{E}) = \mathcal{E}_\perp$. On the other hand, if we have $a\neq 0$ and $b\neq 0$ we can choose $\overline{\boldsymbol{w}} = \begin{bmatrix}1 & 1 & a+1+b & -a+1-b\end{bmatrix}^T$ (assuming that $a+b\neq\pm 1$) thus obtaining

$$\operatorname{alg}(\overline{\boldsymbol{w}}^{-1}\wedge\mathcal{E}) = \operatorname{span}\left\{\begin{bmatrix}1\\1\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\1\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\1\end{bmatrix}\right\}.$$

This example shows that the choice of the effective subspace can affect the size of the reduced model.

VIII. Examples

Example 3. Let consider the HMM provided in [19, Example 3]:

$$P = \begin{bmatrix}1/3 & 1/6 & 1/4 & 1/4 & 0\\ 1/6 & 1/3 & 0 & 1/4 & 1/4\\ 1/3 & 1/6 & 1/4 & 1/4 & 0\\ 1/6 & 1/6 & 1/6 & 0 & 1/2\\ 0 & 1/6 & 1/3 & 1/4 & 1/4\end{bmatrix},\quad \mathcal{S} = \left\{\begin{bmatrix}1/5\\1/5\\1/5\\1/5\\1/5\end{bmatrix}\right\}$$

$$C = \begin{bmatrix}1 & 1 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1\end{bmatrix}.$$

We shall start by studying the single-time marginal problem. We can observe that $\boldsymbol{p}_0\in\mathcal{S}$ is an equilibrium for P and thus $\mathcal{R} = \operatorname{span}\{\mathcal{S}\}$ and also that $\mathcal{N} = \operatorname{span}\{\begin{bmatrix}1 & -2 & 1 & 0 & 0\end{bmatrix}^T\}$. Clearly, the intersection contains only the zero vector, $\mathcal{R}\cap\mathcal{N} = \{\mathbf{0}\}$ and thus the effective subspace can be taken as the reachable one: $\mathcal{E}_\perp = \mathcal{R}$. If we then take $\overline{\boldsymbol{p}} = \boldsymbol{p}_0$ we obtain $\mathscr{A} = \operatorname{alg}(\overline{\boldsymbol{p}}^{-1}\wedge\mathcal{E}) = \operatorname{span}\{\mathbf{1}\}$. The corresponding stochastic reduction and injection matrices are $R = \mathbf{1}^T$ and $J = \overline{\boldsymbol{p}}$ which provide the (trivial) reduced model:

$$\check{P} = \begin{bmatrix}1\end{bmatrix},\quad \check{\mathcal{S}} = \{\begin{bmatrix}1\end{bmatrix}\},\quad \check{C} = \begin{bmatrix}3/5 & 1/5 & 1/5\end{bmatrix}^T.$$

We next focus on the multi-time marginal problem. We have that $\mathcal{N}_\mathcal{C} = \mathcal{N}$, while the conditioned-reachable is equal to:

$$\mathcal{R}_\mathcal{C} = \operatorname{span}\left\{\begin{bmatrix}1/5\\1/5\\1/5\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\1/5\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\0\\1/5\end{bmatrix},\begin{bmatrix}1\\-2\\1\\0\\0\end{bmatrix}\right\}.$$

This implies that the intersection $\mathcal{R}_\mathcal{C}\cap\mathcal{N}_\mathcal{C} = \mathcal{N}_\mathcal{C}$ and thus:

$$\mathcal{E}_{\mathcal{C}\perp} = \operatorname{span}\left\{\begin{bmatrix}1\\1\\1\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}\right\}.$$

Then, we can notice that $\mathcal{E}_\mathcal{C}$ is a unital algebra and by taking $\overline{\boldsymbol{p}} = \mathbf{1}/5$ we obtain the stochastic reduction and injection matrices

$$R = \begin{bmatrix}1 & 1 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1\end{bmatrix},\quad J = \begin{bmatrix}1/3 & 0 & 0\\ 1/3 & 0 & 0\\ 1/3 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

that leads to the reduced model

$$\check{P} = \begin{bmatrix}2/3 & 3/4 & 1/4\\ 1/6 & 0 & 1/2\\ 1/6 & 1/4 & 1/4\end{bmatrix},\quad \check{\mathcal{S}} = \left\{\begin{bmatrix}3/5\\1/5\\1/5\end{bmatrix}\right\}$$

$$\check{C} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}.$$

Example 4. Consider the HMM defined by:

$$P = \begin{bmatrix} 1/2 & 0 & 1/3 & 1/4 \\ 0 & 1/3 & 1/3 & 1/4 \\ 1/2 & 0 & 1/3 & 0 \\ 0 & 2/3 & 0 & 1/2 \end{bmatrix}, \quad \mathcal{S} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right\}$$

$$C = \begin{bmatrix} 1/4 & 1/4 & 1/2 & 7/16 \\ 3/4 & 3/4 & 1/2 & 9/16 \end{bmatrix}.$$

In this case we are only interested in the single-time marginal problem. We can notice that $\mathcal{R} = \mathbb{R}^n$ and thus

$$\mathcal{R} \cap \mathcal{N} = \mathcal{N} = \operatorname{span} \left\{ \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ -3 \\ 4 \end{bmatrix} \right\}.$$

Then we can consider the effective subspace as the orthogonal complement of $\mathcal{N}$, that is

$$\mathcal{E}_\perp = \operatorname{span} \left\{ \begin{bmatrix} 4/9 \\ 4/9 \\ 0 \\ 1/9 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 4/7 \\ 3/7 \end{bmatrix} \right\}.$$

Then we can define $\overline{\boldsymbol{p}} := \mathbf{1}/4$ to obtain

$$\mathscr{A} = \operatorname{alg}(\overline{\boldsymbol{p}}^{-1} \wedge \mathcal{E}_\perp) = \operatorname{span} \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

