Quantum stochastic modelling and tensor networks

Yang, Chengran

2020

Yang, C. (2020). Quantum stochastic modelling and tensor networks. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/144661>

<https://doi.org/10.32657/10356/144661>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 21:11:43 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**QUANTUM STOCHASTIC
MODELLING AND TENSOR
NETWORKS**

CHENGRAN YANG

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

2020

QUANTUM STOCHASTIC MODELLING AND TENSOR NETWORKS

CHENGRAN YANG

SCHOOL OF PHYSICAL AND MATHEMATICAL
SCIENCES

A thesis submitted to the Nanyang Technological
University in partial fulfillment of the requirements for the
degree of Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

29/06/20

.....

Date

Chengran Yang

.....

CHENGRAN YANG

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and the integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

29/06/20

.....

Date

.....

Prof. Mile Gu

Authorship Attribution Statement

This thesis contains material from 3 papers published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 5 is published as [C. Yang, F. C. Binder, V. Narasimhachar, and M. Gu, Phys. Rev. Lett. 121, 260602 \(2018\).](#)

The contributions of the co-authors are as follows:

- Prof Gu and I provided the initial project direction.
- I carry out the principal research, including giving the proof and calculation.
- I prepared the manuscript drafts. The manuscript was revised by Dr Binder, Dr. Narasimhachar and Prof Gu.
- All authors contributed to the research conception, discussions, and final manuscript preparation.

Chapter 6 is published as [T. J. Elliott, C. Yang, F. C. Binder, A. J. Garner, J. Thompson, and M. Gu, arXiv preprint arXiv:1909.02817 \(2019\).](#)

The contributions of the co-authors are as follows:

- T. J. Elliott and I carry out the principal research, contributing equally.
- T. J. Elliott drafted the manuscript. The draft is revised by Prof Gu, me, Dr Binder, Dr Thompson and Dr Garner.
- All authors contributed to the research conception, discussions, and final manuscript preparation.

Chapter 7 is published as [C. Yang, F. C. Binder, M. Gu, and T. J. Elliott Phys. Rev. E 101, 062137 \(2020\)](#)

The contributions of the co-authors are as follows:

- I carry out the principal research.
- I prepare the manuscript. The manuscript is revised by Prof Gu, Dr Binder, and Dr Elliott.
- All authors contributed to the research conception, discussions, and final manuscript preparation.

29/06/20

.....

Date

chengran Yang

.....

CHENGRAN YANG

Abstract

Predicting a stochastic process' future lies at the heart of many scientific areas. A predictive model extracts information from a stochastic process' past and uses it to generate future statistics. There has been significant amount of effort expended towards finding optimal predictive models that minimize the required amount of past information. By taking advantage of quantum resources, quantum models have been shown to reduce the memory requirements beyond classical limits. Meanwhile, tensor networks are an extremely useful mathematical tool for understanding quantum many-body systems.

In this thesis, we explore the connection between tensor networks and quantum predictive models. A particular class of tensor networks, matrix product states (MPS), is utilized to further improve quantum models. First, we establish a connection between matrix product states (MPS) and the optimal classical predictive models. MPS methods offer a systematic method for constructing quantum predictive models and even an improved method for computing the amount of quantum memory. Second, we show that for some families of stochastic processes our method allows us to construct quantum models with unbounded memory advantages over the optimal classical models. Third, we propose a family of divergence measures to quantify the distance between two stochastic processes. This family of divergence measures overcomes certain weaknesses of the existing measures. Moreover, we propose an efficient means of computing the divergence measure using our MPS methods. Finally, we address the problem of constructing quantum predictive models solely from the output of an observed stochastic process. We use machine learning methods to propose a systematic algorithm for learning the quantum model. This provides an alternative method for constructing quantum models with fixed amounts of amount of memory.

Keywords: stochastic processes, predictive models, causal models, tensor networks, machine learning, quantum information, quantum computation

Acknowledgements

Firstly, I would like to thank my supervisor, Mile Gu. His guidance has helped me throughout my Ph.D. I feel extremely lucky to have worked with such a talented physicist.

I would also like to thank members of Mile's Group. I am especially grateful to Felix Binder, and Thomas Elliott. They provide valuable insights and guidance for my projects on quantum stochastic modelling and tensor networks, and feedback on my thesis.

Thanks also to my co-authors, as well as others whom I have discussed with, including Varun Narasimhachar, Andrew Garner, Jayne Thompson, Man-Hong Yung, Nora Tischler, Feiyang Liu and Oscar Dahlsten.

I would also like to thank Paul Riechers for his helpful comments on this thesis.

My Ph.D. would not have been the same without the quantum switch club. I have had a lot of fun playing digital games with them. I especially thank Qing Liu, Ximing Wang, and Minjeong Song for all the good times.

Thanks to my family, especially my parents, my brother, and two lovely nephews for their love and support over the years.

A major thanks to my beloved girlfriend (Mei Niangniang) for her love and support.

Finally, thank you to anyone and everyone who has taken the time to read this thesis.

Contents

Abstract	1
Acknowledgements	2
List of Figures	7
List of Tables	13
Symbols and Acronyms	15
1 Introduction	19
1.1 Problem Statement	19
1.2 Outline of the thesis	22
2 Classical stochastic processes and predictive models	25
2.1 Random variables and entropies	25
2.2 Stochastic processes	26
2.3 Predictive models	27
2.3.1 Relation to generative models	29
2.3.2 An Example: perturbed coin	29
3 Quantum information, quantum computation and quantum models	31
3.1 Quantum information	31
3.1.1 Von Neumann entropy and Entanglement	32
3.1.2 Observable and Measurement	33
3.1.3 Quantum Channel	33
3.2 Quantum computation	34
3.3 Quantum models	34
3.3.1 Experimental realization	36
4 Tensor networks	37
4.1 Tensors	37
4.1.1 Map vectors to matrices	38

4.2	Matrix product states	39
4.2.1	iMPS and boundary	40
4.3	MPS for HMM	40
4.4	MPS for 1D quantum system	41
4.4.1	Transfer matrix	41
4.4.2	The probability of output sequence	42
4.4.3	Evaluating Schmidt coefficients	43
4.4.4	iMPS generated by quantum circuits	43
5	The connection between tensor networks and stochastic processes	45
5.1	Predictive models and MPS representation	45
5.2	Entanglement and memory requirement	51
5.3	Construct quantum models	52
5.4	Applications	55
5.4.1	Nemo processes	55
5.4.2	Discrete renewal processes	57
5.5	Conclusion	59
6	Extreme dimensional compression with quantum modelling	61
6.1	Dual Poisson processes	61
6.2	ε -machine of dual Poisson processes	62
6.3	Unbounded quantum compression advantages	64
6.3.1	Evaluating C_μ and C_q	67
6.4	Conclusion	68
7	Measures of distinguishability between stochastic processes	71
7.1	Criteria for well-behaved measures	72
7.2	Existing measures	73
7.3	A new family of measures	74
7.4	Two specific members	75
7.4.1	Co-emission divergence rate	75
7.4.2	Fidelity divergence rate	79
7.5	Applications	83
7.5.1	Pedagogical example: perturbed coin	84
7.5.2	An artificial three hidden states process	84
7.5.3	Continuous Gaussian processes	86
7.6	Discussion	88
7.7	Conclusion	89
8	Learning quantum models of stochastic processes	91
8.1	The quantum model discovery algorithm	92
8.2	Evaluating the algorithm's performance	96
8.2.1	Hyperparameters	98
8.2.2	Example 1: the asymmetric process	99

8.2.3	Example 2: the quasi-cycle	102
8.2.4	Example 3: the uniform renewal process	104
8.3	Conclusion	107
9	Conclusion and future directions	109

List of Figures

1.1	An illustrative figure of black box. The physical system inside the black box is a random coin. It randomly outputs heads or tails at each time step. The question marks represent unknown future outputs.	20
1.2	Predictive models extract information from the past and use it to predict the future.	20
2.1	The pasts are separated into several classes due to the equivalence relation. Each class is denoted by a causal state s_k .	28
2.2	Transitions between causal states. An arrow represents the transition from one state to another. $x T_{jk}^x$ indicates output x with probability T_{jk}^x	29
2.3	The ε -machine for perturbed coin has two causal states. The output x determines the causal state s_x .	30
3.1	Each qubit is represented by a line. The blue square indicates the unitary operator. The quantum computation is performed from left to right.	35
3.2	A quantum circuit for operating quantum models. The red dashed lines separate the quantum circuit into repeating blocks. In each block, the quantum state is coupled with an ancillary state. Measuring the ancilla in the appropriate basis generates the outputs.	35
4.1	Pictorial representation of tensors. A node with N legs depicts a rank- N tensor. Each leg indicates a corresponding index. Based on this, a scalar has no legs, while a rank-3 tensor has 3 legs.	38
4.2	Rules of tensor networks: (left) Tensor product of two tensors. (right) Summation rule: Linking two legs of two different tensors indicates the summation over this index.	38
4.3	Pictorial representation of vec maps. Vec maps bend a certain leg of a node to the other side.	39
4.4	A pictorial representation for a MPS. The horizontal index is called the bond index. The vertical index is called the physical index.	39
4.5	MPS with translation invariance. $\langle b_l $ and $ b_r \rangle$ are boundary tensors.	40
4.6	MPS representation of a HMM. T_{jk}^x represents the transition probability. The square tensor represents the left boundary, which is the stationary distribution π_i of the hidden states s_i . The triangular tensor is the right boundaries, which is a row vector filled with 1.	41

4.7	Tensor decomposition of sequence probability $P(x_{t:t+L})$. The left boundary indicates that the distribution of states begins with the stationary distribution. The right boundary indicates that the sequence probability $P(x_{0:L})$ equally takes into account all final states s_i^L	41
4.8	The tensor in the red dashed square is the transfer matrix $\mathbb{E}$. Its left two legs are grouped as a row index, and the right two legs are grouped as a column index. η is its leading eigenvalue. V_l and V_r are the corresponding leading left and right eigenvectors.	42
4.9	The expectation value $\langle O \rangle$ of observable O is obtained by contracting all the indices of the whole tensor. The blue nodes, $(\langle b_l \langle b_l ^*$ and $ b_r\rangle b_r\rangle^*)$, represents the left and right boundary of the tensor. The red dashed line indicates the left and right boundary finally converges to the leading left and right eigenvectors of transfer matrix $\mathbb{E}$	43
4.10	Unitary circuit for iMPS A^x	44
5.1	This chapter connects the complexity of stochastic processes and the representational complexity of spin chains. These link (i) ϵ -machines, the provably optimal predictors of a process with the MPS states of a spin chain, (ii) the memory used to simulate a process quantum mechanically with the entanglement of the spin chain and (iii) the Hilbert space dimension of such quantum simulators with the Schmidt rank of the spin chain.	46
5.2	The Nemo process is defined for $0 < p < 1$ and has causal states s_1, s_2 , and s_3 . An edge from s_i to s_j indicates a non-zero transition probability, labelled ' $P(x, s_j s_i) x$ '.	56
5.3	The amount of quantum memory C_q is less than the classical memory C_μ for Nemo processes.	57
5.4	The ϵ -machine of the discrete renewal process. An edge labeled $x p$, indicates that the transition from causal state s_i to s_j generates output x with probability p . The renewal process either emits 1 and returns to s_0 or 0 and evolves further.	58
5.5	The amount of quantum memory C_q converges to a finite value while the optimal classical memory C_μ diverges logarithmically.	58
6.1	Dual Poisson process. Each channel outputs 0 with probability q_i and 1 with probability $1 - q_i$. The red arrow indicates that the transition happens immediately. $\emptyset$ indicates no output.	62
6.2	ϵ -machine for dual Poisson process. Each casual state s_k outputs 0 and goes forward to s_{k+1} . Outputs 1 indicates that all states renew to s_0	63

6.3	Scaling of memory metrics with precision. Both classical memory metrics diverge with increasing precision, wherein the interval Δt is refined. In contrast, the quantum metrics remain finite, evincing an unbounded advantage: C_q tends to a bounded value, while D_q remains constant. Plot shown for $\gamma_1 = 12$, $\gamma_2 = 1$ and $p = 0.9$ (the qualitative features are typical for any non-extremal parameter choice).	64
6.4	Quantum statistical memory of dual Poisson processes. For much of the dual Poisson process family, the information stored by our quantum models is significantly less than one bit, highlighting that significant further compression is possible for ensemble simulators. The information stored is largest when both (i) the emission rates are significantly different, and (ii) the system is more likely to adopt the faster rate, as the state of confidence is frequently reset. $D_q=1$ everywhere except the lines $P=0, 1$ and $\gamma=1$, where $D_q=0$.	68
7.1	(a) Tensor network representation of the transition matrices of P and Q . (b) Tensor network representation of the co-emission probability.	76
7.2	Tensor network representation of (a) the fidelity and (b) boundary vectors.	80
7.3	The example process we consider can be represented by a three state HMM with two variables p and δ . The edge label $x P$ between states s_i and s_j signifies that if the model is in state s_i it will transition to s_j whilst emitting symbol x with probability P .	85
7.4	Comparison of CDR and KL-divergence per symbol for distinguishing between different parameter values of the process at $\delta = 0$. We see that the qualitative behaviour of the two is very similar. As would be expected, both are zero along the line $p = q$ where the processes are equal, and grow as the difference $ p - q $ increases.	86
7.5	The CDR is able to be calculated to ascertain the distinguishability between processes with different output alphabets and representations with different numbers of states, unlike other measures such as the KL-divergence. As expected, we see that the CDR increases as one process becomes increasingly likely to emit a symbol the other cannot. The white region represents an unphysical parameter regime.	87
7.6	The continuous processes are discretized to evaluate the corresponding CDR. The state x with range $(-3, 3)$ and step size $\Delta x = 2$. The parameters of the two processes are set as follows: $\theta = 1, \mu = 0, \sigma = 1$.	88
8.1	Our learning algorithm takes outputs from a stochastic process and for a given memory dimension d_q generates a quantum model which yields approximately correct statistics for the process's future conditioned on its past.	91

8.2	ε -machine for the asymmetric process, parameterized by $p, q \in (0, 1)$. The nodes are the causal states s_0, s_1 and s_2 . The arrows indicate transitions between states, where each label $x P$ indicates output x is produced with probability P	100
8.3	Error in the 2-dimensional models for the asymmetric process across a range of process parameters p, q (see Figure 8.2). The left-hand-side shows the lower bound on the optimal approximate classical model's error; the right-hand-side shows the error of the most accurate quantum model found by our algorithm.	101
8.4	The approximate model of asymmetric example with $p = 0.3, q = 0.8$. s_0 and s_1 of the original asymmetric process are merged into r_0 . The transition probability is minimized by going through the probability vector space. s_2 is mapped to r_1 and its transition probability is unchanged.	101
8.5	Example: Asymmetric process – encoded states. Quantum states produced by the learned $d_q = 2$ encoding map for the asymmetric process with $p = 0.3, q = 0.8$. Each point represents a pure state mapped to by one history (of length 5), and the colour represents the history's associated causal state. For larger points (representing the most probable 99% of histories), the area is proportional to the probability of that history. For smaller points (representing the probable 0.999% of histories), the opacity is proportional to the probability.	102
8.6	ε -machine for the 3-state quasicycle, parameterized by $p \in (0, 1), \delta \in (0, p]$. The nodes are the causal states s_0, s_1 and s_2 . The arrows indicate transitions between states, where each label $x P$ indicates output x is produced with probability P	103
8.7	Error in the 2-dimensional models for the quasi-cycle, across a range of process parameters p, δ (see Figure 8.6). The left-hand-side shows the error of the optimal approximate classical model; the right-hand-side shows the error of the most accurate quantum model found by our algorithm.	103
8.8	The approximate model of asymmetric example with $p = 0.5, \delta = 0.1$. s_0 and s_1 of the original a symmetric process are merged into r_0 . The transition probability is minimized by going through the probability vector space. s_2 is mapped to r_1 . Its transition probability is unchanged.	104
8.9	Quantum states produced by the learned $d_q = 2$ encoding map for a quasicycle with $p = 0.5, \delta = 0.1$. Each point represents the second coordinate of one quantum state as the first coordinate is normalized to be real. The blue circles account for more than 99% probability of the occurrence of the pasts.	104

8.10	ε -machine for a N -step discretization of the uniform renewal process (in which events are spaced by a uniform random interval). The nodes are the causal states $s_0, \dots, s_{N-1}$. The arrows indicate transitions between states, where each label $x P$ indicates output x is produced with probability P	105
8.11	The error of quantum models for the uniform renewal process, as contrasted with N state optimal classical discretizations.	106
8.12	Example: Renewal process – encoded states. Quantum states produced by the learned $d_q = 2$ encoding map for an $N = 3$ discretization of the uniform renewal process, represented on a Bloch sphere. The diagram is interpreted as in Figure 8.5.	107

List of Boxes

6.1	Truncated dual Poisson processes	65
8.1	Recover the completeness relation and evaluate the sequence probability	95
8.2	Encoding map	96
8.3	Quantum model discovery algorithm	97
8.4	Optimal classical approximate models of dimension $d_\mu = 2$	99
8.5	Approximate discrete renewal process.	106

Symbols and Acronyms

Symbols

$\mathcal{H}$	the Hilbert space
$\mathcal{H}_m$	the memory Hilbert space
$\ \cdot\ _2$	the 2-norm of a vector or matrix in the Hilbert space
$\mathcal{E}(\cdot)$	the quantum channel
$\otimes$	the Kronecker (tensor) product
$\langle \cdot, \cdot \rangle$	the inner product of two vectors
X_t	random variable at time t.
x_t	values of the random variable X_t
$\overleftrightarrow{X}$	infinite stochastic process, $\dots X_{-1}, X_0, X_1 \dots$
$\overleftarrow{x}$	values of infinite stochastic process $\overleftrightarrow{X}$
$\overleftarrow{X}_t$	semi-infinite sequences stop at time t, $\dots X_{t-1}, X_t$
$\overleftarrow{x}_t$	values of semi-finite stochastic process $\overleftarrow{X}_t$
$\overrightarrow{X}_t$	semi-infinite stochastic process starting at time t $X_t, X_{t+1} \dots$
$\overrightarrow{x}_t$	values of semi-infinite stochastic process $\overrightarrow{X}_t$
$X_{t:t+L}$	a random variable sequence $X_t, X_{t+1} \dots X_{t+L-1}$.
$x_{t:t+L}$	the values of $X_{t:t+L}$
$P(x)$	probability of outcome x
$P(\overrightarrow{x} \overleftarrow{x})$	Conditional probability of $\overrightarrow{x}$ given $\overleftarrow{x}$
$P(s_k, x s_j)$	the transition probability
T_{jk}^x	the transition matrix
$H(\cdot)$	Shannon entropy
$H_\gamma(\cdot)$	Renyi entropy
$S(\cdot)$	von Neumann entropy
$S_\gamma(\cdot)$	quantum Renyi entropy
C_μ	classical statistical complexity

D_μ	classical topological complexity
C_q	the amount of entropic quantum memory
D_q	the logarithm of the dimension of the quantum model
$\mathbb{I}$	the identity matrix
U	unitary operator
s_k	k^{th} classical causal state
$ \sigma_k\rangle$	the k^{th} quantum state in quantum model
c_i	the complex number
μ	the learning rate
η	the leading eigenvalue
κ	Markov order
∇f	the gradient vector
ρ	the density matrix
ρ_m	the mixed memory state
$\mathcal{N}(\cdot)$	the number of 0s since the last 1
$\mathcal{N}_i$	the index set of the neighbors of agent i

Quantum gates

H	the Hadamard gate $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
σ_x	the Pauli X gate $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
σ_y	the Pauli Y gate $\begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$
σ_z	the Pauli Z gate $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
$\mathbb{I}$	the identity matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$$\text{CNOT gate} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Acronyms

HMM	Hidden Markov Model
MPS	Matrix Product states
iMPS	infinite Matrix product state
i.i.d.	independent and identically distributed

Chapter 1

Introduction

1.1 Problem Statement

Much of science seeks to understand the structure of underlying nature phenomena in order to infer its future behavior. The desire to make predictions is ubiquitous in diverse disciplines, including spin-chains [1], neuroscience [2], stock market analysis [3, 4], and non-equilibrium thermodynamics [5, 6]. However, our nature is fundamentally random, implying future uncertainty when the past is known. However, not all past information carries predictive power, and in performing predictions it is helpful – in terms of memory – to efficiently extract the useful information from the past, and discard the rest [7–10]. As knowing the entire past is not necessary in most cases, a natural question is how much past information is needed for optimal future prediction. This question further motivates compact representation and mathematical models in line with the philosophy of Ockham’s razor [11, 12].

Imagine a physical system hidden in a black box. As the physical system evolves, the black box generates an output at each time step [7–10]. For example, a random coin is put in a black box. At each time, the black box is shaken and then opened to check the state of the coin (heads or tails), as shown in Figure 1.1. The resulting output sequence then reveals information about the enclosed physical system.

Each *predictive model* offers a strategy for extracting information for purposes of future prediction [7–10]. Specifically, such predictive models encode past information into a set of classical states associated with different future behaviors, as shown

FIGURE 1.1: An illustrative figure of black box. The physical system inside the black box is a random coin. It randomly outputs heads or tails at each time step. The question marks represent unknown future outputs.

in Figure 1.2. Different predictive models can have significantly different encoding strategies, which lead to a considerable variation of the entropy (or number) of the distribution over their internal states. This entropy then quantifies the amount of information that the model requires for prediction.

FIGURE 1.2: Predictive models extract information from the past and use it to predict the future.

There has been significant effort dedicated to finding predictive models with the minimal memory [7–10]. Among all predictive models, an ε -machine requires the minimal amount of memory [7, 8]. ε -machines encode past information into so-called causal states. All pasts belonging to a given causal state have statistically indistinguishable future. The entropy of the distribution over causal states then quantifies the minimal memory an ε -machine needs to store relevant past information (see Chapter 2.3 for more details).

Meanwhile, quantum technology is an extremely fast developing area with fruitful outcomes of quantum computers and quantum information [13, 14]. Quantum computers allow for efficient solutions to many problems that are not feasible to

perform with a classical computer [15–18], such as simulation of many-body quantum physical systems [19–22]. Many companies, such as Google and IBM, are joining the competition of quantum technology, building noisy Intermediate-Scale Quantum (NISQ) technology [14]. Quantum information provides the ability to teleport a quantum state over long distances and transmit more information using entanglement [23]. Quantum scientists adapt machine learning methods, a field which has gained notable success in tasks of a similar flavour in the classical case [24–28], to the quantum case [29–33].

All of these exciting developments motivate the research into modeling stochastic processes using quantum resources. Instead of using bits, quantum resources store information in a quantum bit (qubit), which can be in any superposition of 0 and 1. Recent studies reveal that using quantum resources can reduce the amount of memory for predictive models beyond classical limits [34–40, 40–46, 46–48]. Instead of storing a classical state, quantum models encode past information into several quantum states. These quantum states can be highly non-orthogonal, leading to a significant reduction in the quantum states’ entropy compared to a classical mixture. Once a past is encoded in a quantum state, a designed recursive quantum circuit takes such quantum state and sequentially couples it with many ancillary systems. Measuring the ancillary systems generates future statistics according to the given past.

However, there remains several open questions of quantum models. First, is there a physical interpretation of the amount of memory required by a quantum model? Second, can quantum models exhibit dimensional advantages rather than entropic advantages? Third, can we construct quantum models with a fixed amount of memory from a finite amount of data? If the models are not exact, how can we quantify the accuracy of the approximate models?

