Memory-dependent abstractions of stochastic systems through the lens of transfer operators

Banse, Adrien; Delimpaltadakis, Giannis; Laurenti, Luca; Mazo Jr., Manuel; Jungers, Raphaël M.

DOI
10.1145/3716863.3718039

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the 28th International Conference on Hybrid Systems, HSCC 2025

Memory-dependent abstractions of stochastic systems through the lens of transfer operators

Adrien Banse
adrien.banse@uclouvain.be
UCLouvain
Louvain-la-Neuve, Belgium

Giannis Delimpaltadakis
i.delimpaltadakis@tue.nl
Eindhoven University of Technology
Eindhoven, Netherlands

Luca Laurenti
l.laurenti@tudelft.nl
Delft University of Technology
Delft, Netherlands

Manuel Mazo Jr.
m.mazo@tudelft.nl
Delft University of Technology
Delft, Netherlands

Raphaël M. Jungers
raphael.jungers@uclouvain.be
UCLouvain
Louvain-la-Neuve, Belgium

ABSTRACT

With the increasing ubiquity of safety-critical autonomous systems operating in uncertain environments, there is a need for mathematical methods for formal verification of stochastic models. Towards formally verifying properties of stochastic systems, methods based on discrete, finite Markov approximations – *abstractions* – thereof have surged in recent years. These are found in contexts where: either a) one only has partial, discrete observations of the underlying continuous stochastic process, or b) the original system is too complex to analyze, so one partitions the continuous state-space of the original system to construct a handleable, finite-state model thereof. In both cases, the abstraction is an approximation of the discrete stochastic process that arises precisely from the discretization of the underlying continuous process. The fact that the abstraction is Markov and the discrete process is not (even though the original one is) leads to approximation errors. Towards accounting for non-Markovianity, we introduce memory-dependent abstractions for stochastic systems, capturing dynamics with memory effects. Our contribution is twofold. First, we provide a formalism for memory-dependent abstractions based on transfer operators. Second, we quantify the approximation error by upper bounding the total variation distance between the true continuous state distribution and its discrete approximation.

CCS CONCEPTS

• **Computer systems organization** → *Embedded and cyber-physical systems*; • **Mathematics of computing** → **Markov processes**; • **Computing methodologies** → **Uncertainty quantification**.

KEYWORDS

Abstraction, Stochastic system, Transfer operator, Memory Markov model.

HSCC '25, May 6–9, 2025, Irvine, CA, USA

ACM ISBN 979-8-4007-1504-4/2025/05
https://doi.org/10.1145/3716863.3718039

ACM Reference Format:
Adrien Banse, Giannis Delimpaltadakis, Luca Laurenti, Manuel Mazo Jr., and Raphaël M. Jungers. 2025. Memory-dependent abstractions of stochastic systems through the lens of transfer operators. In *28th ACM International Conference on Hybrid Systems: Computation and Control (HSCC '25), May 6–9, 2025, Irvine, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3716863.3718039

1 INTRODUCTION

Autonomous systems operating in uncertain environments are becoming ubiquitous, with applications ranging from autonomous driving, to robots in rescue missions, smart grids, smart buildings, etc. [2, 19]. Towards safe deployment of such autonomous systems, mathematical methods to formally verify if they meet prespecified requirements (e.g., on safety or performance) are needed [10, 15].

To mathematically analyze the aforementioned systems, while accounting for uncertainty, *stochastic models* are often employed. In this context, methods that are based on discrete approximations of stochastic systems, called *abstractions*, have recently surged (see [13] for a survey on abstractions for stochastic systems). Abstractions arise in two different contexts, that nevertheless present many mathematical similarities: a) one has access only to partial, discrete observations of the underlying original stochastic process (this is related to the work on *Markov state models* [18, 20, 21]); b) the original system is too complex to analyze, so one partitions its continuous state-space to construct a finite-state abstraction (this is related to the work on *abstraction-based methods* [1, 5, 6, 12–14, 17]). In both cases, the abstraction takes the form of a finite (sometimes robust) Markov chain.

In virtue of the above, the abstraction is an approximation of the discrete stochastic process that arises from the discretization of the original process. In particular, while the discrete process is not Markov, even though the original continuous process is (see Figure 1 for an example), for computational reasons, the abstraction is generally constructed to be Markov. This leads to approximation errors. To alleviate approximation errors due to non-Markovianity, we introduce memory-dependent abstractions of stochastic systems (so far, only memoryless abstractions have been proposed; see *Related work* below). The introduction of memory aims precisely at capturing memory effects inherent in the non-Markovian discrete process.

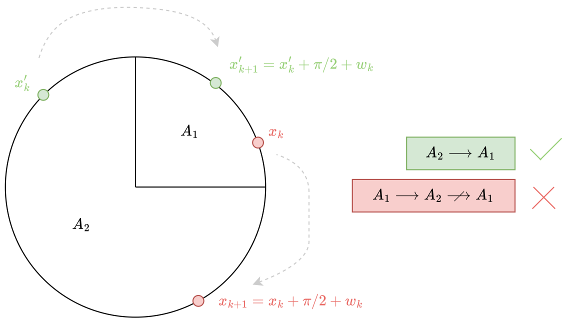

Figure 1: Loss of the Markov property. Consider a stochastic system defined by $x_{k+1} = x_k + \pi/2 + w_k \pmod{2\pi}$, where w_k is a noise whose support is $[0, \pi/10]$. While the original system is Markovian, the discrete process tracking only in which region x_k lies is not. Specifically, one can see that a state x_k can jump from A_2 to A_1 with non-zero probability, i.e. $\mathbb{P}[x_{k+1} \in A_1 | x_k \in A_2] > 0$. However, when a larger memory is considered, we see that, if this state was initially in A_1, it cannot jump successively to A_2 and back to A_1, i.e. $\mathbb{P}[x_{k+2} \in A_1 | x_{k+1} \in A_2, x_k \in A_1] = 0$. The Markov property is therefore lost, since $\mathbb{P}[x_{k+1} \in A_1 | x_k \in A_2] \neq \mathbb{P}[x_{k+2} \in A_1 | x_{k+1} \in A_2, x_k \in A_1]$.

Contributions. In this paper, we develop memory-dependent abstractions of stochastic systems. Inspired by symbolic dynamics [16], we extend the state space of the stochastic system to create a lifted system, where each state represents an ℓ-long sequence of states, ℓ being the considered memory. Further, akin to work on Markov state models [18, 20, 21], we employ transfer-operator theory and construct an ℓ-memory abstraction, through Galerkin approximations of the lifted process's transfer operator. Critically, we provide an upper bound on the total variation distance between the distribution of the original continuous state system and its discrete approximation, enabling formal verification through the abstract model. Finally, we showcase through examples how memory increases approximation accuracy in various situations. This work therefore marks a significant step toward creating smart memory-dependent abstractions for the analysis and control of complex systems.

Related work. Memory-dependent abstractions have been developed for the analysis and control of deterministic systems [3, 4, 22, 23]. However, to the best of our knowledge, such techniques have not been investigated for stochastic systems, where abstractions are memoryless (robust) Markov chains [1, 5, 6, 12, 14, 17]. Arguably, that is because incorporating memory in stochastic abstractions is fundamentally different than the deterministic case. For memory-dependent abstractions of deterministic systems, the *domino rule* is employed, which, deeply rooted in determinism, is simply not applicable for stochastic systems (see Figure 2 for an explanation). Thus, extending memory-dependent abstraction techniques to stochastic systems presents a non-trivial challenge, and requires fundamentally different mathematical tools, which we develop here.

Our work is also deeply related to and inspired by the work on Markov state models [18, 20, 21], which employs Galerkin approximations of the underlying Markov process's transfer operator, to build a finite approximation in a partial-observation scenario. Our contribution w.r.t. [18, 20, 21] is twofold. First, we introduce memory to the discrete approximation, whereas only the memoryless case is studied in those works. Second, although we use intermediate results proven in [21], our bounds - even in the memoryless case - fundamentally differ from those of the latter (see Remark 7 for more details).

Finally, our work is related to partially observable Markov decision processes (POMDPs for short, see [26] for a survey). Indeed, the considered dynamical systems can be framed in the formalism of POMDPs, where the continuous state space of the original system and the discrete cells of the partition are the state and observation spaces, respectively, such as e.g. in [28]. Although the loss of the Markov property for observations is a known phenomenon in POMDP literature [7], our contribution departs from this literature in that we focus on problems that arise in the framework of (safety-critical) abstractions. That is, we study the loss of the Markov property when the state space is continuous and is approximated by cells corresponding to sets of states.

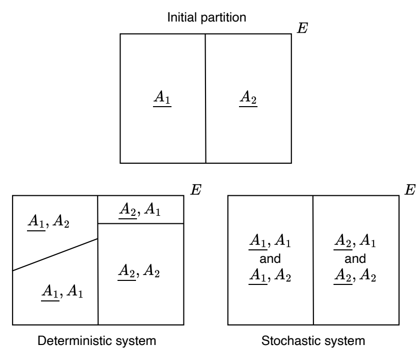

Figure 2: Domino rule for memory-dependent abstractions. Let A_1 and A_2 be two blocks of an initial partition on the state space E. A_1 and A_2 correspond to states of a 1-memory (memoryless) abstraction. Towards a 2-memory abstraction for deterministic systems, the domino rule proceeds as follows. The cell A_1 is divided into A_1, A_1 and A_1, A_2, the sets of states that are in A_1 and that will respectively be either in A_1 or in A_2 at the next timestep (the same happens to subdivide A_2). For stochastic systems, such division is not possible, as generally, even though the system might be in a specific state in A_1, it can visit any of A_1 or A_2 in the next step, due to stochasticity; in other words, there is no set of states in A_1 that deterministically visit either of A_1 and A_2 in the next step (similarly for the states initially in A_2).

Outline. This paper is structured as follows. Section 2 defines the considered family of systems and the studied problem. Section 3 introduces key theoretical concepts, covering probability theory, transfer operators, and Galerkin approximations. Section 4 contains our overall method to abstract systems with ℓ-memory

Markov models. We present the lifted system on which relies our analysis, and we provide a mathematical framework that justifies our method. In Section 5, we provide total variation guarantees. Section 6 provides numerical experiments, and Section 7 concludes with discussions on the approach, its limitation and future directions.