Notice that in this case, the dimension of the algebra is greater than the effective subspace. We thus obtain the stochastic reduction and injection matrices

$$R = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad J = \begin{bmatrix} 1/2 & 0 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

that leads to the reduced model

$$\check{P} = \begin{bmatrix} 5/12 & 2/3 & 1/2 \\ 1/4 & 1/3 & 0 \\ 1/3 & 0 & 1/2 \end{bmatrix}, \quad \check{\mathcal{S}} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\}$$

$$\check{C} = \begin{bmatrix} 1/4 & 1/2 & 7/16 \\ 3/4 & 1/2 & 9/16 \end{bmatrix}.$$

Suppose that, instead of the orthogonal complement, we were to consider the following space as an effective subspace:

$$\mathcal{E} = \operatorname{span} \left\{ \begin{bmatrix} 6 \\ 5/2 \\ 3/2 \\ -1 \end{bmatrix}, \begin{bmatrix} 2 \\ 5/6 \\ 25/2 \\ -25/3 \end{bmatrix} \right\}.$$

We can immediately notice that there is no convex combination of the generators of $\mathcal{E}$ such that it is positive, however, if we consider $\boldsymbol{v} = \begin{bmatrix} 8 & 10/3 & 14 & -28/3 \end{bmatrix}^T$ we have that

$$\mathscr{A} = \operatorname{alg}(\boldsymbol{v}^{-1} \wedge \mathcal{E}) = \operatorname{span} \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \right\}$$

thus showing that a smaller algebra could be found for the reduction, if we were to consider vectors that are not non-negative.

IX. Conclusions and Future Work

In this work, we exploited system-theoretic ideas and algebraic representation of probability spaces to obtain effective reductions of HMMs that preserve the marginals of the original output process, in either the single- or multi-time case. While optimal reductions are explicitly characterized for a class of HMMs, including observable ones, the freedom of choice in the effective subspace makes finding the optimal reductions more challenging in the general case. Nonetheless, we provide an algorithm that produces reduced HMMs of minimal dimension in all considered examples. Based on the analytical and numerical examples we examined, we formulate the following conjecture on the optimality of the natural orthogonal complement.

Conjecture 1. *Let $\mathcal{E}_\perp$ be defined as the (standard) orthogonal complement of $\mathcal{N} \cap \mathcal{R}$ to $\mathcal{R}$, and let $\overline{\boldsymbol{p}}$ be defined as in Corollary 4. Then, given any other choice of $\mathcal{E}$ and $\boldsymbol{w}$* non-negative *it holds that*

$$\dim(\operatorname{alg}(\overline{\boldsymbol{p}}^{-1} \wedge \mathcal{E}_\perp)) \leqslant \dim(\operatorname{alg}(\boldsymbol{w}^{-1} \wedge \mathcal{E})).$$

Remark 11. If the effective subspace is already an algebra with respect to a $\overline{\boldsymbol{p}}$ - inner product then $\dim(\mathcal{E}_\perp) = \dim(\operatorname{alg}(\overline{\boldsymbol{p}}^{-1} \wedge \mathcal{E}_\perp))$, since $\dim(\mathcal{E}^\perp) = \dim(\mathcal{E})$ and $\dim(\mathcal{E}) \leqslant \dim(\operatorname{alg}(\boldsymbol{w}^{-1} \wedge \mathcal{E}))$ by Theorem 3 then the choice of $\mathcal{E}_\perp$ is optimal. Also notice that removing the assumption that $\boldsymbol{w}$ is non-negative makes the statement false. A counterexample is presented at the end of Example 4. However, having $\boldsymbol{w}$ non-negative is necessary in order to obtain a stochastic model.

Proving the conjectured minimality may require novel mathematical ideas: the choice of $\mathcal{E}, \mathcal{E}_C$ that minimize the size of the generated algebras is equivalent to identify the representative of the quotient space that can be described with the least number of indicator vectors, and a way to relate this notion to orthogonality to $\mathscr{N}$ does not seem straightforward to find.

Other natural developments of the proposed framework include a relaxation of the method so that it allows for *approximate preservation of the marginals*, thus yielding reductions in practical situations where noise and partial knowledge might make the exact equivalence we require in this work too stringent, due to the fact that controllable pairs are a dense set [36]. In addition, in many algorithms used to estimate HMMs from data, e.g. [34], the dimension of the "hidden" state space (i.e. n) is assumed to be known. When this is not the case, one could estimate an HMM with a larger than necessary number of hidden variables, and then use an approximate reduction scheme to reduce the estimated model to one of more manageable size. Future work will also be devoted to the adaptation and application of the method to approximate coarse-graining of large-scale systems, to address otherwise untreatable problems [16]–[18].

The algebraic approach also naturally extends to the non-commutative domain, and our method will be extended to quantum systems, in particular quantum walks and open systems in general. Analogies between HMM and quantum walks have been already noted in [37], as well as [38] and [39], which extend the result of [19] to include quantum walks. Lastly, the algebraic viewpoint makes our results potentially interesting towards the solution of outstanding open problems in realization theory and model reduction for positive systems [40].

X. Acknowledgements

T.G. and F.T. wish to thank Lorenzo Finesso, Augusto Ferrante and Lorenza Viola for motivating and stimulating discussions on the topics of this work.

References

[1] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] N. Najkar, F. Razzazi, and H. Sameti, "A novel approach to hmm-based speech recognition systems using particle swarm optimization," *Mathematical and Computer Modelling*, vol. 52, no. 11, pp. 1910–1920, 2010, The BIC-TA 2009 Special Issue, ISSN: 0895-7177.

