IEEE Transactions on Quantum Engineering

Received 29 July 2022; revised 28 June 2023; accepted 14 September 2023; date of publication 2 October 2023; date of current version 2 November 2023.

Digital Object Identifier 10.1109/TQE.2023.3319254

Learning Circular Hidden Quantum Markov Models: A Tensor Network Approach

MOHAMMAD ALI JAVIDIAN[1,2], VANEET AGGARWAL[2,3] (Senior Member, IEEE), AND ZUBIN JACOB[2]

[1]Department of Computer Science, Appalachian State University, Boone, NC 28608 USA
[2]School of Electrical, Computer Engineering, Purdue University, West Lafayette, IN 47907 USA
[3]School of Industrial Engineering, Purdue University, West Lafayette, IN 47907 USA

Corresponding author: Vaneet Aggarwal (e-mail: vaneet@purdue.edu).

This work was supported in part by the Defense Advanced Research Projects Agency Quantum Causality under Grant HR00112010008. This work was presented in part at NeurIPS 2021 Second Workshop on Quantum Tensor Networks in Machine Learning (QTMNL2021), Dec 2021.

ABSTRACT This article proposes circular hidden quantum Markov models (c-HQMMs), which can be applied for modeling temporal data. We show that c-HQMMs are equivalent to a tensor network (more precisely, circular local purified state) model. This equivalence enables us to provide an efficient learning model for c-HQMMs. The proposed learning approach is evaluated on six real datasets and demonstrates the advantage of c-HQMMs as compared to HQMMs and HMMs.

INDEX TERMS Hidden quantum Markov model (HQMM), tensor network.

I. INTRODUCTION

Hidden Markov models (HMMs) are commonly used for modeling temporal data, usually, in cases where the underlying probability distribution is unknown, but certain output observations are known [1], [2]. Hidden quantum Markov models (HQMMs) [3], [4] can be thought of as a reformulation of HMMs in the language of quantum systems (see Section II for formal definition). It has been shown that quantum formalism allows for a more efficient description of a given stochastic process than the classical case [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. This article proposes circular HQMMs and validates them to be more efficient than HQMMs.

We note that circular HMMs (c-HMMs) have been proposed to model HMMs, where the initial and terminal hidden states are connected through the state transition probability [14]. c-HMMs have found application in speech recognition [15], [16], biology and meteorology [17], shape recognition [14], [18], biomedical engineering [19], among others. Given the improved performance of c-HMMs compared to HMMs, it remains open if such an extension can be done for HQMMs, which is the focus of this article.

Even though multiple algorithms for learning HQMMs have been studied, direct learning of the model parameters is inefficient and results in poor local-optimal points [20]. To deal with this challenge, a tensor network-based approach is used to learn HQMMs [13], based on a result that HQMM is equivalent to uniform locally purified states (LPS) tensor network. The model in [13] deals with infinite horizon HQMM, which involves uniform Kraus operators and, thus, uniform LPS. Recently, a different learning approach was proposed in [21] focusing on learning unitary parametrization of HQMMs. In this article, we model a finite sequence of random variables, allowing us to have different Kraus operators at each time instant. Furthermore, we extend the finite-horizon HQMMs to circular HQMMs (c-HQMMs).

To train the parameters of the c-HQMM, we show the equivalence of c-HQMM with a class of tensor networks. To do that, we first define a class of tensor networks called circular LPS (c-LPS). Then, we show that c-HQMM is equivalent to c-LPS. Finally, we propose an algorithm to train c-LPS, thus providing an efficient algorithm for learning c-HQMMs. The results in this article show equivalence of finite-horizon HQMMs and c-HQMMs to the corresponding tensor networks. The key contributions of this work are summarized as follows.

1) We propose c-HQMM for modeling finite-horizon temporal data in Section III-A.

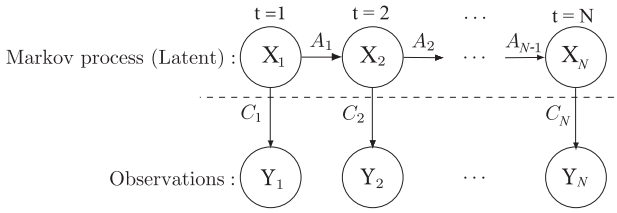


FIGURE 1. Representation of a Moore-type HMM with observation variables Y_i, latent variables X_i, transition matrices $A_i \iff p(X_t|X_{t-1})$, emission matrices $C_i \iff p(Y_t|X_t)$, and prior probability $\pi \iff p(X_1)$.

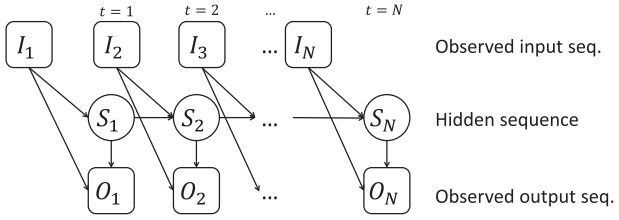


FIGURE 2. Mealy input–output HMM.

2) We propose c-LPS tensor networks in Section III-B, and we show that c-HQMMs are equivalent to c-LPS in Section III-C.
3) The connection between finite horizon HQMMs, which involve nonuniform Kraus operators, and nonuniform LPS is formalized and discussed in Section III-C
4) We provide a learning algorithm for c-HQMM using the tensor network equivalence in Section IV.

In order to validate the proposed framework of c-HQMM and the proposed learning algorithm, we compare with standard HMMs (equivalent to MPS with nonnegative real entries in decomposition) and HQMMs. Numerical evaluation on realistic datasets demonstrates the improved performance of c-HQMMs for modeling temporal data.

II. RELATED WORK AND BACKGROUND

In this section, we briefly review the key related literature on HMMs and tensor networks, with relevant definitions.

A. HIDDEN MARKOV MODELS

HMMs [1], [2] are a class of probabilistic graphical models with the greatest use in problems that enjoy an inherent temporality. These problems consist of a process that unfolds in time, i.e., we have states at time t that are influenced directly by a state at $t-1$. HMMs have found application in such problems, for instance, speech recognition [22], gesture recognition [23], face recognition [24], finance [25], computational biology [26], [27], [28], among others. A hidden Markov model can be defined in one of two ways: 1) Mealy or 2) Moore, depending on the formulation. As shown in Fig. 1, a finite-horizon Moore-type hidden Markov model or HMM consists of a discrete-time, discrete-state Markov chain, with hidden states $X_t \in \{1, \ldots, r\}, t \in \{1, \ldots, N\}$,[1] plus an observation model $p(y_t|x_t)$. The corresponding joint distribution has the form

$$p(X_{1:N}, Y_{1:N}) = p(X_{1:N})p(Y_{1:N}|Y_{1:N}) = p(X_1)p(Y_1|X_1)\Pi_{t=2}^{N} p(X_t|X_{t-1})p(Y_t|X_t). \quad (1)$$

[1]Finite horizon implies that N is finite, and thus the distributions can depend on the time-index. This article focuses on finite horizon probability distributions.

A Mealy input–output HMM [29] is a generalized Mealy finite-state machine, in which the added layer of observed nodes that represents a second observed sequence affects the output sequence through the mediation of a hidden state as shown in Fig. 2. As opposed to Moore-type HMM, the observation no longer depends on current hidden state, but on the previous hidden state and the input token. Note that the expressive power of Moore and Mealy models are the same [30]. But, the Mealy input–output HMMs are more aligned to the definition of HQMM in the following section.