In condensed matter physics, physicists pursue compact representations of many-body systems. The general description of a typical many-body quantum state requires exponential degrees of freedom, unavailable on a classical computer. An efficient and useful solution to this problem is tensor networks, which combine small tensors with network structure [49–52]. A famous form for a 1D system is called matrix product states [53–55], which can exponentially reduce the descriptive space of quantum 1D systems. This description has led to reliable and powerful numerical methods for probing and simulating properties of multi-partite systems [50, 56–58].

In this thesis, we explore the connection between tensor networks and quantum models. These quantum models couple ancillary qubits, the collection of which is regarded as a many-body system. We first show the connection between matrix product states and quantum models [46]. We also construct a family of stochastic processes whose quantum models exhibit unbounded dimensional advantages over their classical ε -machines [59]. In addition, we propose a distance measure between two stochastic processes and use the MPS method to efficiently compute the distance measure [60]. Lastly, we use machine learning methods to develop a systematic algorithm that learns a quantum model with a fixed amount of memory from finite output sequences.

1.2 Outline of the thesis

This thesis is organized as follows:

- Chapters 2, 3, and 4 provide relevant information related to the original results in this thesis. Chapter 2 reviews the theoretical frameworks of information theory, computational mechanics, and the optimal classical predictive models – ε -machines. Chapter 3 provides the background knowledge of quantum information, quantum computation, and quantum models. Chapter 4 introduces the theoretical framework for tensor networks, especially matrix product states (MPS).
- In chapter 5, we explore the connection between matrix product states and quantum models. Each quantum model is associated with an appropriate quantum state of a spin chain, called q-sample. We demonstrate that the ε -machine leads directly to an MPS representation of the corresponding quantum model. Based on this, we demonstrate the equivalence between the amount of memory of the quantum model and the bipartite entanglement in the associated spin chain. Furthermore, we provide a new method of constructing quantum models from ε -machines using the MPS method.
- In chapter 6, we construct a family of stochastic processes, which reset after generating specific outputs. Moreover, we construct quantum models of such processes, which exhibit an unbounded dimensional memory advantages over the optimal classical ε -machine.

- In chapter 7, we propose a family of distance measures between two stochastic processes. These measures satisfy a list of desiderata that a well-behaved measure should satisfy. We consider two specific family members and use MPS methods to provide an efficient algorithm for evaluating these measures.
- In Chapter 8, we address constructing quantum models with a fixed amount of memory from finite output sequences. Here, we use a machine learning framework to develop a systematic algorithm that learns the quantum models from data. We demonstrate our algorithm on three different datasets. We find that allowing for a small error can significantly reduce the required dimension of a quantum model. Moreover, we find approximate quantum models that are provably significantly more accurate than any approximate classical model of the same memory dimension. This proves a quantum model advantage in distortion for a given memory dimension.
- In chapter 9, we conclude the thesis by summarizing all results, discussing potential applications and comment on future research.

Chapter 2

Classical stochastic processes and predictive models

This chapter reviews the basic definitions and properties of random variables, stochastic processes, and entropy measures. With this knowledge in hand, the chapter then discusses predictive models and computational mechanics.

2.1 Random variables and entropies

Randomness is ubiquitous in the natural world and leads to uncertainty. Mathematically such randomness is described in terms of *random variables*. A random variable, denoted by X , denotes a quantity that can take on value x with probability $P(x)$ [61, 62].

A natural question is how much uncertainty a random variable contains. One measure of such uncertainty is the *Shannon entropy* [63, 64]:

$$H(X) := \sum_x -P(x) \log_2 P(x). \quad (2.1)$$

Renyi entropy is a generalization of Shannon entropy, defined by

$$H_\gamma(X) := \frac{1}{1-\gamma} \log_2 \left(\sum_x P(x)^\gamma \right) \quad (2.2)$$

for all non-negative γ . The Renyi entropy $H_\gamma(X)$ becomes Shannon entropy $H(X)$ in the limit as γ goes to 1. An interesting case is $\gamma = 0$ as $H_0(X)$ signifies the number of outputs that can occur.

The *relative entropy* $H(X||\tilde{X})$, also known as Kullback-Leibler divergence (KL-divergence), is a measure of the closeness between two random variables X and $\tilde{X}$ [65, 66], defined as

$$H(X||\tilde{X}) := \sum_x P(x) \log_2 \frac{P(x)}{\tilde{P}(x)} = -H(X) - \sum_x P(x) \log_2 \tilde{P}(x). \quad (2.3)$$

Suppose two joint random variables X and Y generate output (x, y) with *joint probability* $P(x, y)$. The mutual information $I(X : Y)$ can be expressed as the relative entropy of $P(x, y)$ to $P(x)P(y)$

$$I(X : Y) := H(P(x)P(y)||P(x, y)) = H(X) + H(Y) - H(X, Y). \quad (2.4)$$

2.2 Stochastic processes

Stochastic processes are central concept in probability theory. They are widely used as mathematical models of physical systems and observational phenomena, employed in diverse fields ranging from neuron science to stock market analysis [1–6, 67–70]. Specifically, at each time step t , a stochastic processes randomly generates an output x_t , which takes values from some finite alphabet $\mathcal{A}$. Each instance of the output sequence $x_{t:t+L} := x_t x_{t+1} \cdots x_{t+L-1}$ is governed by distribution $P(x_{t:t+L})$. At each time step t the outputs x_t form a random variable. As $t \rightarrow -\infty$ and $L \rightarrow \infty$, the sequence of output random variables $X_{t:t+L} := X_t X_{t+1} \cdots X_{t+L-1}$ becomes bi-infinite $\overleftrightarrow{X} := \cdots X_{-1} X_0 X_1 \cdots$.

There are stochastic processes that exhibit particular features. A stochastic process is *ergodic* if all of its statistical properties can be deduced from a sufficiently long single sequence. A process is *stationary* if the probability of any output sequence is invariant with respect to time:

$$P(x_{t:t+L}) = P(x_{0:L}), \quad \forall t \text{ and } L. \quad (2.5)$$

2.3 Predictive models

Significant effort has been devoted to understanding the underlying patterns and structure of stochastic processes. A predictive model for a given stochastic process uses the process' past $\overleftarrow{x} := x_{-\infty:0}$ to generate predictions of its corresponding future $\overrightarrow{X} := X_{0:\infty}$ such that they obey the corresponding conditional probability distribution $P(\overrightarrow{X}|\overleftarrow{x})$ [9]. To do so, some information about the past needs to be retained in order to correctly predict the future. Storing the entire past, however, is inefficient because different pasts may have the same future, i.e., $P(\overrightarrow{X}|\overleftarrow{x}) = P(\overrightarrow{X}|\overleftarrow{x}')$ for some $\overleftarrow{x}$ and $\overleftarrow{x}'$.

A simple example is a Markov process, the future of which only depends on the last output

$$P(\overrightarrow{x}|\overleftarrow{x}) = P(\overrightarrow{x}|x_{-1}), \forall x_{-1} \text{ and } \overrightarrow{x}. \quad (2.6)$$

A generalization of Markov process is a process, the future of which depends on the last κ outputs, i.e.,

$$P(\overrightarrow{x}|\overleftarrow{x}) = P(\overrightarrow{x}|x_{-\kappa:0}), \forall x_{-\kappa:0} \text{ and } \overrightarrow{x}, \quad (2.7)$$

where κ is called the *Markov order* of the stochastic process. A process is called *Non-Markovian* if its Markov order is larger than 1.

A natural question is how much past information is necessary for optimal prediction. General predictive models compress pasts into a set of classical states $\{r_k\}$, also known as *memory states*. The future is then replicated using the information from these memory states r_k . Suppose each memory state r_k occurs with probability $P(r_k)$. The amount of memory for storing these memory states is quantified by the Renyi entropies of the distribution $P(r_k)$,

$$H_\gamma(P(r_k)) = \frac{1}{1-\gamma} \log_2 \left(\sum_k P(r_k)^\gamma \right). \quad (2.8)$$

Computational mechanics address this problem by finding the optimal encoding strategy [8, 9, 71]. The pasts are regarded as equivalent if and only if their future statistics are identical, which leads to an equivalence relation

$$\overleftarrow{x} \sim \overleftarrow{x}' \text{ iff } P(\overrightarrow{X}|\overleftarrow{x}) = P(\overrightarrow{X}|\overleftarrow{x}'). \quad (2.9)$$

This equivalence relation hence allows for grouping pasts into several equivalence classes of pasts, all of which have identical future statistics, as shown in Figure 2.1. These classes are called *causal states*, labelled s_i . A predictive model encoding pasts $\overleftarrow{x}$ to causal states is called ε -machine, i.e., $\varepsilon(\overleftarrow{x}) = s_i$.

FIGURE 2.1: The pasts are separated into several classes due to the equivalence relation. Each class is denoted by a causal state s_k .

Suppose each casual state s_k occurs with probability π_k . The memory required by an ε -machine is quantified by the Renyi entropies of the causal states, defined by

$$C_\mu^\gamma := H_\gamma(\pi) = \frac{1}{1-\gamma} \log_2 \left(\sum_k \pi_k^\gamma \right). \quad (2.10)$$

ε -machines are provably minimal among all classical predictive models [9]. These Renyi entropies are considered fundamental measures of structure and complexity in the underlying process. There are two particularly meaningful cases. The most well-studied is the *statistical complexity* $C_\mu \equiv C_\mu^1$ [8, 72], which measures the minimal amount of past information (quantified by Shannon entropy) that is required for any classical predictive model. Meanwhile, the *topological complexity* $D_\mu := C_\mu^0$ [73] is the logarithm of the minimal number of classical memory states of a predictive model.

Once the past $\overleftarrow{x}$ is mapped to a causal state s_j , the future is generated from transitions between the causal states. Each transition from causal state s_j to s_k generates an output x with probability $P(x, s_k | s_j)$, as shown in Figure 2.2. The iteration of the transition replicates the future statistics of the stochastic process. The distribution π over causal states is the steady state of the *transition matrix* $T_{jk}^x := P(x, s_k | s_j)$,

$$\pi T = \pi \quad (2.11)$$

where $T = \sum_x T^x$.

FIGURE 2.2: Transitions between causal states. An arrow represents the transition from one state to another. $x|T_{jk}^x$ indicates output x with probability T_{jk}^x

2.3.1 Relation to generative models

Predictive models have an interesting relation to *generative models* [74]. A generative model is a probabilistic model in which the probability of the outputs' distribution is approximated based upon a finite set of samples. Generative models have been applied in a wide range of fields and problems such as visual recognition tasks [26], speech recognition [75], and natural language processing [76]. For example, a hidden Markov model (HMM) [77] uses a set of hidden states to model discrete time series, applied in the areas of thermodynamics [6, 78] and computational molecular biology [79]. Besides, a Boltzmann machine [80] comprises a probabilistic network of binary units with a quadratic energy function, finding applications in image processing [81, 82]. Other deep generative models play an essential role in machine learning [83, 84]. However, the only difference between predictive models and generative models is the specification that the method of generation be causal – the state of the machine at any time contains only information available from the past. Indeed, ε -machines are a special class of hidden Markov models (HMM) with the *deterministic* (or *unifilar*) property – the state of the machine at each time-step is fully determined by past observation.

2.3.2 An Example: perturbed coin

We illustrate the computational mechanics by an example, called perturbed coin [34]. The perturbed coin is a Markov process with two causal states. Its future output depends on the previous output as shown in Figure 2.3. The perturbed coin generates the same output as the one prior with probability $1 - p$ and a different output with probability p .

FIGURE 2.3: The ε -machine for perturbed coin has two causal states. The output x determines the causal state s_x .

The distribution π (see Eq. (2.11)) over the causal states of perturbed coin is $[0.5, 0.5]$ for any p . The amount of memory requirement C_μ is 1 except $p = 0.5$, where C_μ coincides to that of a completely random process.

Chapter 3

Quantum information, quantum computation and quantum models

In this chapter, we review the basic concepts in quantum information, quantum computations, and quantum modeling. We refer the reader to [34, 35, 40, 46, 85] for further details.

3.1 Quantum information

A closed quantum system is described by a pure state $|\psi\rangle$, which is a normalized complex vector in the Hilbert space $\mathcal{H}$ with dimension d .

$$|\psi\rangle = \sum_i c_i |i\rangle \quad (3.1)$$

where c_i is complex number and $\{|i\rangle\}$ forms an orthonormal basis.

Consider a closed composite quantum system AB in the pure state $|\psi_{AB}\rangle$. The state of subsystem A is described by a density matrix ρ_A , obtained by a partial trace over subsystem B

$$\begin{aligned} \rho_A &= \text{Tr}_B(|\psi_{AB}\rangle\langle\psi_{AB}|) \\ &= \sum_{i_B} \langle i_B | \psi_{AB} \rangle \langle \psi_{AB} | i_B \rangle, \end{aligned} \quad (3.2)$$

where $\{|i_B\rangle\}$ is an orthonormal basis in the state space of quantum system B. ρ_A is a $d_A \times d_A$ Hermitian matrix with all eigenvalues being non-negative.

3.1.1 Von Neumann entropy and Entanglement

Akin to how the Shannon entropy describes the uncertainty of a random variable, the *von Neumann entropy* describes the uncertainty of a quantum system ρ . It is given by

$$S(\rho) = \sum_i -\lambda_i \log_2 \lambda_i, \quad (3.3)$$

where λ_i are the eigenvalues of density matrix ρ .

Moreover, analogous to the classical Renyi entropy H_γ (see Eq. (2.2)), we have the quantum Renyi entropy

$$S_\gamma(\rho) = H_\gamma(\lambda_i), \quad (3.4)$$

where λ_i are the eigenvalues of ρ . The quantum Renyi entropy tends toward the von Neumann entropy in the limit as γ goes to 1. For $\gamma = 0$, the quantum Renyi entropy quantifies the rank of the density matrix ρ – the number of non-zero eigenvalues.

Generalizing classical mutual information, the *quantum mutual information* quantifies the correlation between two (quantum) subsystems A and B.

$$I(A : B) := S(\rho_A) + S(\rho_B) - S(\rho_{AB}). \quad (3.5)$$

The *entanglement* is one of the fundamental quantum resources behind speed-ups in quantum computation. For pure composite states $|\psi_{AB}\rangle$, the *entanglement entropy* describes the entanglement between subsystems A and B. It is given by the von Neumann entropy of one of the subsystems

$$E(|\psi_{AB}\rangle) := S(\rho_A) = S(\rho_B). \quad (3.6)$$

Any bipartite pure system has a *Schmidt decomposition* [13], which provides a canonical decomposition of the state into a separable basis of the composite system.

Theorem 3.1.1. There exists an orthonormal basis $\{|i_A\rangle\}$ for subsystem A and an orthonormal basis $\{|i_B\rangle\}$ for subsystem B such that

$$|\psi_{AB}\rangle = \sum_i \alpha_i |i_A\rangle |i_B\rangle, \quad (3.7)$$

where α_i are non-negative values satisfying $\sum_i |\alpha_i|^2 = 1$ and called *Schmidt coefficients*. The number of non-zero Schmidt coefficients α_i is called the *Schmidt rank*.

Then the entanglement of $|\psi_{AB}\rangle$ is

$$E(|\psi_{AB}\rangle) = - \sum_i \alpha_i^2 \log_2 \alpha_i^2. \quad (3.8)$$

3.1.2 Observable and Measurement

Information about a quantum system is extracted by performing a *quantum measurement*. Here, we focus on *projective measurements*, which are described by an orthonormal basis $\{|i\rangle\}$. Performing the measurement on a quantum system ρ collapses the system to state $|i\rangle$ with probability $\langle i|\rho|i\rangle$.

An observable is a Hermitian operator O , which has an eigenvalue decomposition

$$O = \sum_i o_i |i\rangle\langle i|, \quad (3.9)$$

where o_i are the eigenvalues of O and $|i\rangle$ are the corresponding eigenvectors. Note that o_i also corresponds to the possible outcomes of the observable O . Measuring the observable O is equivalent to performing a projective measurement in the basis $\{|i\rangle\}$. The expectation value of the observable is

$$\langle O \rangle_\rho = \sum_i P(i) o_i = \sum_i \langle i|\rho|i\rangle o_i = \text{Tr}(\rho O). \quad (3.10)$$

3.1.3 Quantum Channel

A *quantum channel* represents a discrete evolution of a quantum system. Assume a quantum system A in state ρ is coupled with another system B in the vacuum

state $|0\rangle$ by a unitary transformation U . Then, the quantum state of the coupled system becomes $U\rho \otimes |0\rangle\langle 0| U^\dagger$. Tracing out system B yields

$$\begin{aligned}\mathcal{E}(\rho) &:= \text{Tr}_B(U\rho \otimes |0\rangle\langle 0| U^\dagger) \\ &= \sum_x A^x \rho A^{x'},\end{aligned}\tag{3.11}$$

where $\{|x\rangle\}$ is an orthogonal basis of system B and $\{A^x := \langle x|U|0\rangle\}$ are the *Kraus operators* of the channel, satisfying the *completeness relation*

$$\sum_x A^{x\dagger} A^x = \mathbb{I}.\tag{3.12}$$

In fact, we can always find a unitary to implement any given quantum channel $\mathcal{E}$ with corresponding Kraus operators $\{A^x\}$.

Theorem 3.1.2. For any Kraus operators $\{A^x\}$ with completeness relation, we can find a unitary operator U such that

$$\langle x|U|0\rangle = A^x.\tag{3.13}$$

3.2 Quantum computation

A digital quantum computer uses quantum bits (*qubits*) as its fundamental working unit. A qubit can be implemented by a two level quantum system, described by a complex vector in a Hilbert space $\mathcal{H}$ with dimension 2. A *quantum circuit* of a digital quantum computer requires appropriate manipulation of the qubits, effecting the evolution of a multi-qubit quantum system, as shown in Figure 3.1. A general unitary evolution is performed via a series of elementary quantum gates, which consist of single-qubit gates and the two-qubit controlled-NOT (C-NOT) gate.

3.3 Quantum models

Quantum models use quantum resources to generate predictions more efficiently than classical models. Specifically, quantum models work by encoding a given

FIGURE 3.1: Each qubit is represented by a line. The blue square indicates the unitary operator. The quantum computation is performed from left to right.

past $\overleftarrow{x}$ into a pure quantum state $|\sigma_{\overleftarrow{x}}\rangle \in \mathcal{H}_m$ instead of storing a classical state $\epsilon(\overleftarrow{x})$ [40]. This quantum state $|\sigma_{\overleftarrow{x}}\rangle$, which is stored in the model's memory, is then fed into a quantum circuit, as shown in Figure 3.2. At each time step, the quantum state is coupled with an output register initialized in quantum state $|0\rangle$ by a model-dependent unitary operator U . The output of sequential measurements on the coupled output registers generates a sequence according to the appropriate statistics $P(\vec{X}||\sigma_{\overleftarrow{x}}\rangle) = P(\vec{X}|\overleftarrow{x})$.

FIGURE 3.2: A quantum circuit for operating quantum models. The red dashed lines separate the quantum circuit into repeating blocks. In each block, the quantum state is coupled with an ancillary state. Measuring the ancilla in the appropriate basis generates the outputs.

As each past is encoded into a quantum state $|\sigma_{\overleftarrow{x}}\rangle$, counting all possible pasts yields the mixed quantum state

$$\rho_m = \sum_{\overleftarrow{x}} P(\overleftarrow{x}) |\sigma_{\overleftarrow{x}}\rangle \langle \sigma_{\overleftarrow{x}}|. \quad (3.14)$$

The resource of interest here is the amount of (quantum) memory required. This is quantified by the quantum Renyi entropy of ρ_m :

$$C_q^\gamma := \frac{1}{1-\gamma} \log_2(\text{Tr}[\rho_m^\gamma]). \quad (3.15)$$

Analogous to the classical case, $C_q := C_q^1$ represents the average entropic memory required by the quantum model. It has been proven that such quantum encodings almost always lead to a reduction in entropic memory cost ($C_q < C_\mu$) as these quantum states are non-orthogonal [34, 35, 41, 42, 47]. The quantity $D_q := C_q^0$ corresponds to the logarithm of the dimension of the memory Hilbert space $\mathcal{H}_m$ in which the quantum states $|\sigma_{\underline{x}}\rangle$ lie. There exist certain processes for which the quantum memory states $|\sigma_{\underline{x}}\rangle$ can be highly linearly dependent, such that D_q can be significantly less than D_μ (see Chapter 6) [47, 59].

One systematic mean to construct a model for any discrete-time, discrete-event stationary stochastic relies on first identifying the classical ε -machine [40, 46]. Each classical causal state s_k is mapped to a corresponding state $|\sigma_k\rangle$. The coupling unitary is designed such that

$$U|\sigma_j\rangle|0\rangle = \sum_{x,k} \sqrt{P(s_k, x|s_j)} |\sigma_k\rangle|x\rangle. \quad (3.16)$$

However, this construction does not always result in dimensional advantages. Introducing complex phases before the summands of Eq. (3.16) can further reduce the memory [48, 59], in some cases resulting in potentially unbounded advantage (see Chapter 6).

3.3.1 Experimental realization

Such quantum advantages, both in entropic and dimensional memory, have already been demonstrated in proof-of-concept experiments with small-scale photonic setups [86–88]. However, the current technology is only able to accurately generate a few steps of future outputs. The system’s noise strongly affects further steps’ outputs and the error of the outputs accumulates quickly, which leads to statistical deviation. The possibility to address larger numbers of classical bits with smaller errors would presently allow classical computers to achieve a higher level of precision than current quantum computers. But it still is a promising start towards practical quantum advantages in this area.

Chapter 4

Tensor networks

Quantum systems require exponential descriptions that are intractable for a classical computer. *Tensor networks* address such problem by grouping small tensors with network structure based on certain rules, such as MPS [46, 49, 58, 89], PEPS [90] and MERA [91].

In this chapter, we focus on a specific type of tensor network, called *matrix product states* (MPS), which is an efficient tool for understanding one-dimensional quantum systems [53–55]. We review the basic notions and refer the reader to [92] for details.

4.1 Tensors

A *tensor* can be regarded as a multi-index array. The number of indices is the *rank* of the tensor. For instance, the scalar is a rank-0 tensor, and a vector is a rank-1 tensor. Significant analysis and manipulation of tensor networks can be achieved intuitively and efficiently through pictorial representations, as shown in Figure 4.1.

There are two critical rules of tensor networks. The first one is that grouping two tensors A and B represents the tensor product of these two tensors $C = A \otimes B$, as shown in Figure 4.2. Another critical component of tensor networks is the summation rule, as shown in Figure 4.2. Linking two indices of two different tensors indicates the summation over this index, also known as *contraction*. These two rules together offer the ability to combine small tensors into larger tensors. A

FIGURE 4.1: Pictorial representation of tensors. A node with N legs depicts a rank- N tensor. Each leg indicates a corresponding index. Based on this, a scalar has no legs, while a rank-3 tensor has 3 legs.

tensor network is a large tensor that combines many small tensors through network structure.

FIGURE 4.2: Rules of tensor networks: (left) Tensor product of two tensors. (right) Summation rule: Linking two legs of two different tensors indicates the summation over this index.