[3] P Baldi, Y Chauvin, T Hunkapiller, and M. A. McClure, "Hidden markov models of biological primary sequence information.," *Proceedings of the National Academy of Sciences*, vol. 91, no. 3, pp. 1059–1063, 1994.

[4] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg, "Improved microbial gene identification with glimmer," *Nucleic acids research*, vol. 27, no. 23, pp. 4636–4641, 1999.

[5] A. Krogh, I. S. Mian, and D. Haussler, "A hidden markov model that finds genes in e. coli dna," *Nucleic acids research*, vol. 22, no. 22, pp. 4768–4778, 1994.

[6] M. Vidyasagar, *Hidden Markov Processes: Theory and Applications to Biology*. Princeton University Press, 2014.

[7] M. Cidota and M. Dumitrescu, "A multinomial - hidden markov model for communication systems influenced by external factors," in *2012 7th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, 2012, pp. 235–240.

[8] A. Dainotti, A. Pescapé, P. S. Rossi, F. Palmieri, and G. Ventre, "Internet traffic modeling by means of hidden markov models," *Computer Networks*, vol. 52, no. 14, pp. 2645–2662, 2008, ISSN: 1389-1286.

[9] K. Salamatian and S. Vaton, "Hidden markov modeling for network communication channels," *ACM SIGMETRICS Performance Evaluation Review*, vol. 29, no. 1, pp. 92–101, 2001.

[10] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov models: estimation and control*. Springer Science & Business Media, 2008, vol. 29.

[11] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

[12] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete-time controlled markov processes with average cost criterion: A survey," *SIAM Journal on Control and Optimization*, vol. 31, no. 2, pp. 282–344, 1993.

[13] M Vidyasagar, "(Hidden) Markov Processes: Theory and Applications to Biology," p. 288, 2011.

[14] L. Mevel and L. Finesso, "Bayesian estimation of hidden markov models," in *Proceedings of the Mathematical Theory of Networks and Systems Conference, MTNS-2000*, 2000.

[15] Q. Huang, R. Ge, S. Kakade, and M. Dahleh, "Minimal realization problems for hidden markov models," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1896–1904, 2016.

[16] A. C. Antoulas, *Approximation of large-scale dynamical systems*. SIAM, 2005.

[17] X Cheng and J. Scherpen, "Model reduction methods for complex network systems," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, pp. 425–453, 2021.

[18] H. Sandberg and R. M. Murray, "Model reduction of interconnected linear systems," *Optimal Control Applications and Methods*, vol. 30, no. 3, pp. 225–245, 2009.

[19] H. Ito, S.-I. Amari, and K. Kobayashi, "Identifiability of hidden Markov information sources and their minimum degrees of freedom," en, *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 324–333, Mar. 1992.

[20] J. Von Neumann, *Mathematical foundations of quantum mechanics: New edition*. Princeton university press, 2018.

[21] O. Bratteli and D. W. Robinson, *Operator algebras and quantum statistical mechanics: Volume 1: C*-and W*-Algebras. Symmetry Groups. Decomposition of States*. Springer Science & Business Media, 2012.

[22] P. A. Meyer, *Quantum probability for probabilists*. Springer Science & Business Media, 1995.

[23] T. Tao, *254a, notes 5: Free probability*, 2010. [Online]. Available: https://terrytao.wordpress.com/2010/02/10/245a-notes-5-free-probability/.

[24] A. Schönhuth, "Complete identification of binary-valued hidden markov processes," *arXiv preprint arXiv:1101.3712*, 2011.

[25] J. G. Kemeny and J. L. Snell, *Finite Markov Chains: With a New Appendix "Generalization of a Fundamental Matrix"* (Undergraduate Texts in Mathematics), Reprint. New York, NY Heidelberg Berlin: Springer, 1983.

[26] L. White, R. Mahony, and G. Brushe, "Lumpable hidden markov models-model reduction and reduced complexity filtering," *IEEE Transactions on Automatic Control*, vol. 45, no. 12, pp. 2297–2306, 2000.

[27] N. Ay and J. P. Crutchfield, "Reductions of hidden information sources," *Journal of Statistical Physics*, vol. 120, no. 3, pp. 659–684, 2005.

[28] S. Apers, A. Sarlette, and F. Ticozzi, "Characterizing limits and opportunities in speeding up markov chain mixing," *Stochastic Processes and their Applications*, vol. 136, pp. 145–191, 2021, ISSN: 0304-4149.

[29] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," 1961.

[30] H. Maassen, "Quantum probability," *Quantum Probability Communications: Qp-pq (Volumes 12)*, vol. 12, p. 23, 2003.

[31] G. Marro and G. Basile, *Controlled and Conditioned Invariants in Linear System Theory*. Feb. 1994, vol. 30.

[32] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*.

[33] H. H. Rosenbrock, *State-space and multivariable theory*. Wiley Interscience Division, 1970, vol. 3.

[34] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning hidden markov models," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1460–1480, 2012, JCSS Special Issue: Cloud Computing 2011.

[35] S. S. Ge, Z. Sun, and T. H. Lee, "Reachability and controllability of switched linear discrete-time systems," *IEEE Transactions on Automatic Control*, vol. 46, no. 9, pp. 1437–1441, 2001.

[36] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*. Wiley, New York, 1967.

[37] S. Apers, A. Sarlette, and F. Ticozzi, "Simulation of quantum walks and fast mixing with classical processes," *Phys. Rev. A*, vol. 98, p. 032 115, 3 2018.

[38] U. Faigle and A. Schönhuth, "Efficient tests for equivalence of hidden markov processes and quantum random walks," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1746–1753, 2011.

[39] A. Schonhuth, "Simple and efficient solution of the identifiability problem for hidden markov sources and quantum random walks," in *2008 International Symposium on Information Theory and Its Applications*, 2008, pp. 1–6.