Notations. Given any set A, χ_A is the *indicator function* of A. $\mathbb{R}$ and $\mathbb{C}$ respectively denote the sets of *real* and *complex* numbers. The sets $\mathbb{R}_{>0}$ and $\mathbb{R}_{\geq 0}$ respectively denote the set of positive and non-negative real numbers. For $a \in \mathbb{C}$, $|a|$ denotes the *modulus* of a. For a natural number n, the set $[n]$ denotes $\{1, \ldots, n\}$. Let $E \subseteq \mathbb{R}^d$, we denote by $\mathcal{B}(E)$ the *Borel set of* E, and the couple $(E, \mathcal{B}(E))$ forms a measurable space. All along this work we consider *probability spaces* $(E, \mathcal{B}(E), \lambda)$, where $\lambda : \mathcal{B}(E) \to [0,1]$ is a *probability measure.* A set $(F, \langle \cdot, \cdot \rangle)$ is called a *Hilbert space* if $\langle \cdot, \cdot \rangle$ is a *dot product.* In this paper, given a measure μ on E, we consider the Hilbert space of *square-integrable functions*, noted $L^2(\mu)$, defined as the set of functions $f : E \to \mathbb{R}$ such that $\int_{x \in E} (f(x))^2 \mu(\mathrm{d}x) < +\infty$, and where the dot product is defined as $\langle f, g \rangle = \int_{x \in E} f(x) g(x) \mu(\mathrm{d}x)$. The associated norm is defined as $\|f\|_2 = \langle f, f \rangle^{1/2}$. Let $L^1(\mu)$ be the set of functions such that $\int_{x \in E} |f(x)| \mu(\mathrm{d}x) < +\infty$, we consider that μ is such that $L^2(\mu) \subseteq L^1(\mu)$, and we define $\|f\|_1 = \int_{x \in E} |f(x)| \mu(\mathrm{d}x)$. Let $P : L^2(\mu) \to L^2(\mu)$ be an operator, the *operator norm* of P is defined as $\|P\|_p = \sup_{f \in L^2(\mu) : \|f\|_p \leq 1} \|Tf\|_p$, for $p = 1, 2$. Finally, given a measurable space $(E, \mathcal{F})$, where $\mathcal{F}$ is any σ-algebra, and given two probability measures μ and ν on $(E, \mathcal{F})$, the *total variation distance* between μ and ν is defined as $\mathsf{TV}(\mu, \nu) = \sup_{A \in \mathcal{F}} |\sigma(A) - \nu(A)|$.

2 PROBLEM FORMULATION

2.1 System description

In this work, we consider discrete-time stochastic dynamical systems defined as

$$\begin{cases} x_{k+1} \sim \tau(\cdot | x_k), \\ x_0 \sim \lambda_0, \\ y_k = h(x_k), \end{cases} \tag{1}$$

where $x_k \in E$ is the *state* of the system at time k and the set $E \subseteq \mathbb{R}^d$ is the *state space.* For all $x \in E$, $\tau(\cdot|x) : \mathcal{B}(E) \to [0,1]$ is the *transition kernel.* The probability measure $\lambda_0 : \mathcal{B}(E) \to [0,1]$ is the *initial measure.* $y_k \in F$ is the *output* at time k, and the set F is the *output space*, which is assumed to be finite. Without loss of generality, we consider that $F = \{1, \ldots, n\}$ throughout the paper. The function $h : E \to F$ is called the *output function* and defines a *partition* of the state space E. Indeed, let

$$A_i = \{x \in E : h(x) = i\}, \tag{2}$$

then the collection of sets $A_1, \ldots, A_n$ is such that

$$\text{(covering)} \quad \bigcup_{i=1}^{n} A_i = E,$$
$$\text{(pairwise disjoint)} \quad \forall i \neq j : A_i \cap A_j = \emptyset.$$

A state sequence $x_0, x_1, \ldots$ of system (1) is a realisation of the *state stochastic process*, which is denoted by $(X_k^{\lambda_0})_{k \geq 0}$. When the initial measure is clear from the context, we omit λ_0 from the notation and simply write X_k. Given the definition of system (1), the probability measure associated to the state process is defined as

$$\mathbb{P}[X_0 \in A_0] = \lambda_0(A_0),$$
$$\mathbb{P}[X_k \in A_k] = \int_{x_{k-1} \in E} \cdots \int_{x_0 \in E} \tau(A_k | x_{k-1}) \ldots \tau(\mathrm{d}x_1 | x_0) \lambda_0(\mathrm{d}x_0),$$
$$\mathbb{P}[X_0 \in A_0, \ldots, X_k \in A_k] =$$
$$\int_{x_{k-1} \in A_{k-1}} \cdots \int_{x_0 \in A_0} \tau(A_k | x_{k-1}) \ldots \tau(\mathrm{d}x_1 | x_0) \lambda_0(\mathrm{d}x_0),$$

where $A_0, \ldots, A_k \in \mathcal{B}(E)$. We denote by λ_k the probability measure induced by $\mathbb{P}[X_k \in \cdot]$.

We now recall the definition of an *invariant measure.*

Definition 2.1 (Invariant measure). A measure $\mu : \mathcal{B}(E) \to [0,1]$ is said to be *invariant* for the system (1) if, for all $A \in \mathcal{B}(E)$, it holds that

$$\int_{x \in E} \tau(A|x) \mu(\mathrm{d}x) = \mu(A).$$

Assumption 1 (Existence and uniqueness). System (1) admits a unique invariant measure, denoted by μ.

Assumption 2 (Ergodicity). System (1) converges in total variation to the invariant measure, that is

$$\lim_{k \to \infty} \mathsf{TV}(\lambda_k, \mu) = 0.$$

Assumptions 1 and 2 are standard in the literature [8, 9], and commonly met in many cases of practical interest. We leave the extension to systems that do not satisfy Assumption 1 and Assumption 2 as future work.

The output sequence $y_0, y_1, \ldots$ of system (1) is a realisation of the *output stochastic process*, denoted by $(Y_k^{\lambda_0})_{k \geq 0}$. Again, we omit λ_0 when it is clear from the context. The output process is defined as

$$Y_k = h(X_k) \in \{1, \ldots, n\}.$$

Observe that, although the continuous process $(X_k)_{k \geq 0}$ is a Markov process, the discrete process $(Y_k)_{k \geq 0}$ is generally non-Markovian [3, 21], that is

$$\mathbb{P}[Y_{k+1} = i_{k+1} | Y_k = i_k, \ldots, Y_0 = i_0] \neq \mathbb{P}[Y_{k+1} = i_{k+1} | Y_k = i_k].$$

We invite the reader to see Figure 1 for an illustrative example of this phenomenon.

2.2 Problem statement

In this paper, given an infinitely long output sequence $\{y_i\}_i$, we aim at approximating the continuous Markov process $(X_k)_{k \geq 0}$ with a *discrete, ℓ-memory Markov process*, denoted by $(\tilde{Y}_{\ell,k})_{k \geq 0}$. That is, we construct the discrete process $(\tilde{Y}_{\ell,k})_{k \geq 0}$ such that

$$\mathbb{P}[\tilde{Y}_{\ell,k} = i_k | \tilde{Y}_{\ell,k-1} = i_{k-1}, \ldots, \tilde{Y}_{\ell,0} = i_0]$$
$$= \mathbb{P}[\tilde{Y}_{\ell,k} = i_k | \tilde{Y}_{\ell,k-1} = i_{k-1}, \ldots, \tilde{Y}_{\ell,k-\ell} = i_{k-\ell}].$$

In the literature, ℓ-memory Markov processes are also sometimes referred to as *Markov chains with memory ℓ* (see e.g. [27]), and constitute Markov chains in a lifted state space.

More precisely, we show that the discrete process $(\tilde{Y}_{\ell,k})_{k \geq 0}$ induces a probability measure $\tilde{\lambda}_{\ell,k}$ on the state space E, and we address the following problem.

Problem 1. Given an infinitely long output sequence $\{y_i\}_i$ of system (1), construct an ℓ-memory Markov process $(\tilde{Y}_{\ell,k})_{k\geq 0}$ and compute the distance

$$\mathsf{TV}(\lambda_k, \tilde{\lambda}_{\ell,k}), \tag{3}$$

where $\tilde{\lambda}_{\ell,k}$ is the probability measure on the state space E induced by $(\tilde{Y}_{\ell,k})_{k\geq 0}$.

Remark 1. In practice, as we employ ergodicity of system (1) (see Assumption 2), the given output sequence $\{y_i\}_i$ has to be sufficiently – not infinitely – long, so that the Markov process $(X_k)_{k\geq 0}$ has almost reached its steady state (in the sense that $\mathsf{TV}(\lambda_k, \mu) < \varepsilon$, for small ε). See also Remark 4. ∎

Observe that, through $\tilde{\lambda}_{\ell,k}$ and $\mathsf{TV}(\lambda_k, \tilde{\lambda}_{\ell,k})$, one can derive bounds on probabilistic properties of system (1) (e.g., bounds on the probability of the state X_k landing on some unsafe set, in some finite horizon). Our work is motivated by two distinct settings, which nevertheless present many mathematical similarities:

Case 1 - Partially observable systems. We would like to analyze properties of the underlying process $(X_k)_{k\geq 0}$, but can only observe (samples y_k of) the output process $(Y_k)_{k\geq 0}$. Thus, through observing $(Y_k)_{k\geq 0}$, we construct a discrete, ℓ-memory Markov approximation $(\tilde{Y}_{\ell,k})_{k\geq 0}$. Including memory is done precisely to capture non-Markovian effects inherent to the observed process $(Y_k)_{k\geq 0}$. As it will become evident both from the theory and the numerical examples, in certain cases, increasing memory ℓ leads to tighter approximations. Overall, this is related to earlier works on Markov state models [18, 20, 21].

Case 2 - Finite abstractions. System (1) is fully observable (disregard the output process), but too complex to derive analytic results on its properties. Akin to standard abstraction methods, we discretize the state space to derive a finite partition $\{A_i\}_i$. In contrast, through our method, we start from a coarser partition $\{A_i\}_i$, and, towards tighter approximations, the refinement is performed through increasing the abstraction's memory ℓ; hence, the title of the paper. Alternatively, while standard abstraction methods, towards approximating $(X_k)_{k\geq 0}$, through partitioning the state space, implicitly approximate the non-Markovian process $(Y_k)_{k\geq 0}$ by a (1-memory) Markov process, we approximate $(Y_k)_{k\geq 0}$ by an ℓ-memory Markov process $(\tilde{Y}_{\ell,k})_{k\geq 0}$, aiming precisely at capturing the non-Markov effects introduced exactly by partitioning the state-space in the first place

Approach. To address Problem 1, we rely on the transfer operator of the stochastic system and its spectral properties. In particular, we define a lifted system that describes the evolution of a ℓ-long sliding window of states $(x_k, \dots, x_{k+\ell-1})$. We then construct a memory Markov process so that the transition matrix of the latter is a Galerkin approximation of the transfer operator of the lifted system. We derive two upper bounds on the total variation (3). The first one consists of accumulation of projection errors, and increases with k. The second one is a consequence of the convergence of both models to the invariant measure, and decreases with k. For any memory ℓ, the combination of these two bounds therefore provides a computable upper bound on the total distance.

3 PRELIMINARIES

As explained in Section 2, our results rely on the theory of transfer operators, and their Galerkin approximation. This section formally introduces these concepts.