B. HIDDEN QUANTUM MARKOV MODEL

HQMM was introduced in [3] to model evolution from one quantum state to another while generating classical output symbols. To produce an output symbol, a measurement or Kraus operation [31] is performed on the machine's internal state. One can use an auxiliary quantum system called *ancilla* to implement a Kraus operation. In every time step, the internal state of the HQMM interacts with its ancilla, which is then read out by a projective measurement. After every measurement, the ancilla is reset into its initial state, while the internal state of the HQMM remains hidden [4]. As in the classical case, an HQMM can be composed by the repeated application of the quantum sum rule (plays the role of transition matrices in HMMs) and quantum Bayes rule (plays the role of emission matrices in HMMs) [32] encoded using the sets of Kraus operators $\{K_{t,w}\}$ (where the subscripts w coincide with the output symbols of the HQMM machine) and $\{K_{t,x}\}$ (where the subscripts x coincide with the given observations), respectively, for $t \in \{1, \ldots, N\}$

$$\rho_t' = \sum_w K_{t,w}\rho_{t-1}K_{t,w}^\dagger \qquad \text{(quantum sum rule)}$$

$$\rho_t = \frac{K_{t,x}\rho_t' K_x^\dagger}{\text{tr}(\sum_x K_{t,x}\rho_t' K_{t,x}^\dagger)}. \qquad \text{(quantum Bayes rule)}$$

We can condense these two expressions into a single term for a given observation x by setting $K_{t,x,w} = K_{t,x}K_{t,w}$, for $t = 1, \ldots, N$, as follows:

$$\rho_{t|_x} = \frac{\sum_w K_{t,x,w}\rho_{t-1|_x}K_{t,x,w}^\dagger}{\text{tr}(\sum_w K_{t,x,w}\rho_{t-1|_x}K_{t,x,w}^\dagger)}. \qquad \text{(state update rule)}$$

IEEE Transactions on Quantum Engineering

We now formally define HQMMs using the Kraus operator-sum representation (the definition is modified from [20] to account for finite N).

Definition 1 (HQMM): An N-horizon d-dimensional Hidden Quantum Markov Model with a set of discrete observations $\mathcal{O}$ is a tuple $(\mathbb{C}^{d\times d}, \{K_{i,x,w_x}\})$, where the initial state $\rho_0 \in \mathbb{C}^{d\times d}$ and the Kraus operators $\{K_{i,x,w_x}\} \in \mathbb{C}^{d\times d}$, for all $x \in \mathcal{O}$, $i \in \{1,\ldots,N\}$, $w_x \in \mathbb{N}$, satisfy the following constraints:

1) ρ_0 is a density matrix of arbitrary rank;
2) the full set of Kraus operators across all observables provide a quantum operation,[2] i.e., $\sum_{x,w_x} K^\dagger_{i,x,w_x} K_{i,x,w_x} = I$ for all $i \in \{1,\ldots,N\}$.

The joint probability of a given sequence is given by

$$p(x_1,\ldots,x_N) = \vec{I}^T \left(\sum_{w_{x_N}} K^\dagger_{N,x_N,w_{x_N}} \otimes K_{N,x_N,w_{x_N}}\right) \cdots \left(\sum_{w_{x_1}} K^\dagger_{1,x_1,w_{x_1}} \otimes K_{1,x_1,w_{x_1}}\right) \vec{\rho_0} \quad (2)$$

where $\vec{I}^T$, $\vec{\rho_0}$ indicate vectorization (column-first convention) of identity matrix and ρ_0, respectively. This model is illustrated in Fig. 4(a).

This representation was used in [33, Algorithm 1] to develop HQMMs by constructing quantum analogues of classical operations on graphical models, and show that HQMMs are a more expressive model class compared to HMMs. This fact that HQMMs can give more efficient descriptions of classical stochastic processes was noted in the literature before as in [5] and [6]. HQMMs enable us to generate more complex random output sequences than HMMs, even when using the same number of internal states [4], [33]. In other words, HQMMs are strictly more expressive than classical HMMs [13].

C. TENSOR NETWORK

Tensor Network is a set of tensors (high-dimensional arrays), where some or all of its indices are contracted according to some pattern [34], [35]. They have been used to study many-body quantum systems [36], [37]. Further, they have been adopted for supervised learning in machine learning [38], [39], [40]. Some of the classes of tensor networks we use in this work include variants of matrix product states (MPSs) and locally purified states (LPSs).

One class of tensor networks is MPS, where an order-N tensor $T_{d\times\cdots\times d}$, with rank r has entry $(x_1,\ldots,x_N)$ ($x_i \in \{1,\ldots,d\}$) given as

$$T_{x_1,\ldots,x_N} = \sum_{\{\alpha_i\}_{i=0}^N=1}^{r} A_0^{\alpha_0} A_{1,x_1}^{\alpha_0,\alpha_1} A_{2,x_2}^{\alpha_1,\alpha_2} \cdots A_{N-1,x_{N-1}}^{\alpha_{N-2},\alpha_{N-1}} A_{N,x_N}^{\alpha_{N-1},\alpha_N} A_N^{\alpha_N} \quad (3)$$

where $A_k, k \in \{0, N+1\}$, is a vector of dimension r, and element (α_k) is denoted as $A_k^{\alpha_k}$. Further, $A_k, k \in \{1,\ldots,N\}$ is an order-3 tensor of dimension $d \times r \times r$, where element (x, α_L, α_R) is denoted as $A_{k,x}^{\alpha_L,\alpha_R}$.

Another class of tensor network that is studied in this article is the LPS, where an order-N tensor $T_{d\times\cdots\times d}$, with purirank r and purification dimension μ has entry $(x_1,\ldots,x_N)$ ($x_i \in \{1,\ldots,d\}$) given as:

$$T_{x_1,\ldots,x_N} = \sum_{\{\alpha_i,\alpha_i'\}_{i=0}^N=1}^{r} \sum_{\{\beta_i\}_{i=1}^N=1}^{\mu} A_0^{\alpha_0,\alpha_0'} A_{1,x_1}^{\beta_1,\alpha_0,\alpha_1} A^{\dagger\beta_1,\alpha_0',\alpha_1'}_{1,x_1} A_{2,x_2}^{\beta_2,\alpha_1,\alpha_2} A^{\dagger\beta_2,\alpha_1',\alpha_2'}_{2,x_2} \cdots A_{N-1,x_{N-1}}^{\beta_{N-1},\alpha_{N-2},\alpha_{N-1}} A^{\dagger\beta_{N-1},\alpha_{N-2},\alpha_{N-1}}_{N-1,x_{N-1}} A_{N,x_N}^{\beta_N,\alpha_{N-1},\alpha_N} A^{\dagger\beta_N,\alpha_{N-1}',\alpha_N'}_{N,x_N} A_{N+1}^{\alpha_N,\alpha_N'} \quad (4)$$

where $A_k, k \in \{0, N+1\}$, is an $r \times r$ matrix, where the element (α_k, α_k') is denoted as $A_k^{\alpha_k,\alpha_k'}$. Further, $A_k, k \in \{1,\ldots,N\}$, is an order-4 tensor of dimension $d \times \mu \times r \times r$, where the element $(x, \beta, \alpha_L, \alpha_R)$ is denoted as $A_{k,x}^{\beta,\alpha_L,\alpha_R}$, and elements belong to $\mathbb{R}$ or $\mathbb{C}$, as defined based on the context. The use of $\dagger$ refers to the Hermitian conjugate.

Tensor networks can be represented using tensor diagrams, where boxes represent tensors, and indices in the tensors are represented by lines emerging from the boxes. The lines connecting tensors correspond to contracted indices, whereas lines that do not go from one tensor to another correspond to open indices [35]. The tensor diagrams corresponding to tensor networks MPS and LPS can be seen in Fig. 3(b) and 3(c), respectively.

D. RELATION BETWEEN HMMS AND TENSOR NETWORKS

As noted in [3], [6], [13], [41], [42], tensor networks have direct correspondence with HMMs. Note that this connection was made (though not explicitly by name) in [43]. In particular, nonnegative MPSs are HMMs [42], and uniform locally purified states are HQMMs [13]. Note that equivalence assumes that the tensor networks are normalized as the probabilities, while we will not explicitly normalize the tensor networks in the proofs, it will be accounted in the learning.

E. LEARNING OF HQMM

Two state-of-the-art algorithms for learning HQMMs were proposed in [20] and [33]. Both algorithms use an iterative maximum-likelihood algorithm to learn Kraus operators to model sequential data using an HQMM. The proposed algorithm in [33] is slow, and there is no theoretical guarantee

[2] If $\sum_w \mathcal{K}_w^\dagger \mathcal{K}_w = I$, then $\mathcal{K} = \sum_w \mathcal{K}_w \rho \mathcal{K}_w^\dagger$ is called a quantum channel. However, if $\sum_w \mathcal{K}_w^\dagger \mathcal{K}_w < I$, then $\mathcal{K}$ is called a stochastic quantum operation.