[40] L. Benvenuti and L. Farina, "A tutorial on the positive realization problem," *IEEE Transactions on automatic control*, vol. 49, no. 5, pp. 651–664, 2004.

Tommaso Grigoletto received an M.Sc. degree in automation engineering from the University of Padua in 2020.

He is currently a Ph.D. student at the University of Padua and a visiting research scholar at Dartmouth College, Hanover, NH. His research interests include quantum and classical control and model reduction of quantum and classical dynamics.

Francesco Ticozzi received the "Laurea" degree in management engineering and a Ph.D. in automatic control and operations research from the University of Padua, Italy, in 2002 and 2007, respectively. Since 2007 he has been with the Department of Information Engineering at the University of Padova, first as a Research Associate and Assistant Professor, and currently as an Associate Professor. During 2005–2010 he held visiting appointments at the Physics and Astronomy Dept. of Dartmouth College, Hanover, New Hampshire, where he has held adjunct positions since 2011. Dr. Ticozzi's research interests include modeling and control of stochastic classical and quantum systems, quantum codes and error correction, networked dynamics and information-theoretic approaches to control systems.

Appendix A
A reduction result for switching autonomous systems

This appendix is dedicated to introducing a general condition ensuring exact model reduction for switching autonomous systems. Both the single-time and the multi-time marginals can be described by the dynamics of this type. Consider a discrete-time, autonomous, switching, linear system

$$\begin{cases} \boldsymbol{x}(t+1) = F_i \boldsymbol{x}(t) \\ \boldsymbol{y}(t) = H\boldsymbol{x}(t) \\ \boldsymbol{x}(0) \in \mathcal{I} \end{cases}$$

denoted by the triplet $(\{F_i\}, H, \mathcal{I})$. The evolution at any time clearly depends on the sequence of evolutions F_i activated. Let us denote with $\boldsymbol{y}(s_{0:l})$ the output of the system at time l associated to a sequence $s_{0:l} = s_l, \dots, s_0$ of length l of selected evolution F_{s_k}. The output at any time $l > 0$ can be computed as $\boldsymbol{y}(s_{0:l}) = H \prod_{j=l}^{0} F_{s_j} \boldsymbol{x}_0$ while for $t = 0$ we have $\boldsymbol{y}(0) = H\boldsymbol{x}_0$.

Let $\mathcal{R} \subseteq \mathbb{R}^n$ be a linear subspace such that $\mathcal{I} \subseteq \mathcal{R}$ and is F_i-invariant, i.e. $F_i\mathcal{R} \subseteq \mathcal{R}$, for all i. Let $\tilde{\mathcal{N}} \subseteq \mathbb{R}^n$ be a linear subspace such that $\tilde{\mathcal{N}} \subseteq \ker H$ and is F_i-invariant, i.e. $F_i\tilde{\mathcal{N}} \subseteq \tilde{\mathcal{N}}$, for all i. Let then define $\mathcal{E}$ to be any completion of $\mathcal{R} \cap \tilde{\mathcal{N}}$ to $\mathcal{R}$, i.e. $\mathcal{R} = (\mathcal{R} \cap \tilde{\mathcal{N}}) \oplus \mathcal{E}$.

Theorem 4. *Consider any subspace $\mathcal{V}$ such that $\mathcal{E} \subseteq \mathcal{V}$ with $m = \dim(\mathcal{V})$ and let $\Pi_{\mathcal{V}}$ be the orthogonal projection onto $\mathcal{V}$ with respect to an inner product $\langle \cdot, \cdot \rangle$. Assume that $\Pi_{\mathcal{V}}(\mathcal{R} \cap \tilde{\mathcal{N}}) \subseteq \mathcal{R} \cap \tilde{\mathcal{N}}$, and let $R : \mathbb{R}^n \to \mathbb{R}^m$ and $J : \mathbb{R}^m \to \mathcal{V}$ be two factors of the orthogonal projection, $\Pi_{\mathcal{V}} = JR$.*

Let consider the reduced model $(\{\check{F}_i\}, \check{H}, \check{\mathcal{I}}) = (\{RF_iJ\}, HJ, R\mathcal{I})$. Then the reduced model reproduces the same output as the original model, i.e.

$$H \prod_{j=l}^{0} F_{s_j} \boldsymbol{x}_0 = \check{H} \prod_{j=l}^{0} \check{F}_{s_j} \check{\boldsymbol{x}}_0$$

for any sequence $s_{0:l}$ and any initial condition $\boldsymbol{x}_0 \in \operatorname{span}\{\mathcal{I}\}$ and the relative $\check{\boldsymbol{x}}_0 = R\boldsymbol{x}_0$.

Proof. Let $\mathcal{W}_1$ be the completion of $\mathcal{R} \cap \tilde{\mathcal{N}}$ to $\tilde{\mathcal{N}}$, i.e. $\tilde{\mathcal{N}} = \mathcal{W}_1 \oplus (\mathcal{R} \cap \tilde{\mathcal{N}})$; let $\mathcal{W}_2$ be the completion of $(\mathcal{R} \cap \tilde{\mathcal{N}}) \oplus \mathcal{E} \oplus \mathcal{W}_1$ to $\mathbb{R}^n$, i.e. $\mathbb{R}^n = \mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{E} \oplus (\mathcal{R} \cap \tilde{\mathcal{N}})$; let $\mathcal{T}$ the remainder sub-space, such that $\mathbb{R}^n = \mathcal{V} \oplus \mathcal{T}$ and thus $\mathcal{T} \subseteq (\tilde{\mathcal{N}} \cap \mathcal{R}) \oplus \mathcal{W}_1 \oplus \mathcal{W}_2$. Let us also denote with $\Pi_{\mathcal{T}}$ the orthogonal projector onto $\mathcal{T}$ with respect to the considered inner product $\langle \cdot, \cdot \rangle$.