4.1.1 Map vectors to matrices

We define the left and right *vec map* $\mathcal{V}_l$ and $\mathcal{V}_r$ that map a vector to a matrix.

$$\mathcal{V}_l (|j\rangle\langle i|) := |i\rangle\langle j|, \quad \mathcal{V}_r (|i\rangle\langle j|) := |i\rangle\langle j|. \quad (4.1)$$

We specify a matrix's left leg as a row index and its right leg as a column index, as shown in Figure 4.3. A row vector is a matrix with one row and thus only has the column index. Similarly, a column vector only has the row index.

FIGURE 4.3: Pictorial representation of vec maps. Vec maps bend a certain leg of a node to the other side.

4.2 Matrix product states

Matrix product states (MPS) are a type of tensor network. They provide efficient approximation of 1D systems, which satisfy so-called area law (see [93] for further details). Suppose $c_{x_0, x_1 \dots x_{L-1}}$ is a given tensor with L indices $\{x_0, x_1 \dots x_{L-1}\}$. MPS sequentially decomposes

$$c_{x_0 x_1 \dots x_{L-1}} = A^{[0]x_0} A^{[1]x_1} \dots A^{[L-1]x_{L-1}}, \quad (4.2)$$

where each $A^{[k]x_k}$ is an $m_k \times m_k$ complex matrix, called *site matrix*, and $[k]$ denotes the k^{th} site. The pictorial representation of Eq. (4.2) is shown in Figure 4.4.

Assume $d_{\max}$ is the maximal dimension of the site matrices $A^{[k]x_k}$. As each site matrix has at most $d_{\max}^2$ elements, MPS representations exponentially reduce the amount of memory requirement for a 1D quantum system to $O(|\mathcal{A}|Ld_{\max}^2)$ from $O(|\mathcal{A}|^L)$ where $|\mathcal{A}|$ is the size of output alphabet (or dimension of physical index).

FIGURE 4.4: A pictorial representation for a MPS. The horizontal index is called the bond index. The vertical index is called the physical index.

4.2.1 iMPS and boundary

Consider a specific type of MPS with translation invariance. Namely all of the local tensors are identical— $A^{[k]} = A$ for all k —so that the site matrices are all square matrices of the same size. The *boundary* of an MPS is introduced to contract the sequential production of site matrices to a scalar-tensor. We focus on the open boundary which is described by two vectors $\langle b_l|$ and $|b_r\rangle$. Then we can express a tensor in terms of an MPS and its open boundary, as shown in Figure 4.5.

$$c_{x_0 x_1 \dots x_{L-1}} = \langle b_l| A^{x_0} A^{x_1} \dots A^{x_{L-1}} |b_r\rangle \quad (4.3)$$

FIGURE 4.5: MPS with translation invariance. $\langle b_l|$ and $|b_r\rangle$ are boundary tensors.

The MPS is also known as *infinite MPS (iMPS)* as L goes to infinity.

4.3 MPS for HMM

Although MPS was introduced for quantum states, it can be applied to classical stochastic processes [46, 60, 94]. Consider stochastic processes with HMM representations, which consist of several hidden states s_i and the transition probability $T_{jk}^x := P(s_k, x|s_j)$ (see Section 2.3). The transition probability has three indices, and thus is a rank-3 tensor, as shown in Figure 4.6. The left boundary is the stationary distribution of the HMM states π_i , represented by the square. Moreover, the right boundary is a column vector filled with 1s, represented by the triangle.

The probability of a particular sequence $x_{0:L}$ being generated by a HMM is given by

$$P(x_{0:L}) = \sum_{s_i^0, s_i^1 \dots s_i^{L-1}, s_i^L} \pi_i^0 P(s_i^1, x_0 | s_i^0) \dots P(s_i^L, x_{L-1} | s_i^{L-1}). \quad (4.4)$$

where s_i^t represents the i^{th} hidden state at time step t . Eq. (4.4) can be represented by an MPS, as shown in Figure 4.7.

FIGURE 4.6: MPS representation of a HMM. T_{jk}^x represents the transition probability. The square tensor represents the left boundary, which is the stationary distribution π_i of the hidden states s_i . The triangular tensor is the right boundaries, which is a row vector filled with 1.

FIGURE 4.7: Tensor decomposition of sequence probability $P(x_{t:t+L})$. The left boundary indicates that the distribution of states begins with the stationary distribution. The right boundary indicates that the sequence probability $P(x_{0:L})$ equally takes into account all final states s_i^L .

4.4 MPS for 1D quantum system

Consider a translationally invariant quantum state of a 1D spin-chain with L sites, expressed as

$$|\psi\rangle = \sum_{x_0 x_1 \dots x_{L-1}} c_{x_0 x_1 \dots x_{L-1}} |x_0 x_1 \dots x_{L-1}\rangle, \quad (4.5)$$

where $\{|x_k\rangle\}$ is an orthonormal basis of each local Hilbert space. The description of the coefficients $c_{x_0 x_1 \dots x_{L-1}}$ can be described by an MPS A^x and its boundary $\langle b_l|$ and $|b_r\rangle$. Then the quantum state can be expressed in terms of MPS representation.

$$|\psi\rangle = \sum_{x_0 x_1 \dots x_{L-1}} \langle b_l | A^{x_0} A^{x_1} \dots A^{x_{L-1}} | b_r \rangle |x_0 x_1 \dots x_{L-1}\rangle. \quad (4.6)$$

4.4.1 Transfer matrix

Suppose a quantum state $|\psi\rangle$ has iMPS representation A^x . Many of its properties are determined by the transfer matrix

$$\mathbb{E} = \sum_x A^x \otimes (A^x)^*, \quad (4.7)$$

where $(A^x)^*$ is the complex conjugate of A^x . The transfer matrix has the leading eigenvalue η , and the leading left and right eigenvectors $\langle v_l|$ and $|v_r\rangle$.

Applying the vec maps in Eq. (4.1) to the transfer matrix $\mathbb{E}$ gives two corresponding *completely positive (CP) maps* $\mathcal{E}_l$ and $\mathcal{E}_r$,

$$\begin{aligned}\mathcal{E}_l(|j\rangle\langle i|) &:= \mathcal{V}_l(\langle j|\langle i|\mathbb{E}) = \sum_x (A^x)^\dagger |i\rangle\langle j| A^x \\ \mathcal{E}_r(|j\rangle\langle i|) &:= \mathcal{V}_r(\mathbb{E}|j\rangle\langle i|) = \sum_x A^x |j\rangle\langle i| (A^x)^\dagger.\end{aligned}\tag{4.8}$$

Analogously, the vec maps turn the leading left and right eigenvectors $\langle v_l|$ and $|v_r\rangle$ into matrices V_l and V_r , as shown in Figure 4.8.

FIGURE 4.8: The tensor in the red dashed square is the transfer matrix $\mathbb{E}$. Its left two legs are grouped as a row index, and the right two legs are grouped as a column index. η is its leading eigenvalue. V_l and V_r are the corresponding leading left and right eigenvectors.

4.4.2 The probability of output sequence

Consider an iMPS $\{A^x\}$ with left and right boundaries $\langle b_l|$ and $|b_r\rangle$. The probability of observing the sequence $x_{t:t+L}$ is the expectation value of the observable $O = |x_{t:t+L}\rangle\langle x_{t:t+L}|$, i.e., $P(x_{t:t+L}) = |\langle\psi|x_{t:t+L}\rangle|^2$, as shown in Figure 4.9. If the transfer matrix of the iMPS has a unique leading eigenvalue, the sequence probability is given by

$$P_{\text{MPS}}(x_{t:t+L}) = \text{Tr} [A^{x_{t+L-1}} \dots A^{x_t} V_l A^{x_t} \dots A^{x_{t+L-1}} V_r], \tag{4.9}$$

where V_l and V_r satisfy $\text{Tr}(V_l V_r) = 1$.

FIGURE 4.9: The expectation value $\langle O \rangle$ of observable O is obtained by contracting all the indices of the whole tensor. The blue nodes, $(\langle b_l| \langle b_l|^*$ and $|b_r\rangle |b_r\rangle^*)$, represents the left and right boundary of the tensor. The red dashed line indicates the left and right boundary finally converges to the leading left and right eigenvectors of transfer matrix $\mathbb{E}$.

4.4.3 Evaluating Schmidt coefficients

The MPS representation allows us to evaluate the Schmidt coefficients of quantum state $|\psi\rangle$ across left/right bi-partition [50].

Theorem 4.4.1. Assume the iMPS' transfer matrix has leading left and right eigenvectors V_l and V_r . Then the Schmidt coefficients are the square root of eigenvalues of the density matrix

$$\rho := W_l V_r W_l^\dagger, \quad (4.10)$$

where $W_l^\dagger W_l = V_l$ and $\text{Tr}(V_l V_r) = 1$.

4.4.4 iMPS generated by quantum circuits

Any iMPS can be prepared through a recursive quantum circuit, as shown in Figure 4.10 [89]. That is, we can always find an equivalent iMPS composed of Kraus operators of a quantum channel.

Theorem 4.4.2. Assume the transfer matrix of the iMPS B^x has leading eigenvalue η and leading left eigenvector V_l . There exists an equivalent iMPS A^x such that

$$A^x = \frac{1}{\sqrt{\eta}} W_l B^x W_l^{-1} \quad (4.11)$$

where $W_l^\dagger W_l = V_l$. Moreover, A^x are Kraus operators that satisfy the completeness relation.

FIGURE 4.10: Unitary circuit for iMPS A^x .

According to Theorem 4.4.2, we assume MPS matrices $(A^x)^\dagger$ are Kraus operators. We then can always find a unitary operator U such that $\langle x|U|0\rangle = (A^x)^\dagger$. Suppose L quantum channels are sequentially concatenated, as shown in Figure 4.10. Memory system $|b_l\rangle$ (or boundary state) and output system $|0\rangle^{\otimes L}$ are fed into the circuit. The quantum state of memory system and output system becomes

$$\begin{aligned}
 U(L)|b_r\rangle|0\rangle^{\otimes L} &= \sum_{x_0} |x_0\rangle\langle x_0| U(L)|b_l\rangle|0\rangle^{\otimes L} \\
 &= \sum_{x_{L-1}} U(L-1)(A^{x_0})^\dagger|b_l\rangle|0\rangle^{\otimes L-1}|x_0\rangle \\
 &= \sum_{x_0x_1\cdots x_{L-1}} (A^{x_0}\cdots A^{x_{L-1}})^\dagger|b_l\rangle|x_0x_1\cdots x_{L-1}\rangle,
 \end{aligned} \tag{4.12}$$

where $U(L)$ denotes the concatenated unitary operator. The last step is to decouple the memory system by performing a projective measurement on the memory system and post-select the result $|b_r\rangle$.

$$\begin{aligned}
 |\psi\rangle &= \sum_{x_0x_1\cdots x_{L-1}} \langle b_r|(A^{x_0}\cdots A^{x_{L-1}})^\dagger|b_l\rangle|x_0x_1\cdots x_{L-1}\rangle \\
 &= \sum_{x_0x_1\cdots x_{L-1}} \langle b_l|A^{x_0}\cdots A^{x_{L-1}}|b_r\rangle|x_0x_1\cdots x_{L-1}\rangle.
 \end{aligned} \tag{4.13}$$

Chapter 5

The connection between tensor networks and stochastic processes

This chapter contains the first original results of this thesis, where we develop a connection between ε -machines and the MPS representation, shown in Figure 5.1. We begin by associating each stochastic process with a suitable quantum state of a spin chain called q-sample, measurement of which generates the corresponding stochastic process. We then show that the classical ε -machine of the process leads to a systematic MPS representation of the q-sample. Conversely, applying MPS methods to this state allows us to construct a quantum model for the associated stochastic process – the best known quantum model [35, 36, 40]. Lastly, we show that the entanglement across the past-future bipartition of the q-sample coincides exactly with the quantum memory requirements of the quantum model. These results thus provide a direct means of using MPS methods to study the resource requirements of quantum stochastic simulation, extending their relevance to the field of predictive modeling.

5.1 Predictive models and MPS representation

Here we present our main results—the connection between predictive models and the MPS formalism. This is done by associating a stochastic process with a large

FIGURE 5.1: This chapter connects the complexity of stochastic processes and the representational complexity of spin chains. These link (i) ϵ -machines, the provably optimal predictors of a process with the MPS states of a spin chain, (ii) the memory used to simulate a process quantum mechanically with the entanglement of the spin chain and (iii) the Hilbert space dimension of such quantum simulators with the Schmidt rank of the spin chain.

entangled 1D quantum state

$$|P(\overleftarrow{x})\rangle = \sum_{\overleftarrow{x}} \sqrt{P(\overleftarrow{x})} |\overleftarrow{x}\rangle, \quad (5.1)$$

where $\{|\overleftarrow{x}\rangle\}$ is an orthonormal basis corresponding to the outputs $\overleftarrow{x} := \cdots x_{-1}x_0 \cdots$. Measuring the quantum state in the basis $\{|\overleftarrow{x}\rangle\}$ generates the stochastic process. $|P(\overleftarrow{x})\rangle$ are known as q-samples [95, 96], and lie at the heart of many quantum sampling problems [16, 97–100]. The q-sample $|P(\overleftarrow{x})\rangle$ has iMPS representation obtained from the ϵ -machine of the stochastic process $P(\overleftarrow{x})$.

Theorem 5.1.1. A stochastic process $\overleftarrow{X}$, represented by an ϵ -machine with transition matrix T , is fully represented by the iMPS with site matrix

$$A_{kj}^x = \sqrt{T_{kj}^x}. \quad (5.2)$$

The process is ergodic if and only if the transfer matrix of the iMPS is non-degenerate.

The proof of the above theorem relies on several lemmas.

Lemma 5.1.2. The leading eigenvalue of the transfer matrix $\mathbb{E}$ of the iMPS in Eq. (4.7) is 1.

Proof. First note that the CP map $\mathcal{E}_l$ given by Eq. (4.8) has the same eigenvalues as the transfer matrix $\mathbb{E}$. It hence suffices to show that the largest eigenvalue of $\mathcal{E}_l$ is 1. For the proof we use the matrix norm $\|\cdot\|_1$ which is defined by $\|B\|_1 := \sum_{ij} |B_{ij}|$ for a matrix $B = \sum_{i,j} b_{ij} |i\rangle\langle j|$ in an orthonormal basis $\{|i\rangle\}$. For any rank-1 matrix $|j\rangle\langle k|$,

$$\begin{aligned} & \|\mathcal{E}_l(|j\rangle\langle k|)\|_1 \\ &= \sum_{mn} |\langle m|\mathcal{E}_l(|j\rangle\langle k|)|n\rangle| \\ &= \sum_{mn} \sum_x \sqrt{P(x, s_m|s_j)P(x, s_n|s_k)}. \end{aligned} \tag{5.3}$$

Since the causal state at the next step is determined by the output symbol and the previous causal state, we obtain

$$\begin{aligned} \sum_{mn} \sum_x \sqrt{P(x, s_m|s_j)P(x, s_n|s_k)} &= \sum_x \sqrt{P(x|s_j)P(x|s_k)} \\ &\leq 1. \end{aligned} \tag{5.4}$$

The last step follows from the Cauchy-Schwarz inequality. Equality holds iff $k = j$. For any matrix B as above

$$\|\mathcal{E}_l(B)\|_1 \leq \sum_{ij} |b_{ij}| \|\mathcal{E}_l(|i\rangle\langle j|)\|_1 \leq \sum_{ij} |b_{ij}| = \|B\|_1. \tag{5.5}$$

Therefore any eigenvalue λ of $\mathcal{E}_l$ is less or equal to 1.

In general, for any diagonal density matrix $\rho_d = \sum p_k |k\rangle\langle k|$

$$\begin{aligned} \mathcal{E}_l(\rho_d) &= \sum_x (A^x)^\dagger \rho_d A^x \\ &= \sum_x (A^x)^\dagger \sum_k p_k |k\rangle\langle k| A^x \\ &= \sum_j \sum_k p_k \sum_x P(x, s_j|s_k) |j\rangle\langle j|. \end{aligned} \tag{5.6}$$

For the stationary distribution in particular we define

$$V_l = \sum_k \pi_k |k\rangle\langle k|, \quad (5.7)$$

which has the following property

$$\mathcal{E}_l(V_l) = V_l. \quad (5.8)$$

This can be seen from Eq. (5.6) by noting that $\sum_{x,k} P(x, s_j | s_k) \pi_k = \pi_j$ due to stationarity. In other words, V_l is an eigenvector of $\mathcal{E}_l$ with eigenvalue 1 – hence, a leading eigenvector.

□

Lemma 5.1.3. The transfer matrix has only one eigenvalue, of magnitude one, iff the corresponding process is ergodic.

Proof. First, we assume that the process is ergodic and that V_l is a leading eigenvector of $\mathcal{E}_l$. Assume that V_l has non-zero off-diagonal elements. Since $\|\mathcal{E}_l(|j\rangle\langle k|)\|_1 < 1$ for all $j \neq k$,

$$\|\mathcal{E}_l(V_l)\|_1 \leq \sum_{jk} |(V_l)_{jk}| \|\mathcal{E}_l(|j\rangle\langle k|)\|_1 < \|V_l\|_1, \quad (5.9)$$

which contradicts $\|\mathcal{E}_l(V_l)\|_1 = \|V_l\|_1$. Thus, V_l only has non-zero diagonal elements. Convergence of the process then ensures that there is only one diagonal matrix that is the leading eigenvector of the CP map $\mathcal{E}_l$.

The inverse case is also true. We consider again Eq. (5.6). Assuming that the transfer matrix's largest-magnitude eigenvalue is unique, ρ_d always converges to the matrix $\sum_k \pi_k |k\rangle\langle k|$. Hence, any distribution under transition converges to the stationary distribution. □

Lemma 5.1.4. The leading left and right eigenvectors have the following expressions

$$V_l = \sum_k \pi_k |k\rangle\langle k|, \quad (5.10)$$

$$V_r = \sum_{k,j} \langle \sigma_k | \sigma_j \rangle |k\rangle\langle j|, \quad (5.11)$$

where $|\sigma_k\rangle$ are the internal states of the corresponding quantum model of the stochastic process $P(\overleftrightarrow{X})$ (see Eq. (3.16)).

Proof. It is easy to check that V_l is the fixed point of the CP map $\mathcal{E}_l(\cdot)$,

$$\mathcal{E}_l(V_l) = V_l. \quad (5.12)$$

For the leading right eigenvector, we show that V_r (Eq. (5.11)) is a fixed point of the CP map $\mathcal{E}_r(\rho) = \sum_x A^x \rho (A^x)^\dagger$. Applying $\mathcal{E}_r$ to V_r , we obtain

$$\begin{aligned} \langle k | \mathcal{E}_r(V_r) | j \rangle &= \sum_x \langle k | A^x V_r A^{x\dagger} | j \rangle \\ &= \sum_{m,n,x} \sqrt{T_{km}^x T_{jn}^x} \langle \sigma_m | \sigma_n \rangle. \end{aligned} \quad (5.13)$$

Now we use the fact that $U|\sigma_k\rangle|0\rangle = \sum_{x,m} \sqrt{T_{km}^x} |\sigma_m\rangle|x\rangle$, which results in

$$\langle \sigma_k | \sigma_j \rangle = \sum_{m,n,x} \sqrt{T_{km}^x T_{jn}^x} \langle \sigma_m | \sigma_n \rangle. \quad (5.14)$$

Thus, $\langle k | \mathcal{E}_r(V_r) | j \rangle = \langle \sigma_k | \sigma_j \rangle$. In other words, V_r is the leading right eigenvector. $\square$

With the above lemma in hands, we show the proof of Theorem 5.1.1.

Proof. –First, we construct a finite quantum state

$$|\psi\rangle = \sum_{x_{t:t+L}} \langle 0 | A^{x_t} A^{x_{t+1}} \dots A^{x_{t+L-1}} | 0 \rangle | x_{t:t+L} \rangle \quad (5.15)$$

with a boundary $|0\rangle$ and $x_{t:t+L} := x_t x_{t+1} \dots x_{t+L-1}$. In the limit $L \rightarrow \infty, t \rightarrow -\infty$, this yields the iMPS. Since each element of each A^x is non-negative the amplitudes generated by the iMPS are as well. In Lemma 5.1.3 we prove that the largest-magnitude eigenvalue of the transfer matrix of the iMPS given by Eq. (5.2) is nondegenerate if and only if the process is ergodic. Hence, the boundary is irrelevant.

Thus, we only need to show that the amplitudes correspond to the classical distribution $P(\overleftrightarrow{X})$. As shown in Chapter 4 Eq. (4.9), the probability distribution

generated by the iMPS is

$$P_{\text{MPS}}(X_{t:t+L} = x_{t:t+L}) = \text{Tr}(A^{x_{t+L-1}} \dots A^{x_t} V_l A^{x_t} \dots A^{x_{t+L-1}} V_r). \quad (5.16)$$

The leading left eigenvector V_l and leading right eigenvector V_r have the following form

$$V_l = \sum_k \pi_k |k\rangle \langle k|, \quad V_r = \sum_{k,j} \langle \sigma_k | \sigma_j \rangle |k\rangle \langle j|, \quad (5.17)$$

where π_k are the stationary probabilities of causal states s_k and $\{|j\rangle\}$ is an orthonormal local site basis (see Lemma 5.1.4). Inserting these expressions into Eq. (5.16), we obtain the distribution of measurement outcomes of an iMPS

$$\begin{aligned} P_{\text{MPS}}(x_{t:L}) &= \text{Tr}(A^{x_{t+L-1}} \dots A^{x_t} V_l A^{x_t} \dots A^{x_{t+L-1}} V_r) \\ &= \sum_n \langle n | A^{x_{t+L-1}} \dots A^{x_t} V_l A^{x_t} \dots A^{x_{t+L-1}} \sum_m |m\rangle \langle m | V_r | n \rangle \\ &= \sum_{n,m,i} \pi_j \sqrt{P(x_{t:t+L}, s_n | s_i) P(x_{t:t+L}, s_m | s_i)} \langle \sigma_m | \sigma_n \rangle. \end{aligned} \quad (5.18)$$

We note that the next causal state is identified by the output symbol x and the previous causal state. In other words, there exists only one causal state s_n for which $P(x_{t:t+L}, s_n | s_i)$ is non-zero.

$$\begin{aligned} P_{\text{MPS}}(x_{t:t+L}) &= \sum_{i,n} \pi_j P(x_{t:t+L}, s_n | s_i) \langle \sigma_n | \sigma_n \rangle \\ &= P(x_{t:t+L}) \end{aligned} \quad (5.19)$$

In the limit $L \rightarrow \infty$, $t \rightarrow -\infty$, this distribution becomes $P(\overleftrightarrow{X})$. $\square$

This theorem demonstrates that a classical ε -machine of a stochastic process gives direct rise to an iMPS. This representation has a clear operational meaning: each element of the site matrix corresponds to the transition probability between classical causal states.

Note that the iMPS in Eq. (5.2) is guaranteed to generate the correct statistics due to ε -machines being predictive models. This follows from Theorem 1, and does not generally hold for nonpredictive HMMs. In addition, ergodicity of the process

implies the nondegeneracy of the iMPS. Thus, the q -sample of the process is always the unique ground state of some local Hamiltonian [101].

5.2 Entanglement and memory requirement

Now, we make use of this iMPS representation to point out the relation between the amount of memory requirement of the corresponding quantum model (see Eq. (3.15)) and the entanglement of a process' q -sample across any bipartition.

Theorem 5.2.1. The entanglement across any bipartition of a stochastic process' q -sample equals the entropic quantum memory C_q . Its Schmidt rank is equal to the dimension of the memory Hilbert space.