3.1 Probability theory

We recall that μ denotes the unique invariant measure of system (1). We first give a definition of *μ-weighted probability density functions.*

Definition 3.1 (μ-weighted probability density function [20]). A function $v : E \to \mathbb{R}$ is called a *μ-weighted probability density function* if it is such that

$$\int_{x\in E} v(x)\mu(\mathrm{d}x) = 1,$$

and $v(x) \geq 0$ μ-almost surely.

The following remark gives an interpretation of μ-weighted probability density functions with respect to usual probability density functions.

Remark 2. Let $\sigma : \mathcal{B}(E) \to [0, 1]$ be any probability measure on the state space. If there exists $v : E \to [0, 1]$ such that

$$\sigma(A) = \int_{x\in A} v(x)\mu(\mathrm{d}x),$$

for all $A \in \mathcal{B}(E)$ and if the latter is uniquely defined up to μ-null sets, then v is the so-called *Radon-Nikodym derivative* of σ with respect to the invariant measure, and is denoted by $\mathrm{d}\sigma/\mathrm{d}\mu$ (see e.g. [25] for more details). By Definition 3.1, v is also a μ-weighted probability density function. Note that, if it exists, the Radon-Nikodym derivative of σ with respect to the Lebesgue measure, denoted by p, is known as the usual probability density function, and satisfies

$$\sigma(A) = \int_{x\in A} p(x)\mathrm{d}x$$

for all $A \in \mathcal{B}(E)$. ∎

In this paper, since we assume the existence of a unique invariant distribution, we only work with μ-weighted probability density functions and simply refer to them as probability density functions.

Remark 3. Since $\mu(A) = \int_{x\in A} \mu(\mathrm{d}x)$, the probability density function corresponding to the invariant measure is the constant function $\mathbb{1}(x) = 1$ for all $x \in E$. ∎

Finally, consider two μ-weighted probability density functions v_1 and v_2 that respectively correspond to two probability measures λ_1 and λ_2, then, the identity

$$\mathsf{TV}(\lambda_1, \lambda_2) = \frac{1}{2}\|v_1 - v_2\|_1 = \frac{1}{2}\int_{x\in E} |v_1(x) - v_2(x)|\mu(\mathrm{d}x)$$

holds [8].

3.2 Transfer operator

The *transfer operator* corresponding to a transition kernel τ is defined as follows (see e.g. [21, 24]).

Definition 3.2 (Transfer operator). Given a state space E and a kernel τ, the *transfer operator* is the operator $T : L^2(\mu) \to L^2(\mu)$ such that

$$\int_{x\in E} \tau(A|x)v(x)\mu(\mathrm{d}x) = \int_{y\in A} (Tv)(y)\mu(\mathrm{d}y)$$

for all functions $v \in L^2(\mu)$ and all sets $A \in \mathcal{B}(E)$.

If the measure $\tau(\cdot|x)$ admits a Radon-Nikodym derivative $t(\cdot|x) := \mathrm{d}\tau(\cdot|x)/\mathrm{d}\mu$ for all $x \in E$, then the transfer operator is explicitly defined as

$$(Tv)(y) = \int_{x\in E} t(y|x)v(x)\mu(\mathrm{d}x) \tag{4}$$

for all functions $v \in L^2(\mu)$.

Intuitively, the operator T propagates square-integrable functions of the state space in time, including probability density functions in $L^2(\mu)$. Let $v_0 = \mathrm{d}\lambda_0/\mathrm{d}\mu \in L^2(\mu)$ be the probability density of λ_0 (the initial measure of System 1) then one can verify that

$$\lambda_k(A) := \mathbb{P}[X_k \in A] = \int_{x\in A} (T^k v_0)(x)\mu(\mathrm{d}x). \tag{5}$$

We therefore write $v_k := T^k v_0$ in the rest of this paper. Note that v_k is the μ-weighted probability density function of the measure λ_k. Furthermore, as pointed out in [21], since $\mathbb{1}$ is the invariant density (see Remark 3), T satisfies $T\mathbb{1} = \mathbb{1}$, and $\mathbb{1}$ is the only fixed point of T. Finally we define the spectrum of T.

Definition 3.3 (Transfer operator spectrum). Let $e \in \mathbb{C}$ and $u \in L^2(\mu)$. If

$$Tu = eu,$$

then e and u form a pair of *eigenvalue* and *eigenfunction* of T. All such pairs form the *spectrum* of T. Moreover, a set of eigenfunctions $\{u_1, \ldots, u_m\}$ is said to be *orthonormal* if

$$\langle u_i^\ell, u_j^\ell \rangle = 0$$

for all $i, j = 1, \ldots, m$ such that $i \neq j$, and

$$\|u_i^\ell\|_2 = \sqrt{\langle u_i^\ell, u_i^\ell \rangle} = 1$$

for all $i = 1, \ldots, m$, where $\langle \cdot, \cdot \rangle$ is the inner product defined as

$$\langle f, g \rangle = \int_{x\in E} f(x)g(x)\mu(\mathrm{d}x). \tag{6}$$

Note that the invariant density $\mathbb{1}$ is an eigenfunction of T with eigenvalue 1, since $T\mathbb{1} = \mathbb{1}$.

3.3 Galerkin methods

In this work, inspired by works on Markov state models [18, 20, 21], we use a Galerkin method to approximate a lifted transfer operator.

Definition 3.4 (Projection operator). Let $(H, \langle \cdot, \cdot \rangle)$ be a Hilbert space, and let $\|\cdot\|$ be its associated norm. For a closed subspace $D \subset H$, the surjective map $Q : H \to D$ is an orthogonal projection onto D if $Q^2 = Q$ and $\sup_{f\in H:\|f\|=1} \|Qf\| = 1$.

Consider $(L^2(\mu), \langle \cdot, \cdot \rangle)$, the Hilbert space of square-integrable functions, where $\langle \cdot, \cdot \rangle$ is the usual dot product (6), together with a closed subspace $D \subset L^2(\mu)$ generated by a finite set of functions, that is

$$D = \mathrm{span}(\{\phi_1, \ldots, \phi_n\})$$

for n functions $\phi_i \in L^2(\mu)$, and let $Q : L^2(\mu) \to D$ be the unique projection operator defined above. Then the operator

$$P := QTQ : D \to D$$

is called the *Galerkin approximation* of T. Since $P : D \to D$ is a finite-dimensional operator, it admits a *matrix representation* $\mathbf{P} \in \mathbb{R}^{n\times n}$, defined as follows. For all $f \in D$, let $\mathbf{f} \in \mathbb{R}^n$ be the vector representation of f in the basis $\{\phi_i\}_{i\in[n]}$, that is

$$f = \sum_{i=1}^n \mathbf{f}_i \phi_i.$$

Then, if $f' = Pf$ with vector representation $\mathbf{f}'$, it holds that

$$\mathbf{f}' = \mathbf{P}^\top \mathbf{f}.$$

In the following, we use boldface symbols to denote vector/matrix representations of functions/operators.

4 MEMORY-DEPENDENT ABSTRACTIONS

In this section, we show how to approximate the true densities v_k of system (1) with the densities $\tilde{v}_{\ell,k}$ obtained from an ℓ-memory Markov abstraction, itself derived by a Galerkin approximation of the transfer operator of a lifted system. Finally, we prove the correctness of our approach by upper bounding $\mathsf{TV}(\lambda_k, \tilde{\lambda}_{\ell,k})$.

4.1 Overall method

Our approach to approximate v_k is summarized in Algorithm 1. We recall that, in what follows, χ_A denotes the indicator function of set A. Also, we consider transition matrices of ℓ-memory Markov models, denoted $\mathbf{P}_\ell \in \mathbb{R}^{n^\ell \times n^\ell}$, where each row and column of such a matrix is labeled by ℓ-long sequences of outputs $i_0 i_1 \ldots i_{\ell-1}$, where $i_j \in \{1, \ldots, n\}$. Moreover, the matrix is such that

$$i_1 \ldots i_{\ell-1} \neq j_1 \ldots j_{\ell-1} \implies (\mathbf{P}_\ell)_{i_0\ldots i_{\ell-1}, j_1\ldots, j_\ell} = 0.$$

Indeed the state $i_0 \ldots i_{l-1}$ of the ℓ-memory Markov model represents the event $Y_0 = i_0, \ldots, Y_{\ell-1} = i_{\ell-1}$, and the state $j_1, \ldots, j_\ell$ represents the event $Y_1 = j_1, \ldots, Y_\ell = j_\ell$. Thus naturally there can be no transition from $i_0, \ldots, i_{\ell-1}$ to $j_1, \ldots, j_\ell$ if $i_1 \ldots i_{\ell-1} \neq j_1 \ldots j_{\ell-1}$. As a consequence, the matrix $\mathbf{P}_\ell$ contains only $n^{\ell+1}$ possibly non-zero entries. Each considered vector $\mathbf{v}^\ell \in \mathbb{R}^{n^\ell}$ also has entries labeled with $i_0 \ldots i_{\ell-1}$, such that the matrix-vector product $\mathbf{P}_\ell \mathbf{v}^\ell$ is well defined.

Algorithm 1 proceeds as follows:

1. An ℓ-memory Markov chain is built, based on the steady-state dynamics (the dynamics when on the invariant measure). Each entry $(\mathbf{P}_\ell)_{i_0\ldots i_{\ell-1}, i_1\ldots i_\ell}$ contains the probability to go to the blocks A_{i_ℓ} knowing the ℓ last blocks were $A_{i_0}, \ldots, A_{\ell-1}$. It directly follows from (7) that this matrix is stochastic.
2. The initial probability vector on the n^ℓ output sequences $i_0 \ldots i_{\ell-1}$ is computed. Again, it follows from (8) that this vector sums to 1.
3. This probability is propagated $k - \ell + 1$ times with the ℓ-Markov chain transition matrix $\mathbf{P}_\ell$. The vector now contains entries labeled $i_{k-\ell+1} \ldots i_k$ containing the approximated output joint probability from time $k - \ell + 1$ to time k.
4. The joint probability is marginalized so that the vector $(\tilde{\mathbf{v}}_{\ell,k})_{i_k}$ contains the approximated probabilities at time k, from which one may compute $\tilde{v}_{\ell,k}$.