$$\begin{aligned}\left[\prod_{j=l}^{0} F_{s_j} - \Pi_{\mathcal{V}} \prod_{j=l}^{0} F_{s_j} \Pi_{\mathcal{V}}\right] \boldsymbol{x}_0 &= \left[(\Pi_{\mathcal{V}} + \Pi_{\mathcal{T}}) F_{s_l} (\Pi_{\mathcal{V}} + \Pi_{\mathcal{T}}) \prod_{j=l-1}^{0} F_{s_j} - \Pi_{\mathcal{V}} \prod_{j=l}^{0} F_{s_j} \Pi_{\mathcal{V}}\right] \boldsymbol{x}_0 \\ &= \left[[\Pi_{\mathcal{V}} F_{s_l} \Pi_{\mathcal{V}} + \Pi_{\mathcal{T}} F_{s_l} \Pi_{\mathcal{V}} + F_{s_l} \Pi_{\mathcal{T}}] \prod_{j=l-1}^{0} F_{s_j} - \Pi_{\mathcal{V}} F_{s_l} \Pi_{\mathcal{V}} \Pi_{\mathcal{V}} \prod_{j=l-1}^{0} F_{s_j} \Pi_{\mathcal{V}}\right] \boldsymbol{x}_0 \\ &= \Pi_{\mathcal{V}} F_{s_l} \Pi_{\mathcal{V}} \underbrace{\left(\prod_{j=l-1}^{0} F_{s_j} - \Pi_{\mathcal{V}} \prod_{j=l-1}^{0} F_{s_j} \Pi_{\mathcal{V}}\right) \boldsymbol{x}_0}_{\boldsymbol{v}} + \left[[\Pi_{\mathcal{T}} F_{s_l} \Pi_{\mathcal{V}} + F_{s_l} \Pi_{\mathcal{T}}] \prod_{j=l-1}^{0} F_{s_j}\right] \boldsymbol{x}_0 \\ &= \Pi_{\mathcal{V}} F_{s_l} \Pi_{\mathcal{V}} \boldsymbol{v} + \Pi_{\mathcal{T}} F_{s_l} \Pi_{\mathcal{V}} \prod_{j=l-1}^{0} F_{s_j} \boldsymbol{x}_0 + F_{s_l} \Pi_{\mathcal{T}} \prod_{j=l-1}^{0} F_{s_j} \boldsymbol{x}_0 \end{aligned} \tag{15}$$

We can notice that, for any sequence $s_{0:l}$ we have that $\breve{H} \prod_{j=l}^{0} \breve{F}_{s_j} \breve{\boldsymbol{x}}_0 = H \Pi_{\mathcal{V}} \prod_{j=l}^{0} F_{s_j} \Pi_{\mathcal{V}} \boldsymbol{x}_0$ and thus the statement can be also be put in the form

$$H \left[\prod_{j=l}^{0} F_{s_j} - \Pi_{\mathcal{V}} \prod_{j=l}^{0} F_{s_j} \Pi_{\mathcal{V}}\right] \boldsymbol{x}_0 = 0$$

for any sequence $s_{0:l}$ and for any $\boldsymbol{x}_0 \in \mathcal{I}$. To prove the statement we will thus show that for any sequence $s_{0:l}$ and for any initial condition $\boldsymbol{x}_0 \in \mathcal{I}$ it holds

$$\left[\prod_{j=l}^{0} F_{s_j} - \Pi_{\mathcal{V}} \prod_{j=l}^{0} F_{s_j} \Pi_{\mathcal{V}}\right] \boldsymbol{x}_0 \in \tilde{\mathcal{N}} \cap \mathcal{R}.$$

We will prove this statement by induction.

Let then consider the case of $t = 0$. We have to prove $[I - \Pi_{\mathcal{V}}]\boldsymbol{x}_0 \in \tilde{\mathcal{N}} \cap \mathcal{R}$. Then by noticing that, $(I - \Pi_{\mathcal{V}})\boldsymbol{x}_0 = \Pi_{\mathcal{T}} \boldsymbol{x}_0$ and that $\Pi_{\mathcal{T}} \boldsymbol{x}_0 \in \tilde{\mathcal{N}} \cap \mathcal{R}$, the statement is proved in the case $t = 0$.

Assume then that

$$\boldsymbol{v} := \left[\prod_{j=l-1}^{0} F_{s_j} - \Pi_{\mathcal{V}} \prod_{j=l-1}^{0} F_{s_j} \Pi_{\mathcal{V}}\right] \boldsymbol{x}_0 \in \tilde{\mathcal{N}} \cap \mathcal{R}$$

and we want to prove that

$$\left[\prod_{j=l}^{0} F_{s_j} - \Pi_{\mathcal{V}} \prod_{j=l}^{0} F_{s_j} \Pi_{\mathcal{V}}\right] \boldsymbol{x}_0 \in \tilde{\mathcal{N}} \cap \mathcal{R}.$$

By rewriting this as in Equation (15) we can observe that it is equal to the sum of three parts. We can then notice that:

- $\boldsymbol{v} \in \tilde{\mathcal{N}} \cap \mathcal{R}$, $\Pi_{\mathcal{V}} \boldsymbol{v} \in \tilde{\mathcal{N}} \cap \mathcal{R}$ by assumption, thus, $P \Pi_{\mathcal{V}} \boldsymbol{v} \in \tilde{\mathcal{N}} \cap \mathcal{R}$ and also $\Pi_{\mathcal{V}} P \Pi_{\mathcal{V}} \boldsymbol{v} \in \tilde{\mathcal{N}} \cap \mathcal{R}$;
- $\prod_{j=l-1}^{0} F_{s_j} \boldsymbol{x}_0 \in \mathcal{R}$ by hypothesis, $\Pi_{\mathcal{T}} \prod_{j=l-1}^{0} F_{s_j} \boldsymbol{x}_0 \in \tilde{\mathcal{N}} \cap \mathcal{R}$ and $F_{s_l} \Pi_{\mathcal{T}} \prod_{j=l-1}^{0} F_{s_j} \boldsymbol{x}_0 \in \tilde{\mathcal{N}} \cap \mathcal{R}$;
- $\prod_{j=l-1}^{0} F_{s_j} \boldsymbol{x}_0 \in \mathcal{R}$ by hypothesis, $\Pi_{\mathcal{V}} \prod_{j=l-1}^{0} F_{s_j} \boldsymbol{x}_0 \in \mathcal{R}$, $F_{s_l} \Pi_{\mathcal{V}} \prod_{j=l-1}^{0} F_{s_j} \boldsymbol{x}_0 \in \mathcal{R}$ and $\Pi_{\mathcal{R}} F_{s_j} \Pi_{\mathcal{V}} \prod_{j=l-1}^{0} F_{s_j} \boldsymbol{x}_0 \in \tilde{\mathcal{N}} \cap \mathcal{R}$.

Finally, since all three summands belong to $\tilde{\mathcal{N}} \cap \mathcal{R}$, their sum also belongs to the same subspace, and the statement is proved. □

In order to apply the result to our multi-time problem, we need a straightforward extension.

Corollary 5. *Under the assumptions of Theorem 4, let us further assume that F_i are factorized as $F_i = D_i A$, that $\mathcal{R}$ and $\tilde{\mathcal{N}}$ are A-invariant and D_i-invariant for all i and also that $\mathcal{I} = \bigcup_i D_i \mathcal{S}$ for some set $\mathcal{S}$.*

Then the matrices $\{\breve{F}_i\}$ of the reduced model can be taken to be $\breve{F}_i = \breve{D}_i \breve{A}$ with $\breve{D}_i = R D_i J$, $\breve{A} = RAJ$. .

The proof of this corollary follows exactly that of Theorem 4, where $H \Pi_{\mathcal{V}} \prod_{j=l}^{0} (\Pi_{\mathcal{V}} D_{s_j} A \Pi_{\mathcal{V}}) \Pi_{\mathcal{V}} D_i \boldsymbol{x}_0$ is substituted by $H \Pi_{\mathcal{V}} \prod_{j=l}^{0} (\Pi_{\mathcal{V}} D_{s_j} \Pi_{\mathcal{V}} A \Pi_{\mathcal{V}}) \Pi_{\mathcal{V}} D_i \Pi_{\mathcal{V}} \boldsymbol{x}_0$, and in the induction we leverage the fact that $\tilde{\mathcal{N}}, \mathcal{R}$ and thus $\tilde{\mathcal{N}} \cap \mathcal{R}$ are invariant for $\Pi_{\mathcal{V}}$, A and D_i, for all i.

Appendix B
Proofs

This Appendix collects some proofs that were not included in the main text to improve readability.

Proof of Proposition 1. Let start with the first part of the statement. The fact that $\mathscr{A}$ is closed under linear combinations and $\mathbf{1} \in \mathscr{A}$ follows directly from the definition of $\mathscr{A}$. The closure of $\mathscr{A}$ under element-wise product follows from the closure of $\mathcal{F}$ under the same operation. In particular let consider $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{A}$, then $\boldsymbol{x} \wedge \boldsymbol{y} = \sum_{i,j} x_i y_j \boldsymbol{f}_i \wedge \boldsymbol{f}_j$ and , since $\boldsymbol{f}_i \in \mathcal{F}$ for all i, $\boldsymbol{f}_i \wedge \boldsymbol{f}_j \in \mathcal{F}$ and thus $\boldsymbol{x} \wedge \boldsymbol{y} \in \mathscr{A}$. So $\mathscr{A}$ is an algebra, namely the set of $\mathcal{F}$-measurable random variables, and it is the minimal one by construction.

We can then consider the second part of the statement. First of all, notice that the vectors that are idempotent for the element-wise product are composed only of zeros and ones. Let then consider a general element $\boldsymbol{x} \in \mathscr{A}$ and let $x_{i*} = \max_{i=1,\dots,n} |x_i|$. We can then compute $\boldsymbol{x}' = \boldsymbol{x}/x_{i*} \in \mathscr{A}$ that will have value 1 in the positions where $\boldsymbol{x}$ has value x_{i*}, possibly values -1 in the position where $\boldsymbol{x}$ has value $-x_{i*}$ and values in the range $(-1, 1)$ in all the others positions. We can then define $\boldsymbol{x}'' = 0.5(\boldsymbol{x}' + \boldsymbol{x}' \wedge \boldsymbol{x}') \in \mathscr{A}$ that will have have value 1 in the positions where $\boldsymbol{x}$ has value x_{i*} and values in the range $(-1, 1)$ in all the others positions. Finally the first idempotent element of the desired set is $\boldsymbol{f}_1 = \lim_{n \to \infty} (\boldsymbol{x}'')^n \in \mathscr{A}$ with element-wise power. Notice that $\boldsymbol{f}_1$ will have 1 in the same positions as $\boldsymbol{x}'$ and zeros in all the others. This implies that $\boldsymbol{f}_1$ is idempotent. By iterating