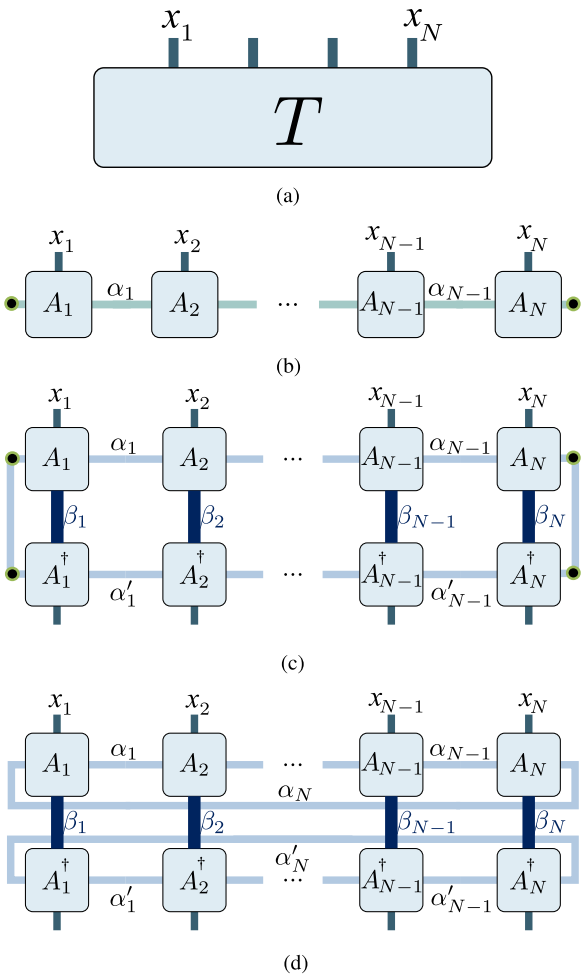


FIGURE 3. Tensor diagrams corresponding to different tensor networks. Black end dots indicate boundary vectors. (a) An order-N Tensor T. (b) An order-N MPS (Tensor Train) T (c) An order-N LPS T (d) An order-N c-LPS T.

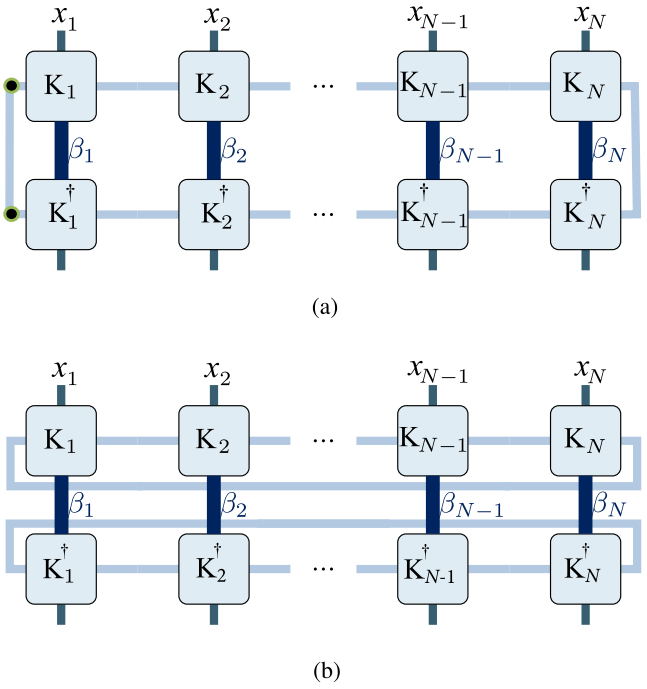


FIGURE 4. HQMM and c-HQMM: with observation variables x_i, where K_i are $= K_{i,x_i,w_{x_i}}$ and $\beta_i = |w_{x_i}|$ is determined by the Kraus-rank. (a) A HQMM. The rightmost connecting line at the boundary represents the application of the identity. Black end dots indicate boundary vectors. (b) A c-HQMM.

that the algorithm steps towards the optimum at every iteration [20]. The proposed algorithm in [20], however, uses a gradient-based algorithm. Although the proposed algorithm in [20] can learn an HQMM that outperforms the corresponding HMM, this comes at the cost of a rapid scaling in the number of parameters. To deal with this issue, equivalence between HQMMs and tensor networks have been considered to achieve efficient learning [13], [42].

III. PROPOSED MODELS

A. PROPOSED C-HQMM

In this section, we propose a circular HQMM (c-HQMM) for modeling temporal data.

Definition 2 (c-HQMM): An N-horizon d-dimensional c-HQMM with a set of discrete observations $\mathcal{O}$ is a tuple $(\mathbb{C}^{d\times d}, \{K_{i,x,w_x}\})$, where the Kraus operators are given as $\{K_{i,x,w_x}\} \in \mathbb{C}^{d\times d}$, for all $x \in \mathcal{O}, i \in \{1,\ldots,N\}, w_x \in \mathbb{N}$. The full set of Kraus operators across all observables provide a quantum operation, i.e., $\sum_{x,w_x} K^\dagger_{i,x,w_x} K_{i,x,w_x} = I$, for all $i \in \{1,\ldots,N\}$. The joint probability of a given sequence is given by

$$p(x_1,\ldots,x_N) = \mathrm{tr}\left(\left(\sum_{w_{x_N}} K^\dagger_{N,x_N,w_{x_N}} \otimes K_{N,x_N,w_{x_N}}\right) \cdots \left(\sum_{w_{x_1}} K^\dagger_{1,x_1,w_{x_1}} \otimes K_{1,x_1,w_{x_1}}\right)\right) \quad (5)$$

where $\mathrm{tr}(\cdot)$ indicates the trace of the resulting matrix. This model is illustrated in Fig. 4(b).

B. PROPOSED C-LPS MODEL

In this section, we introduce circular LPS (c-LPS) as a tensor ring extension of LPS, where an order-N with d-dimensional indices, puri-rank r, and purification dimension μ has entries given as

$$T_{x_1,\ldots,x_N} = \sum_{\{\alpha_i,\alpha'_i\}_{i=1}^N=1}^{r} \sum_{\{\beta_i\}_{i=1}^N=1}^{\mu} A^{\beta_1,\alpha_N,\alpha_1}_{1,x_1} A^{\dagger\beta_1,\alpha'_N,\alpha'_1}_{1,x_1} \cdots A^{\beta_N,\alpha_{N-1},\alpha_N}_{N,x_N} A^{\dagger\beta_N,\alpha'_{N-1},\alpha'_N}_{N,x_N} \quad (6)$$

where $A_k, k \in \{1,\ldots,N\}$, is an order-4 tensor of dimension $d \times \mu \times r \times r$, as shown in Fig. 3(d), where the element $(x,\beta,\alpha_L,\alpha_R)$ is denoted as $A^{\beta,\alpha_L,\alpha_R}_{k,x}$.

For each x_i, $i \in \{1, \ldots, N\}$, the index contraction of A_i and ${A^\dagger}_i$ over $\beta_i \in \{1, \ldots, \mu\}$ results in an order-4 tensor of dimension $r \times r \times r \times r$. This 4-D array can be reshaped as an $r^2 \times r^2$ matrix and rewritten as $\tau_{x_i} = \sum_{\beta_i=1}^{\mu} {B^\dagger}_{x_i,\beta_i} \otimes B_{x_i,\beta_i}$, where $\{B_{x_i,\beta_i}\}_{x_i=1}^{d} \in \mathbb{C}^{r\times r}$. Similar to the MPS transfer matrix (or transfer operator) [44], we call $\tau = \sum_{x_i=1}^{d} \tau_{x_i}$ the c-LPS transfer operator.

C. C-HQMM MODELS ARE C-LPS

We first note that nonnegative MPS (denoting by $\text{MPS}_{\mathbb{R}_{\geq 0}}$) are HMM [42]. In other words, any HMM can be mapped to an MPS with nonnegative elements, and any $\text{MPS}_{\mathbb{R}_{\geq 0}}$ can be mapped to a HMM. Similarly, local quantum circuits with ancillas are locally purified states [42]. [13] recently considered an infinite time model of HQMM, where Kraus operators do not depend on time and showed the equivalence of these HQMMs with the uniform LPS with a positive definite matrix structure. However, our work considers a nonuniform finite-time structure by having Kraus operators depend on time. We note that the equivalent tensor structure corresponding to c-HQMM is open, which is studied in this section. The next result describes the relation between c-HQMM and c-LPS:

Theorem 1: c-HQMM models and c-LPS models are equivalent and have the same expressive power. Formally, for a given c-HQMM model with the joint probability p there exist a c-LPS structure T such that $p(x_1, \ldots, x_N) = T_{x_1,\ldots,x_N}$ for all entries $(x_1, \ldots, x_N)$, where the decomposition entries $A_{i,x}^{b,a_1,a_2} \in \mathbb{C}$. Also, for a given c-LPS model T with the above mentioned properties, there exist a c-HQMM model with the joint probability p such that $p(x_1, \ldots, x_N) = T_{x_1,\ldots,x_N}$ for all entries $(x_1, \ldots, x_N)$.