The returned function in Algorithm 1 is a piecewise constant μ-weighted probability density function, denoted $\tilde{v}_{\ell,k}$. It is such that

$$\tilde{\lambda}_{\ell,k}(A) = \int_{x\in A} \tilde{v}_{\ell,k}(x)\mu(\mathrm{d}x) \approx \int_{x\in A} v_k(x)\mu(\mathrm{d}x) = \lambda_k(A)$$

Algorithm 1 Compute $\tilde{v}_{\ell,k}$, the ℓ-memory approximation of v_k

1: Compute the ℓ-memory transition probabilities of the output process $(Y_k^\mu)_{k\geq 0}$ of system (1), initialized at the invariant measure

$$(\mathbf{P}_\ell)_{i_0\dots i_{\ell-1},i_1\dots i_\ell} := \mathbb{P}\left[Y_\ell^\mu = i_\ell \middle| Y_0^\mu = i_0,\dots,Y_{\ell-1}^\mu = i_{\ell-1}\right]. \quad (7)$$

2: Compute the initial ℓ-long joint probabilities

$$(\tilde{\mathbf{v}}_0^\ell)_{i_0\dots i_{\ell-1}} := \mathbb{P}\left[Y_0^{\lambda_0} = i_0,\dots,Y_{\ell-1}^{\lambda_0} = i_{\ell-1}\right] \quad (8)$$

3: Propagate the l-long joint probabilities with the ℓ-memory Markov model

$$\tilde{\mathbf{v}}_{k-\ell+1}^\ell = \left(\mathbf{P}_\ell^{k-\ell+1}\right)^\top \tilde{\mathbf{v}}_0^\ell \quad (9)$$

4: Marginalize

$$(\tilde{\mathbf{v}}_{\ell,k})_{i_k} := \sum_{i_{k-\ell+1}=1}^{n}\cdots\sum_{i_{k-1}=1}^{n}(\tilde{\mathbf{v}}_{k+\ell-1}^\ell)_{i_{k-\ell+1}\dots i_{k-1}i_k}$$

return $\tilde{v}_{\ell,k}(x_k) = \sum_{i_k=1}^n \frac{(\tilde{\mathbf{v}}_{\ell,k})_{i_k}}{\mathbb{P}[Y_k^\mu = i_k]}\chi_{A_{i_k}}(x_k)$.

for all $A \in \mathcal{B}(E)$, where $v_k = T^k v_0$ is the true μ-weighted probability density function at time k, and $\approx$ denotes that we use $\tilde{\lambda}_{\ell,k}$ (or $\tilde{v}_{\ell,k}$) as an approximation of λ_k (resp. v_k). This will be further explained in the next subsections.

Remark 4. In practice, computing the invariant output probabilities (7) can be done in at least two ways. Either one samples a sufficiently large number of $(\ell+1)$-long output traces initialized at the invariant distribution. Or, employing ergodicity and Birkhoff's theorem (see e.g. [25]), one samples a very large output trace initialized at any initial distribution. ■

In the rest of this section, we introduce the mathematical formalism surrounding the construction of the abstraction.

4.2 Lifted system

Our approach is based on the study of the lifted state process $(X_k,\dots,X_{k+\ell-1})_{k\geq 0}$ and output process $(Y_k,\dots,Y_{k+\ell-1})_{k\geq 0}$. In the following subsections, we show that the abstraction constructed in Algorithm 1 is a Galerkin approximation of the transfer operator of this lifted process. In this section, we formally define it along with its invariant distribution. We then conclude by making the link with the original system 1.

The lifted system is defined as

$$\begin{cases}(x_{k+1},\dots,x_{k+\ell}) \sim \tau^\ell(\cdot|x_k,\dots,x_{k+\ell-1}),\\ (x_0,\dots,x_{\ell-1}) \sim \lambda_0^\ell,\\ (y_k,\dots,y_{k+\ell-1}) = (h(x_k),\dots,h(x_{k+\ell-1})).\end{cases} \quad (10)$$

In the definition above, for all $A_1,\dots,A_\ell \in \mathcal{B}(E)$ and all $x_0,\dots,x_{\ell-1} \in E$, the lifted kernel τ^ℓ is defined as

$$\tau^\ell(A_1\times\cdots\times A_\ell|x_0,\dots,x_{\ell-1}) = \begin{cases}\tau(A_\ell|x_{\ell-1}) & \text{if } x_1 \in A_1,\dots,x_{\ell-1}\in A_{\ell-1},\\ 0 & \text{otherwise.}\end{cases} \quad (11)$$

For all sets $A_0,\dots,A_{\ell-1} \in \mathcal{B}(E)$, the initial measure λ_0^ℓ is defined as

$$\lambda_0^\ell(A_0\times\cdots\times A_{\ell-1}) = \int_{x_0\in A_0}\cdots\int_{x_{\ell-1}\in A_{\ell-1}}\tau(\mathrm{d}x_{\ell-1}|x_{\ell-2})\dots\tau(\mathrm{d}x_1|x_0)\lambda_0(\mathrm{d}x_0).$$

Owing to Assumption 1, the lifted system admits a unique invariant measure μ^ℓ (see e.g. [8, Equation (4.1)]), defined as

$$\mu^\ell(A_0\times\cdots\times A_{\ell-1}) = \int_{x_0\in A_0}\cdots\int_{x_{\ell-1}\in A_{\ell-1}}\tau(\mathrm{d}x_{\ell-1}|x_{\ell-2})\dots\tau(\mathrm{d}x_1|x_0)\mu(\mathrm{d}x_0). \quad (12)$$

Lifted system (10) admits a transfer operator T_ℓ, according to Definition 3.2. The initial measure λ_0^ℓ admits a μ^ℓ-weighted probability density function, denoted $v_0^\ell(x_0,\dots,x_{\ell-1})$, which is a *joint probability density function* on the first ℓ states. These joint densities are propagated with the lifted transfer operator, and, for all $k \geq \ell-1$,

$$v_{k-\ell+1}^\ell(x_{k-\ell+1},\dots,x_k) = (T_\ell^{k-\ell+1}v_0)(x_{k-\ell+1},\dots,x_k)$$

is the joint probability density on the states $x_{k-\ell+1}$ to x_k. The corresponding measure is denoted by $\lambda_{k-\ell+1}^\ell$.

Remark 5 (Notations). Study of joint measures and joint probability density functions are at the center of this work. We therefore draw the reader's attention on the fact that, all along the paper, we note joint measure (resp. μ^ℓ-weighted density) on E^ℓ with a superscript λ^ℓ (resp. v^ℓ), whereas measures (resp. μ-weighted densities) on E are without any superscript λ (resp. v). In contrast, superscripts on operators, e.g., P^k or T^k, denote powers (or recursive applications of the operator). ■

4.3 Abstraction

In this section, we show that the transition matrix of the ℓ-memory Markov chain constructed in Algorithm 1 corresponds to a Galerkin approximation of the transfer operator T_ℓ. In particular, we specify the basis of functions with which we project T_ℓ, and re-interpret Algorithm 1 in terms of functions and operators. Doing this will allow us to derive bounds in Section 5 on the total variation distance between v_k and the approximated function $\tilde{v}_{\ell,k}$ given by Algorithm 1.

Given the output partition $A_1,\dots,A_n$ on the original state space E (as defined in (2)), we consider the subspace of piecewise constant functions $D_n^\ell \subset L^2(\mu^\ell)$, defined as

$$D_n^\ell := \operatorname{span}\left(\left\{\psi_{i_1\dots i_\ell}\right\}_{i_1,\dots,i_\ell\in[n]}\right),$$

where

$$\psi_{i_1\dots i_\ell}(x_1,\dots,x_\ell) := \frac{\chi_{A_{i_1}}(x_1)\dots\chi_{A_{i_\ell}}(x_\ell)}{\mathbb{P}[X_1^\mu\in A_{i_1},\dots,X_\ell^\mu\in A_{i_\ell}]}, \quad (13)$$

where we recall that χ_A denotes the indicator function of A. D_n^ℓ is therefore a set of piecewise constant functions on E^ℓ.

In this work, we make the assumption that the denominator in (13) is positive (as formally stated below in Assumption 3). We claim that this assumption is not restrictive for two main reasons. First, it holds in many practical cases such as unbounded noise. Second, it also suffices to assume that $\mathbb{P}[X_1^\mu\in A_{i_1},\dots,X_\ell^\mu\in A_{i_\ell}] = 0$ implies

$\mathbb{P}[X_1^{\lambda_0} \in A_{i_1}, \dots, X_\ell^{\lambda_0} \in A_{i_\ell}] = 0$. That is, zero measure steady state correspond to zero measure initial conditions. For the sake of brevity and simplicity, we leave this extension for further work.

Assumption 3. The system (1) is such that its invariant measure μ satisfies

$$\mathbb{P}[X_1^\mu \in A_{i_1}, \dots, X_\ell^\mu \in A_{i_\ell}] > 0 \tag{14}$$

for all sequence $A_{i_1}, \dots, A_{i_\ell}$ of blocks of the output partition.

The following proposition shows that $\mathbf{P}_\ell$, the matrix built in Algorithm 1, is the matrix representation of the Galerkin approximation of T_ℓ. For the sake of readability, all proofs can be found in Appendix A.

Proposition 4.1. *Let $Q_\ell : L^2(\mu^\ell) \to D_n^\ell$ be a projection operator as defined in Definition 3.4, and let*

$$P_\ell \coloneqq Q_\ell T_\ell Q_\ell$$

be the Galerkin approximation of T_ℓ on D_n^ℓ. Then it holds that

$$\begin{aligned} &P_\ell \psi_{i_0 \dots i_{\ell-1}} \\ &= \sum_{i_\ell=1}^{n} \mathbb{P}\left[Y_\ell^\mu = i_\ell \middle| Y_0^\mu = i_0, \dots, Y_{\ell-1}^\mu = i_{\ell-1}\right] \psi_{i_1 \dots i_\ell} \end{aligned} \tag{15}$$

for all $i_0, \dots, i_{\ell-1} \in [n]$. Therefore $\mathbf{P}_\ell$, as defined in (7), is the matrix representation of P_ℓ.

We can therefore re-interpret Algorithm 1 through the lens of transfer operators and their Galerkin approximations. Figure 3 summarizes this interpretation. First, we compute P_ℓ, the Galerkin approximation of T_ℓ, whose matrix representation is $\mathbf{P}_\ell$, as computed in (7) in Algorithm 1. Second, we compute the piecewise approximate initial density

$$\begin{aligned} v_0^\ell(x_0, \dots, x_{\ell-1}) &\approx \tilde{v}_0^\ell(x_0, \dots, x_{\ell-1}) \\ &= (Q_\ell v_0^\ell)(x_0, \dots, x_{\ell-1}), \end{aligned}$$

whose vector representation is $\tilde{\mathbf{v}}_0^\ell$, computed in (8) in Algorithm 1. Third, we approximate the joint density from $x_{k-\ell+1}$ to x_k with

$$\begin{aligned} v_{k-\ell+1}^\ell(x_{k-\ell+1}, \dots, x_k) &\approx \tilde{v}_{k-\ell+1}^\ell(x_{k-\ell+1}, \dots, x_k) \\ &= (P_\ell^{k-\ell+1} v_0^\ell)(x_{k-\ell+1}, \dots, x_k). \end{aligned} \tag{16}$$

Since $P_\ell = Q_\ell T_\ell Q_\ell$, it holds that

$$P_\ell^{k-\ell+1} v_0^\ell = P_\ell^{k-\ell+1}(Q_\ell v_0^\ell) = P_\ell^{k-\ell+1} \tilde{v}_0^\ell,$$

and therefore that the vector representation of $\tilde{v}_{k-\ell+1}^\ell$ is $\tilde{\mathbf{v}}_{k-\ell+1}^\ell$, as computed in (9). Finally, we marginalize $\tilde{v}_{k-\ell+1}^\ell(x_{k-\ell+1}, \dots, x_k)$ to get $\tilde{v}_{\ell,k}(x_k)$, whose vector representation in D_n is $\tilde{\mathbf{v}}_{\ell,k}$.