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2023.3279209

the procedure on $\boldsymbol{x} - x_{i*}\boldsymbol{f}_1$, and so on, we obtain the whole set of idempotent elements $\{\boldsymbol{f}_i\} \subset \mathscr{A}$ such that $\boldsymbol{x} = \sum_i x_i \boldsymbol{f}_i$ up to a reordering of the coefficients x_i. We shall denote with $\mathrm{idem}(\boldsymbol{x})$ the function that, given an element $\boldsymbol{x}$, returns the set of idempotent elements $\{\boldsymbol{f}_i\}$ that generate $\boldsymbol{x}$. We then have that $\mathrm{idem}(\mathscr{A}) \supseteq \cup_{\boldsymbol{x} \in \mathscr{A}} \mathrm{idem}(\boldsymbol{x})$ by definition, while to prove $\mathrm{idem}(\mathscr{A}) \subseteq \cup_{\boldsymbol{x} \in \mathscr{A}} \mathrm{idem}(\boldsymbol{x})$ it suffice to notice that each element of $\mathrm{idem}(\mathscr{A})$ is also an element of $\mathscr{A}$. This implies $\mathrm{idem}(\mathscr{A}) = \cup_{\boldsymbol{x} \in \mathscr{A}} \mathrm{idem}(\boldsymbol{x})$. Then, by construction, it holds that $\mathrm{span}\{\mathrm{idem}(\mathscr{A})\} = \mathscr{A}$.

We shall then notice that $\mathcal{F} = \mathrm{idem}(\mathscr{A})$ contains the elements, $\boldsymbol{0}, \boldsymbol{1} \in \mathcal{A}$, and is closed under the operations $\wedge$, $\vee$ and $\neg$. This shows that $\mathrm{idem}(\mathscr{A})$ is a σ-algebra. Then, since $\mathscr{A} = \mathrm{span}\{\mathrm{idem}(\mathscr{A})\}$ then any element in $\mathscr{A}$ is $\mathrm{idem}(\mathscr{A})$-measurable. Moreover, $\mathrm{idem}(\mathscr{A})$ is minimal because subtracting any element from it would make that element (seen as a r.v.) non-measurable. We thus have $\mathcal{F} = \mathrm{idem}(\mathscr{A})$.

Finally, $\mathrm{res}(\mathcal{F}) \subset \mathcal{F}$ is such that $\boldsymbol{f}_i \wedge \boldsymbol{f}_j = \boldsymbol{0}$, for all $\boldsymbol{f}_i, \boldsymbol{f}_j \in \mathrm{res}(\mathcal{F})$ $i \neq j$. This implies that $\langle \boldsymbol{f}_i, \boldsymbol{f}_j \rangle = \delta_{i-j}$ which means that is a set of orthogonal vectors. Moreover $\boldsymbol{f} = \vee_{\boldsymbol{f}_j \in S} \boldsymbol{f}_j$ with $S \subseteq \mathrm{res}(\mathcal{F})$ for all $\boldsymbol{f} \in \mathcal{F} \backslash \{\boldsymbol{0}\}$ or, equivalently, $\boldsymbol{f} = \sum_j c_j \boldsymbol{f}_j$ with $c_j \in \{0, 1\}$ for all $\boldsymbol{f} \in \mathcal{F}$. This implies that $\mathscr{A} = \mathrm{span}\{\mathrm{res}(\mathcal{A})\}$ and thus $\mathrm{res}(\mathscr{A})$ is an orthogonal basis for $\mathscr{A}$ and $\dim(\mathscr{A}) = |\mathrm{res}(\mathscr{A})|$. □

Proof of Lemma 1. First of all, note that the modified inner product can be written in many equivalent forms: $\langle \boldsymbol{v}, \boldsymbol{w} \rangle_{\boldsymbol{p}} = \mathbb{E}_{\boldsymbol{p}}[\boldsymbol{v} \wedge \boldsymbol{w}] = \langle \boldsymbol{p}, \boldsymbol{v} \wedge \boldsymbol{w} \rangle = \langle \boldsymbol{p} \wedge \boldsymbol{v}, \boldsymbol{w} \rangle = \langle \boldsymbol{v}, \boldsymbol{p} \wedge \boldsymbol{w} \rangle$.

Let us then consider $\mathbb{E}_{|\mathscr{A},\boldsymbol{p}}$. We can notice that $\mathrm{image}(\mathbb{E}_{|\mathscr{A},\boldsymbol{p}}) = \mathscr{A}$ and that $\mathbb{E}^2_{|\mathscr{A},\boldsymbol{p}} = \mathbb{E}_{|\mathscr{A},\boldsymbol{p}}$. It remains to be proven the fact that $\mathbb{E}_{|\mathscr{A},\boldsymbol{p}}$ is self adjoint with respect to the inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{p}}$, that is $\left\langle \boldsymbol{v}, \mathbb{E}_{|\mathscr{A},\boldsymbol{p}} \boldsymbol{w} \right\rangle_{\boldsymbol{p}} = \left\langle \mathbb{E}_{|\mathscr{A},\boldsymbol{p}} \boldsymbol{w}, \boldsymbol{v} \right\rangle_{\boldsymbol{p}}$. Such an equality can be rewritten, using equivalent forms of the modified inner product above, as $\left\langle \boldsymbol{v}, \boldsymbol{p} \wedge \boldsymbol{p} \wedge \mathbb{E}_{|\mathscr{A},\boldsymbol{p}} \boldsymbol{w} \right\rangle = \left\langle \boldsymbol{p} \wedge \boldsymbol{p} \wedge \mathbb{E}_{|\mathscr{A},\boldsymbol{p}} \boldsymbol{v}, \boldsymbol{w} \right\rangle$. That is equivalent to prove that $\boldsymbol{p} \wedge \mathbb{E}_{|\mathscr{A},\boldsymbol{p}}$ is self-adjoint with respect to the standard inner product, which can be verified by simply computing it.