Proof. –First, the leading left and right eigenvectors can always be decomposed as $V_l = W_l^\dagger W_l$ and $V_r = W_r W_r^\dagger$. The Schmidt coefficients are the positive square roots of the eigenvalues of the density matrix $\rho = W_l V_r W_l^\dagger$ (see Chapter 4). In line with Eq. (5.17) we choose $W_l = \sum_k \sqrt{\pi_k} |k\rangle \langle k|$. Then we construct a quantum state

$$|\psi_{AB}\rangle = \sum_k \sqrt{\pi_k} |k\rangle_A |\sigma_k\rangle_B^*, \quad (5.20)$$

where $|\sigma_k\rangle^*$ is the complex conjugate of $|\sigma_k\rangle$. Partial trace over system B yields

$$\rho_A = \sum_{k,j} \sqrt{\pi_k \pi_j} \langle \sigma_k | \sigma_j \rangle |k\rangle \langle j| = W_l V_r W_l^\dagger. \quad (5.21)$$

Thus, the Schmidt coefficients α_k of the q -sample are the square roots of the density matrix ρ_A 's eigenvalues. On the other hand, partial trace over system A gives

$$\rho_B = \sum_k \pi_k |\sigma_k\rangle \langle \sigma_k|. \quad (5.22)$$

Note that the complex conjugate of ρ_B is equal to ρ_m [Eq. (3.14)]. Hence, since ρ_A and ρ_B must have the same spectrum, ρ_A also has the same spectrum as ρ_m . Thus, the Schmidt coefficients are the square roots of the eigenvalues of ρ_m .

This implies that any Renyi entropy of the squared Schmidt coefficients and the corresponding entropy of ρ_m are also equal

$$H_\gamma(\alpha_k^2) = H_\gamma(\rho_m). \quad (5.23)$$

For the special cases $\gamma = 1$ and $\gamma = 0$, we obtain Theorem 5.2.1. $\square$

This connection illustrates an interesting link between the resource costs of generating a time sequence—a property with a temporal direction, with quantum correlations between spatial systems—a property without temporal direction.

5.3 Construct quantum models

Our final result presents a systematic means of constructing a stochastic process' quantum model from the iMPS of its q -sample.

Theorem 5.3.1. The quantum model for a stochastic process [40] can be systematically constructed from its iMPS representation according to Theorem 5.1.1:

- The internal quantum states of the quantum model are

$$|\sigma_j\rangle := W_r^\dagger |j\rangle, \quad (5.24)$$

where W_r is a decomposition $W_r W_r^\dagger = V_r$ such that the image of $W_r^\dagger$ is the same as the image of V_r .

- The stepwise unitary interaction of the quantum model is given by

$$\langle x|U|0\rangle := (W_r^{-1} A^x W_r)^\dagger, \quad (5.25)$$

where W_r^{-1} is defined to be the inverse matrix of W_r on the memory Hilbert space $\mathcal{H}_m = \text{span}\{|\sigma_k\rangle\}$.

There are two aspects of Theorem 5.3.1. First, U defined in Eq. (5.25) is a unitary operator. More importantly, U couples the internal quantum states with the ancillary system such that correct statistics are generated, i.e., $U|\sigma_k\rangle|0\rangle = \sum_{x,j} \sqrt{T_{kj}^x} |\sigma_j\rangle |x\rangle$.

For the first part, as U is partly defined on the space $\mathcal{H}_m \otimes |0\rangle$, we show that U is an isometry which can be extended to a unitary operator.

Lemma 5.3.2. U in Eq. (5.25) is an isometry such that

$$\langle 0|U^\dagger U|0\rangle = \mathbb{I} \quad (5.26)$$

Proof.

$$\begin{aligned} \langle 0|U^\dagger U|0\rangle &= \sum_x \langle 0|U^\dagger|x\rangle \langle x|U|0\rangle \\ &= \sum_x W_r^{-1} A^x V_r (A^x)^\dagger (W_r^{-1})^\dagger \\ &= W_r^{-1} V_r (W_r^{-1})^\dagger \\ &= \mathbb{I}, \end{aligned} \quad (5.27)$$

where $\mathbb{I}$ is the identity matrix. □

Before showing that U correctly couples the internal quantum states and the ancillary system, we state two lemmas that discuss the image of the operators A^x and W_r .

Lemma 5.3.3. A^x maps a quantum state in memory Hilbert $\mathcal{H}_m$ space to memory Hilbert space, i.e. $A^x|\psi\rangle \in \mathcal{H}_m$ if $|\psi\rangle \in \mathcal{H}_m$.

Proof. Assume the leading right eigenvector has the eigenvalue decomposition

$$V_r = \sum_k \lambda_k |\nu_k\rangle \langle \nu_k|, \quad (5.28)$$

where all λ_k are positive eigenvalues and $|\nu_k\rangle$ is the eigenvector. Assume the lemma is wrong. This implies that there exists a quantum state $|\psi\rangle$ that is not in the memory Hilbert space but satisfies

$$\langle \psi|A^x|\nu_s\rangle \neq 0 \quad (5.29)$$

for some $|\nu_s\rangle$ since the basis $\{|\nu_k\rangle\}$ spans the memory Hilbert space. Hence

$$\begin{aligned}
\langle\psi|V_r|\psi\rangle &= \langle\psi|\sum_x A^x V_r (A^x)^\dagger |\psi\rangle \\
&= \sum_{x,s} \lambda_s |\langle\psi|A^x|\nu_s\rangle|^2 \\
&\geq \lambda_s |\langle\psi|A^x|\nu_s\rangle|^2 \\
&> 0
\end{aligned} \tag{5.30}$$

However, $\langle\psi|V_r|\psi\rangle = 0$ contradicts the above inequality. Thus, the lemma is correct. $\square$

Then, we prove that

Lemma 5.3.4. W_r maps any quantum state into the memory Hilbert space.

Proof. W_r has the singular value decomposition

$$W_r = \sum_i c_i |u_i\rangle\langle v_i|, \tag{5.31}$$

where c_i is the singular value. $\{|u_i\rangle\}$ and $\{\langle v_i|\}$ are orthogonal bases. $\{|u_i\rangle\}$ spans the image space of W_r . Thus $V_r = \sum_i c_i^2 |u_i\rangle\langle u_i|$ has the same image space as W_r . And the memory Hilbert space, which is the image of map V_r , is equal to the image of W_r . $\square$

Then we show that U of Theorem 5.3.1 provides the correct statistics

Proof. Combining Lemma 5.3.3 and Lemma 5.3.4, we have

$$\mathcal{P}_{\mathcal{H}_m} A^x W_r = A^x W_r, \tag{5.32}$$

where $\mathcal{P}_{\mathcal{H}_m}$ is a projector onto the memory Hilbert space $\mathcal{H}_m$.

Applying U on the quantum state and the ancillary system, we have

$$\begin{aligned}
U|\sigma_k\rangle|0\rangle &= \sum_x (W_r^{-1} A^x W_r)^\dagger |\sigma_k\rangle |x\rangle \\
&= \sum_x (P_{\mathcal{H}_m} A^x W_r)^\dagger |k\rangle |x\rangle \\
&= \sum_{x,j} (W_r)^\dagger |j\rangle \langle j| (A^x)^\dagger |k\rangle |x\rangle \\
&= \sum_{x,j} \sqrt{T_{kj}^x} |\sigma_j\rangle |x\rangle.
\end{aligned} \tag{5.33}$$

□

The approach makes use of existing results in MPS literature. There, it was found that each MPS representation of a quantum state gives rise to a means of synthesizing the state through a unitary quantum circuit [58, 89]. Here, we adapt these methodologies for the specific class of iMPS in (Eq. (5.2)) such that the resulting quantum circuits are also valid predictive models. That is, the resulting circuit allows a systematic means to encode any given past $\overleftarrow{x}$ into a suitable quantum state $|\sigma_k\rangle$. A prescription for the exact circuit on $|\sigma_k\rangle$ then allows generation of correct conditional future statistics $P(\overrightarrow{X}|\overleftarrow{x})$. In addition, such models feature smaller memory dimension whenever V_r is not full rank.

5.4 Applications

In this section, we apply our results to two different stochastic processes— Nemo process and discrete renewal process. Consider a process' statistics $P(\overleftarrow{X}, \overrightarrow{X})$ and its q-sample $|\psi\rangle$. Our results allow (1) direct construction of an MPS representation for $|\psi\rangle$, (2) the use of iMPS methods to find the quantum model for the stochastic process, and (3) evaluation the corresponding quantum advantage of simulating such a process quantum mechanically.

5.4.1 Nemo processes

The first example comes with Nemo process [35], as defined by its ϵ -machine in Figure 5.2. This process has infinite Markov order because it requires infinite

past outputs to determine the causal state in the worst case. This property made such processes challenging to model prior to the development of computational mechanics.

FIGURE 5.2: The Nemo process is defined for $0 < p < 1$ and has causal states s_1 , s_2 , and s_3 . An edge from s_i to s_j indicates a non-zero transition probability, labelled ' $P(x, s_j|s_i)|x$ '.

First, according to Theorem 5.2.1, the iMPS site matrix for $|\psi\rangle$ is given by

$$A^0 = \begin{bmatrix} \sqrt{p} & 0 & 0 \\ 0 & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 \end{bmatrix}, \quad A^1 = \begin{bmatrix} 0 & \sqrt{1-p} & 0 \\ 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & 0 & 0 \end{bmatrix}. \quad (5.34)$$

Using Theorem 5.2.1, we can compute the quantum memory C_q by using the MPS to quantify the entanglement in $|\psi\rangle$. C_q is plotted, along with C_μ , in Figure 5.3. Clearly, the quantum model require less memory than provably optimal counterparts for all allowed values p .

The internal quantum states and the unitary operator U are also straightforward to obtain, by virtue of Theorem 5.3.1. They are obtained by operating $W_r^\dagger$ on the basis $\{|j\rangle\}$

$$\begin{aligned} |\sigma_0\rangle &= c \left(\sqrt{2p}|e_0\rangle + \sqrt{1-p}|e_1\rangle \right), \\ |\sigma_1\rangle &= c \left(\sqrt{p}|e_1\rangle + |e_2\rangle \right), \\ |\sigma_2\rangle &= c \left(|e_0\rangle + \sqrt{p}|e_2\rangle \right), \end{aligned} \quad (5.35)$$

with $c \equiv (1+p)^{-\frac{1}{2}}$ and $\{|e_i\rangle\}$ forms an orthonormal basis.

FIGURE 5.3: The amount of quantum memory C_q is less than the classical memory C_μ for Nemo processes.

According to Theorem 5.3.1, the unitary operator U is characterized by

$$\begin{aligned}
 U|e_0\rangle|0\rangle &= \sqrt{p}|e_0\rangle|0\rangle + \xi|e_1\rangle|0\rangle + \xi|e_1\rangle|1\rangle, \\
 U|e_1\rangle|0\rangle &= |e_2\rangle|1\rangle, \\
 U|e_2\rangle|0\rangle &= |e_0\rangle|1\rangle,
 \end{aligned} \tag{5.36}$$

where $\xi = \sqrt{\frac{1-p}{2}}$.

The matrix form of U , written in the basis $\{|e_0\rangle|0\rangle, |e_0\rangle|1\rangle, |e_1\rangle|0\rangle, |e_1\rangle|1\rangle, |e_2\rangle|0\rangle, |e_2\rangle|1\rangle\}$ is

$$\begin{bmatrix}
 \sqrt{p} & \# & 0 & \# & 0 & \# \\
 0 & \# & 0 & \# & 1 & \# \\
 \sqrt{\frac{1-p}{2}} & \# & 0 & \# & 0 & \# \\
 \sqrt{\frac{1-p}{2}} & \# & 0 & \# & 0 & \# \\
 0 & \# & 0 & \# & 0 & \# \\
 0 & \# & 1 & \# & 0 & \#
 \end{bmatrix} \tag{5.37}$$

where $\#$ denotes undetermined terms which may be assigned by a Gram-Schmidt procedure for the matrix columns.

5.4.2 Discrete renewal processes

We illustrate these results using the discrete renewal process with uniform emission probability [42, 102], as defined by its ε -machine in Figure 5.4. First, according to

Theorem 5.1.1, the iMPS site matrix for $|\psi\rangle$ is given by

$$A_{k,k+1}^0 = \sqrt{\frac{N-k-1}{N-k}}, \quad A_{k,0}^1 = \sqrt{\frac{1}{N-k}}. \quad (5.38)$$

FIGURE 5.4: The ε -machine of the discrete renewal process. An edge labeled $x|p$, indicates that the transition from causal state s_i to s_j generates output x with probability p . The renewal process either emits 1 and returns to s_0 or 0 and evolves further.

With Theorem 5.2.1, we can compute the quantum memory C_q by using the MPS to quantify the entanglement in $|\psi\rangle$. In Figure 5.5, we see that the quantum model requires only a bounded memory while the classical optimal memory scales as $O(\log N)$ [42]. In fact, unbounded memory advantage for C_q^0 can also occur [47].

FIGURE 5.5: The amount of quantum memory C_q converges to a finite value while the optimal classical memory C_μ diverges logarithmically.

The internal quantum states and the unitary operator U are straightforward to obtain, by virtue of Theorem 5.3.1. First, note that $W_r^\dagger$ in the basis $\{|j\rangle\}$ is given

by the matrix

$$\begin{bmatrix} \sqrt{\frac{1}{N}} & 0 & \cdots & 0 \\ \sqrt{\frac{1}{N}} & \sqrt{\frac{1}{N-1}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{1}{N}} & \sqrt{\frac{1}{N-1}} & \cdots & 1 \end{bmatrix}. \quad (5.39)$$

Thus, the signal state is

$$|\sigma_k\rangle = W_r^\dagger |k\rangle = \sum_{j \geq k} \sqrt{\frac{1}{N-k}} |j\rangle. \quad (5.40)$$

This leads to the desired elements of the unitary operator U :

$$\begin{aligned} \langle 0|U|0\rangle &:= (W_r^{-1} A^0 W_r)^\dagger = \sum_{j=0}^{N-2} |j+1\rangle \langle j| = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \\ \langle 1|U|0\rangle &:= (W_r^{-1} A^1 W_r)^\dagger = |\sigma_0\rangle \langle N| = \begin{bmatrix} 0 & 0 & 0 & \cdots & \sqrt{\frac{1}{N}} \\ 0 & 0 & 0 & \cdots & \sqrt{\frac{1}{N}} \\ 0 & 0 & 0 & \cdots & \sqrt{\frac{1}{N}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{\frac{1}{N}} \end{bmatrix}. \end{aligned}$$

5.5 Conclusion

In this chapter, we have connected two previously distinct notions of complexity: the memory cost of prediction and the representational complexity of spin chains—allowing results in one field to catalyze new insights in the other. By associating each stochastic process with a suitable quantum state of a spin chain, we demonstrated that predictive models constructed in complexity science lead directly to the iMPS representation of a general classical stochastic process. Based on this, optimization using standard iMPS techniques allows systematic construction of quantum models—the current state of the art for predictive modeling. Moreover,

we established that the memory requirement of such quantum models exactly coincides with the bipartite entanglement in the associated spin chain state.

Chapter 6

Extreme dimensional compression with quantum modelling

Quantum models can drastically reduce the amount of entropic memory C_q for modelling stochastic processes [34, 35, 39, 42, 47, 103–105]. In this chapter, we show quantum models can gain unbounded dimensional memory advantages, which quantifies the amount of resource required for single-shot application. We consider a family of such processes where the memory dimension required of a faithful classical model diverges with precision. However, we identify quantum models that compress all pasts into two dimensions. This allows for all relevant history to be stored in a single two-state quantum system (*qubit*), evincing an extreme quantum advantage that scales without bound. Moreover, our protocol requires only a single probe qubit to extract the future statistics. This turns a problem from the converse scenario – that tracking a finite quantum system can require infinite classical resources [106–109] – into a useful tool.

6.1 Dual Poisson processes

Consider a system that undergoes a series of Poissonian decay events through one of two channels with rates γ_1 and γ_2 . After each event, the decay channel for the next emission is chosen randomly, with probability p or $\bar{p} = 1 - p$ respectively, as shown in Figure 6.1. The choice of channel is hidden internally in the system, such that an external observer can only see when the decay events occur. Specifically,

we consider an observer operating on discrete timesteps Δt , recording a 1 when an event occurs, and 0 otherwise. We call the resultant stochastic process a dual Poisson process, and it manifests as a series of 1s separated by strings of 0s.

FIGURE 6.1: Dual Poisson process. Each channel outputs 0 with probability q_i and 1 with probability $1 - q_i$. The red arrow indicates that the transition happens immediately. $\emptyset$ indicates no output.

Note that the probabilistic choice of channel occurs only after events (1s), and remains unchanged across non-events (0s). The probability that a contiguous string of 0s (bookended by 1s) is of at least length n is given by the so-called survival probability $\Phi(k)$:

$$\Phi(k) = pq_1^k + \bar{p}q_2^k, \quad (6.1)$$

where $q_i = \exp(-\gamma_i \Delta t)$ for some choice of γ_i . We shall now look at the scaling of the memory metrics of causal models for such processes as the temporal precision Δt is refined – making the process increasingly non-Markovian. With arbitrary $\Phi(k)$ this framework describes general renewal processes [110].

6.2 ε -machine of dual Poisson processes

The construction of ε -machine relies on identifying the equivalent pasts (see Section 2.3). Since the observer is unaware of the choice of decay channel, the information they must track reflects their confidence in the chosen rate based on the time since last emission. That is, a causal model tracks back the number of 0s $\mathcal{N}(\overleftarrow{x})$ since the last 1 for any given past $\overleftarrow{x}$. As the output is governed by the

survival probability, the future for a given past $\overleftarrow{x}$ is

$$\begin{aligned} P(0|\overleftarrow{x}) &= \frac{\Phi(k+1)}{\Phi(k)}, \\ P(1|\overleftarrow{x}) &= 1 - \frac{\Phi(k+1)}{\Phi(k)}. \end{aligned} \quad (6.2)$$

where $k = \mathcal{N}(\overleftarrow{x})$. As a result, the equivalence relation is

$$\overleftarrow{x} \sim \overleftarrow{x'} \quad \text{iff} \quad \mathcal{N}(\overleftarrow{x}) = \mathcal{N}(\overleftarrow{x'}). \quad (6.3)$$

Therefore, the causal state s_k represent the class consisted of pasts with $\mathcal{N}(\overleftarrow{x}) = k$ (e.g. $s_2 = \{\dots 100\}$). Combining Eq. (6.2) and Eq. (6.3), the transition probability between causal states is

$$\begin{aligned} P(s_{k+1}, 0|s_k) &= \frac{\Phi(k+1)}{\Phi(k)}, \\ P(s_0, 1|s_k) &= 1 - \frac{\Phi(k+1)}{\Phi(k)}. \end{aligned} \quad (6.4)$$

The ε -machine of dual Poisson process is plotted in Figure 6.2.

FIGURE 6.2: ε -machine for dual Poisson process. Each casual state s_k outputs 0 and goes forward to s_{k+1} . Outputs 1 indicates that all states renew to s_0

According to the above construction of ε -machine, we can immediately identify that C_μ and D_μ are infinite in the continuum limit ($\Delta t \rightarrow 0$), as the number of causal states grows to ∞ . However, the differences between conditional probabilities become increasingly small for states s_k with large k and the probability of reaching such states is very small. We hence introduce a truncated form of the model, where after sufficiently large $k\Delta t$ the causal states are all merged together – and study the associated complexities $\tilde{C}_\mu$ and $\tilde{D}_\mu$ of this model (see Box. 6.1). Their scaling with increasing precision (i.e. decreasing Δt) for $\gamma_1 = 12$, $\gamma_2 = 1$ and $p = 0.9$ is

FIGURE 6.3: Scaling of memory metrics with precision. Both classical memory metrics diverge with increasing precision, wherein the interval Δt is refined. In contrast, the quantum metrics remain finite, evincing an unbounded advantage: C_q tends to a bounded value, while D_q remains constant. Plot shown for $\gamma_1 = 12$, $\gamma_2 = 1$ and $p = 0.9$ (the qualitative features are typical for any non-extremal parameter choice).

shown in Figure 6.3. Note that the qualitative features of this plot are typical for any non-extremal choice of parameters (i.e. $P \neq 0, 1$ and $\gamma_1 \neq \gamma_2$).

6.3 Unbounded quantum compression advantages

We now show that this scaling divergence is a purely classical phenomenon, and need not persist in the quantum regime. We construct quantum models of such processes for which the memory metrics are finite at any level of precision. These quantum models exhibit unbounded quantum advantages in compression, forming our main result.

Theorem 6.3.1. A quantum model with $C_q \leq 1$ and $D_q \leq 1$ exists for any dual Poisson process at any level of precision Δt .

Proof. The quantum models work by encoding the memory into one qubit, and using another to probe it (see Section 3). Specifically, the causal state s_k is encoded into

$$|\sigma_k\rangle = \frac{\sqrt{pq_1^k} + ig\sqrt{\bar{p}q_2^k}}{\sqrt{\Phi(k)}}|0\rangle + \frac{i\sqrt{(1-g^2)\bar{p}q_2^k}}{\sqrt{\Phi(k)}}|1\rangle, \quad (6.6)$$

where

$$g = \frac{\sqrt{(1-q_1)(1-q_2)}}{1 - \sqrt{q_1q_2}}. \quad (6.7)$$

As noted above, D_μ remains infinite at any level of discretisation, as there is no maximum k that the processes cannot exceed, and no merging of different s_k into the same causal state. However, keeping this infinite overhead of states provides very little additional predictive power at large $k\Delta t$, as the probability of the process reaching these states is small, and the difference in conditional probabilities between s_k and s_{k+1} is monotonically decreasing towards zero with increasing k . We therefore introduce a truncated form of the process, where there is a designated ‘terminal’ state at $\bar{k}$, such that all states at $k > \bar{k}$ are merged down into this state. This terminal state must have transition probabilities that are a weighted average of all the merged states; when an event happens the model transitions to s_0 as before, but now on non-events the system remains in the $s_{\bar{k}}$ state, as opposed to advancing further.

There is a level of subjective choice in how $\bar{k}$ is selected. Appropriate methods can be for example based on the fidelity of the conditional distributions at larger k , or on the probabilities of reaching such states. For concreteness, we pick a straightforward criterion:

$$\bar{k} := \min_{k \in \mathbb{N}} k | \Phi(k) \leq \delta(1 - \Phi(1)). \quad (6.5)$$

That is, $\bar{k}$ is the first state for which the probability of reaching said state is less than a fraction δ of the probability of decay in the first timestep. We here use $\delta = 0.01$.