5 TOTAL VARIATION GUARANTEES

In this section we upper bound $\mathsf{TV}(v_k, \tilde{v}_{\ell,k})$, the total variation between the true density v_k and $\tilde{v}_{\ell,k}$, the approximated density given by Algorithm 1, thereby providing formal guarantees on the correctness of our approach. To derive our bounds, we require certain assumptions on the spectrum of the lifted transfer operator T_ℓ, which are equivalent to those considered in [21] and formally defined in Assumption 4.

Assumption 4. The lifted transfer operator T_ℓ admits m real eigenvalues $e_{\ell,1}, \dots, e_{\ell,m}$ such that $1 = e_{\ell,0} > e_{\ell,1} \geq \dots e_{\ell,m}$, together with an orthonormal set of eigenfunctions $\mathbb{1} = u_0^\ell, u_1^\ell, \dots, u_m^\ell$ (see Definition 3.3). Moreover, for some $r_\ell < |e_{\ell,m}|$, all remaining eigenvalues $e_{\ell,i} \in \mathbb{C}$ for $i > m$ are such that $|e_{\ell,i}| < r_\ell$. Finally let $\Pi_\ell : L^2(\mu^\ell) \to L^2(\mu^\ell)$ be the operator defined by

$$\Pi_\ell f^\ell = \sum_{i=1}^{m} \langle f^\ell, u_i^\ell \rangle u_i^\ell. \tag{17}$$

We assume that $\Pi_\ell T_\ell = \Pi_\ell T_\ell \Pi_\ell$.

Remark 6. Assumption 4 is related but not the same as Assumption 1 and Assumption 2. As discussed in [21], sufficient conditions for Assumption 4 to hold are *reversibility* and *sufficient ergodicity* (such as defined in [21, Remark 2.1]), which are stronger assumptions than Assumption 1 and Assumption 4. As stated in [11, 21], reversibility and sufficient ergodicity are natural and satisfied for a large class of dynamical systems. ■

Similarly as in [21], our bound relies on the quantity

$$\delta_\ell \coloneqq \max_{i=1,\dots,m} \|Q_\ell u_i^\ell - u_i^\ell\|_2, \tag{18}$$

which quantifies the maximal projection error on the spectrum.

Our total variation bound consists of two components. The first component increases with k and arises from the cumulative projection errors, becoming more conservative as k grows. On the other hand, the second component is characterized by the convergence of both the true density and the approximated one towards the invariant density $\mathbb{1}$, and decreases with k. Unlike the first component, it is initially conservative but tightens progressively over time. The increasing and decreasing components are studied respectively in Theorem 5.1 and Theorem 5.2 for the joint densities, and the final bound is given for the marginalized densities in Corollary 5.3. We stress that both Theorem 5.1 and Theorem 5.2 bound the same quantity. However, these bounds are complementary, as one is producing better bounds for small k, while the other for large k.

Theorem 5.1 (Increasing). *For any memory $\ell \geq 1$, horizon $k \geq \ell$ and initial joint density $v_0^\ell \in L^2(\mu^\ell)$, let $\lambda_{k-\ell+1}^\ell$ and $\tilde{\lambda}_{k-\ell+1}^\ell$ be the joint measures respectively defined by*

$$\begin{aligned} v_{k-\ell+1}^\ell &= T_\ell^{k-\ell+1} v_0^\ell, \\ \tilde{v}_{k-\ell+1}^\ell &= P_\ell^{k-\ell+1} v_0^\ell, \end{aligned}$$

and similarly for $\lambda_{k-\ell}^\ell$ and $\tilde{\lambda}_{k-\ell}^\ell$. Then, if Assumption 4 is satisfied, it holds that

$$\begin{aligned} \mathsf{TV}(\lambda_{k-\ell+1}^\ell, \tilde{\lambda}_{k-\ell+1}^\ell) \leq{}& \mathsf{TV}(\lambda_{k-\ell}^\ell, \tilde{\lambda}_{k-\ell}^\ell) \\ &+ \frac{1}{2}\left(m e_{\ell,1} \delta_\ell + r_\ell\right) e_{\ell,1}^{k-\ell} \|v_0^\ell\|_2, \end{aligned} \tag{19}$$

where δ_ℓ is defined in (18).

Theorem 5.2 (Decreasing). *For any memory $\ell \geq 1$, horizon $k \geq \ell$ and initial joint density $v_0^\ell \in L^2(\mu^\ell)$, let $\lambda_{k-\ell+1}^\ell$ and $\lambda_{k-\ell+1}^\ell$ be the joint measures respectively defined as in (19). Then, if Assumption 4 is satisfied, it holds that*

$$\mathsf{TV}(\lambda_{k-\ell+1}^\ell, \tilde{\lambda}_{k-\ell+1}^\ell) \leq e_{\ell,1}^{k-\ell+1} \|v_0^\ell\|_2.$$

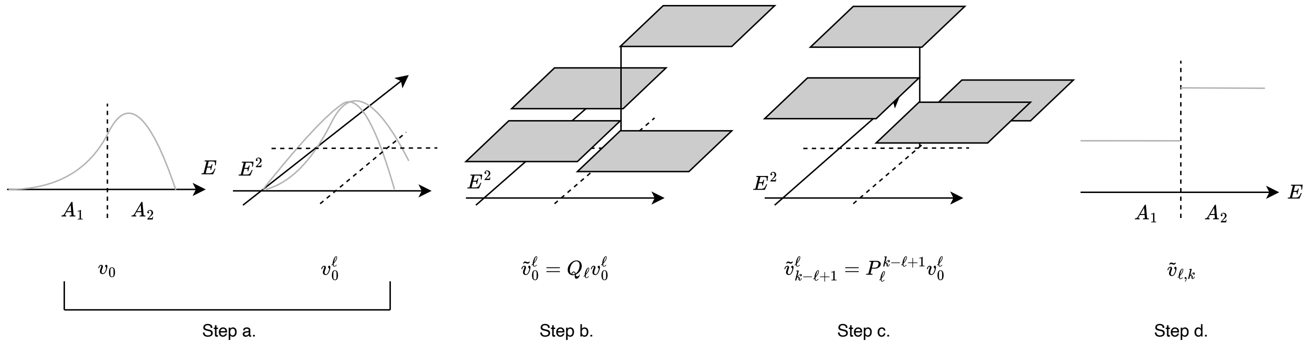

Figure 3: Summary of the method described in Algorithm 1, from left to right. Step a) We consider the lifted process defined in (10)**. Step b) We project the x_0-to-$x_{\ell-1}$ joint distribution on a finitely generated space D_n^ℓ. Step c) We propagate it with the Galerkin approximation P_ℓ to get the approximate $x_{k-\ell+1}$-to-x_k joint distribution. Step d) We marginalize it to retrieve the approximate density at x_k.**

Note that the following corollary is a result on the final marginalized densities, and not the joint densities such as in Theorem 5.1 and Theorem 5.2.

Corollary 5.3. *For any memory $\ell \geq 1$, horizon $k \geq \ell$ and initial joint density $v_0^\ell \in L^2(\mu^\ell)$, let*

$$\overline{\mathsf{TV}}_{inc} := \mathsf{TV}(\lambda_0^\ell, \tilde{\lambda}_0^\ell) + \left(me_{\ell,1}\delta_\ell + r_\ell\right) \frac{1 - e_{\ell,1}^{k-\ell+1}}{1 - e_{\ell,1}} \|v_0^\ell\|_2,$$
$$\overline{\mathsf{TV}}_{dec} := e_{\ell,1}^{k-\ell+1} \|v_0^\ell\|_2,$$

where δ_ℓ is defined in (18)*. Also, let λ_k and $\tilde{\lambda}_{\ell,k}$ be the measures respectively defined by*

$$v_k = T^k v_0,$$
$$\tilde{v}_{\ell,k} \text{ output of Algorithm 1.}$$

Then it holds that

$$\mathsf{TV}(\lambda_k, \tilde{\lambda}_{\ell,k}) \leq \min\left\{\overline{\mathsf{TV}}_{inc}, \overline{\mathsf{TV}}_{dec}\right\}.$$

Remark 7. We should stress that, although we rely on similar tools, our bounds differ from those of [21], even in the memoryless case ($\ell = 1$). Indeed the authors of [21] consider the operator norm

$$\|Q_1 T_1^k Q_1 - (Q_1 T_1 Q_1)^k\|_2$$

as the error, which makes the assumption that the ground truth probability is $v_k = Q_1 T_1^k Q_1 v_0$. The setting of [21] therefore does not correspond to the setting of this paper, as we consider that the ground truth is $v_k = T_1^k v_0$. Taking into account this continuous ground truth makes our error larger, and consists in a supplementary technical challenge than a simple extension of [21]. ∎

6 NUMERICAL EXPERIMENTS

In this section we motivate the method described in Section 4.1 for the two cases described in Section 2. In both cases, we consider the following dynamical system.

Example 6.1. Consider the 2-dimensional linear system of the form

$$\begin{cases} x_{k+1} = Ax_k + w_k, \\ w_k \sim \mathcal{N}(m_w, \Sigma_w), \\ x_0 \sim \mathcal{N}(m_0, \Sigma_0), \end{cases}$$

with

$$A = \begin{pmatrix} 0.995 & 0.005 \\ 0 & 0.98 \end{pmatrix} \tag{20}$$

and $m_w = (0,0)^\top, \Sigma_w = 0.07I_2, m_0 = (-0.4, -0.4)^\top, \Sigma_0 = 0.3I_2$, with I_2 the 2-dimensional identity matrix. Since A is stable, this system converges in total variation to a unique invariant distribution $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$, where $m_\mu = (0,0)^\top$ and

$$\Sigma_\mu \approx \begin{pmatrix} 7.36896 & 0.347856 \\ 0.347856 & 1.76768 \end{pmatrix}. \tag{21}$$

The latter was computed by solving the Riccati equation $\Sigma_\mu = A\Sigma_\mu A^\top + \Sigma_w$ with the `MatrixEquations.jl` package.[1] ∎

In the experiments below, the matrices $\mathbf{P}_\ell$ have been computed with one very long trajectory $\{y_i\}_{i=1,\ldots,10^5}$ (see Remark 4), and the initial vectors $\mathbf{v}_0^\ell$ have been computed with $10^5/\ell$ samples of length ℓ. More details about how $\mathsf{TV}(\lambda_k, \tilde{\lambda}_{\ell,k})$ has been computed in practice can be found in Appendix B.