We need the following lemmas to prove the theorem (the detailed proof of lemmas are provided in the Appendix). The following lemma shows that a c-HQMM operator can be mapped to a c-LPS operator that computes the same function.

Lemma 1: From a given c-HQMM model with Kraus operators of rank β and the joint probability p, one can construct a c-LPS model, T, of the same puri-rank such that $p(x_1, \ldots, x_N) = T_{x_1,\ldots,x_N}$ for all entries $(x_1, \ldots, x_N)$.

The following lemma shows that the c-LPS transfer operator can be rescaled and transformed into a trace-preserving map.

Lemma 2: For the given c-LPS with the transfer operator $\tau = \sum_{x=1}^{d} \tau_x$, where $\tau_x = \sum_{\beta_i=1}^{\mu} {B^\dagger}_{x,\beta_i} \otimes B_{x,\beta_i}$, $x \in X_i$, $i \in \{1, \ldots, N\}$, there exist an invertible matrix M and a scalar α such that by replacing B_{x,β_i} with $\alpha \mathbf{vec}(M)^{1/2} B_{x,\beta_i} \mathbf{vec}(M)^{-1/2}$, where $\mathbf{vec}(M)$ is the vectorization of M, we obtain a transfer operator for this c-LPS that is a trace preserving map.

Now, we are ready to provide the proof of Theorem 1.

Proof of Theorem 1: In order to show the equivalence of two models, we need to show that c-LPS can be mapped to a c-HQMM that computes the same function, and vice versa. First, we show that a c-LPS can be mapped to a c-HQMM that computes the same function. From the definition of the transfer operator for c-LPS models and the operators defined for the c-HQMMs based on Kraus operators, it's not difficult to see that both models have operators of the same form. Lemma 2 shows that the transfer operators in a c-LPS can be rescaled and similarity transformed into one that is trace-preserving. So, a c-LPS of order-N with d-dimensional indices, puri-rank r, and purification dimension μ can be rescaled and similarity transformed into a c-HQMM of N-horizon with d-dimensional and the Kraus-rank μ.

Second, we show that a c-HQMM can be mapped to a c-LPS that computes the same function. Since both models have operators of the same form, Lemma 1 shows that the operators in a c-HQMM can be mapped into operators in a c-LPS that computes the same function. So, a c-HQMM of N-horizon with d-dimensional and the Kraus-rank β can be mapped into a c-LPS of order-N with d-dimensional indices, puri-rank d, and purification dimension β. ■

The proof structure can be directly specialized to HQMM, where we can obtain the following result:

Lemma 3: For a given HQMM model with the joint probability p there exist a LPS structure T such that $p(x_1, \ldots, x_N) = T_{x_1,\ldots,x_N}$ for all entries $(x_1, \ldots, x_N)$ where the decomposition entries $A_{i,x}^{b,a_1,a_2} \in \mathbb{C}$ and the evaluation functional is restricted to the full Kraus rank $\vec{I}^T$. Note that the boundary matrices ($i = 1, N$) in LPS are vectors and we need them to have nonnegative real elements, i.e., $A_{i,x}^{b,a_1,a_2} \in \mathbb{R}_{\geq 0}$. Also, for a given LPS structure T with the above mentioned properties there exist a HQMM model with the joint probability p such that $p(x_1, \ldots, x_N) = T_{x_1,\ldots,x_N}$ for all entries $(x_1, \ldots, x_N)$.

Note that nonterminating uniform LPS (uLPS) are equivalent to HQMM [13]. In uLPS, boundary vectors originate from density matrices of arbitrary rank. As shown in [13], the evaluation functional of uLPS can be rescaled and transformed into a one that will converge to $\vec{I}^T$ when $N \to \infty$. To have the equivalency of finite-horizon HQMM and LPS, we need to restrict LPS models to the evaluation function $\vec{I}^T$, as stated in Lemma 3. It should be noted that HQMMs with $|w| = 1$ can also be represented by MPS (as opposed to LPS).

IV. LEARNING ALGORITHM FOR CIRCULAR LPS MODELS

In this section, we propose an algorithm for learning c-LPS Models as in Theorem 1 via a maximum likelihood estimation (MLE) approach. The proposed algorithm is a modification of the algorithm proposed in [42] for learning LPS models, except that we take into account the cyclic structure. Note that a similar maximization likelihood algorithm is proposed for learning MPS in [45].

Problem 1 (MLE for Distribution Approximation): Assume that $\{x_i = (x_1^i, \ldots, x_N^i)\}_{i=1}^{n}$ is a sample of size n from an experiment with N discrete random variables. To estimate this discrete multivariate distribution, we use c-LPS model

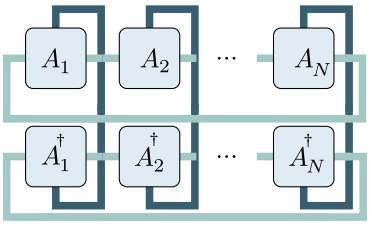


FIGURE 5. Contraction of tensor ring to compute: $\mathcal{Z}_T = \sum_{x_1,\ldots,x_N} T_{x_1,\ldots,x_N}$.

as defined in Section III-B. So, we have:

$$p(x_1,\ldots,x_N) \approxeq \sum_{\{\alpha_i,\alpha_i'\}_{i=1}^N=1}^{r} \sum_{\{\beta_i\}_{i=1}^N=1}^{\mu} A_{1,x_1}^{\beta_1,\alpha_N,\alpha_1} A_{1,x_1}^{\dagger\beta_1,\alpha_N',\alpha_1'} \cdots A_{N,x_N}^{\beta_N,\alpha_{N-1},\alpha_N} A_{N,x_N}^{\dagger\beta_N,\alpha_{N-1}',\alpha_N'}.$$

Our objective here is to estimate tensor elements of the c-LPS, i.e., $w = A_{i,x_i}^{\beta_i,\alpha_{L_i},\alpha_{R_i}}$, for $i = 1,\ldots,N$, where α_{L_i}, α_{R_i} refer to the left and right indices respectively. For this purpose, we minimize the negative log-likelihood

$$\mathcal{L} = -\sum_i \log \frac{T_{x_i}}{\mathcal{Z}_T} \tag{7}$$

where T_{x_i} is obtained using the contraction of circular LPS, and $\mathcal{Z}_T = \sum_{x_i} T_{x_i}$ is a normalization factor.

To find the optimal solution, we calculate the derivative of the log-likelihood with respect to w as follows:

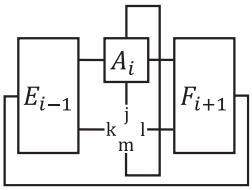
$$\partial_w \mathcal{L} = -\sum_i \frac{\partial_w T_{x_i}}{T_{x_i}} - \frac{\partial_w \mathcal{Z}_T}{\mathcal{Z}_T} \tag{8}$$

We use a minibatch gradient descent algorithm to minimize the negative log-likelihood. First, we compute the sum over a batch of training instances at each optimization step. Then, we update parameters in the tensor network by a small step in the inverse direction of the gradient. Since we use tensors with complex elements ($\in \mathbb{C}$) in our experiments, we use Wirtinger derivatives [46] regarding the conjugated tensor elements. Now, we explain how to compute $\partial_w \mathcal{Z}_T$, $\mathcal{Z}_T$, $\partial_w T_{x_i}$, and T_{x_i} in (8). For a c-LPS of puri-rank r, we compute the normalization factor $\mathcal{Z}_T$ by contracting the tensor network

$$\mathcal{Z}_T = \sum_{x_1,\ldots,x_N} T_{x_1,\ldots,x_N}. \tag{9}$$

As shown in Fig. 5, we perform this contraction from left to right by contracting at each step the two vertical indices (corresponding to d_i and $d_i^\dagger$ with respect to the supports of X_i and $X_i^\dagger$) and then, each of the two horizontal indices (with respect to α_is and α_i's, respectively). Finally, we trace out the indices corresponding to the rings. In this contraction, we store intermediate results from the contraction of the first i tensors in E_i, and we repeat the same procedure from the right by storing intermediate results of the contraction of the last $N-i$ tensors in F_{i+1}. We compute the derivatives of the normalization for each tensor as follows:

$$\frac{\partial \mathcal{Z}_T}{\partial A_{i,m}^{\dagger j,k,l}} = 2 \times$$



We can use a similar approach to compute T_{x_i} and $\partial_w T_{x_i}$ for a training sample, except that we set the contracted index corresponding to an observed variable to its observed value. We note that a similar approach to learn the model can be used for HQMM and HMM structures.

a) Asymptotic cost of evaluating c-LPS vs LPS: The time complexity of evaluation of an order-N c-LPS with d-dimensional indices, puri-rank r, and purification dimension μ is of order $O(d\mu Nr^5)$ but for the LPS of the same characteristics is of order $O(d\mu Nr^3)$. Note that, based on our experiments, the rank in c-LPS could be chosen smaller. Further, our experiments show improved results in c-LPS with same number of parameters.