Box. 6.1: Truncated dual Poisson processes

Clearly, all quantum states $|\sigma_k\rangle$ lie in a Hilbert space with dimension 2. We assign

$$\begin{aligned} |\phi_1\rangle &= |0\rangle, \\ |\phi_2\rangle &= g|0\rangle + \sqrt{1-g^2}|1\rangle. \end{aligned} \quad (6.8)$$

The quantum states can be expressed in terms of $\{|\phi_1\rangle, |\phi_2\rangle\}$

$$|\sigma_k\rangle = \frac{\sqrt{pq_1^k}}{\sqrt{\Phi(k)}} |\phi_1\rangle + \frac{i\sqrt{pq_2^k}}{\sqrt{\Phi(k)}} |\phi_2\rangle. \quad (6.9)$$

The next step is to show there exists unitary U satisfying

$$U|\sigma_k\rangle|0\rangle = \sqrt{\frac{\Phi(k+1)}{\Phi(k)}}|\sigma_{k+1}\rangle|0\rangle + \sqrt{1 - \frac{\Phi(k+1)}{\Phi(k)}}|\sigma_0\rangle|1\rangle. \quad (6.10)$$

As the ancillary system is $|0\rangle$, the unitary operator can be represented by Kraus operators A^x . Inserting A^x into Eq. (6.10) provides

$$U|\sigma_k\rangle|0\rangle = A^0|\sigma_k\rangle|0\rangle + A^1|\sigma_k\rangle|1\rangle. \quad (6.11)$$

Then the Kraus operators are required to satisfy

$$\begin{aligned} A^0|\sigma_k\rangle &= \sqrt{\frac{\Phi(k+1)}{\Phi(k)}}|\sigma_{k+1}\rangle, \\ A^1|\sigma_k\rangle &= \sqrt{1 - \frac{\Phi(k+1)}{\Phi(k)}}|\sigma_0\rangle. \end{aligned} \quad (6.12)$$

We define A^x in terms of $|\phi_i\rangle$,

$$\begin{aligned} A^0|\phi_1\rangle &:= \sqrt{q_1}|\phi_1\rangle, \\ A^0|\phi_2\rangle &:= \sqrt{q_2}|\phi_2\rangle, \\ A^1|\phi_1\rangle &:= \sqrt{1 - q_1}(\sqrt{p}|\phi_0\rangle + i\sqrt{\bar{p}}|\phi_1\rangle), \\ A^1|\phi_2\rangle &:= \sqrt{1 - q_2}(\sqrt{p}|\phi_0\rangle + i\sqrt{\bar{p}}|\phi_1\rangle). \end{aligned} \quad (6.13)$$

Clearly, A^x satisfy the equality (Eq. (6.12)). Besides, operators A^x defined in Eq. (6.13) are well defined Kraus operators. The completeness relation of A^x is seen from the fact that for any quantum state $|\tilde{\phi}\rangle = c_1|\phi_1\rangle + c_2|\phi_2\rangle$, we have

$$\begin{aligned} \langle \tilde{\phi} | \sum_x A^{x\dagger} A^x | \tilde{\phi} \rangle &= \sum_{i,j,x} c_i^* c_j \langle \phi_i | A^{x\dagger} A^x | \phi_j \rangle \\ &= |c_1|^2 + |c_2|^2 + c_1^* c_2 \langle \phi_1 | \phi_2 \rangle + c_1 c_2^* \langle \phi_2 | \phi_1 \rangle \\ &= \langle \tilde{\phi} | \tilde{\phi} \rangle \\ &= 1. \end{aligned} \quad (6.14)$$

Note that $\sqrt{q_1 q_2} \langle \phi_1 | \phi_2 \rangle + \sqrt{(1 - q_1)(1 - q_2)} = \langle \phi_1 | \phi_2 \rangle$. □

Crucially, Theorem 6.3.1 evinces that the memory states can be encoded into a

single qubit, guaranteeing $C_q \leq 1$ and $D_q \leq 1$. Moreover, since the process is generically not memoryless, and a binary system is the smallest possible memory, we can conclude that our model is (single-shot) minimal and that $D_q = 1$ is the quantum topological complexity. In Figure 6.3 we compare the scaling of the quantum memory metrics with those of the minimal classical model. We thus observe the unbounded scaling of the quantum compression advantage in both ensemble and single-shot settings.

6.3.1 Evaluating C_μ and C_q

We calculate the quantum statistical memory C_q of our models in the continuum limit ($\Delta t \rightarrow 0$) for the whole dual Poisson process family. Due to the timescale invariance of C_q for renewal processes [42], the entire process family can be characterised by two parameters: $\gamma = \gamma_1/\gamma_2$ and p . Moreover, due to symmetries of the processes, $\{\gamma, p\}$ yields the same model as $\{\gamma^{-1}, \bar{p}\}$.

Prior work on the computational mechanics of renewal processes [111] has established that the steady-state probabilities of the causal states are proportional to their survival probabilities. That is, $\pi_k = \mu \Phi(k)$, where $\mu^{-1} := \sum_{k=0}^{\infty} \Phi(k)$. For dual Poisson processes,

$$\mu = \frac{(1 - q_1)(1 - q_2)}{p(1 - q_2) + \bar{p}(1 - q_1)}. \quad (6.15)$$

Thus, we have that $C_\mu = -\sum_{k=0}^{\infty} \pi_k \log_2 \pi_k$ and

$$\tilde{C}_\mu = -\sum_{k=0}^{\bar{k}-1} \pi_k \log_2 \pi_k - \pi_{\geq \bar{k}} \log_2 \pi_{\geq \bar{k}}, \quad (6.16)$$

where $\pi_{\geq \bar{k}} = \sum_{k=\bar{k}}^{\infty} \pi_k$. Further, using Eq. (3.14) with Eq. (6.6), we obtain

$$\rho_m = \mu \begin{bmatrix} \frac{p}{1-q_1} + \frac{g^2 \bar{p}}{1-q_2} & \frac{g\sqrt{(1-g^2)\bar{p}}}{1-q_2} - \frac{i\sqrt{(1-g^2)p\bar{p}}}{1-\sqrt{q_1 q_2}} \\ \frac{g\sqrt{(1-g^2)\bar{p}}}{1-q_2} + \frac{i\sqrt{(1-g^2)p\bar{p}}}{1-\sqrt{q_1 q_2}} & \frac{(1-g^2)\bar{p}}{1-q_2} \end{bmatrix}, \quad (6.17)$$

which may be straightforwardly diagonalised to find the two eigenvalues $\{\lambda_1, \lambda_2\}$, and hence calculate $C_q = -\lambda_1 \log_2(\lambda_1) - \lambda_2 \log_2(\lambda_2)$.

FIGURE 6.4: **Quantum statistical memory of dual Poisson processes.** For much of the dual Poisson process family, the information stored by our quantum models is significantly less than one bit, highlighting that significant further compression is possible for ensemble simulators. The information stored is largest when both (i) the emission rates are significantly different, and (ii) the system is more likely to adopt the faster rate, as the state of confidence is frequently reset. $D_q = 1$ everywhere except the lines $P = 0, 1$ and $\gamma = 1$, where $D_q = 0$.

In Figure 6.4 we plot the continuum limit C_q for a broad parameter range. Interestingly, we see that for most choices of parameters $C_q \ll D_q$, suggesting yet further significant memory savings can be achieved in the ensemble regime. We see that C_q is largest when the two rates are significantly different (since the choice of decay channel has larger impact on the future) and the system predominantly picks the faster rate (since the confidence in the choice of channel is frequently reset).

6.4 Conclusion

In this chapter, we have shown an unbounded dimensional scaling advantage for the dual Poisson process. Both our quantum model and the optimal classical model provide an approximation of the fully-continuous process for finite timesteps Δt . Interestingly, a decrease in timestep size is not accompanied by an increase in memory size for the quantum case. The quantum model memory size D_q is entirely independent of the timestep, and does not exhibit the classical scaling of memory with precision. This indicates that the limiting factor in the accuracy of quantum models of such processes is not the available memory, but the accuracy with which

it can be addressed. Our results already indicate a robustness of the quantum advantage in some sense. The errors in the implementation of the quantum model can be accounted for by limiting the precision Δt to not exceed that achievable by the experiment. And the truncated process mitigate the problem of noise exceeding the difference in future statistics for large k . Ultimately, while it would not be possible to witness the scaling difference all the way up to the continuum limit, it can still be shown up to the best achievable precision.

Chapter 7

Measures of distinguishability between stochastic processes

How alike are the behaviours of two systems? How similar are the trajectories of two stock prices? Much of the physical world can be described as a collection of interacting stochastic processes; understanding how distinguishable two processes are allows us to answer such questions. These problems are of universal relevance – for example, quantifying how closely a model replicates its target has applications in fields such as protein homology [112] and speech recognition [113, 114]. Meanwhile, understanding how much external noise or perturbations impact upon the behaviour of a system is a central task in studies of open systems [115], quantum computation [14], and machine learning [116, 117].

Many measures have been proposed for this task [65, 118–125]. However, a number of these are measures tailored for quantifying distances between distributions. Though they work well for quantifying distances between finite strings, they typically do not behave well in the context of processes where infinite strings of observational data arise as a process continues to run. Particularly, they fail to quantify *how* different processes are [118, 119, 121–123]. Others, such as the Kullback-Leibler (KL)-divergence [65, 120], require intensive computational resources to evaluate and can behave pathologically in seemingly innocuous situations such as processes with different output alphabets. Finally, measures that possess an intrinsic dependence on a representation of a process rather than the process itself

will often fail by misidentifying different models of the same process with identical observable behaviour as being distinguishable [125, 126].

In this chapter, we suggest several requirements a measure of process distinguishability should satisfy. We then propose a family of measures that satisfy all of these properties. We focus on two specific members of this family and develop an efficient method to compute the exact value of this measure. Furthermore, we illustrate our proposal by applying it to a set of example scenarios designed to highlight where other measures either cannot be applied or behave pathologically.

7.1 Criteria for well-behaved measures

With the above questions as motivation, we suggest that a good measure of distinguishability between stochastic processes $R(P, Q)$ should satisfy the following criteria:

1. *Non-negativity*: $R(P, Q) \geq 0$.
2. *Symmetry*: $R(P, Q) = R(Q, P)$.
3. *Identity of indiscernibles*: $R(P, Q) = 0 \Leftrightarrow P = Q$, i.e., $R(P, Q) = 0$ iff the processes are identical.

Together these three conditions define a semimetric distance. Note that as with many proposed measures of distance between processes, we do not demand the triangle inequality $R(P, Q) \leq R(P, G) + R(Q, G) \forall G$ be satisfied, and so the measure will not necessarily be a metric.

4. *Model-independence*: $R(P, Q)$ should depend only on observable properties (i.e., the outputs) of the processes, and not any underlying models.

This condition enforces that the measure should be identical when calculated relative to any representation of the processes.

5. *Continuity*: Suppose stochastic process Q depends on a continuous parameter δ . Continuity mandates that $\lim_{\delta \rightarrow \delta_0} R(P, Q(\delta)) = R(P, Q(\delta_0))$.

Smooth deformations in the parameters defining a process will smoothly change the distinguishability between it and other processes – this condition enforces that this is reflected in the measure.

7.2 Existing measures

With these requirements at hand, we are able to assess the behavior of existing measures that have been proposed for quantifying how distinct two stochastic processes are. One such widely-used measure, often serving as a cost function for many machine learning works, is the aforementioned KL-divergence (see Eq. (2.3))

$$D_{KL}(P||Q) = \sum_{x_{t:t+L}} P(x_{t:t+L}) \log_2 \left[\frac{P(x_{t:t+L})}{Q(x_{t:t+L})} \right]. \quad (7.1)$$

It can be seen however, that the KL-divergence does not satisfy the continuity criterion. It becomes singular when two stochastic processes contain different sets of possible output sequences, no matter how small the probability is of these unique sequences occurring. Moreover, while not a violation of any of the above criteria, a further drawback of the KL-divergence is its high computational cost, as it requires a calculation over all output sequences, the number of which grows exponentially with their length L .

Other measures, such as the Jensen-Shannon-divergence [123], though free from singularities, can still fail the continuity criterion in the context of processes where the output sequence lengths are infinite. This problem is endemic to measures based on distances between distributions, such as trace norms and the Bures distance [119] – and highlights a crucial key difference between distributions over finite sequences and processes. Specifically, these measures asymptotically saturate to their maximal value as the length of the output sequences increases. Because any two different processes can be asymptotically distinguished for sufficiently long output sequences. As a result, these measures are either 0 or maximal – identifying whether P and Q are different processes, but not how different they are. A good measure of process distinguishability should be equipped to handle this distinction.

Finally, the model independence criterion rules out other measures [125, 126] that are explicitly based on the structure of a particular model. That is, for such measures two models of the same stochastic process may be identified as having non-zero distance between them despite exhibiting identical observable behaviour. Furthermore, such model-dependent measures may also be impossible to evaluate for some pairs of models with sufficiently distinct structures.

7.3 A new family of measures

In light of such issues with commonly-used measures, we seek a measure that satisfies all of our criteria. We propose a family of measures of process distinguishability, called divergence rates, which measure how quickly the observed behaviour of two processes becomes distinguishable. That is, they quantify how much the distance between output sequence probability distributions grows with their length.

Consider a metric distance measure between distributions $D(P, Q)$ that is normalised such that $0 \leq D(P, Q) \leq 1$. We introduce the notion of *similarity* S_D , that can be thought of as the complement to the distance, satisfying

$$S_D(P, Q) := \sqrt{1 - D(P, Q)^2}. \quad (7.2)$$

We then define the D divergence rate as

$$R_D(P, Q) := - \lim_{L \rightarrow \infty} \frac{1}{L} \log[S_D(P, Q; L)], \quad (7.3)$$

where $S_D(P, Q; L)$ and similarly $D(P, Q; L)$ are used to denote these quantities evaluated for distributions formed from sequences of length L output by the processes. The D divergence rate can be seen to parameterise the rate at which the similarity (according to the distance D) of the two processes decays once many symbols have been observed. The limitation of $L \rightarrow \infty$ accounts for strings with infinite length, which suppresses the distance induced by different initial points. This parallels the notion of entropy rates [127], which similarly capture long-term behaviour by averaging over long sequences.

Theorem 7.3.1. Suppose a continuous, normalised metric distance $D(P, Q; L)$ scales with L as $D(P, Q; L) \sim 1 - \alpha \exp(-\eta L)$ for non-negative $\alpha \leq 1$ and non-negative real η that continuously depends on the stochastic processes. $R_D(P, Q)$ induced by the distance $D(P, Q; L)$ fulfils the above requirements on a measure of process distinguishability. For such a distance, we have that $R_D(P, Q) = \eta/2$.

Proof. Conditions 1, 2 and 4 immediately follow from the definition of $R_D(P, Q)$ and the properties of $D(P, Q)$ as a metric. By directly inserting the scaling $D(P, Q; L) = 1 - \alpha \exp(-\eta L)$ into Eq. (7.3) it can be seen that $R_D(P, Q) = \eta/2$. As $D(P, Q)$ is a metric it follows that if two processes are identical we must have

$\eta = 0$, and conversely $\eta = 0$ indicates that two processes have identical long-term behaviour – and hence condition 3 is satisfied. Note that α is irrelevant, and depends only on transient behaviour resulting from the initial configuration of the two processes. Finally, condition 5 follows from the equality $R_D = \eta/2$ and the continuity of the decaying rate η . $\square$

The above theorem holds only for specific metric distances, which continuously depends on the stochastic processes and scales as desired. Furthermore, whenever $D(P, Q)$ exhibits the required scaling, we see that $R_D(P, Q)$ is infinite iff $D(P, Q; L)$ becomes exactly 1 within finite L , rather than just asymptotically approaching it. Qualitatively, this can be understood as the measure being infinite iff the two processes can be discriminated with certainty by observing a sufficiently long yet finite sequence of outputs.

7.4 Two specific members

Theorem 7.3.1 demonstrate that a divergence measure is well-behaved (satisfy all the criteria) whenever it decays properly. There are many choices of the metric distance $D(\cdot, \cdot)$. However, only certain metric distances satisfy the requirements of theorem 7.3.1. In this section, we discuss two specific members of the divergence rate that is well-behaved. Moreover, we develop an efficient way to compute the each member of the divergence rate.

7.4.1 Co-emission divergence rate

We now consider the case where the distance used is the l_2 norm, given by

$$D_{l_2}(P, Q) := \frac{1}{\sqrt{2}} \|\hat{P} - \hat{Q}\|_2, \quad (7.4)$$

where $\hat{P} = P/\|P\|_2$ and $\hat{Q} = Q/\|Q\|_2$. Noting that $\|\hat{P} - \hat{Q}\|_2^2 = 2 - 2\langle \hat{P}, \hat{Q} \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product, this distance can be expressed in terms of so-called co-emission probabilities [122] $C(P, Q; L) = \sum_{x_{0:L}} P(x_{0:L})Q(x_{0:L})$. Using this, we obtain

$$R_C(P, Q) = - \lim_{L \rightarrow \infty} \frac{1}{2L} \log \left[\frac{C(P, Q; L)}{\sqrt{C(P, P; L)C(Q, Q; L)}} \right], \quad (7.5)$$

which we call the *co-emission divergence rate* (CDR).

Theorem 7.4.1. The CDR satisfies all of the requirements specified for a good measure of process distinguishability.

As D_{l_2} is a continuous metric distance, we need only to show that the measure obeys the specified scaling.

Lemma 7.4.2. $D_{l_2}(P, Q; L)$ scales as $1 - \alpha \exp(-\eta L)$ with the continuity of η .

In proving Lemma 7.4.2, we provide an efficient method for evaluating CDR between stochastic processes with general HMM representation, which is not required to be deterministic.

Corollary 7.4.2.1. Given a HMM representation of process P with transition probabilities $P(s_j, x|s_i)$ and of process Q with $Q(\tilde{s}_n, x|\tilde{s}_m)$, the CDR between them is given by

$$R_C(P, Q) = -\frac{1}{2} \log \left[\frac{\mu_{PQ}}{\sqrt{\mu_P} \sqrt{\mu_Q}} \right], \quad (7.6)$$

where μ_P , μ_Q and μ_{PQ} are the leading eigenvalues of the transfer matrices $\mathbb{E}_{PP}$, $\mathbb{E}_{QQ}$ and $\mathbb{E}_{PQ}$, defined as

$$(\mathbb{E}_{PQ})_{im,jn} := \sum_x P(s_j, x|s_i) Q(\tilde{s}_n, x|\tilde{s}_m). \quad (7.7)$$

Proof. First note that every stationary stochastic process has a HMM representation [9]; we consider two stationary stochastic processes P and Q with HMMs $T_{ij}^x := P(s_j, x|s_i)$ and $\tilde{T}_{mn}^x := Q(\tilde{s}_n, x|\tilde{s}_m)$, where s_i and $\tilde{s}_m$ are the corresponding hidden states. The corresponding pictorial representations are shown in Figure 7.1(a) (see Section 4.3).

FIGURE 7.1: (a) Tensor network representation of the transition matrices of P and Q . (b) Tensor network representation of the co-emission probability.

The co-emission probability is

$$C(P, Q; L) = \sum_{x_{t:t+L}} P(x_{t:t+L}) Q(x_{t:t+L}), \quad (7.8)$$

obtained by contracting the output indices $x_{t:t+L}$ over tensors T and $\tilde{T}$, as shown in the Figure 7.1(b). The tensor structure in the dashed square, which repeatedly appears in the network, has four legs, i.e., is a rank 4 tensor. Combining the left two legs together as a row index, and the right two legs as a column index, this becomes a matrix – the transfer matrix:

$$(\mathbb{E}_{PQ})_{im,jn} := \sum_x P(s_j, x|s_i) Q(\tilde{s}_n, x|\tilde{s}_m). \quad (7.9)$$

The leftmost and rightmost tensors represent the left and right boundaries, respectively. The left boundary $\langle bl|$ is a row vector with elements $v_{ij} = \pi_i \tilde{\pi}_j$. The right boundary $|br\rangle$ is a column vector filled with 1s, such that the hidden states at the last step are equally weighted, i.e.,

$$P(x_{t:t+L}) Q(x_{t:t+L}) = \sum_{s_j, s_n} P(x_{t:t+L}, s_j) Q(x_{t:t+L}, s_n). \quad (7.10)$$

If $\mathbb{E}_{PQ}$ is diagonalisable, it has an eigenvalue decomposition

$$\mathbb{E}_{PQ} = \sum_i \mu_i |r_i\rangle \langle l_i|, \quad (7.11)$$

where μ_i are eigenvalues of $\mathbb{E}_{PQ}$, sorted in order of decreasing magnitude, and $|r_i\rangle$ and $\langle l_i|$ are the associated right- and left-eigenvectors. Consequently, we have

$$\mathbb{E}_{PQ}^L = \mu_1^L (|r_1\rangle \langle l_1|) + \sum_{i \neq 1} \left(\frac{\mu_i}{\mu_1} \right)^L |r_i\rangle \langle l_i|. \quad (7.12)$$

As $\mathbb{E}_{PQ}$ is constructed from probabilities, it is non-negative. Its left- and right-leading eigenvectors are then non-negative according to the Perron-Frobenius theorem [128, 129]. Thus, the left- and right-boundary vectors have non-zero overlap with the associated leading left- and right-eigenvectors of the matrix $\mathbb{E}_{PQ}$, and therefore the co-emission probability has the following scaling

$$C(P, Q; L) = \langle bl | \mathbb{E}_{PQ}^L | br \rangle = \mu_{PQ}^L (\alpha_{PQ} + O((\frac{\mu_2}{\mu_{PQ}})^L)) \sim \alpha_{PQ} \mu_{PQ}^L, \quad (7.13)$$

where $\mu_{PQ} := \mu_1$ and $\alpha_{PQ} = \langle bl | r_1 \rangle \langle l_1 | br \rangle$ is positive. This scaling holds even if $\mathbb{E}_{PQ}$ is not diagonalisable; this can be proved using the Jordan form of the matrix.

Using the same argument, we also have

$$C(P, P; L) \sim \alpha_P \mu_P^L \quad \text{and} \quad C(Q, Q; L) \sim \alpha_Q \mu_Q^L. \quad (7.14)$$

where μ_P is the leading eigenvalue of the transfer matrix $\mathbb{E}_{PP}$ and μ_Q is the leading eigenvalue of the transfer matrix $\mathbb{E}_{QQ}$. Then, we have

$$D_{l_2}(P, Q; L)^2 = 1 - \frac{C(P, Q; L)}{\sqrt{C(P, P; L)C(Q, Q; L)}} \sim 1 - \alpha \left(\frac{\mu_{PQ}}{\sqrt{\mu_P \mu_Q}} \right)^L \quad (7.15)$$

where $\alpha = \alpha_{PQ} / \sqrt{\alpha_P \alpha_Q}$. Thus, the distance has the desired scaling. The continuity of the decaying rate follows from the continuity of the leading eigenvalues, which continuously depends on the coefficients of the characteristic equations. Thus, we proving Lemma 7.4.2, and in turn Theorem 7.4.1. Taking $L \rightarrow \infty$ leads to

$$\lim_{L \rightarrow \infty} \frac{1}{L} \log \left(\sum_{x_{t:t+L}} P(x_{t:t+L}) Q(x_{t:t+L}) \right) = \log \mu_{PQ}, \quad (7.16)$$

and analogously,

$$\lim_{L \rightarrow \infty} \frac{1}{L} \log \left(\sum_{x_{t:t+L}} P(x_{t:t+L}) P(x_{t:t+L}) \right) = \log \mu_P, \quad (7.17)$$

$$\lim_{L \rightarrow \infty} \frac{1}{L} \log \left(\sum_{x_{t:t+L}} Q(x_{t:t+L}) Q(x_{t:t+L}) \right) = \log \mu_Q. \quad (7.18)$$

Therefore,

$$R_C(P, Q) = -\frac{1}{2} \log \frac{\mu_{PQ}}{\sqrt{\mu_P \mu_Q}}. \quad (7.19)$$

□

The computational complexity of calculating these eigenvalues (and hence the CDR) depends only on the number of hidden states $|S|$ in our HMM representations of P and Q , scaling polynomially with both. We need only calculate the leading eigenvalue of a $|S_P||S_Q| \times |S_P||S_Q|$ matrix for μ_{PQ} , and similarly for μ_P and μ_Q . Moreover, as we only need the leading eigenvalues, we can use tools such as the power method [130] rather than full spectral decomposition. Crucially, there is no scaling of complexity with the length of sequences considered (the $L \rightarrow \infty$ limit is implicitly accounted for). And unlike Monte Carlo methods used to estimate e.g. KL-divergences, the result is exact.