Case 1 - Partially observable systems. In this case, the system is only partially observable, and we only have access to the outputs. The state space is discretized as follows: each dimension is partitioned into $(-\infty, -1)$, $(1, \infty)$, and the interval $[-1, 1]$ is further partitioned into p subintervals of equal size. Thus the partition contains $n = (p+2)^2$ cells. In this case, we fix $p = 3$, leading to a 25 cells partition, and we approximate the discrete process $(Y_k)_{k\geq 0}$ with the process $(\tilde{Y}_{\ell,k})$ as defined in Section 4, for $\ell = 1, 2, 3$. More precisely, we compute $\tilde{v}_{\ell,k}$ with Algorithm 1, and compute $\mathsf{TV}(\lambda_k, \tilde{\lambda}_{\ell,k})$ for $k \in \{0, \ldots, 100\}$.

The results are in Figure 4. One can see that increasing memory reduces $\mathsf{TV}(\lambda_k, \tilde{\lambda}_{\ell,k})$ for most horizons k, **thereby increasing the approximation quality**. Moreover, one can see that the observed

[1] See https://github.com/andreasvarga/MatrixEquations.jl.

bounds follow the theoretical setting of Theorem 5.1 and Theorem 5.2, as the bounds seem to follow two regimes, first increasing and then decreasing.

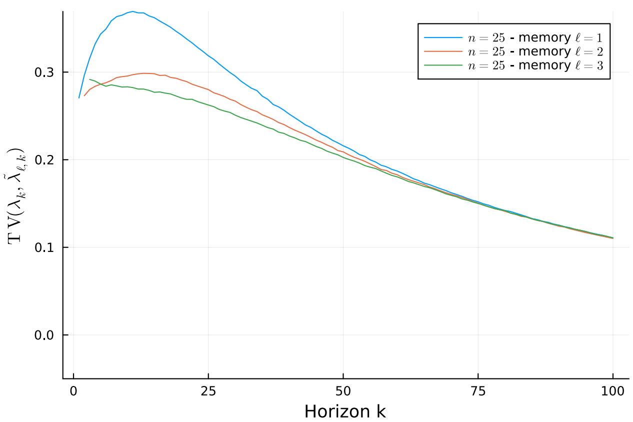

Figure 4: Approximation quality of ℓ-memory Markov models for partially observable systems. Introducing memory improves approximation quality.

Case 2 - Finite abstractions. In this case, the state process of the system is available, but the state space is discretized in order to abstract the system. We consider two comparable settings:

(1) (Classical Markov chain) The state space is discretized in $n = 729$ blocks in the same fashion as above (with $p = 25$). Memory $\ell = 1$ is considered, leading to $729^2 = 531441$ transition probabilities $(\mathbf{P}_1)_{i_1,i_2}$ with $i_1, i_2 \in \{1, \dots, 729\}$.

(2) (2-memory Markov model) The state space is discretized in the same way (uniformly in the square $[-1, 1]^2$) in $n = 81$ blocks ($p = 7$). Memory $\ell = 2$ is considered, also leading to $81^3 = 531441$ transition probabilities $(\mathbf{P})_{i_1 i_2, i_2 i_3}$ with $i_1, i_2, i_3 \in \{1, \dots, 81\}$.

The two settings lead to discrete objects of the same size, since one only needs to store 531441 values to save them. For these two settings, we compute $\tilde{v}_{\ell,k}$ with Algorithm 1, and compute $\mathrm{TV}(\lambda_k, \tilde{\lambda}_{\ell,k})$ for $k \in \{0, \dots, 100\}$. The results are in Figure 5. We observe that, even though the initial partition is coarser, larger memory leads to a better approximation, showcasing the fact that memory allows to construct **smarter abstractions than classical approaches**.

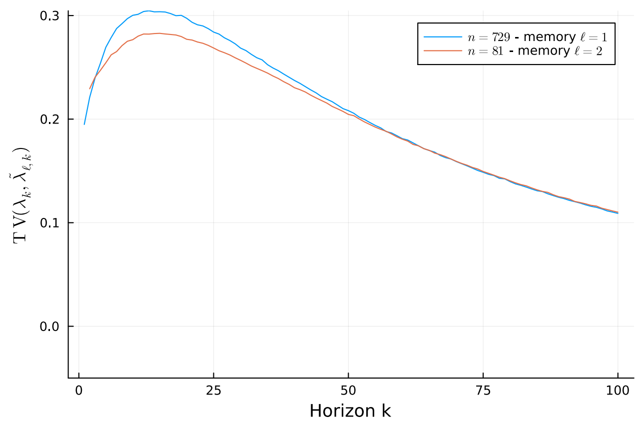

Figure 5: Approximation quality of memory-dependent abstractions of an observable system. One can see that, for the same number of transition probabilities ($n^{\ell+1} = 531441$ in both cases), starting from a coarser partition and increasing the memory leads to better approximations.

7 CONCLUSIONS AND FURTHER WORK

In summary, in this work, we have introduced memory-dependent abstractions for stochastic systems. Our formalism, based on Galerkin approximations of lifted transfer operators, provides a theoretical framework for studying these abstractions. We have also upper bounded the approximation error, that we define as the total variation distance between the true distribution on the state space and the one of the memory-dependent approximation. We showed that this error consists of two regimes, one increasing (because of the accumulation of projection errors), and one decreasing (thanks to ergodicity). Through numerical experiments, we have demonstrated that increasing memory reduces the approximation error in various scenarios, highlighting how memory-dependent abstractions effectively address the issue of non-Markovianity of the discrete process induced by the discretization.

There are many interesting directions for future work. First, our numerical experiments suggest that for partially observable systems with a fixed partition, increasing memory allows to improve the approximation quality. Identifying the class of systems for which increasing memory guarantees a better approximation is an interesting direction for further research. Second, we plan to extend this work to the data-driven setting, and exploring aspects such as sample complexity as a function of the number of blocks and memory. Third, while our current bounds are valid, they suffer from conservatism and often exceed 1, the maximal value of any total variation distance. Therefore, we aim to investigate alternative approaches that directly rely on intermediate results on the 1-norm, rather than relying on 2-norm results as we do here.

REFERENCES

[1] Alessandro Abate, Joost-Pieter Katoen, John Lygeros, and Maria Prandini. 2010. Approximate Model Checking of Stochastic Hybrid Systems. *European Journal of Control* 16, 6 (Jan. 2010), 624–641. https://doi.org/10.3166/ejc.16.624-641

[2] Mazen Alamir. 2022. Learning against uncertainty in control engineering. *Annual Reviews in Control* 53 (2022), 19–29. https://doi.org/10.1016/j.arcontrol.2022.03.007

[3] Adrien Banse, Licio Romao, Alessandro Abate, and Raphael Jungers. 2023. Data-driven memory-dependent abstractions of dynamical systems. In *Proceedings of The 5th Annual Learning for Dynamics and Control Conference (Proceedings of Machine Learning Research, Vol. 211)*, Nikolai Matni, Manfred Morari, and George J. Pappas (Eds.). PMLR, 891–902. https://proceedings.mlr.press/v211/banse23a.html

[4] Adrien Banse, Licio Romao, Alessandro Abate, and Raphaël M. Jungers. 2023. Data-driven Abstractions via Adaptive Refinements and a Kantorovich Metric. In *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE. https://doi.org/10.1109/cdc49753.2023.10383513

[5] Giannis Delimpaltadakis, Luca Laurenti, and Manuel Mazo. 2024. Formal Analysis of the Sampling Behavior of Stochastic Event-Triggered Control. *IEEE Trans. Automat. Control* 69, 7 (July 2024), 4491–4505. https://doi.org/10.1109/tac.2023.3333748

[6] Eduardo Figueiredo, Andrea Patane, Morteza Lahijanian, and Luca Laurenti. 2024. Uncertainty Propagation in Stochastic Systems via Mixture Models with Error Quantification. https://doi.org/10.48550/ARXIV.2403.15626 arXiv:2403.15626 [eess.SY]

[7] Jiacheng Guo, Minshuo Chen, Huan Wang, Caiming Xiong, Mengdi Wang, and Yu Bai. 2023. Sample-Efficient Learning of POMDPs with Multiple Observations In Hindsight. arXiv:2307.02884 [cs.LG] https://arxiv.org/abs/2307.02884
[8] Martin Hairer. 2006. Ergodic Properties of Markov Processes. https://api.semanticscholar.org/CorpusID:51990013
[9] Anatole Katok and Boris Hasselblatt. 1995. *Introduction to the Modern Theory of Dynamical Systems*. Cambridge University Press. https://doi.org/10.1017/cbo9780511809187
[10] John C. Knight. 2002. Safety critical systems: challenges and directions. In *Proceedings of the 24th international conference on Software engineering - ICSE '02 (ICSE '02)*. ACM Press, 547. https://doi.org/10.1145/581339.581406
[11] I. Kontoyiannis and S. P. Meyn. 2011. Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probability Theory and Related Fields* 154, 1–2 (June 2011), 327–339. https://doi.org/10.1007/s00440-011-0373-4
[12] Morteza Lahijanian, Sean B. Andersson, and Calin Belta. 2015. Formal Verification and Synthesis for Discrete-Time Stochastic Systems. *IEEE Trans. Automat. Control* 60, 8 (Aug. 2015), 2031–2045. https://doi.org/10.1109/tac.2015.2398883
[13] Abolfazl Lavaei, Sadegh Soudjani, Alessandro Abate, and Majid Zamani. 2022. Automated verification and synthesis of stochastic hybrid systems: A survey. *Automatica* 146 (Dec. 2022), 110617. https://doi.org/10.1016/j.automatica.2022.110617
[14] Abolfazl Lavaei, Sadegh Soudjani, and Majid Zamani. 2020. Compositional abstraction of large-scale stochastic systems: A relaxed dissipativity approach. *Nonlinear Analysis: Hybrid Systems* 36 (May 2020), 100880. https://doi.org/10.1016/j.nahs.2020.100880
[15] Edward A. Lee and Sanjit A. Seshia. 2016. *Introduction to Embedded Systems A Cyber-Physical Systems Approach*. MIT Press. 568 pages.
[16] Douglas Lind and Brian Marcus. 1995. *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press. https://doi.org/10.1017/cbo9780511626302
[17] Yiming Meng and Jun Liu. 2023. Robustly Complete Finite-State Abstractions for Control Synthesis of Stochastic Systems. *IEEE Open Journal of Control Systems* 2 (2023), 235–248. https://doi.org/10.1109/ojcsys.2023.3294829
[18] Adam Nielsen, Konstantin Fackeldey, and Marcus Weber. 2013. *On a Generalized Transfer Operator*. Technical Report 13-74. ZIB, Takustr. 7, 14195 Berlin.
[19] Eric Pairet, Juan David Hernandez, Marc Carreras, Yvan Petillot, and Morteza Lahijanian. 2022. Online Mapping and Motion Planning Under Uncertainty for Safe Navigation in Unknown Environments. *IEEE Transactions on Automation Science and Engineering* 19, 4 (Oct. 2022), 3356–3378. https://doi.org/10.1109/tase.2021.3118737
[20] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. 2011. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics* 134, 17 (May 2011). https://doi.org/10.1063/1.3565032
[21] Marco Sarich, Frank Noé, and Christof Schütte. 2010. On the Approximation Quality of Markov State Models. *Multiscale Modeling and Simulation* 8, 4 (Jan. 2010), 1154–1177. https://doi.org/10.1137/090764049
[22] Anne-Kathrin Schmuck and Jörg Raisch. 2014. Asynchronous l-complete approximations. *Systems and Control Letters* 73 (Nov. 2014), 67–75. https://doi.org/10.1016/j.sysconle.2014.08.005
[23] Anne-Kathrin Schmuck, Paulo Tabuada, and Jörg Raisch. 2015. Comparing asynchronous l-complete approximations and quotient based abstractions. In *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 6823–6829. https://doi.org/10.1109/cdc.2015.7403294
[24] Ch. Schütte, W. Huisinga, and P. Deuflhard. 2001. *Transfer Operator Approach to Conformational Dynamics in Biomolecular Systems*. Springer Berlin Heidelberg, 191–223. https://doi.org/10.1007/978-3-642-56589-2_9
[25] Albert N. Shiryaev. 2016. *Probability-1: Volume 1*. Springer New York. https://doi.org/10.1007/978-0-387-72206-1
[26] Marnix Suilen, Thom Badings, Eline M. Bovy, David Parker, and Nils Jansen. 2025. *Robust Markov Decision Processes: A Place Where AI and Formal Methods Meet*. Springer Nature Switzerland, Cham, 126–154. https://doi.org/10.1007/978-3-031-75778-5_7
[27] Sheng-Jhih Wu and Moody T. Chu. 2017. Markov chains with memory, tensor formulation, and the dynamics of power iteration. *Appl. Math. Comput.* 303 (June 2017), 226–239. https://doi.org/10.1016/j.amc.2017.01.030
[28] Enlu Zhou, Michael C. Fu, and Steven I. Marcus. 2010. Solving Continuous-State POMDPs via Density Projection. *IEEE Trans. Automat. Control* 55, 5 (2010), 1101–1116. https://doi.org/10.1109/TAC.2010.2042005