V. NUMERICAL EVALUATIONS: MAXIMUM LIKELIHOOD ESTIMATION ON REAL DATA

To evaluate the performance of the proposed algorithm for learning c-HQMMs, we used the same datasets as used in [42] and learn HMM, HQMM (nonuniform), and c-HQMM using their respective tensor representations. HMM is equivalent to $\text{MPS}_{\mathbb{R}_{\geq 0}}$, and is a baseline for other structures. We show the equivalence of c-HQMMs to tensor networks and also the equivalence of LPS and HQMM for finite N with nonuniform Kraus operators, inspired by findings in the literature [13], [42]. It is worth noting that although the HMM is an important statistical tool for modeling data with sequential correlations in neighboring samples, such as in time series data, it also has been used for modeling nonsequence data via tensor decomposition in the literature [42], [47]. We compare the performance of training HMM, HQMM, and c-HQMM using equivalent tensor representations (i.e., MPS, LPS, and c-LPS, respectively) on six different real nonsequence data of categorical variables, where the following parameters are used.

1) Bond dimension/rank of the tensor networks: $r = 2, 3, 4, 5$, and 6.
2) Learning rate was chosen using a grid search on powers of 10 going from 10^{-5} to 10^5.
3) Batch size, i.e., the number of training samples per minibatch, was set to 20.
4) Number of iterations was set to a maximum of 1000.
5) The dimension of the purification index, i.e., μ for LPS and c-LPS was set to 2.

We briefly introduce the following real datasets used in our experiments here.

1) The **biofam** dataset was constructed by Müller et al. [48] from the data of the retrospective biographical survey carried out by the Swiss Household Panel in 2002. In includes only individuals who were at least 30 years old at the time of the survey for whom we consider sequences of family life states between ages 15 and 30. The biofam dataset describes thus family life courses of 2000 individuals born between 1909 and 1972. Biofam data includes 16 state variables: The states numbered from 0 to 7 are defined from the combination of five basic states, namely Living with parents (Parent), Left home (Left), Married (Marr), Having Children (Child), Divorced; The covariates are: Sex, birthyr (birth year), nat_1_02 (first nationality), plingu02 (language of questionnaire), p02r01 (religion), p02r04 (religious participation), cspfaj (father's social status), cspmoj (mother's social status).
2) The **lymphography** data was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia [49]. All attribute values in the database have been entered as numeric values corresponding to their index in the list of attribute values for that attribute domain as given here: class (normal find, metastases, malign lymph, fibrosis), lymphatics (normal, arched, deformed, displaced), block of affere (no, yes), block of lymph. c (no, yes), block of lymph. s (no, yes), by pass (no, yes), extravasates (no, yes), regeneration of (no, yes), early uptake in (no, yes), lym.nodes dimin (0-3), lym.nodes enlar (1-4), changes in lym. (bean, oval, round), defect in node (no, lacunar, lac. marginal, lac. central), changes in node (no, lacunar, lac. margin, lac. central), changes in structure (no, grainy, drop-like, coarse, diluted, reticular, stripped, faint), special forms (no, chalices, vesicles), dislocation of (no, yes), exclusion of no (no, yes), no. of nodes in (0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, >=70).
3) **SPECT Heart** data is a dataset on cardiac Single Proton Emission Computed Tomography (SPECT) images. Each patient classified into two categories: 1) Normal and 2) abnormal [49].
4) 1984 United States **Congressional Voting Records** includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition) [49].
5) **Primary Tumor** dataset was provided by Ljubljana Oncology Institute [49]. All attribute values in the database have been entered as numeric values corresponding to their index in the list of attribute values for that attribute domain as given here: class (lung, head & neck, esophasus, thyroid, stomach, duoden & sm.int, colon, rectum, anus, salivary glands, pancreas, gallblader, liver, kidney, bladder, testis, prostate, ovary, corpus uteri, cervix uteri, vagina, breast), age (<30, 30-59, >=60), sex (male, female), histologic-type (epidermoid, adeno, anaplastic), degree-of-diffe (well, fairly, poorly), and 13 binary (yes, no) variables: Bone, bone-marrow, lung, pleura, peritoneum, liver, brain, skin, neck, supraclavicular, axillar, mediastinum, and abdominal.
6) **Solar Flare** dataset contains 3 potential classes. Each class attribute counts the number of solar flares of a certain class that occur in a 24 h period [49].

TABLE I Characteristics of Real Datasets Used in Experiments

Data set	Sample size	Number of attributes	Source
biofam data	2000	16	[48]
Lymphography	148	18	[49]
SPECT Heart	187	23	[49]
Congressional Voting Records	435	17	[49]
Primary Tumor	339	17	[49]
Solar Flare	1065	13	[49]

Each data point reported here is the lowest negative log-likelihood (7) obtained from 10 trials with different (random) initialization of tensors[3]. We used 75% of each dataset for training (for this purpose, we fit the model to the data and then we evaluate the negative log-likelihood of the fitted model) and 25% for the testing case (for this purpose, we use the fitted model to the training data directly to evaluate the negative log-likelihood of the trained model for testing data). A summary of data characteristics for our evaluations listed in Table I.

Key Highlights w.r.t. Results: The obtained results, summarized in Fig. 6, show the following.

1) The tensor representations can be used to learn different HMMs.
2) We observe that despite the different algorithm choices, c-LPS and LPS lead to better modeling of the data distribution for the same rank compared to MPS on almost all datasets.
3) The results indicate that c-HQMM outperforms HQMM and HMM.
4) In many cases, the performance difference for training datasets between LPS and c-LPS for ranks 5 and 6 is more significant than those with ranks 2 and 3. Further, the improvement depends on the dataset. We also note that we plot negative of log likelihoods, so the gap in the likelihoods is larger. The results suggest that in generic settings HQMM and c-HQMM should be

[3]A user manual that explains how to create a model, load a dataset and train the model on the dataset is available at supplementary material: https://github.com/majavid/Circular-Hidden-Quantum-Markov-Models