Identical reasoning can be done for $\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$. Note that $\mathrm{image}(\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}) = \boldsymbol{p} \wedge \mathscr{A}$, that $(\mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}})^2 = \mathbb{E}^T_{|\mathscr{A},\boldsymbol{p}}$ and that $\boldsymbol{p}^{-1} \wedge \mathbb{E}_{|\mathscr{A},\boldsymbol{p}}$ is self adjoint with respect to the standard inner product and the statement is proved. □

Proof of Proposition 3. Both $\ker C \supseteq \mathcal{N}$ and $\mathcal{S} \subseteq \mathcal{R}$ are well-known properties. We have to prove that $\mathcal{N} \supseteq \mathcal{N}_C$ and $\mathcal{R} \subseteq \mathcal{R}_C$.

Regarding the reachable space we have that $\mathrm{span}\{P^{y_{0:l}} \boldsymbol{p}_0, \forall y_{0:l} \text{ s.t. } l = k, \forall \boldsymbol{p}_0 \in \mathcal{S}\} \supseteq \mathrm{span}\{P^k \boldsymbol{p}_0, \forall \boldsymbol{p}_0 \in \mathcal{S}\}$ for all $k \geqslant 0$. This is proven directly using lemma 5.

For the non-observable subspace it holds that

$$\begin{bmatrix} \boldsymbol{1}^T \mathrm{diag}(\boldsymbol{e}_0) P P_C^{y_{0:l-1}} \\ \vdots \\ \boldsymbol{1}^T \mathrm{diag}(\boldsymbol{e}_m) P P_C^{y_{0:l-1}} \end{bmatrix} = C P P_C^{y_{0:l-1}} \boldsymbol{p}_0.$$

Then, we have that $\ker[C P P_C^{y_{0:l-1}}] = \ker[C P^l]$ for all $y_{0:l-1}$ of length l, for any length l. Once again this is proved by using Lemma 5. Consider a vector $\boldsymbol{v} \in \ker[C P P_C^{y_{0:l-1}}]$ for all $y_{0:l-1}$. Then it holds that $C P P_C^{y_{0:l-1}} \boldsymbol{v} = 0$ summing both sides of this equation over all sequences $y_{0:l-1}$ of length $l - 1$ and using the Lemma above we obtain $C P^{|y_{0:l-1}|+1} \boldsymbol{v} = 0$, thus proving the statement.

The statement on the effective subspaces follows directly from the other two. □

Proof of Lemma 6. We shall start by constructing a vector $\widetilde{\boldsymbol{w}}$ such that $\mathrm{supp}(\widetilde{\boldsymbol{w}}) = \mathrm{supp}(\mathcal{W})$. Starting from it we then construct a vector $\overline{\boldsymbol{w}}$ such that it is a linear combination of every generator.

By definition of support of a vector space, for each $\boldsymbol{e}_i \in \mathrm{supp}(\mathcal{W})$ there exists a vector $\boldsymbol{x}_i \in \mathcal{W}$ such that $\boldsymbol{e}_i^T \boldsymbol{x}_i \neq 0$, forming a set $\{\boldsymbol{x}_i\}$. Without loss of generality, we assume $i = 0, \dots, m$ with $m = \dim(\mathrm{supp}(\mathcal{W}))$. We can then define $\widetilde{\boldsymbol{w}}_0 = \boldsymbol{x}_0$ and iteratively compute $\widetilde{\boldsymbol{w}}_i = \widetilde{\boldsymbol{w}}_{i-1} + \lambda_i \boldsymbol{x}_i$ with $\lambda_i \notin \{-\boldsymbol{e}_j^T \widetilde{\boldsymbol{w}}_{i-1} / \boldsymbol{e}_j^T \boldsymbol{x}_i, \forall j | \boldsymbol{e}_j^T \boldsymbol{x}_i \neq 0\} \cup \{0\}$. Since this set is finite, it is always possible to choose a suitable $\lambda_i \in \mathbb{R}$ for each i. At the end of the iteration process, we obtain $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}_m = \sum_{i=0}^m \lambda_i \boldsymbol{x}_i \in \mathcal{W}$. To prove that $\mathrm{supp}(\widetilde{\boldsymbol{w}}) = \mathrm{supp}(\mathcal{W})$ we can simply observe that $\boldsymbol{e}_j^T \widetilde{\boldsymbol{w}} = \sum_{i=0}^m \lambda_i \boldsymbol{e}_j^T \boldsymbol{x}_i \neq 0$ by construction for all $\boldsymbol{e}_j \in \mathrm{supp}(\mathcal{W})$. On the other hand, for every $\boldsymbol{e}_j \notin \mathrm{supp}(\mathcal{W})$, $\boldsymbol{e}_j^T \boldsymbol{x}_i = 0$ for all i and thus $\boldsymbol{e}_j^T \widetilde{\boldsymbol{w}} = 0$.

This $\widetilde{\boldsymbol{w}}$ must be described as a linear combination of some of the generators, say $\widetilde{\boldsymbol{w}} = \sum_{i \in S} \lambda_i \boldsymbol{w}_i$, for some set of indices S. We can then use the same procedure as before: take $\boldsymbol{w}_i$ such that $i \notin S$, by choosing any $\lambda_i \notin \{-\boldsymbol{e}_j^T \widetilde{\boldsymbol{w}} / \boldsymbol{e}_j^T \boldsymbol{w}_i, \forall j | \boldsymbol{e}_j^T \boldsymbol{w}_i \neq 0\} \cup \{0\}$ we have $\mathrm{supp}(\widetilde{\boldsymbol{w}} + \lambda_i \boldsymbol{w}_i) = \mathrm{supp}(\mathcal{W})$. Iterating this procedure on the remaining vectors $\{\boldsymbol{w}_i | i \notin S\}$ we obtain $\overline{\boldsymbol{w}} = \widetilde{\boldsymbol{w}} + \sum_{i \notin S} \lambda_i \boldsymbol{w}_i$ such that $\mathrm{supp}(\overline{\boldsymbol{w}}) := \mathrm{supp}(\mathcal{W})$. □