A PROOFS

Our results rely on the Hölder's inequality, that we recall hereinafter.

Lemma A.1 (Hölder's inequality). *Given a measurable space $(E, \mathcal{F})$, together with a measure μ, and $p, q \in [0, +\infty]$ such that $1/p + 1/q = 1$. Then, for all $f : E \to \mathbb{R}$ and $g : E \to \mathbb{R}$, it holds that*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

A direct consequence of Hölder's inequality is that, for all function $f \in L^2(\mu)$, it holds that

$$\|f\|_1 \leq \|f\|_2.$$

Our proofs also rely on the following lemma, that holds under Assumption 4.

Lemma A.2 ([21, Lemma 2.2]). *For all initial densities $v_0^\ell \in L^2(\mu^\ell)$, and $k \geq \ell - 1$ it holds that*

$$\|P_\ell^{k-\ell+1} v_0^\ell - \mathbb{1}\|_2 \leq \|(T_\ell Q_\ell)^{k-\ell+1} v_0^\ell - \mathbb{1}\|_2 \leq e_{\ell,1}^{k-\ell+1} \|v_0^\ell\|_2.$$

Finally, we will also need the fact that all transfer operator has a unitary norm.

Lemma A.3. *Given a state space E and a kernel τ, the transfer operator $T : L^2(\mu) \to L^2(\mu)$ is such that $\|T\|_1 = 1$.*

Proof. We first recall the operator norm

$$\|T\|_1 = \sup_{f \in L^2(\mu) : \|f\|_1 \leq 1} \|Tf\|_1.$$

First we prove that $\|T\|_1 \geq 1$. Take any nonnegative function $f : E \to \mathbb{R}_{\geq 0}$ such that $\|f\|_1 = 1$. Then it holds that

$$\|Tf\|_1 = \int_{x \in E} |(Tf)(x)| \mu(\mathrm{d}x).$$

By definition of the transfer operator, if f is nonnegative, then Tf is also nonnegative. Therefore,

$$\begin{aligned} \|Tf\|_1 &= \int_{x \in E} (Tf)(x) \mu(\mathrm{d}x) \\ &= \int_{x \in E} f(x) \mu(\mathrm{d}x) \\ &= 1, \end{aligned}$$

where the first equality holds by (4), and the second by assumption. By definition of sup, this proves the first claim.

Now we prove that $\|T\|_1 \leq 1$. Any function $f : E \to \mathbb{R}$ can be written as $f = f^+ - f^-$, where $f^+ : E \to \mathbb{R}_{\geq 0}$ and $f^- : E \to \mathbb{R}_{\geq 0}$ are respectively the positive and negative parts of f. Therefore, for all functions f such that $\|f\|_1 \leq 1$, it holds that

$$\begin{aligned}\|Tf\|_1 &= \int_{x\in E} |(Tf)(x)|\mu(\mathrm{d}x)\\ &= \int_{x\in E} |(T(f^+ - f^-))(x)|\mu(\mathrm{d}x)\\ &= \int_{x\in E} |(Tf^+)(x) - (Tf^-)(x)|\mu(\mathrm{d}x)\\ &= \int_{x\in E} (Tf^+)(x) + (Tf^-)(x)\mu(\mathrm{d}x)\\ &= \int_{x\in E} (Tf^+)(x)\mu(\mathrm{d}x) + \int_{x\in E} (Tf^-)(x)\mu(\mathrm{d}x)\\ &= \int_{x\in E} f^+(x)\mu(\mathrm{d}x) + \int_{x\in E} f^-(x)\mu(\mathrm{d}x)\\ &= \int_{x\in E} f^+(x) + f^-(x)\mu(\mathrm{d}x)\\ &= \int_{x\in E} |f^+(x) - f^-(x)|\mu(\mathrm{d}x)\\ &= \int_{x\in E} |f(x)|\mu(\mathrm{d}x)\\ &\leq 1,\end{aligned}$$

which concludes the proof. □

A.1 Proof of Proposition 4.1

First we define an orthonormal basis for D_n^ℓ, given by $\{\phi_{i_1\ldots i_\ell}\}_{i_1,\ldots,i_\ell\in[n]}$:

$$\phi_{i_1\ldots i_\ell}(x_1,\ldots,x_\ell) = \frac{\chi_{A_{i_1}}(x_1)\ldots\chi_{A_{i_\ell}}(x_\ell)}{\sqrt{\mathbb{P}[X_1^\mu \in A_{i_1},\ldots,X_\ell^\mu \in A_{i_\ell}]}}.$$

The orthogonality of the basis follows, since for all $i_1,\ldots,i_\ell \in [n]$,

$$\begin{aligned}&\langle \phi_{i_1\ldots i_\ell}, \phi_{i_1\ldots i_\ell}\rangle\\ &\quad= \frac{\int_{x_1\in A_{i_1}}\cdots\int_{x_\ell\in A_{i_\ell}} \tau(\mathrm{d}x_\ell|x_{\ell-1})\ldots\tau(\mathrm{d}x_1|x_0)\mu(\mathrm{d}x_0)}{\mathbb{P}[X_1^\mu \in A_{i_1},\ldots,X_\ell^\mu \in A_{i_\ell}]}\\ &\quad= 1,\end{aligned}$$

and for all $i_1\ldots i_\ell \neq j_1\ldots j_\ell$, $\langle\phi_{i_1\ldots i_\ell}, \phi_{j_1\ldots j_\ell}\rangle = 0$. As a consequence, by definition of Q_ℓ, it holds that

$$Q_\ell v^\ell = \sum_{i_1\ldots i_\ell} \langle v^\ell, \phi_{i_1\ldots i_\ell}\rangle\phi_{i_1\ldots i_\ell} \tag{22}$$

for all $v^\ell \in L^2(\mu^\ell)$.

Now, by definition of P_ℓ and by (22), for all $i_0,\ldots,i_{\ell-1} \in [n]$,

$$\begin{aligned}&P_\ell\psi_{i_0,\ldots,i_{\ell-1}}\\ &\quad= Q_\ell T_\ell Q_\ell \psi_{i_0,\ldots,i_{\ell-1}}\\ &\quad= Q_\ell\left(T_\ell\psi_{i_0,\ldots,i_{\ell-1}}\right)\\ &\quad= \sum_{j_1,\ldots,j_\ell\in[n]} \langle T_\ell\psi_{i_0,\ldots,i_{\ell-1}}, \phi_{j_1,\ldots,j_\ell}\rangle\phi_{j_1,\ldots,j_\ell} \qquad (23)\\ &\quad= \sum_{j_1,\ldots,j_\ell\in[n]} \frac{\left\langle T_\ell\left(\chi_{A_{i_0}}\cdots\chi_{A_{i_{\ell-1}}}\right), \chi_{A_{j_1}}\cdots\chi_{A_{j_\ell}}\right\rangle}{\mathbb{P}[X_0^\mu \in A_{i_0},\ldots,X_{\ell-1}^\mu \in A_{i_{\ell-1}}]}\psi_{j_1,\ldots,j_\ell}.\end{aligned}$$

By (11) , the dot product above equals zero if $j_1\ldots j_{\ell-1} \neq i_1\ldots i_{\ell-1}$, and, otherwise, is equal to

$$\begin{aligned}&\int_{x_0\in A_{i_0}}\cdots\int_{x_{\ell-1}\in A_{i_{\ell-1}}} \tau(A_{j_\ell}|x_{\ell-1})\tau(dx_{\ell-1}|x_{\ell-2})\ldots\tau(\mathrm{d}x_1|x_0)\mu(\mathrm{d}x_0)\\ &\quad= \mathbb{P}[X_0^\mu \in A_{i_0},\ldots,X_\ell^\mu \in A_{j_\ell}].\end{aligned}$$