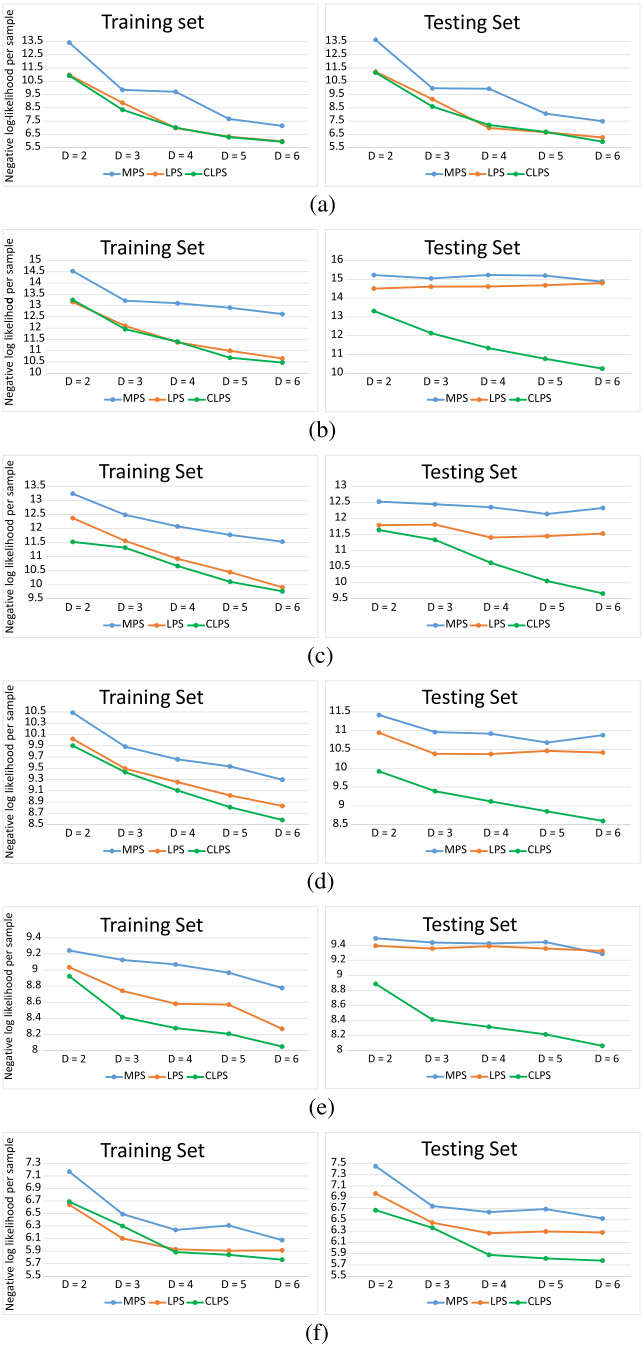


FIGURE 6. **Maximum likelihood estimation with tensor networks MPS, LPS, and c-LPS for learning HMM, HQMM, and c-HQMM, respectively, from the data on different datasets: (a) biofam dataset of family life states from the Swiss Household Panel biographical survey [48]; datasets from the UCI Machine Learning Repository [49]: (b) Lymphography, (c) SPECT Heart, (d) Congressional Voting Records, (e) Primary Tumor, and (f) Solar Flare for the training and testing datasets, respectively.**

preferred over both HMM models. Further, c-HQMM gives the best performance among the considered models.

5) Fig. 6, also, shows the result for different tensor network representations on the testing sets for different data, indicating consistency with the observations (1)–(4) and confirm them. However, the main advantage of using c-LPS (c-HQMMs) over LPS (HQMM) and MPS (HHMs) comes to the scene in testing sets and the improvement of performance is pronounced in these cases, indicating the stability and nonoverfitting of c-LPS (c-HQMMs) in practice. The similarity of c-LPS performance behavior (see data points and curves) across training and testing sets confirms this observation.

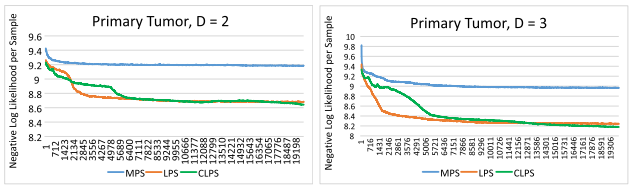


FIGURE 7. **Convergence plots for the Primary Tumor dataset for $D = 2, 3$, and 20 000 iterations.**

6) c-LPS (c-HQMM) has slightly more parameters than LPS (HQMM) due to the existence of the cycle. This cyclic structure allows for better structural learning with a single rank parameter. For noncyclic structures, the corner ranks being smaller and center ranks being larger limit the generalization. Due to better structural learning, the test error with c-LPS is significantly better than the baselines. Such an improved generalization characteristic with cyclic structures have been seen in other problems for c-MPS, for instance in data completion and data classification problems [39], [40].

To see the convergence rate of MPS, LPS, and c-LPS, we depicted the convergence plots for the Primary Tumor dataset for $D = 2, 3$, and 20 000 iterations in Fig. 7. Note that the approaches converge within 20 000 iterations. Although MPS converges much faster than others, LPS and c-LPS converge to better optimal points.

VI. CONCLUSION

This article proposes a new class of HMMs, that we called c-HQMMs. c-HQMMs can be used to model temporal data. We proved that c-HQMMs are equivalent to c-LPS models. Leveraging this result, we proposed an MLE-based algorithm for learning c-HQMMs from data via c-LPS. We evaluated the proposed learning approach on six real nonsequence datasets, demonstrating the advantage of c-HQMMs on multiple datasets compared to HQMMs and HMMs. Testing the performance of the proposed algorithm on temporal data is left as further work.

Code and data for reproducing our results is available at: https://github.com/majavid/Circular-Hidden-Quantum-Markov-Models.

IEEE Transactions on Quantum Engineering

APPENDIX
PROOFS

In this Appendix, we will prove the different Lemmas that help proving the main Theorem. The following lemma shows that a c-HQMM operator can be mapped to a c-LPS operator that computes the same function.

Lemma 4: From a given c-HQMM model with Kraus operators of rank β and the joint probability p, one can construct a c-LPS model, T, of the same puri-rank such that $p(x_1, \ldots, x_N) = T_{x_1,\ldots,x_N}$ for all entries $(x_1, \ldots, x_N)$.

Proof: From the definition of the transfer operator for c-LPS models and the operators defined for the c-HQMMs based on Kraus operators, it's not difficult to see that both models have operators of the same form. Now, we map $\sum_{x_i,w_{x_i}} {K^\dagger}_{i,x_i,w_{x_i}} \otimes K_{i,x_i,w_{x_i}}, i \in \{1, \cdots, N\}$ to tensors $A_i, A_i^\dagger, i \in \{1, \cdots, N\}$ in an c-LPS. For this purpose, we set $\beta_i = |w_{x_i}|$, i.e., the Kraus-rank in c-HQMM plays the role of purification dimension in c-LPS. For fixed values x_i and w_{x_i}, $K_{i,x_i,w_{x_i}}$ and ${K^\dagger}_{i,x_i,w_{x_i}}$ are $r \times r$ matrices that play the role of tensors A_i and ${A^\dagger}_i$ in the c-LPS, respectively.

The following lemma shows that the c-LPS transfer operator can be rescaled and transformed into a trace-preserving map. ■

Lemma 5: For the given c-LPS with the transfer operator $\tau = \sum_{x=1}^{d} \tau_x$, where $\tau_x = \sum_{\beta_i=1}^{\mu} {B^\dagger}_{x,\beta_i} \otimes B_{x,\beta_i}, x \in X_i, i \in \{1, \ldots, N\}$, there exist an invertible matrix M and a scalar α such that by replacing B_{x,β_i} with $\alpha \mathbf{vec}(M)^{1/2} B_{x,\beta_i} \mathbf{vec}(M)^{-1/2}$, where $\mathbf{vec}(M)$ is the vectorization of M, we get a transfer operator for this c-LPS that is a trace preserving map.

Proof: To show that the c-LPS transfer operator is trace preserving is equivalent to show that the identitiy $\vec{I}$ is the fixed point of $\tau^\dagger$, i.e., $\tau^\dagger \vec{I} = \vec{I}$. Assume that τ is not trace preserving. Without loss of generality assume that the two eigenvalues of τ with greatest magnitude, λ_1, λ_2, satisfy $|\lambda_1| > |\lambda_2|$. If we replace $B_{x_i,\beta_i}, x_i \in \{1, \ldots, d\}$ with $B_{x_i,\beta_i}/\sqrt{\lambda_1}$, we obtain a transfer operator τ' that enjoys the leading eigenvalue of magnitude 1. Note that this rescaling leaves the joint probability distributions unchanged. In this case, the quantum Perron-Frobenius theorem [50, chapter 16] implies that $\lambda_1 = 1$ and ${\tau'}^\dagger$ has a unique fixed-point operator $\vec{\sigma}_*$ (i.e., ${\tau'}^\dagger \vec{\sigma}_* = \vec{\sigma}_*$) which is the vectorization of a full-rank (and so, invertible) positive density matrix.