Corollary 7.4.2.1 can be applied to any stationary, ergodic processes, since any stationary stochastic process can be described in terms of HMMs. Although the number of their hidden states can be infinite, called infinite HMM, CDR for infinite HMM can be approximated by CDR for approximated finite HMM as the number of hidden states increases.

7.4.2 Fidelity divergence rate

Here, we present another member of the divergence rate family that also satisfies the desired properties of a process distinguishability measure. This divergence rate is called fidelity divergence rate (FDR), as the associated distance is the Bures/Hellinger distance $D_B(P, Q; L) = \sqrt{1 - F(P, Q; L)}$, which is expressed in terms of the fidelity $F(P, Q; L) := \sum_{x_{t:t+L}} \sqrt{P(x_{t:t+L})Q(x_{t:t+L})}$ [119]. Then $S_{D_B}(P, Q; L) = \sqrt{1 - D_B(P, Q; L)^2} = \sqrt{F(P, Q; L)}$, and the FDR is

$$R_F(P, Q) = -\frac{1}{2} \lim_{L \rightarrow \infty} \frac{1}{L} \log(F(P, Q; L)). \quad (7.20)$$

Theorem 7.4.3. The FDR satisfies all the proposed requirements for a measure of process distinguishability.

Similar to the CDR, we will demonstrate that D_B exhibits the required scaling for the FDR to satisfy our requirements, and provide an efficient method for its evaluation given deterministic HMM representations of the processes. A deterministic (or unifilar) HMM is one for which the current hidden state can always be deduced with certainty given the previous state and output (see Section 2.3). This means that each state only has at most one outgoing edge for each symbol. Every stationary stochastic process has a deterministic HMM representation [9].

Consider two stationary stochastic processes, P and Q , with deterministic HMM representations $P(s_j, x|s_i)$ and $Q(\tilde{s}_n, x|\tilde{s}_m)$, respectively, and associated transfer matrix $(\mathbb{E}_{PQ}^F)_{im,jn} = \sum_x \sqrt{P(s_j|x, s_i)Q(\tilde{s}_n, x|\tilde{s}_m)}$. Similarly, we show that D_B has the desired scaling property.

Lemma 7.4.4. $D_B(P, Q; L)$ scales as $1 - \alpha \exp(-\eta L)$.

In proving Lemma 7.4.4, we obtain an efficient method for evaluating FDR between stochastic processes with deterministic HMMs.

Corollary 7.4.4.1. Given deterministic HMM representations $P(s_j, x|s_i)$ and $Q(\tilde{s}_n, x|\tilde{s}_m)$ of processes P and Q , the FDR is given by

$$R_F(P, Q) = -\frac{1}{2} \log_2 \mu_{PQ}, \quad (7.21)$$

where μ_{PQ} is the leading eigenvalue of operator $\mathbb{E}_{PQ}^F = \sum_x \sqrt{P(s_j|x, s_i)Q(\tilde{s}_n, x|\tilde{s}_m)}$.

Proof. The fidelity $\sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i)Q(x_{0:L}|\tilde{s}_m)}$, conditioned on starting in hidden states $(s_i, \tilde{s}_m)$ has a pictorial representation as shown in Figure 7.2(a) (see Theorem 5.1.1). The left boundary represents the i^{th} and m^{th} standard basis vectors $|i\rangle$ and $|m\rangle$ in the corresponding space, while the right boundary, denoted by $|br\rangle$, is the column vector filled with 1s, as shown in Figure 7.2(b).

FIGURE 7.2: Tensor network representation of (a) the fidelity and (b) boundary vectors.

The tensor structure in the dashed square is the transfer matrix $\mathbb{E}_{PQ}^F$, which acts repeatedly on the left boundary $|i, m\rangle$. Similar to the proof for the co-emission, if $\mathbb{E}_{PQ}^F$ is diagonalisable we have the eigenvalue decomposition

$$\mathbb{E}_{PQ}^F = \sum_i \mu_i |r_i\rangle \langle l_i|, \quad (7.22)$$

where μ_i are the eigenvalues of $\mathbb{E}_{PQ}^F$ sorted in order of decreasing magnitude, and $|r_i\rangle$ and $\langle l_i|$ are the associated right- and left-eigenvectors. Consequently,

$$(\mathbb{E}_{PQ}^F)^L = \mu_1^L (|r_1\rangle \langle l_1| + \sum_{i \neq 1} \left(\frac{\mu_i}{\mu_1}\right)^L |r_i\rangle \langle l_i|), \quad (7.23)$$

where $\alpha = \langle bl|r_1\rangle\langle l_1|br\rangle$ is the overlap between the left vector and the leading eigenvector of $\mathbb{E}_{PQ}^F$. Because $\mathbb{E}_{PQ}^F$ is non-negative matrix, its left- and right-leading eigenvectors are non-negative according to the Perron-Frobenius theorem [128, 129].

Since $\langle i, m|$ spans the whole space, there always exists a vector $\langle i, m|$ such that it has non-zero overlap with leading left eigenvector of transfer operator $\mathbb{E}_{PQ}^F$, i.e., $\langle i, m|r_1\rangle > 0$. As with the co-emission, the above scaling still holds when $\mathbb{E}_{PQ}^F$ is not diagonalisable, as can be shown using the Jordan form. Thus we have

$$\lim_{L \rightarrow \infty} \frac{1}{L} \log \sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i) Q(x_{0:L}|\tilde{s}_m)} = \log \mu_{PQ} \quad (7.24)$$

for certain boundary vectors $\langle i, m|$. As μ_{PQ} is the leading eigenvalue, it is easy to check that

$$\lim_{L \rightarrow \infty} \max_{i,m} \frac{1}{L} \log \sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i) Q(x_{0:L}|\tilde{s}_m)} = \log \mu_{PQ}. \quad (7.25)$$

The above quantity is the fidelity conditional on certain past $(s_i, \tilde{s}_m)$. We now bound the non-conditioned fidelity in the following.

$$\begin{aligned} F(P, Q; L) &:= \sum_{x_{0:L}} \sqrt{P(x_{0:L}) Q(x_{0:L})} \\ &= \sum_{x_{0:L}} \sqrt{\sum_i \pi_i P(x_{0:L}|s_i)} \times \sqrt{\sum_m \tilde{\pi}_m Q(x_{0:L}|\tilde{s}_m)}. \end{aligned} \quad (7.26)$$

The inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ implies

$$\begin{aligned} F(P, Q; L) &\leq \sum_{x_{0:L}, i, m} \sqrt{\pi_i \tilde{\pi}_m} \times \sqrt{P(x_{0:L}|s_i) Q(x_{0:L}|\tilde{s}_m)} \\ &\leq \max_{i,m} \left(\sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i) Q(x_{0:L}|\tilde{s}_m)} \right). \end{aligned} \quad (7.27)$$

Therefore, we have

$$\begin{aligned} R_F(P, Q) &= -\frac{1}{2} \lim_{L \rightarrow \infty} \frac{1}{L} \log_2 F(P, Q; L) \\ &\geq -\frac{1}{2} \lim_{L \rightarrow \infty} \frac{1}{L} \log_2 \max_{i,m} \left(\sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i) Q(x_{0:L}|\tilde{s}_m)} \right) \\ &= -\frac{1}{2} \log_2 \mu_{PQ}. \end{aligned} \quad (7.28)$$

On the other hand, using inequality $\sqrt{\sum_i p_i x_i} \geq \sum_i p_i \sqrt{x_i}$, we have

$$\begin{aligned} F(P, Q; L) &\geq \sum_{i,m} \pi_i \tilde{\pi}_m \sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i) Q(x_{0:L}|\tilde{s}_m)} \\ &\geq \max_{i,m} \pi_i \tilde{\pi}_m \sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i) Q(x_{0:L}|\tilde{s}_m)}. \end{aligned} \quad (7.29)$$

Similarly, we also obtain

$$R_F(P, Q) \leq -\frac{1}{2} \log_2 \mu_{PQ}. \quad (7.30)$$

Thus, we see that the fidelity decays exponentially with the length of the sequence, and thus D_B exhibits the required scaling. The continuity of decaying rate also follows from the continuity of leading eigenvalues, thus proving Lemma 7.4.4 and Theorem 7.4.3. $\square$

The fidelity divergence rate can thus be obtained by evaluating the leading eigenvalue of the transfer matrix $\mathbb{E}_{PQ}^F$. The computational complexity of this method depends only polynomially on the number of hidden states in the deterministic HMM representations of each process, and thus can be efficiently computed.

Finally, we provide upper and lower bounds for the FDR that can be calculated even when we do not have deterministic representations of the processes, from their statistics alone.

Theorem 7.4.5. Suppose two stochastic processes P, Q have finite Markov order and the larger one is κ . Then the fidelity divergence rate has the following upper and lower bounds

$$\begin{aligned} R^\downarrow &:= \min_{x_{-\kappa:0}} (-\log F[P(x|x_{-\kappa:0}), Q(x|x_{-\kappa:0})]) \leq 2R_F(P, Q) \\ R^\uparrow &:= \max_{x_{-\kappa:0}} (-\log F[P(x|x_{-\kappa:0}), Q(x|x_{-\kappa:0})]) \geq 2R_F(P, Q). \end{aligned} \quad (7.31)$$

First, having Markov order κ implies that

$$\begin{aligned} P(x_{0:L+\kappa+1}) &= P(x_{L+\kappa}|x_{0:L+\kappa})P(x_{0:L+\kappa}) \\ &= P(x_{L+\kappa}|x_{L:L+\kappa})P(x_{0:L+\kappa}). \end{aligned} \quad (7.32)$$

From this, we find that

$$P(x_{0:L+\kappa+1}) \leq P(x_{0:L+\kappa}) \max_{x_{L:L+\kappa}} P(x_{L+\kappa}|x_{L:L+\kappa}). \quad (7.33)$$

Thus, we have

$$\begin{aligned} F(P, Q; L + \kappa + 1) &= \sum_{x_{0:L+\kappa+1}} \sqrt{P(x_{0:L+\kappa+1})Q(x_{0:L+\kappa+1})} \\ &= \sum_{x_{0:L+\kappa}} \sqrt{P(x_{0:L+\kappa})Q(x_{0:L+\kappa})} \times \sum_{x_{L+\kappa}} \sqrt{P(x_{L+\kappa}|x_{L:L+\kappa})Q(x_{L+\kappa}|x_{L:L+\kappa})} \\ &\leq \sum_{x_{0:L+\kappa}} \sqrt{P(x_{0:L+\kappa})Q(x_{0:L+\kappa})} \times \max_{x_{L:L+\kappa}} \sum_{x_{L+\kappa}} \sqrt{P(x_{L+\kappa}|x_{L:L+\kappa})Q(x_{L+\kappa}|x_{L:L+\kappa})} \\ &= F(P, Q; L + \kappa) \times \max_{x_{L:L+\kappa}} \sum_{x_{L+\kappa}} \sqrt{P(x_{L+\kappa}|x_{L:L+\kappa})Q(x_{L+\kappa}|x_{L:L+\kappa})}. \end{aligned} \quad (7.34)$$

Substituting the above into the definition of FDR leads to

$$\begin{aligned} 2R_F(P, Q) &= \lim_{L \rightarrow \infty} -\frac{1}{L} \log_2(F(P, Q; L)) \\ &\leq \lim_{L \rightarrow \infty} -\frac{1}{L} (\log_2 F(P, Q; \kappa) + (L - \kappa)R^\uparrow) \\ &= R^\uparrow. \end{aligned} \quad (7.35)$$

The lower bound $2R_F(P, Q) \leq R^\downarrow$ can similarly be obtained by replacing maximisations with minimisations and reversing the directions of the inequalities.

7.5 Applications

In this section, we apply CDR to several different examples and demonstrate the capability of CDR for quantifying the similarity between different stochastic processes. The reason for using CDR is that CDR can be efficiently evaluated between processes with any HMMs while FDR requires those HMMs to be deterministic.

7.5.1 Pedagogical example: perturbed coin

As a pedagogical example of how our efficient method for computing the CDR works, we study the distinguishability between two versions of the perturbed coin process [34] (see its HMM in Figure 2.3) with different parameters. This is a Markov process, as output 0 indicates the hidden state is s_0 and output 1 indicates the hidden state s_1 . Consider two perturbed coin processes P and Q with parameters p and q respectively. Then the transfer matrices are

$$\begin{aligned} \mathbb{E}_{PP} &= \begin{bmatrix} (1-p)^2 & 0 & 0 & p^2 \\ (1-p)p & 0 & 0 & (1-p)p \\ (1-p)p & 0 & 0 & (1-p)p \\ p^2 & 0 & 0 & (1-p)^2 \end{bmatrix} & \mathbb{E}_{QQ} &= \begin{bmatrix} (1-q)^2 & 0 & 0 & q^2 \\ (1-q)q & 0 & 0 & (1-q)q \\ (1-q)q & 0 & 0 & (1-q)q \\ q^2 & 0 & 0 & (1-q)^2 \end{bmatrix} \\ \mathbb{E}_{PQ} &= \begin{bmatrix} (1-p)(1-q) & 0 & 0 & pq \\ (1-p)q & 0 & 0 & p(1-q) \\ p(1-q) & 0 & 0 & (1-p)q \\ pq & 0 & 0 & (1-p)(1-q) \end{bmatrix} \end{aligned} \quad (7.36)$$

Evaluating the leading eigenvalues of these matrices, we obtain

$$\mu_P = p^2 + (1-p)^2, \quad \mu_Q = q^2 + (1-q)^2, \quad \mu_{PQ} = pq + (1-p)(1-q). \quad (7.37)$$

Therefore, the CDR is

$$R_C(P, Q) = -\frac{1}{2} \log_2 \frac{pq + (1-p)(1-q)}{\sqrt{(p^2 + (1-p)^2)(q^2 + (1-q)^2)}}. \quad (7.38)$$

Clearly, $R(P, Q) = 0$ iff the two processes are identical, i.e., $p = q$.

7.5.2 An artificial three hidden states process

We present an illustrative example using a highly tunable process that highlights several scenarios in which our measure can be employed, where other previously proposed measures of process distinguishability break down. The most general form of this example process can be represented by a HMM with three hidden states, as illustrated in Figure 7.3. The model has two variable parameters p and

δ ; we use $G(p, \delta)$ to represent the process generated by the model for a particular set of parameters

FIGURE 7.3: The example process we consider can be represented by a three state HMM with two variables p and δ . The edge label $x|P$ between states s_i and s_j signifies that if the model is in state s_i it will transition to s_j whilst emitting symbol x with probability P .

First, we show that the CDR exhibits qualitatively similar behaviour to the (symmetric) KL-divergence (per symbol) in a scenario where the latter can be applied. Note that we must utilise Monte-Carlo methods [66] to estimate the KL-divergence, due to its computationally-intensive nature. Let process $P = G(p, 0)$ for $p \in [0.1, 0.9]$, and similarly process $Q = G(q, 0)$ for $q \in [0.1, 0.9]$. We calculate the CDR using the method described in Corollary 7.4.2.1, while the symmetric KL-divergence per symbol is estimated using the Monte Carlo method for sequences of length $L = 1000$ and a sampling set size $M = 50$.

From Figure 7.4, we see that the CDR and KL-divergence exhibit similar behavior. We emphasise that we are able to efficiently compute the exact CDR, while we are only able to estimate the KL-divergence as it requires exponentially growing resources with sequence length. Moreover, as the processes considered have infinite Markov order (i.e. their behaviour is conditioned on outputs from infinitely far back into the past), no measure based on sequences with finite L can capture the full behaviour of the processes exactly.

Second, we consider a scenario where the KL-divergence cannot be suitably used. Consider the case where again process $P = G(p, 0)$ for $p \in [0.1, 0.9]$, but now $Q = G(q_1, q_2)$ for $q_1 \in [0.1, 0.9]$ and $q_2 \in [0, 1 - q_1]$. When $\delta = 0$ the hidden state s_2 cannot be reached, and so the symbol 2 is never emitted – thus for any $q_2 \neq 0$, P

FIGURE 7.4: Comparison of CDR and KL-divergence per symbol for distinguishing between different parameter values of the process at $\delta = 0$. We see that the qualitative behaviour of the two is very similar. As would be expected, both are zero along the line $p = q$ where the processes are equal, and grow as the difference $|p - q|$ increases.

and Q have different output alphabets and so exhibit infinite KL-divergence (per symbol). Nevertheless, the CDR varies smoothly with the parameters, and we are still able to efficiently calculate it, as shown in Figure 7.5 for the plane defined by $p = q_1$. Furthermore, since the HMM representations of the processes have different numbers of accessible states for $q_2 \neq 0$, a number of other measures based on the model topology cannot be properly applied [77].

7.5.3 Continuous Gaussian processes

We also apply CDR to continuous processes that generate continuous outputs and continuously depend on time. Here, we consider Markov processes described in terms of the transition probability $P(x, t|x', t')$, which is the probability of finding process at state x at time t given previous state x' at t' . For instance, the Ornstein-Uhlenbeck (OU) process is a potential description of the Brownian motion with transition probability given by

$$P(x, \Delta t|x_0, 0) = \mathcal{N}\left(x_0 e^{-\theta \Delta t}, \sqrt{\frac{D}{\theta}(1 - e^{-2\theta \Delta t})}\right), \quad (7.39)$$

where $\mathcal{N}(x_0 e^{-\theta \Delta t}, \sigma)$ represents the Gaussian distribution with mean $x_0 e^{-\theta \Delta t}$ and the variance $\sigma = \sqrt{D/\theta(1 - e^{-2\theta \Delta t})}$. We apply CDR to compare an OU process

FIGURE 7.5: The CDR is able to be calculated to ascertain the distinguishability between processes with different output alphabets and representations with different numbers of states, unlike other measures such as the KL-divergence. As expected, we see that the CDR increases as one process becomes increasingly likely to emit a symbol the other cannot. The white region represents an unphysical parameter regime.

with a totally random Gaussian process with null correlation and the transition probability

$$P(x, \Delta t | x_0, 0) = \mathcal{N}(\mu, \sigma). \quad (7.40)$$

The numerical results is shown in Figure 7.6. In Figure 7.6(A), exact CDR is calculated using Corollary 7.4.2.1. The yellow dotted line and the red dashed line are evaluated using a variation of Eq. (7.6), i.e.,

$$R_C(P, Q) = \lim_{L \rightarrow \infty} g(L+1) - g(L) \quad (7.41)$$

where

$$g(L) = \frac{1}{2} \log \left[\frac{C(P, Q; L)}{\sqrt{C(P, P; L)C(Q, Q; L)}} \right]. \quad (7.42)$$

(A) CDR for different initial points. $\Delta t = 1$. The OU process is realized with two different initial points. One is steady state, the other is random state.

(B) CDR for different Δt .

FIGURE 7.6: The continuous processes are discretized to evaluate the corresponding CDR. The state x with range $(-3, 3)$ and step size $\Delta x = 2$. The parameters of the two processes are set as follows: $\theta = 1, \mu = 0, \sigma = 1$.

$R_C(P, Q)$ is approximated by $g(L + 1) - g(L)$. Figure 7.6(B), the exact CDR for different time step size is plotted. This obeys our intuition that the OU process and totally random Gaussian process is almost identical when Δt is large.

7.6 Discussion

Though our efficient method for computing the CDR relies on having HMM representations of the processes considered, the measure itself does not rely on this.

In lieu of HMM representations, Monte Carlo methods can be employed on the sequence probabilities to calculate the CDR, as with the KL-divergence. We also note that while we have here considered stationary processes, a modified form of the CDR can be applied to non-stationary processes, where instead of taking $L \rightarrow \infty$ we take the longest sequence possible; the measure will then yield the average decay of process similarity per symbol.

A further generalisation that can be considered is the effect of relabelling the alphabet, such that each output is ascribed to a different symbol. In general, the observed statistics after such a relabelling will be different, and as such the CDR between the original and relabelled processes will generally be non-zero. However, from another perspective, one can argue that the two processes still exhibit statistically identical behaviour – with merely a different nomenclature for the events. To remedy this, one can consider an alphabet-symmetrised form of the divergence rate, where it is minimised over all permutations of the alphabet for one of the processes. This would then identify a process and its relabelled version as having zero CDR.

7.7 Conclusion

To summarise, we have discussed a set of conditions we believe a good measure of process distinguishability should satisfy, and proposed a family of divergence rates that satisfy them. We focussed on a particular example of this family, the CDR, and developed an efficient method for its computation. We illustrate the advantages of our measure relative to previously proposed measures by applying it to example scenarios where other measures behave pathologically. Finally, we discussed a number of possible generalisations of the measure.

Chapter 8

Learning quantum models of stochastic processes

Whilst to-date the focus has been on exact quantum models, the limited memory available in practical scenarios might not even allow the realization of exact models. This necessitates approximate models that operate with a given amount of memory. For approximate models a key metric of success is the accuracy. It remains an open question whether there is a quantum advantage for accuracy, when comparing quantum models to classical with the same amount of memory. A key hurdle to answer this question is to construct such quantum models.

FIGURE 8.1: Our learning algorithm takes outputs from a stochastic process and for a given memory dimension d_q generates a quantum model which yields approximately correct statistics for the process's future conditioned on its past.

In this chapter, we present a systematic method—an algorithm—for determining quantum encodings with optimized accuracy, given the data of a classical stochastic process and the desired model memory dimension, as shown in Figure 8.1. We take inspiration from automated classical machine learning [24–28], and how such methods can often be adapted successfully to the quantum case [29, 30], e.g. for quantum state representation [131], quantum state tomography [31], classifying different phases [32, 33], quantum neural nets [132] and other sampling tasks [133–140]. Of particular relevance here is the maximum likelihood paradigm [141] and Adam optimisation [142]. We also use the matrix product state formalism [46, 50, 58, 89, 143] (see Chapter 4), for compact representations of the quantum objects being trained.

Our algorithm solely uses the output data of a stochastic process. The tolerated error for the model can be set to zero, or a finite value. To our knowledge, this is the first method for finding approximate quantum compressions of these quantum models. We demonstrate the power of our algorithm by applying it to datasets sampled from three different stochastic processes, showing numerically that the resulting quantum models successfully compress past information and accurately predict future statistics. We find through our algorithm that these quantum models can have significantly lower dimension than their zero-error counterparts. Moreover, we demonstrate examples where the approximate quantum model has significantly less error compared with any approximate classical model of the same dimension.

Our algorithm adds value beyond a complementary approach of replicating classical models with a smaller quantum memory [40, 144]. Those approaches yield quantum models with the same error as the classical machine without a control over the dimension of the quantum memory, whereas our algorithm demonstrably finds an advantage in accuracy over the classical model. Moreover our algorithm does not necessitate the identification of a classical model of the stochastic process firstly, but rather takes samples of the stochastic process as inputs.

8.1 The quantum model discovery algorithm

Our algorithm takes a sufficiently long sequence $x_{0:L}$ from a stationary stochastic process and our desired memory dimension d_q , and generates a quantum model

(see Chapter 3) that approximately replicates the stochastic process's statistical behavior.