Therefore, inserting this in (23) yields

$$P_\ell\psi_{j_0,\ldots,j_{\ell-1}} = \sum_{j_1,\ldots,j_\ell\in[n]} \frac{\mathbb{P}[X_0^\mu \in A_{i_0},\ldots,X_\ell^\mu \in A_{j_\ell}]}{\mathbb{P}[X_0^\mu \in A_{i_0},\ldots,X_{\ell-1}^\mu \in A_{i_{\ell-1}}]}\psi_{j_1,\ldots,j_\ell},$$

which is (15) by definition of the output process, and the proof is completed. □

A.2 Proof of Theorem 5.1

First, by the triangular inequality, it holds that

$$\mathsf{TV}(\lambda_{k-\ell+1}^\ell, \tilde\lambda_{k-\ell+1}^\ell) \leq \mathsf{TV}(\lambda_{k-\ell+1}^\ell, \overline\lambda_{k-\ell+1}^\ell) + \mathsf{TV}(\overline\lambda_{k-\ell+1}^\ell, \tilde\lambda_{k-\ell+1}^\ell),$$

where $\lambda_{k-\ell+1}^\ell$, $\overline\lambda_{k-\ell+1}^\ell$ and $\tilde\lambda_{k-\ell+1}^\ell$ are respectively the measures corresponding to

$$v_{k-\ell+1}^\ell = T_\ell^{k-\ell+1}v_0^\ell,$$

$$\overline v_{k-\ell+1}^\ell = T_\ell P_\ell^{k-\ell}v_0^\ell,$$

and

$$\tilde v_{k-\ell+1}^\ell = P_\ell^{k-\ell+1}v_0^\ell.$$

First, we show that

$$\mathsf{TV}(\lambda_{k-\ell+1}^\ell, \overline\lambda_{k-\ell+1}^\ell) \leq \mathsf{TV}(\lambda_{k-\ell}^\ell, \tilde\lambda_{k-\ell}^\ell).$$

It holds that $v_{k-\ell+1}^\ell = T_\ell v_{k-\ell}^\ell$, and that $\overline v_{k-\ell+1}^\ell = T_\ell\tilde v_{k-\ell}^\ell$. Therefore, by definition of the total variation distance and by Lemma A.3,

$$\begin{aligned}\mathsf{TV}(\lambda_{k-\ell+1}^\ell, \overline\lambda_{k-\ell+1}^\ell) &= \frac{1}{2}\left\|T_\ell\left(v_{k-\ell}^\ell - \tilde v_{k-\ell}^\ell\right)\right\|_1\\ &\leq \frac{1}{2}\|T_\ell\|_1\|v_{k-\ell}^\ell - \tilde v_{k-\ell}^\ell\|_1\\ &= \frac{1}{2}\|v_{k-\ell}^\ell - \tilde v_{k-\ell}^\ell\|_1\\ &= \mathsf{TV}(\lambda_{k-\ell}^\ell, \tilde\lambda_{k-\ell}^\ell).\end{aligned}$$

Second, we show that

$$\mathsf{TV}(\overline\lambda_{k-\ell+1}^\ell, \tilde\lambda_{k-\ell+1}^\ell) \leq \frac{1}{2}(me_{\ell,1}\delta_\ell + r_\ell)e_{\ell,1}^{k-\ell}\|v_0^\ell\|_2.$$

It holds that $\overline v_{k-\ell+1}^\ell = T_\ell\tilde v_{k-\ell}^\ell$, and that $\tilde v_{k-\ell+1}^\ell = P_\ell\tilde v_{k-\ell}^\ell = Q_\ell T_\ell\tilde v_{k-\ell}^\ell$. Therefore, by definition of the total variation distance, and by Hölder's inequality, it holds that

$$\begin{aligned}\mathsf{TV}(\overline\lambda_{k-\ell+1}^\ell, \tilde\lambda_{k-\ell+1}^\ell) &= \frac{1}{2}\|(T_\ell - Q_\ell T_\ell)\tilde v_{k-\ell}^\ell\|_1\\ &\leq \frac{1}{2}\|(T_\ell - Q_\ell T_\ell)\tilde v_{k-\ell}^\ell\|_2.\end{aligned}$$

Now we follow a similar reasoning as in the proof of [21, Theorem 3.1]. Let $Q_\ell^\perp = \mathrm{Id} - Q_\ell$, where Id the identity operator. Since

$Q_\ell^\perp \mathbb{1} = 0$, it holds that

$$\begin{aligned}\|(T_\ell - Q_\ell T_\ell)\tilde{v}^\ell_{k-\ell}\|_2 &= \|Q_\ell^\perp T_\ell \tilde{v}^\ell_{k-\ell}\|_2 \\ &= \|Q_\ell^\perp T_\ell Q_\ell (T_\ell Q_\ell)^{k-\ell}\tilde{v}^\ell_0\|_2 \\ &= \|Q_\ell^\perp T_\ell Q_\ell ((T_\ell Q_\ell)^{k-\ell}\tilde{v}^\ell_0 - \mathbb{1})\|_2 \\ &\leq \|Q_\ell^\perp T_\ell Q_\ell\|_2 \|(T_\ell Q_\ell)^{k-\ell}\tilde{v}^\ell_0 - \mathbb{1}\|_2.\end{aligned}$$

Following [21, Eq. (63)], the first factor is such that

$$\|Q_\ell^\perp T_\ell Q_\ell\|_2 \leq m e_{\ell,1}\delta_\ell,$$

and, by Lemma A.2, the second is such that

$$\|(T_\ell Q_\ell)^{k-\ell}\tilde{v}^\ell_0 - \mathbb{1}\|_2 \leq e_{\ell,1}^{k-\ell}\|v^\ell_0\|_2,$$

which concludes the proof. □

A.3 Proof of Theorem 5.2

By the triangular inequality, it holds that

$$\mathsf{TV}(\lambda^\ell_{k-\ell+1}, \tilde{\lambda}^\ell_{k-\ell+1}) \leq \mathsf{TV}(\lambda^\ell_{k-\ell+1}, \mu^\ell) + \mathsf{TV}(\mu^\ell, \tilde{\lambda}^\ell_{k-\ell+1}),$$

where μ^ℓ is the lifted invariant measure as defined in (12). By Hölder's inequality and by Assumption 4, it holds that

$$\mathsf{TV}(\lambda^\ell_{k-\ell+1}, \mu^\ell) \leq \frac{1}{2}\|v^\ell_{k-\ell+1} - \mathbb{1}\|_2 \leq \frac{1}{2} e_{\ell,1}^{k-\ell+1}\|v^\ell_0\|_2.$$

By Lemma A.2, we can follow the exact same reasoning as above to get

$$\mathsf{TV}(\mu^\ell, \tilde{\lambda}^\ell_{k-\ell+1}) \leq \frac{1}{2} e_{\ell,1}^{k-\ell+1}\|v^\ell_0\|_2,$$

which concludes the proof. □

A.4 Proof of Corollary 5.3

First, by Theorem 5.1, it holds that

$$\begin{aligned}&\mathsf{TV}(\lambda^\ell_{k-\ell+1}, \tilde{\lambda}^\ell_{k-\ell+1}) \\ &\quad\leq \mathsf{TV}(\lambda^\ell_0, \tilde{\lambda}^\ell_0) + \sum_{i=0}^{k-\ell} \frac{1}{2}\left(m e_{\ell,1}\delta_\ell + r_\ell\right) e^i_{\ell,1}\|v^\ell_0\|_2 \\ &\quad= \mathsf{TV}(\lambda^\ell_0, \tilde{\lambda}^\ell_0) + \frac{1}{2}\left(m e_{\ell,1}\delta_\ell + r_\ell\right)\left(\sum_{i=0}^{k-\ell} e^i_{\ell,1}\right)\|v^\ell_0\|_2 \\ &\quad= \mathsf{TV}(\lambda^\ell_0, \tilde{\lambda}^\ell_0) + \frac{1}{2}\left(m e_{\ell,1}\delta_\ell + r_\ell\right)\frac{1 - e_{\ell,1}^{k-\ell+1}}{1 - e_{\ell,1}}\|v^\ell_0\|_2.\end{aligned}$$

Now it remains to show that

$$\mathsf{TV}(\lambda_k, \tilde{\lambda}_{\ell,k}) \leq \mathsf{TV}(\lambda^\ell_{k-\ell+1}, \tilde{\lambda}^\ell_{k-\ell+1}).$$

By definition of the total variation distance,

$$\begin{aligned}&\mathsf{TV}(\lambda^\ell_{k-\ell+1}, \tilde{\lambda}^\ell_{k-\ell+1}) \\ &\quad= \sup_{A_{k-\ell+1},\ldots,A_k \in \mathcal{B}(E)} \left| \begin{matrix}\lambda^\ell_{k-\ell+1}(A_{k-\ell+1} \times \cdots \times A_k) \\ \quad - \lambda^\ell_{k-\ell+1}(A_{k-\ell+1} \times \cdots \times A_k)\end{matrix}\right| \\ &\quad\geq \sup_{A_k \in \mathcal{B}(E)} \left| \begin{matrix}\lambda^\ell_{k-\ell+1}(E \times \cdots \times E \times A_k) \\ \quad - \lambda^\ell_{k-\ell+1}(E \times \cdots \times E \times A_k)\end{matrix}\right| \\ &\quad= \sup_{A_k \in \mathcal{B}(E)} |\lambda_k(A_k) - \lambda_{\ell,k}(A_k)| \\ &\quad= \mathsf{TV}(\lambda_k, \lambda_{\ell,k}),\end{aligned}$$

and the proof is completed. □

B DETAILS ABOUT THE NUMERICAL EXPERIMENTS

In order to compute $\mathsf{TV}(\lambda_k, \tilde{\lambda}_{\ell,k})$ from Example 6.1 and $\tilde{v}_{\ell,k}$, we proceed as follows. First, we know that

$$\lambda_k = \mathcal{N}(m_k, \Sigma_k),$$

where m_k and Σ_k satisfy the recurrence

$$m_{k+1} = A m_k, \quad \Sigma_{k+1} = A\Sigma_k A^\top + \Sigma_w.$$

The μ-weighted probability density function $v_k \in L^2(\mu)$ is given by

$$v_k = \frac{\mathrm{d}\lambda_k}{\mathrm{d}\mu} = \frac{\mathrm{d}\lambda_k}{\mathrm{d}\lambda^*}\frac{\mathrm{d}\lambda^*}{\mathrm{d}\mu} = \frac{\mathrm{d}\lambda_k}{\mathrm{d}\lambda^*}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda^*}\right)^{-1},$$

where λ^* is the Lebesgue measure. Therefore, v_k is defined as

$$v_k(x) = \frac{f_{\mathcal{N}(m_k,\Sigma_k)}(x)}{f_{\mathcal{N}(m_\mu,\Sigma_\mu)}(x)},$$

where $f_{\mathcal{N}(m,\Sigma)}$ is the usual probability density function of Gaussian distributions with respect to the Lebesgue measure. Finally, we compute the total variation with a Monte-Carlo approximation, that is

$$\begin{aligned}\mathsf{TV}(\lambda_k, \tilde{\lambda}_{\ell,k}) &= \frac{1}{2}\int_{x\in E} |v_k(x) - \tilde{v}_{\ell,k}(x)|\mu(\mathrm{d}x) \\ &\approx \frac{1}{2}\sum_{i=1}^{10^4} |v_k(x_i) - \tilde{v}_{\ell,k}(x_i)|,\end{aligned}$$

where x_i are i.i.d. samples from the invariant measure.