Now, we replace $B_{x_i,\beta_i}, x_i \in \{1, \ldots, d\}$ with $B'_{x_i,\beta_i} = \sigma_*^{1/2} B_{x_i,\beta_i} \sigma_*^{-1/2}$. We show that $\tau'' = \sum_{x_i=1}^{N} \tau'_{x_i}$ is a trace preserving map. Note that since $\sigma_*, \sigma_*^{-1/2}$, and $\sigma_*^{1/2}$ are Hermitian, we have ${\sigma_*}^\dagger = (\sigma_*)^T$, ${\sigma^\dagger}_*^{-1/2} = (\sigma_*^{-1/2})^T$, and ${\sigma^\dagger}_*^{1/2} = (\sigma_*^{1/2})^T$, respectively. For three matrices X, Y, Z that $W = XYZ$, we have $(Z^T \otimes X)\vec{Y} = \vec{W}$. Using these properties, we have

$$\tau''^\dagger \vec{I} = \left(\sum_{x_i=1}^{d} \tau'^\dagger_{x_i}\right) \vec{I} = \left(\sum_{x_i=1}^{d} \left(\sum_{\beta_i=1}^{\mu} {B'^\dagger}_{x_i,\beta_i} \otimes B'^\dagger_{x_i,\beta_i}\right)\right) \vec{I}$$

$$= \left(\sum_{x_i=1}^{d} \left(\sum_{\beta_i=1}^{\mu} \left(\sigma_*^{-1/2}\right)^T {B^\dagger}^\dagger_{x_i,\beta_i} \left(\sigma_*^{1/2}\right)^T \otimes \sigma_*^{-1/2} \times B^\dagger_{x_i,\beta_i} \sigma_*^{1/2}\right)\right) \vec{I}$$

$$= \left(\left(\sigma_*^{-1/2}\right)^T \otimes \sigma_*^{-1/2}\right) \left(\sum_{x_i=1}^{d} \sum_{\beta_i=1}^{\mu} {B^\dagger}^\dagger_{x_i,\beta_i} \otimes B^\dagger_{x_i,\beta_i}\right) \times \left(\left(\sigma_*^{1/2}\right)^T \otimes \sigma_*^{1/2}\right) \vec{I}$$

$$= \left(\left(\sigma_*^{-1/2}\right)^T \otimes \sigma_*^{-1/2}\right) \tau'^\dagger_{x_i} \vec{\sigma}_*$$

$$= \left(\left(\sigma_*^{-1/2}\right)^T \otimes \sigma_*^{-1/2}\right) \vec{\sigma}_* = \vec{I}.$$

So, the transfer operator τ can be rescaled and similarity transformed into one that is trace-preserving. ■

Lemma 6: For a given HQMM model with the joint probability p there exist a LPS structure T such that $p(x_1, \ldots, x_N) = T_{x_1,\ldots,x_N}$ for all entries $(x_1, \ldots, x_N)$ where the decomposition entries $A_{i,x}^{b,a_1,a_2} \in \mathbb{C}$ and the evaluation functional is restricted to the full Kraus rank $\vec{I}^T$. Note that the boundary matrices ($i = 1, N$) in LPS are vectors and we need them to have nonnegative real elements, i.e., $A_{i,x}^{b,a_1,a_2} \in \mathbb{R}_{\geq 0}$. Also, for a given LPS structure T with the abovementioned properties there exist a HQMM model with the joint probability p such that $p(x_1, \ldots, x_N) = T_{x_1,\ldots,x_N}$ for all entries $(x_1, \ldots, x_N)$.

Proof: The given LPS, as defined in Section II, has entries of the form

$$T_{x_1,\ldots,x_N} = \sum_{\{\alpha_i,\alpha'_i\}_{i=0}^{N}=1}^{r} \sum_{\{\beta_i\}_{i=1}^{N}=1}^{\mu} A_0^{\alpha_0,\alpha'_0} A_{1,x_1}^{\beta_1,\alpha_0,\alpha_1} {A^\dagger}_{1,x_1}^{\beta_1,\alpha'_0,\alpha'_1} \times A_{2,x_2}^{\beta_2,\alpha_1,\alpha_2} {A^\dagger}_{2,x_2}^{\beta_2,\alpha'_1,\alpha'_2} \cdots A_{N-1,x_{N-1}}^{\beta_{N-1},\alpha_{N-2},\alpha_{N-1}} {A^\dagger}_{N-1,x_{N-1}}^{\beta_{N-1},\alpha_{N-2},\alpha_{N-1}} A_{N,x_N}^{\beta_N,\alpha_{N-1},\alpha_N} \times {A^\dagger}_{N,x_N}^{\beta_N,\alpha'_{N-1},\alpha'_N} A_{N+1}^{\alpha_N,\alpha'_N}.$$

Since both models have operators of the same form, we can write and manipulate the joint probability of a sequence of N observations as the unnormalized probability mass function over N discrete random variables $\{X_i\}_{i=1}^{N}$ as follows:

$$p(x_1, \ldots, x_n) = \vec{I}^T \left(\sum_{\beta_N=1}^{\mu} {B^\dagger}_{x_N,\beta_N} \otimes B_{x_N,\beta_N}\right) \cdots \left(\sum_{\beta_1=1}^{\mu} {B^\dagger}_{x_1,\beta_1} \otimes B_{x_1,\beta_1}\right) \vec{\rho}_0 \quad (10)$$

where $\vec{I}$ is the right boundary (also called evaluation functional) is the vectorized version of the identity matrix, i.e., $A_{N+1} = I$, and $\vec{\rho_0}$ is the vectorized version of the initial state, i.e., $A_0 = \rho_0$. So, LPS models and HQMMs differ only in two things: 1) while LPS can have an arbitrary Kraus-rank evaluation functional, HQMMs are restricted to the identity evaluation functional $\vec{I}^T$ of full Kraus rank; and (2) HQMMs operators are trace-preserving.

Following the same approach used in the proof of Theorem 1, the transfer operator of an LPS can be rescaled and similarity transformed into one that is trace-preserving. For the uniform LPS models, the evaluation functional of this transformed model will then converge to $\vec{I}^T$ [13]. However, in a finite-horizon LPS this may not be the case.

Therefore, HQMM model is equivalent to a LPS structure where the decomposition entries $A_{i,x}^{b,a_1,a_2} \in \mathbb{C}$, where the evaluation functional is restricted to the vectorized of the identity matrix, i.e., $\vec{I}^T$. ■

REFERENCES

[1] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986, doi: 10.1109/MASSP.1986.1165342.

[2] W. Zucchini and I. L. MacDonald, *Hidden Markov Models for Time Series: An Introduction Using R*, Boca Raton, FL, USA: Chapman and Hall/CRC, 2009, doi: 10.1201/b20790.

[3] A. Monras, A. Beige, and K. Wiesner, "Hidden quantum Markov models and non-adaptive read-out of many-body states," *arXiv:1002.2337*, 2010, doi: 10.48550/arXiv.1002.2337.

[4] L. A. Clark, W. Huang, T. M. Barlow, and A. Beige, "Hidden quantum Markov models and open quantum systems with instantaneous feedback," in *Proc. Interdiscipl. Symp. Complex Syst.*, 2015, pp. 143–151, doi: 10.1007/978-3-319-10759-2_16.

[5] M. Gu, K. Wiesner, E. Rieper, and V. Vedral, "Quantum mechanics can reduce the complexity of classical models," *Nature Commun.*, vol. 3, no. 1, pp. 1–5, 2012, doi: 10.1038/ncomms1761.

[6] A. Monras and A. Winter, "Quantum learning of classical stochastic processes: The completely positive realization problem," *J. Math. Phys.*, vol. 57, no. 1, 2016, Art. no. 015219, doi: 10.1063/1.4936935.

[7] J. R. Mahoney, C. Aghamohammadi, and J. P. Crutchfield, "Occam's quantum strop: Synchronizing and compressing classical cryptic processes via a quantum channel," *Sci. Rep.*, vol. 6, no. 1, pp. 1–11, 2016, doi: 10.1038/srep20495.

[8] C. Aghamohammadi, S. P. Loomis, J. R. Mahoney, and J. P. Crutchfield, "Extreme quantum memory advantage for rare-event sampling," *Phys. Rev. X*, vol. 8, no. 1, 2018, Art. no. 011025, doi: 10.1103/PhysRevX.8.011025.

[9] Q. Liu, T. J. Elliott, F. C. Binder, C. Di Franco, and M. Gu, "Optimal stochastic modeling with unitary quantum dynamics," *Phys. Rev. A*, vol. 99, no. 6, 2019, Art. no. 062110, doi: 10.1103/PhysRevA.99.062110.

[10] S. P. Loomis and J. P. Crutchfield, "Strong and weak optimizations in classical and quantum models of stochastic processes," *J. Stat. Phys.*, vol. 176, no. 6, pp. 1317–1342, 2019, doi: 10.1007/s10955-019-02344-x.

[11] T. J. Elliott, C. Yang, F. C. Binder, A. J. Garner, J. Thompson, and M. Gu, "Extreme dimensionality reduction with quantum modeling," *Phys. Rev. Lett.*, vol. 125, no. 26, 2020, Art. no. 260501, doi: 10.1103/PhysRevLett.125.260501.