The quantum model functions through repeated application of the single unitary operator U (see Figure 3.2) followed by a measurement on the output register, and so one key aim of the learning algorithm is to determine the appropriate parameter set A that specifies U . To learn these from the sample sequence $x_{0:L}$, we randomly initialize A , and then train by minimizing the cost function:

$$C = -\log_2 \hat{P}(x_{0:L}; A), \quad (8.1)$$

where $\hat{P}(x_{0:L}; A)$ is the probability of outputting the sequence $x_{0:L}$ generated by the quantum model with parameter A .

It has been proven that in the limit of $L \rightarrow \infty$, the cost function of producing $x_{0:L}$ using a hidden Markov model is maximized if and only if the model exactly replicates the target stochastic process's behavior [145, 146]. It hence follows that training U to produce $x_{0:L}$ with maximum likelihood is asymptotically equivalent to training it as a quantum model. (In Eq. (8.1), the negative sign allows for minimization, rather than maximization; the logarithm, as a monotonic function, does not affect the extremal points.)

There is typically a tradeoff between the allowed resources, here d_q , and the achievable accuracy, i.e. the minimum cost function value. Practically, one might vary the d_q of a quantum model to find a sweet-spot where the resulting model does not lose much information.

The parameters are updated using gradient-based methods. The essential idea is to update A by numerical calculation of the element-wise partial derivatives for each parameter to determine ∇C , and then move along the direction of gradient descent:

$$A' = A - \mu \nabla C, \quad (8.2)$$

where μ is a positive value, called the *learning rate*. Taylor expansion of the cost function shows that this will not increase the cost function if the learning rate is chosen properly. There are variants of gradient descent, such as stochastic gradient descent with momentum [147], RMSProp [148] and Adam [142]. The Adam method computes individual adaptive learning rates for different parameters from estimates

of the first and second moments of the gradients [142] and we have empirically found it to work well for our purposes here. We then output A once the gradient descent has effectively terminated (when $|\nabla C|$ is below a chosen threshold value). Since certain initial parameters A can become trapped in local minima, the entire process is repeated several times with different randomly initialized parameters, and we choose the parameters with the minimal cost.

It remains to discuss how the parameter set A specifies the value of the cost function, and ultimately the unitary U . Naïvely, one might specify two real parameters for each element of the complex matrix U – but because the output register is initially always in the state $|0\rangle$, most of these entries do not affect the cost function. It is then sufficient to perform the gradient descent on a subset of these parameters, corresponding to the elements of the Kraus operators $\{A^x := \langle x|U|0\rangle\}$ – a set of $|\mathcal{A}| d_q \times d_q$ complex matrices, where $|\mathcal{A}|$ denotes the alphabet size.

The space of $d_q \times d_q$ unitary matrices is a constrained submanifold of $\mathbb{R}^{2d_q^2}$ (i.e., not every complex matrix is unitary), and on the Kraus operators $\{A^x\}$ this constraint manifests as the completeness relation $\sum_x (A^x)^\dagger A^x = \mathbb{I}$. Rather than performing a constrained descent, we instead descend over an unconstrained space of $|\mathcal{A}|$ general $d_q \times d_q$ complex matrices $\{B^x\}$ and use tensor network techniques [46, 50, 58, 89, 89, 143] to infer from these a set of complete matrices $\{A^x\}$ (see Box. 8.1).

This optimization over an unconstrained space is much easier to implement than over a constrained one. Moreover, the cost function can be calculated directly from these $\{B^x\}$ (see Box. 8.1) and so the actual $\{A^x\}$ are determined only after the appropriate parameter set has been learned. The parameter set A hence corresponds to the $2|\mathcal{A}|d_q^2$ real parameters needed to specify the elements of $\{B^x\}$.

Finally, we can construct the unitary U from $\{A^x\}$ (see Theorem 3.1.2). This can be done systematically [40], and there are many free choices in this step that have no impact on the cost function – for instance, we may choose the U that uses the least experimental resources, such as the smallest number of CNOT gates).

These A^x also inform us how to define our encoding map from pasts onto quantum memory states – as required to initialize the quantum model. The form of each A^x indicates that whenever x is observed in the output register, the memory state updates as $\rho_m \mapsto (A^x)^\dagger \rho_m A^x$. Then, the observation of a past $\overleftarrow{x}$ implies applying a sequence of Kraus operators $A^{\overleftarrow{x}} := A^{x_0} A^{x_{-1}} \dots$ on the initial quantum memory

Any iMPS $\{B^x\}$ can be converted to iMPS $\{A^x\}$ with completeness relation. First, consider the linear map $\varepsilon_l^B(\cdot) := \sum_x (B^x)^\dagger \cdot B^x$. By construction, this is a completely positive map (see Eq. (4.8)), so its leading eigenvalue η is a positive real number, and the associated eigenmatrix V is positive semi-definite. Thus, V admits a decomposition $V = W^\dagger W$ (where W are invertible). We can then define:

$$A^x := W B^x W^{-1} / \sqrt{\eta}. \quad (8.3)$$

Every set $\{A^x\}$ formed in this way satisfies the completeness relation, since (noting $(W^{-1})^\dagger \equiv (W^\dagger)^{-1}$):

$$\begin{aligned} \sum_x (A^x)^\dagger A^x &= \sum_x \frac{1}{\eta} (W^{-1})^\dagger B^{x\dagger} W^\dagger W B^x W^{-1} \\ &= \frac{1}{\eta} (W^\dagger)^{-1} \left(\sum_x B^{x\dagger} W^\dagger W B^x \right) W^{-1} \\ &= (W^\dagger)^{-1} W^\dagger W W^{-1} = \mathbb{I}. \end{aligned} \quad (8.4)$$

We can write the likelihood directly in terms of B^x (see Chapter 4.9)

$$\hat{P}(x_{0:L}; B^x) = \text{Tr}(B^{x_{0:L}} \tilde{\rho} B^{x_{0:L}\dagger} V) / \eta^L, \quad (8.5)$$

where $B^{x_{0:L}} := B^{x_{L-1}} \cdots B^{x_0}$ and $\tilde{\rho} := W^{-1} \rho_0 W^{-1\dagger}$ is the leading eigenmatrix of the completely positive map $\varepsilon^B(\cdot) = \sum_x B^x \cdot B^{x\dagger}$.

Box. 8.1: Recover the completeness relation and evaluate the sequence probability

state $|\sigma_0\rangle$. The past $\overleftarrow{x}$ should therefore be encoded as the quantum state

$$\varepsilon(\overleftarrow{x}) := |\sigma_{\overleftarrow{x}}\rangle := A^{\overleftarrow{x}} |\sigma_0\rangle / \|A^{\overleftarrow{x}} |\sigma_0\rangle\|_2, \quad (8.7)$$

where $\|\cdot\|_2$ denotes the l_2 norm. Here, the initial quantum state $|\sigma_0\rangle$ is chosen to be the leading eigenvector of one of the Kraus operators (e.g. A^0)—see Box. 8.2 for details.

We summarize the learning algorithm Box. 8.3

The encoding map ε should map the past $\overleftarrow{x}$ onto a pure quantum state $|\sigma_{\overleftarrow{x}}\rangle$, which is used as the initial configuration of the quantum model. Assume first that the process has finite Markov order κ such that only the latest κ steps of a given past $\overleftarrow{x}$ affect future statistics, i.e., $\hat{P}(\overrightarrow{X}|\overleftarrow{x}) \equiv \hat{P}(\overrightarrow{X}|x_{-\kappa:0})$. For the purpose of the model, any given past $\overleftarrow{x}$ is equivalent to the concatenation $0^\infty, x_{-\kappa:0}$, where $0^\infty := \dots 0, 0, 0$. That is,

$$\hat{P}(\overrightarrow{x}|\overleftarrow{x}) = \hat{P}(\overrightarrow{x}|0^\infty, x_{-\kappa:0}). \quad (8.6)$$

Recall that an output of 0^∞ corresponds to applying Kraus operator A^0 infinitely many times on the initial quantum state. As a result, any initial quantum state converges to its leading eigenvector $|\sigma_0\rangle$ of A^0 after applying A^0 infinite times. The quantum state transitions to $|\sigma_{x_{-\kappa:0}}\rangle := A^{x_{-\kappa:0}}|\sigma_0\rangle$ after outputting $x_{-\kappa:0}$.

What if the process has infinite Markov order? Then, the perfect initialization of the machine would require an observation of the entire process history. However, one can approximate by taking a long but finite string, which will have exponentially diminishing error with the length of the string [149, 150]. Then, the above argument about the choice of $|\sigma_0\rangle$ holds.

Box. 8.2: Encoding map

8.2 Evaluating the algorithm's performance

Once the quantum models are obtained, we can evaluate its performance compared to the original statistical models by the KL-divergence (see also Eq. (7.1)),

$$D_{KL}(P||\hat{P}) := \sum_x P(x) \log_2 \frac{P(x)}{\hat{P}(x)}, \quad (8.8)$$

a prevalent measure in information theory and machine learning that quantifies the discrepancy that a distribution P has from a training distribution $\hat{P}$. In particular, to quantify the error in prediction, we evaluate

$$D(\overleftarrow{x}) := \lim_{L \rightarrow \infty} \frac{1}{L} D_{KL}(P(X_{0:L}|\overleftarrow{x}) || \hat{P}(X_{0:L} | |\sigma_{\overleftarrow{x}}\rangle)), \quad (8.9)$$

where $P(X_{0:L}|\overleftarrow{x})$ is the true conditional distribution of the stochastic process and $\hat{P}(X_{0:L} | |\sigma_{\overleftarrow{x}}\rangle)$ is the conditional distribution generated by the quantum model. Taking into account all possible pasts, we have the expectation value of the overall

Inputs:

- d_q – the desired model memory dimension d_q .
- $x_{0:L}$ – a length L sequence from the process.

Outputs:

- U – the unitary dynamics of the quantum model.
- ε – the encoding map from observed histories to d_q -dimensional quantum memory states.

Algorithm:

1. To learn the unitary operator U :
 - (a) Randomly initialize parameter set A corresponding up to completeness to the set of Kraus operator matrices $\{A^x := \langle x|U|0\rangle\}$.
 - (b) Evaluate the cost function $C := -\log \hat{P}(x_{0:L}; A)$ and its gradient ∇C .
 - (c) Update the Kraus operators $\{A^x\}$ using gradient-descent based method (such as Adam).
 - (d) Repeat (b)-(c) until the cost function gradient is sufficiently small.
 - (e) Save final A as candidate parameters.
 - (f) Repeat (a)-(e) as desired, to avoid local minima. Choose the final A with the lowest C .
 - (g) For that A , recover the completeness relation for $\{A^x\}$.
 - (h) Construct a unitary operator U from these $\{A^x\}$.
2. To learn the encoding ε :
 - (a) Compute the leading eigenvector $|\sigma_0\rangle$ of one of the Kraus operators (A^0).
 - (b) ε is then defined as

$$\varepsilon(\overleftarrow{x}) := A^{\overleftarrow{x}}|\sigma_0\rangle / \|A^{\overleftarrow{x}}|\sigma_0\rangle\|_2.$$

Box. 8.3: Quantum model discovery algorithm

error of our quantum model:

$$D_e(P, \hat{P}) = \mathbb{E}_{\overleftarrow{x}} D(\overleftarrow{x}). \quad (8.10)$$

KL-divergence $D_{KL}(P||\hat{P})$ has strong operational meanings in different fields, such as quantifying the information gain in machine learning [151, 152] and also the decay rate of error probabilities in hypothesis testing [127, 153]. However, it becomes singular in many cases, including stochastic processes with different alphabet

(see Chapter 7). In such a case, alternative distance measures (e.g. CDR in Chapter 7) could be applied [60]. For the learned quantum models in this chapter, this problem of infinite divergence does not occur. This is because the learned models have some (even if low) probability of producing every possible sequence for a given alphabet and hence cover all possible strings produced by the true process.

Finally, we will compare the error of our quantum models with the error of optimal classical predictive models, including cases where the classical memory dimension is less than the minimal zero-error value d_μ (i.e. with compression error). We identify the optimal approximate classical predictive model of a given d_μ for examples 1 and 2 below (see Box. 8.4). For the highly non-Markovian process of example 3 where an optimal model cannot be identified in a similar manner, we instead construct an approximate ε -machine of the true process by merging its classical causal states. The essential construction principle is to merge those states with little contribution to the process while keeping the remaining parts unchanged.

8.2.1 Hyperparameters

The hyperparameters are parameters that are defined before training. The length of the sample in the first two examples was 1000, while the third example's was 5000. The third example – the renewal process – has more casual states, and training quantum models with higher d_q also require more data. For all of the three examples, the learning rate μ is 0.1 and the optimizer is Adam [142]. We repeat the training 3 times and choose the quantum model of minimal cost.

The evaluation of the error $D_e(P, \hat{P})$ relies on computing the KL-divergence

$$D_{KL}(P(X_{0:L}|\overleftarrow{x})||\hat{P}(X_{0:L}|\epsilon(\overleftarrow{x}))). \quad (8.13)$$

Here, the past $\overleftarrow{x}$ is truncated to be $x_{-5:0}$ and the length of the future L is 1. This amounts to assuming that the Markov order of the process is less than or equal to 5, which is true by construction for our examples.

For a classical model of dimension, d_μ , the encoding map ε of such models must map each $\overleftarrow{x}$ to one of d_μ classical states $\{r_j\}$. Each state is associated with certain future statistics $\hat{P}(\overrightarrow{x}|r_j = \varepsilon(\overleftarrow{x}))$. However, only specific probabilistic assignments maintain deterministic property, which is required for a predictive model. We quantify “optimal” here by minimizing the expected KL-divergence stochastic processes.

$$D_e(P, \hat{P}) = \mathbb{E}_{\overleftarrow{x}} \frac{1}{L} D_{KL}(P(X_{0:L}|\overleftarrow{x}) || \hat{P}(X_{0:L}|\varepsilon(\overleftarrow{x}))). \quad (8.11)$$

Assume two pasts $\overleftarrow{x}$ and $\overleftarrow{x}'$ belongs to an identical causal state s_i of the original stochastic process and they are mapped to different state by the approximate classical model ε i.e., $\varepsilon(\overleftarrow{x}) = r_i$ and $\varepsilon(\overleftarrow{x}') = r_j$. One of the states has lower distance than the other (without loss of generality say r_i):

$$D_{KL}(P(X_{0:L}|s_i) || \hat{P}(X_{0:L}|r_i)) \leq D_{KL}(P(X_{0:L}|s_i) || \hat{P}(X_{0:L}|r_j)). \quad (8.12)$$

We can then construct an alternative model ε' identical to ε every way, except it now maps $\overleftarrow{x}'$ to r_i instead of r_j . As D_e takes account into all pasts, the model ε' obtains a better distance.

Therefore, splitting causal states into many does not give a better model. The optimal model can be obtained by merging the casual states. Note that merging the causal states may cause the model to lose deterministic property (see Section 2.3), which is required by a predictive model.

However, if the process is Markovian, (such as in examples 1 and 2), the deterministic property is preserved for all combinations, because a single output is sufficient to determine the causal state. Thus, the single output determines the memory state r_i (merged casual state). In the case of $d_\mu = 2$, we minimize D_e by going through all possible combinations and evaluating their associated future statistics $\hat{P}(x_0|r_i)$ for the model's next output.

Box. 8.4: Optimal classical approximate models of dimension $d_\mu = 2$

8.2.2 Example 1: the asymmetric process

We now describe the asymmetric process and its relevance, and use our learning algorithm to find a quantum model and evaluate its performance. Here, we consider the simple asymmetric process [47], whose optimal classical predictive model is depicted in Figure 8.2. The dimension of the classical memory $d_\mu = 3$, but the optimal quantum model has $d_q = 2$ [47]. The asymmetric process is the first process where quantum dimensional advantage was ever discovered and part of a

landmark experimental demonstration [86]. The asymmetric process also suggests an interesting perspective on the origin of the arrow of time [47, 154].

We test our algorithm with $d_q = 2$ on a sequence of data generated from the process. To benchmark the performance of the quantum model we firstly evaluate the error $D_e(P, \hat{P})$ of an optimal approximate classical model with $d_\mu = 2$, as shown in Figure 8.3.

We then evaluate the same error quantity $D_e(P, \hat{P})$ for the quantum model, depicted in the same figure. Typically the error of the quantum model is below 0.01 and that of the classical model around 0.05. Thus our algorithm has identified an approximate quantum model which yields a lower error than any approximate classical model of the same dimension. To our knowledge it is the first time such an example has been identified.

FIGURE 8.2: ε -machine for the asymmetric process, parameterized by $p, q \in (0, 1)$. The nodes are the causal states s_0 , s_1 and s_2 . The arrows indicate transitions between states, where each label $x|P$ indicates output x is produced with probability P .

We present the optimal approximate classical model (see Box. 8.4) for asymmetric process with $p = 0.3, q = 0.8$ in Figure 8.4. For the resulting quantum model we found the Kraus operators:

$$\begin{aligned} A^0 &= \begin{bmatrix} 0.676 & 0.317 \\ 0.316 & 0.150 \end{bmatrix}, \\ A^1 &= \begin{bmatrix} -0.264 & 0.534 \\ -0.336 & 0.729 \end{bmatrix}, \\ A^2 &= \begin{bmatrix} -0.241 & -0.095 \\ 0.449 & 0.227 \end{bmatrix}. \end{aligned} \tag{8.14}$$

FIGURE 8.3: Error in the 2-dimensional models for the asymmetric process across a range of process parameters p, q (see Figure 8.2). The left-hand-side shows the lower bound on the optimal approximate classical model's error; the right-hand-side shows the error of the most accurate quantum model found by our algorithm.

FIGURE 8.4: The approximate model of asymmetric example with $p = 0.3, q = 0.8$. s_0 and s_1 of the original asymmetric process are merged into r_0 . The transition probability is minimized by going through the probability vector space. s_2 is mapped to r_1 and its transition probability is unchanged.

The quantum states $|\sigma_{\mathbf{x}}\rangle$ associated with encodings of pasts of length 5 are plotted in Figure 8.5. The more likely quantum states lie in three clusters. This coincides with our expectation from the classical ε -machine, which has 3 causal states— but here, this has been discovered directly from the sample during training. Moreover, the location of these states aligns with our expectation from the theoretical optimal quantum states ascertained in [47]: two clusters (pink and blue) approximately correspond to orthogonal quantum states, while the third (green) lies biased between the two.

FIGURE 8.5: **Example: Asymmetric process – encoded states.** Quantum states produced by the learned $d_q = 2$ encoding map for the asymmetric process with $p = 0.3, q = 0.8$. Each point represents a pure state mapped to by one history (of length 5), and the colour represents the history's associated causal state. For larger points (representing the most probable 99% of histories), the area is proportional to the probability of that history. For smaller points (representing the probable 0.999% of histories), the opacity is proportional to the probability.

8.2.3 Example 2: the quasi-cycle

The second dataset we use to test our algorithm is the quasi-cycle, whose optimal classical predictive model is shown in Figure 8.6. Certain quasicycles are valid thermal operations in that they preserve the Gibb's thermal state, but interestingly without respecting detailed balance [155]. It is known that the classical model requires $d_\mu = 3$ for all choices of process parameters (p, δ) , and likewise the optimal exact quantum model requires $d_q = 3$ for most parameters [48]. However, there is a particular one-dimensional subspace of (p, δ) (defined in Figure 8.6), where we can analytically find an exact quantum model with $d_q = 2$ (this reduction is achieved through applying phase terms into the summands in Eq. (3.16) [48]).

Here, we set $d_q = 2$ and test our algorithm on the dataset. We present the error of our quantum model together with the error of an optimal classical counterpart with $d_\mu = 2$ (see Box. 8.4), as shown in Figure 8.7. Clearly, a broad area of (p, δ) can be accurately approximated by a compressed quantum model, which has much smaller error than any classical predictive model with the same dimension.

FIGURE 8.6: ε -machine for the 3-state quasicycle, parameterized by $p \in (0, 1), \delta \in (0, p]$. The nodes are the causal states s_0, s_1 and s_2 . The arrows indicate transitions between states, where each label $x|P$ indicates output x is produced with probability P .

FIGURE 8.7: Error in the 2-dimensional models for the quasi-cycle, across a range of process parameters p, δ (see Figure 8.6). The left-hand-side shows the error of the optimal approximate classical model; the right-hand-side shows the error of the most accurate quantum model found by our algorithm.

For the quasi-cycle, we show the case where $p = 0.5, \delta = 0.1$. The optimal cmodel is shown in Figure 8.8. The Kraus operators found for $d_q = 2$ are

$$\begin{aligned}
 A^0 &= \begin{bmatrix} 0.021 - 0.077i & -0.104 + 0.084i \\ -0.277 + 0.236i & 0.656 + 0.098i \end{bmatrix}, \\
 A^1 &= \begin{bmatrix} 0.508 + 0.135i & -0.233 + 0.465i \\ 0.279 - 0.171i & 0.073 + 0.277i \end{bmatrix}, \\
 A^2 &= \begin{bmatrix} 0.371 + 0.271i & 0.304 + 0.010i \\ -0.209 + 0.473i & 0.055 + 0.308i \end{bmatrix}.
 \end{aligned} \tag{8.15}$$

The quantum states $|\sigma_{\underline{x}}\rangle$ are plotted in Figure 8.9. As shown in Figure 8.9, most of the states (weighted by probability of occurrence) lie in three distinct clusters – but here, no pair of clusters is orthogonal.

FIGURE 8.8: The approximate model of asymmetric example with $p = 0.5, \delta = 0.1$. s_0 and s_1 of the original a symmetric process are merged into r_0 . The transition probability is minimized by going through the probability vector space. s_2 is mapped to r_1 . Its transition probability is unchanged.

FIGURE 8.9: Quantum states produced by the learned $d_q = 2$ encoding map for a quasicycle with $p = 0.5, \delta = 0.1$. Each point represents the second coordinate of one quantum state as the first coordinate is normalized to be real. The blue circles account for more than 99% probability of the occurrence of the pasts.

8.2.4 Example 3: the uniform renewal process

Renewal processes are a generalization of Poisson processes, describing a series of events spaced by a random time period (see Chapter 6). Discrete-time predictive models approximating renewal processes (such as shown in Figure 8.10) consist of one causal state per fixed length time-step, with some probability of transitioning to the next state emitting the “no event” output (0), and some probability of

transitioning back to the initial state on the emission of the “event” output (1). These probabilities can be calculated from a probability density function Φ and the number N of time steps.

For most Φ , the future behaviour of such a process has a strong dependence of its past: the causal states is determined by counting the number of 0s output since the last 1 and hence $d_\mu = N$ for the exact classical model.

FIGURE 8.10: ε -machine for a N -step discretization of the uniform renewal process (in which events are spaced by a uniform random interval). The nodes are the causal states $s_0, \dots, s_{N-1}$. The arrows indicate transitions between states, where each label $x|P$ indicates output x is produced with probability P .

For a range of classical discretizations $N = 2, 3, 4, 5$, we used our algorithm to train quantum models with $d_q = 2, 3, 4$. For comparison purposes, we also construct classical models with $d_\mu = 2, 3, 4$ (see Box. 8.5). The error of the resulting quantum models and classical models are compared with the optimal classical model of the discretization for a range of N (shown in the right graph of Figure 8.11). For comparison purposes, we calculate the error for the classical approximate models with $d_\mu < N$ (shown in the left graph of Figure 8.11).