[12] T. J. Elliott, "Memory compression and thermal efficiency of quantum implementations of nondeterministic hidden Markov models," *Phys. Rev. A*, vol. 103, no. 5, 2021, Art. no. 052615, doi: 10.1103/PhysRevA.103.052615.

[13] S. Adhikary, S. Srinivasan, J. Miller, G. Rabusseau, and B. Boots, "Quantum tensor networks, stochastic processes, and weighted automata," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2080–2088. [Online]. Available: https://proceedings.mlr.press/v130/adhikary21a

[14] N. Arica and F. Y. Vural, "A shape descriptor based on circular hidden Markov model," in *Proc. 15th Int. Conf. Pattern Recognit.*, 2000, pp. 924–927, doi: 10.1109/ICPR.2000.905592.

[15] Y.-C. Zheng and B.-Z. Yuan, "Text-dependent speaker identification using circular hidden Markov models," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 1988, pp. 580–581, doi: 10.1109/ICASSP.1988.196651.

[16] I. Shahin, "Enhancing speaker identification performance under the shouted talking condition using second-order circular hidden Markov models," *Speech Commun.*, vol. 48, no. 8, pp. 1047–1055, 2006, doi: 10.1016/j.specom.2006.01.005.

[17] H. Holzmann, A. Munk, M. Suster, and W. Zucchini, "Hidden Markov models for circular and linear-circular time series," *Environ. Ecological Statist.*, vol. 13, no. 3, pp. 325–347, 2006, doi: 10.1007/s10651-006-0015-7.

[18] J. Cai, M. Ee, and R. Smith, "Image retrieval using circular hidden Markov models with a garbage state," in *Proc. Image Vis. Comput. Conf.*, 2007, pp. 115–120. [Online]. Available: https://eprints.qut.edu.au/11229/

[19] D. A. Coast, R. M. Stern, G. G. Cano, and S. A. Briller, "An approach to cardiac arrhythmia analysis using hidden Markov models," *IEEE Trans. Biomed. Eng.*, vol. 37, no. 9, pp. 826–836, Sep. 1990, doi: 10.1109/10.58593.

[20] S. Adhikary, S. Srinivasan, G. Gordon, and B. Boots, "Expressiveness and learning of hidden quantum Markov models," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 4151–4161. [Online]. Available: http://proceedings.mlr.press/v108/adhikary20a

[21] V. Markov, V. Rastunkov, A. Deshmukh, D. Fry, and C. Stefanski, "Implementation and learning of quantum hidden Markov models," 2022, *arXiv:2212.03796*, doi: 10.48550/arXiv.2212.03796.

[22] B. H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991, doi: 10.1080/00401706.1991.10484833.

[23] A. D. Wilson and A. F. Bobick, "Parametric hidden Markov models for gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 884–900, Sep. 1999, doi: 10.1109/34.790429.

[24] A. V. Nefian and M. H. Hayes, "Hidden Markov models for face recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1998, pp. 2721–2724, doi: 10.1109/ICASSP.1998.678085.

[25] R. S. Mamon and R. J. Elliott, *Hidden Markov Models in Finance*, vol. 4. Berlin, Germany: Springer, 2007, doi: 10.1007/978-1-4899-7442-6.

[26] A. Siepel and D. Haussler, "Combining phylogenetic and hidden Markov models in biosequence analysis," *J. Comput. Biol.*, vol. 11, no. 2/3, pp. 413–428, 2004, doi: 10.1089/1066527041410472.

[27] T. Koski, *Hidden Markov Models for Bioinformatics*, vol. 2. Berlin, Germany: Springer Science & Business Media, 2001. [Online]. Available: https://link.springer.com/book/9781402001352

[28] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling," *J. Mol. Biol.*, vol. 235, no. 5, pp. 1501–1531, 1994, doi: 10.1006/jmbi.1994.1104.

[29] Y. Bengio and P. Frasconi, "An input output HMM architecture," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 427–434. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1994/hash/8065d07da4a77621450aa84fee5656d9-Abstract.html

[30] B. Vanluyten, J. C. Willems, and B. De Moor, "Equivalence of state representations for hidden Markov models," *Syst. Control Lett.*, vol. 57, no. 5, pp. 410–419, 2008, doi: 10.1016/j.sysconle.2007.10.004.

[31] K. Kraus, A. Böhm, J. D. Dollard, and W. Wootters, "States, effects, and operations: Fundamental notions of quantum theory. Lectures in mathematical physics at the University of Texas at Austin," in *Lecture Notes in Physics*, vol. 190, Berlin, Germany: Springer, 1983, pp. 103–149, doi: 10.1007/35401273.

[32] S. Adhikary, S. Srinivasan, and B. Boots, "Learning quantum graphical models using constrained gradient descent on the Stiefel manifold," 2019, *arXiv:1903.03730*, doi: 10.48550/arXiv.1903.03730.

[33] S. Srinivasan, G. Gordon, and B. Boots, "Learning hidden quantum Markov models," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 1979–1987. [Online]. Available: http://proceedings.mlr.press/v84/srinivasan18a.html

[34] I. V. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2295–2317, 2011, doi: 10.1137/09075228.

[35] R. Orús, "A practical introduction to tensor networks: Matrix product states and projected entangled pair states," *Ann. Phys.*, vol. 349, pp. 117–158, 2014, doi: 10.1016/j.aop.2014.06.013.

[36] R. Orús, "Tensor networks for complex quantum systems," *Nature Rev. Phys.*, vol. 1, no. 9, pp. 538–550, 2019, doi: 10.1038/s42254-019-0086-7.

[37] S. Montangero, E. Montangero, and Evenson, *Introduction to Tensor Network Methods*, Berlin, Germany: Springer, 2018, doi: 10.1007/978-3-030-01409-4.

[38] E. M. Stoudenmire and D. J. Schwab, "Supervised learning with quantum-inspired tensor networks," 2016, *arXiv:1605.05775*, doi: 10.48550/arXiv.1605.05775.

[39] W. Wang, V. Aggarwal, and S. Aeron, "Efficient low rank tensor ring completion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5697–5705, doi: 10.1109/ICCV.2017.607.

[40] W. Wang, Y. Sun, B. Eriksson, W. Wang, and V. Aggarwal, "Wide compression: Tensor ring nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9329–9338, doi: 10.1109/CVPR.2018.00972.

[41] C. Yang, F. C. Binder, V. Narasimhachar, and M. Gu, "Matrix product states for quantum stochastic modeling," *Phys. Rev. Lett.*, vol. 121, Dec. 2018, Art. no. 260602, doi: 10.1103/PhysRevLett.121.1260602.

[42] I. Glasser, R. Sweke, N. Pancotti, J. Eisert, and I. Cirac, "Expressive power of tensor-network factorizations for probabilistic modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, Art. no. 134. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/b86e8d03fe992d1b0e19656875ee557c-Paper.pdf

[43] C. R. Shalizi and J. P. Crutchfield, "Computational mechanics: Pattern and prediction, structure and simplicity," *J. Stat. Phys.*, vol. 104, no. 3, pp. 817–879, 2001, doi: 10.1023/A:1010388907793.

[44] V. Zauner et al., "Transfer matrices and excitations with matrix product states," *New J. Phys.*, vol. 17, no. 5, 2015, Art. no. 053002, doi: 10.1088/1367-2630/17/5/053002.

[45] Z.-Y. Han, J. Wang, H. Fan, L. Wang, and P. Zhang, "Unsupervised generative modeling using matrix product states," *Phys. Rev. X*, vol. 8, no. 3, 2018, Art. no. 031012, doi: 10.1103/PhysRevX.8.031012.

[46] R. Mortini and R. Rupp, *Extension Problems and Stable Ranks: A Space Odyssey*. Berlin, Germany: Springer, 2021, doi: 10.1007/978-3-030-73872-3.

[47] T.-K. Huang and J. Schneider, "Learning hidden Markov models from non-sequence data via tensor decomposition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 333–341. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/cb70ab375662576bd1ac5aaf16b3fca4-Abstract.html

[48] N. S. Müller, M. Studer, and G. Ritschard, "Classification de parcours de vie á l'aide de l'optimal matching," in *Proc. XIVe Rencontre Soc. Francophone Classification (SFC)*, 2007, pp. 157–160.

[49] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[50] J. C. Baez and J. D. Biamonte, *Quantum Techniques in Stochastic Mechanics*. Singapore: World Scientific, 2018, doi: 10.1142/9789813226951_others01.