Note that the error of quantum model must be non-zero, as the training data is finite. The error of the quantum model is due to the limitation of memory and the limitation of data. The error of the classical models is due to the limitation of the number of the states according to our construction. The error of these quantum models are significantly less than the error of classical models for most of the level of discretization N . Models with larger d_q or d_μ maintain smaller error. The quantum models with larger d_q can have smaller error, but they require a larger dataset to train. In the case of $N = 4$, there is a significant gap between the errors of $d_q = 3, 4$ and the error of $d_q = 2$. Thus the algorithm could here be used to identify a sweet-spot for the choice of d_q : setting $d_q = 3$ gives a good accuracy for a small dimension.

FIGURE 8.11: The error of quantum models for the uniform renewal process, as contrasted with N state optimal classical discretizations.

We now consider d_μ -dimensional approximations of the N -states discretizations of the renewal process. As mentioned above, not every combination of causal states will yield a valid deterministic model. However, there is one way of merging states in this context that does not break determinism. Namely, we may combine the final $N - d_\mu$ states in the chain, as shown in the following figure.

For this combined state, an output of 0 unambiguously means that remain in that state, and 1 means we reset back to the first state s_0 . This amounts to approximating the process as an *eventually Poissonian* model [102] (so called, because once state $N - d_\mu$ is eventually reached, the process assumes the Poissonian property where the probability of an event occurring in a given interval no longer depends on the time already spent waiting in that interval). However, if the chance of outputting 0 is nonzero, then there is a finite probability that this approximation can produce strings with longer times between events (i.e. more 0s between successive 1s) than is ever possible in the reference model.

Box. 8.5: Approximate discrete renewal process.

we show the results for $N = 3, d_q = 2$. The Kraus operators found are

$$\begin{aligned} A^0 &= \begin{bmatrix} 0.064 + 0.170i & 0.043 + 0.246i \\ -0.196 + 0.825i & 0.499 - 0.079i \end{bmatrix}, \\ A^1 &= \begin{bmatrix} 0.490 - 0.053i & 0.304 + 0.753i \\ 0.005 - 0.068i & 0.048 - 0.142i \end{bmatrix} \end{aligned} \quad (8.16)$$

The associated encoded states are plotted in Figure 8.12.

FIGURE 8.12: **Example: Renewal process – encoded states.** Quantum states produced by the learned $d_q = 2$ encoding map for an $N = 3$ discretization of the uniform renewal process, represented on a Bloch sphere. The diagram is interpreted as in Figure 8.5.

8.3 Conclusion

We developed an algorithm to learn quantum models with a limited amount of memory solely from samples of stochastic processes. These quantum models are capable of compressing a given observed past into a pure quantum state and use it to predict the future. Our algorithm goes far beyond previous construction of quantum models: not only does it produce models with lower memory requirements than the classical case, it also provides a method to approximate the original process.

The success of our algorithm is demonstrated by the results on three different example stochastic processes. The resulting quantum models successfully capture the target stochastic processes' behavior. Our results show that many processes

can be approximated to a high accuracy using much less quantum memory than its exact quantum model. Therefore, for these examples, we see that many dimensions within a quantum model contribute little statistically to modelling stochastic processes. Moreover the results show that quantum approximate models can have significantly higher accuracy than classical approximate models of the same dimension. This also demonstrates that employing our learning algorithm is better than learning classical models and then deriving quantum models that replicate them.

These quantum models could also be regarded as Born machines, the output probability of which is determined by Born's rule. Recent studies prove that certain Born machines cannot be simulated efficiently by any classical algorithm up to a multiplicative error in output probabilities [156]. Our work thus complements other works on developing learning algorithm based on Born machine for modeling unknown distributions [138–140]. Many of these works aim to produce quantum generative models (that do not need to be initialized deterministically according to an observed process history), whereas our work focus on learning quantum models for predictive tasks where the future statistics are generated according to the observation of a given past.

Chapter 9

Conclusion and future directions

Here we summarize the main results and provide an outline of future directions.

- In Chapter 5, we discussed the connection between tensor networks and quantum models. We applied a specific class of tensor networks, MPS, to study quantum models. These methods revealed that the entanglement in an MPS state can correspond to the amount of memory in a quantum model. Furthermore, we provided an alternative way to construct such quantum models.

These results open a number of promising avenues for future research. From a foundational perspective, the memory requirements of a quantum model describe a task that involves a clear temporal direction—using past information to generate future predictions. In the classical setting, this resource cost can vary drastically when time is reversed [154], such that costs of prediction and retrodiction differ. Meanwhile, this research indicates this asymmetry is drastically reduced when quantum models are allowed. The connection to entanglement provides a new perspective to understand this divergence [47]. Moreover, noting that entanglement is calculated rather differently from complexity, the proven equivalence between the two opens the door to identifying more efficient numerical methods for computing the latter.

Tensor networks also offer sophisticated techniques to reduce the dimension of quantum states – and correspondingly, representations of stochastic processes – at the cost of introducing small inaccuracies. Such methods can now be applied to quantum modeling, offering approximate models with drastically reduced dimensional requirements. Meanwhile, there has been growing

interest in generalizing computational mechanics to higher dimensions [157]. It would indeed be interesting to relate such works to higher dimensional variants of tensor networks, such as MERA [91] and PEPS [158].

- In Chapter 6, we highlighted a family of stochastic processes that can be modeled efficiently using quantum models. That is, quantum models can exactly replicate the future statistics of this process family using only one qubit for its memory, while any classical model requires an unbounded number of bits.

Our findings motivate future work to investigate the typicality of processes that admit such single-shot advantages as well as approximate compression techniques for those that do not. Indeed, the behaviour of a broader range of renewal processes can be captured by generalising Eq. (6.9) to have different amplitudes and include additional states. A clear direction to extend these techniques for scaling advantages is to other continuous parameters such as spatial co-ordinates [104], or more abstract settings such as continuous belief spaces [159].

- In Chapter 7, we introduced measures quantifying the distance between stochastic processes. We proposed several criteria that a well-behaved measure should satisfy, and then presented a family of measures that satisfy them. We focussed on two members of this family, using MPS methods to provide efficient algorithms for their evaluation. These measures were then applied to certain examples to demonstrate their suitability for quantifying the similarity of stochastic processes.

Our measures have potential applications in a broad range of areas, especially those involving stochastic processes such as HMMs [77], computational mechanics [8, 9] and quantum stochastic modelling [34, 35, 40, 42, 48, 105, 160]. Other areas of application include measuring the accuracy of machine learning models and evaluating the performance of time-series inference protocols.

- In Chapter 8, we developed an algorithm that learns quantum models from output sequences. The algorithm utilizes machine learning methods to tune the model parameters in order to minimize a given cost function. Numerical results demonstrated that our algorithm can produce quantum models with high accuracy at a fixed amount of memory, and outperform the optimal classical predictive models with the same amount of memory.

The algorithm enables a systematic means of finding quantum models for real-world stochastic processes. It can be implemented either with classical supercomputers, or a small number of qubits in current and near-term noisy quantum computers. Future directions also include investigating how to mitigate noise when implementing quantum models experimentally. Moreover, it strongly hints that it is possible to minimize the number of pure quantum states used in the encoding by clustering them.

In summary, the thesis pioneered a number of advancements in modelling stochastic processes with quantum resources. We developed a connection between matrix product states and quantum stochastic modelling. We also developed these methods to find better quantum models and evaluate the amount of quantum memory needed to model a given process. Moreover, these methods inspired us to generalize prior quantum models by introducing complex amplitudes. This led to the discovery of a family of stochastic processes for which there are unbounded quantum memory advantages over the optimal classical models. We developed a family of divergence measures to quantify the distance between stochastic processes, which allows us to benchmark the performance of approximate classical and quantum models. Our MPS representation of the stochastic processes provides an efficient way to evaluate these measures. Finally, we developed a systematic algorithm to learn quantum models with a fixed amount of memory solely from the output of a stochastic process. We show that our learning algorithm provides a direct means of finding quantum models – circumventing the need to find an interim classical model to quantize.

As our final thought, we look to the connection between tensor networks and stochastic processes, which has drawn sizable attention within and beyond this thesis. Significant effort has been devoted towards bringing tensor networks into constructing better generative models in the field of machine learning [135, 136, 161, 162]. Although we focused on MPS in this thesis, there are other types of tensor networks for systems with more than one dimension, such as PEPS or MERA. Akin to the connection between the MPS and stochastic processes, we believe other types of tensor networks can provide new perspectives into the modeling of general Bayesian networks using quantum resources. Like the quantum models of stochastic processes, we expect quantum models of general Bayesian networks to maintain memory advantages. From another perspective, machine learning methods are

also showing notable success in the quantum domain. Tensor networks serve as a powerful tool to parameterize quantum circuits for optimization during learning procedures. With this in mind, it is reasonable to look forward to even more fruitful discoveries of the relationships between tensor networks, machine learning, and quantum information.

Bibliography

- [1] David P Feldman and James P Crutchfield. Measures of statistical complexity: Why? *Physics Letters A*, 238(4-5):244–252, 1998. ISSN 03759601. doi: 10.1016/S0375-9601(97)00855-4. URL <http://www.sciencedirect.com/science/article/pii/S0375960197008554>. 19, 26
- [2] Robert Haslinger, Kristina Lisa Klinkner, and Cosma Rohilla Shalizi. The Computational Structure of Spike Trains. *Neural Computation*, 22(1): 121–157, jan 2010. ISSN 0899-7667. doi: 10.1162/neco.2009.12-07-678. URL <http://arxiv.org/abs/1001.0036><http://dx.doi.org/10.1162/neco.2009.12-07-678><http://www.mitpressjournals.org/doi/10.1162/neco.2009.12-07-678>. 19
- [3] Jae Suk Yang, Wooseop Kwak, Taisei Kaizoji, and In Mook Kim. Increasing market efficiency in the stock markets. *European Physical Journal B*, 61(2):241–246, 2008. ISSN 14346028. doi: 10.1140/epjb/e2008-00050-0. URL <http://arxiv.org/abs/physics/0701179><http://dx.doi.org/10.1140/epjb/e2008-00050-0>. 19
- [4] Joongwoo Brian Park, Jeong Won Lee, Jae Suk Yang, Hang Hyun Jo, and Hie Tae Moon. Complexity analysis of the stock market. *Physica A: Statistical Mechanics and its Applications*, 379(1):179–187, 2007. ISSN 03784371. doi: 10.1016/j.physa.2006.12.042. 19
- [5] Alexander B. Boyd, Dibyendu Mandal, and James P. Crutchfield. Identifying Functional Thermodynamics in Autonomous Maxwellian Ratchets. *New Journal of Physics*, 18(2):023049, jul 2015. ISSN 13672630. doi: 10.1088/1367-2630/18/2/023049. URL <http://stacks.iop.org/1367-2630/18/i=2/a=023049?key=crossref.a1366b790634421f37e1d1fa8dbb89bb><http://arxiv.org/abs/1507.01537><http://dx.doi.org/10.1088/1367-2630/18/2/023049>. 19
- [6] Andrew J.P. Garner, Jayne Thompson, Vlatko Vedral, and Mile Gu. Thermodynamics of complexity and pattern manipulation. *Physical Review E*, 95(4):042140, 2017. ISSN 24700053. doi: 10.1103/PhysRevE.95.042140. 19, 26, 29
- [7] Peter Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25(9):907–938, 1986. ISSN 00207748. doi: 10.1007/BF00668821. 19, 20

- [8] James P. Crutchfield and Karl Young. Inferring statistical complexity. *Physical Review Letters*, 63(2):105–108, 1989. ISSN 00319007. doi: 10.1103/PhysRevLett.63.105. [20](#), [27](#), [28](#), [110](#)
- [9] Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3-4):817–879, 2001. ISSN 00224715. doi: 10.1023/A:1010388907793. [27](#), [28](#), [76](#), [79](#), [110](#)
- [10] James P. Crutchfield. Between order and chaos. *Nature Physics*, 8(1):17–24, 2012. ISSN 17452473. doi: 10.1038/nphys2190. URL <http://www.nature.com/doifinder/10.1038/nphys2190>. [19](#), [20](#)
- [11] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam’s razor. *Information processing letters*, 24(6):377–380, 1987. [19](#)
- [12] Carl Edward Rasmussen and Zoubin Ghahramani. Occam’s razor. In *Advances in neural information processing systems*, pages 294–300, 2001. [19](#)
- [13] Michael A. Nielsen and Isaac Chuang. Quantum computation and quantum information. *American Journal of Physics*, 70(5):558–559, 2002. doi: 10.1119/1.1463744. URL <https://doi.org/10.1119/1.1463744>. [20](#), [32](#)
- [14] John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018. [20](#), [21](#), [71](#)
- [15] P. W. Shor. Algorithms for quantum computation: discrete logarithms and factoring. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pages 124–134, Nov 1994. doi: 10.1109/SFCS.1994.365700. [21](#)
- [16] Scott Aaronson and Alex Arkhipov. The computational complexity of linear optics. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, page 333. ACM, 2011. ISBN 9781450306911. doi: 10.1145/1993636.1993682. [46](#)
- [17] Aram W Harrow and Ashley Montanaro. Quantum computational supremacy. *Nature*, 549(7671):203–209, 2017.
- [18] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019. [21](#)
- [19] Seth Lloyd. Universal quantum simulators. *Science*, pages 1073–1078, 1996. [21](#)

- [20] Yudong Cao, Jonathan Romero, Jonathan P Olson, Matthias Degroote, Peter D Johnson, Mária Kieferová, Ian D Kivlichan, Tim Menke, Borja Peropadre, Nicolas PD Sawaya, et al. Quantum chemistry in the age of quantum computing. *arXiv preprint arXiv:1812.09976*, 2018.
- [21] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017.
- [22] Iulia M Georgescu, Sahel Ashhab, and Franco Nori. Quantum simulation. *Reviews of Modern Physics*, 86(1):153, 2014. [21](#)
- [23] Sheng-Kai Liao, Wen-Qi Cai, Wei-Yue Liu, Liang Zhang, Yang Li, Ji-Gang Ren, Juan Yin, Qi Shen, Yuan Cao, Zheng-Ping Li, et al. Satellite-to-ground quantum key distribution. *Nature*, 549(7670):43–47, 2017. [21](#)
- [24] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006. [21](#), [92](#)
- [25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.
- [26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. [29](#)
- [27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [28] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013. [21](#), [92](#)
- [29] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195, 2017. [21](#), [92](#)
- [30] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018. [92](#)
- [31] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 14(5):447, 2018. [92](#)
- [32] Juan Carrasquilla and Roger G Melko. Machine learning phases of matter. *Nature Physics*, 13(5):431, 2017. [92](#)

- [33] Lei Wang. Discovering phase transitions with unsupervised learning. *Physical Review B*, 94(19):195105, 2016. [21](#), [92](#)
- [34] Mile Gu, Karoline Wiesner, Elisabeth Rieper, and Vlatko Vedral. Quantum mechanics can reduce the complexity of classical models. *Nature Communications*, 3:762, 2012. ISSN 20411723. doi: 10.1038/ncomms1761. URL <http://www.nature.com/doifinder/10.1038/ncomms1761>. [21](#), [29](#), [31](#), [36](#), [61](#), [84](#), [110](#)
- [35] John R. Mahoney, Cina Aghamohammadi, and James P. Crutchfield. Occam’s Quantum Strop: Synchronizing and Compressing Classical Cryptic Processes via a Quantum Channel. *Scientific Reports*, 6(1):20495, 2016. ISSN 20452322. doi: 10.1038/srep20495. URL <http://www.nature.com/articles/srep20495>. [31](#), [36](#), [45](#), [55](#), [61](#), [110](#)
- [36] Paul M. Riechers, John R. Mahoney, Cina Aghamohammadi, and James P. Crutchfield. Minimized state complexity of quantum-encoded cryptic processes. *Physical Review A*, 93(5):052317, 2016. ISSN 24699934. doi: 10.1103/PhysRevA.93.052317. URL <http://arxiv.org/abs/1510.08186>. [45](#)
- [37] Whei Yeap Suen, Jayne Thompson, Andrew J. P. Garner, Vlatko Vedral, and Mile Gu. The classical-quantum divergence of complexity in modelling spin chains. *Quantum*, 1:25, 2017. ISSN 2521-327X. doi: 10.22331/q-2017-08-11-25. URL <http://arxiv.org/abs/1511.05738>{%}0A<http://dx.doi.org/10.22331/q-2017-08-11-25>.
- [38] Cina Aghamohammadi, John R. Mahoney, and James P. Crutchfield. The ambiguity of simplicity in quantum and classical simulation. *Physics Letters, Section A: General, Atomic and Solid State Physics*, 381(14):1223–1227, 2017. ISSN 03759601. doi: 10.1016/j.physleta.2016.12.036. URL <http://dx.doi.org/10.1016/j.physleta.2016.12.036>.
- [39] Cina Aghamohammadi, John R. Mahoney, and James P. Crutchfield. Extreme Quantum Advantage when Simulating Classical Systems with Long-Range Interaction. *Scientific Reports*, 7(1):6735, sep 2017. ISSN 20452322. URL <http://arxiv.org/abs/1609.03650>. [61](#)
- [40] Felix C Binder, Jayne Thompson, and Mile Gu. Practical Unitary Simulator for Non-Markovian Complex Processes. *Physical Review Letters*, 120(24):240502, 2018. ISSN 1079-7114. doi: 10.1103/PhysRevLett.120.240502. URL <https://doi.org/10.1103/PhysRevLett.120.240502>. [21](#), [31](#), [35](#), [36](#), [45](#), [52](#), [92](#), [94](#), [110](#)
- [41] C. Aghamohammadi, S. P. Loomis, J. R. Mahoney, and J. P. Crutchfield. Extreme Quantum Advantage for Rare-Event Sampling. *Physical Review X*, 8(1):11025, 2017. ISSN 2160-3308. doi: 10.1103/PhysRevX.8.011025. URL <http://arxiv.org/abs/1707.09553>. [36](#)

- [42] Thomas J. Elliott and Mile Gu. Superior memory efficiency of quantum devices for the simulation of continuous-time stochastic processes. *npj Quantum Information*, 4(1):18, 2018. ISSN 2056-6387. doi: 10.1038/s41534-018-0064-4. URL <http://www.nature.com/articles/s41534-018-0064-4>. 36, 57, 58, 61, 67, 110
- [43] Matthew S. Palsson, Mile Gu, Joseph Ho, Howard M. Wiseman, and Geoff J. Pryde. Experimentally modeling stochastic processes with less memory by the use of a quantum processor. *Science Advances*, 3(2):e1601302, 2017. ISSN 23752548. URL <http://advances.sciencemag.org/content/3/2/e1601302>.
- [44] Andrew J P Garner, Qing Liu, Jayne Thompson, Vlatko Vedral, and Mile Gu. Unbounded memory advantage in stochastic simulation using quantum mechanics. *New Journal of Physics*, 19(10):103009, sep 2016. ISSN 1367-2630. URL <http://stacks.iop.org/1367-2630/19/i=10/a=103009?key=crossref.cb08b7a84aded24b5c21f1796dafd36e><http://arxiv.org/abs/1609.04408><http://dx.doi.org/10.1088/1367-2630/aa82df>.
- [45] Farzad Ghafari Jouneghani, Mile Gu, Joseph Ho, Jayne Thompson, Whei Yeap Suen, Howard M. Wiseman, and Geoff J. Pryde. Observing the ambiguity of simplicity via quantum simulations of an Ising spin chain. *arXiv preprint*, 2017. URL <http://arxiv.org/abs/1711.03661>.
- [46] Chengran Yang, Felix C. Binder, Varun Narasimhachar, and Mile Gu. Matrix product states for quantum stochastic modeling. *Phys. Rev. Lett.*, 121:260602, Dec 2018. doi: 10.1103/PhysRevLett.121.260602. URL <https://link.aps.org/doi/10.1103/PhysRevLett.121.260602>. 21, 22, 31, 36, 37, 40, 92, 94
- [47] Jayne Thompson, Andrew J. P. Garner, John R. Mahoney, James P. Crutchfield, Vlatko Vedral, and Mile Gu. Causal Asymmetry in a Quantum World. *Physical Review X*, 8(3):31013, 2018. ISSN 2160-3308. doi: 10.1103/PhysRevX.8.031013. URL <http://arxiv.org/abs/1712.02368>. 36, 58, 61, 99, 100, 101, 109
- [48] Qing Liu, Thomas Elliott, Felix Binder, Carlo Di Franco, Mile Gu, et al. Optimal stochastic modelling with unitary quantum dynamics. *arXiv preprint arXiv:1810.09668*, 2018. 21, 36, 102, 110
- [49] David Perez-Garcia, Frank Verstraete, Michael M Wolf, and J Ignacio Cirac. Matrix product state representations. *arXiv preprint quant-ph/0608197*, 2006. 21, 37
- [50] R. Orús and G. Vidal. Infinite time-evolving block decimation algorithm beyond unitary evolution. *Physical Review B - Condensed Matter and Materials Physics*, 78(15):155117, 2008. ISSN 10980121. doi: 10.1103/PhysRevB.78.155117. URL <http://scholar.google.com/scholar?hl=en{%}7B{%}&{%}7DbtnG=Search{%}7B{%}&{%}7Dq=intitle:>

- The+iTEBD+algorithm+beyond+unitary+evolution{%}7B{%}{%}7D1. 21, 43, 92, 94
- [51] Román Orús. A practical introduction to tensor networks: Matrix product states and projected entangled pair states. *Annals of Physics*, 349:117–158, 2014. ISSN 1096035X. doi: 10.1016/j.aop.2014.06.013.
 - [52] Jacob C Bridgeman and Christopher T Chubb. Hand-waving and Interpretive Dance: An Introductory Course on Tensor Networks. *arXiv:1603.03039 [cond-mat, physics:hep-th, physics:quant-ph]*, 2016. URL <http://arxiv.org/abs/1603.03039>. 21
 - [53] Ian Affleck, Tom Kennedy, Elliott H. Lieb, and Hal Tasaki. Rigorous results on valence-bond ground states in antiferromagnets. *Physical Review Letters*, 59(7):799–802, 1987. ISSN 00319007. doi: 10.1103/PhysRevLett.59.799. 21, 37
 - [54] A. Klümper, A. Schadschneider, and J. Zittartz. Matrix-product-groundstates for one-dimensional spin-1 quantum antiferromagnets. *Epl*, 24(4):293–297, jul 1993. ISSN 12864854. doi: 10.1209/0295-5075/24/4/010. URL <http://dx.doi.org/10.1209/0295-5075/24/4/010>.
 - [55] Tomaz Prosen. Open XXZ spin chain: Nonequilibrium steady state and a strict bound on ballistic transport. *Physical Review Letters*, 106(21):217206, 2011. ISSN 00319007. doi: 10.1103/PhysRevLett.106.217206. 21, 37
 - [56] Steven R. White. Density matrix formulation for quantum renormalization groups. *Physical Review Letters*, 69(19):2863–2866, 1992. ISSN 00319007. doi: 10.1103/PhysRevLett.69.2863. URL <http://link.aps.org/doi/10.1103/PhysRevLett.69.2863>. 21
 - [57] F. Verstraete, D. Porras, and J. I. Cirac. Density matrix renormalization group and periodic boundary conditions: A quantum information perspective. *Physical Review Letters*, 93(22):26–29, 2004. ISSN 00319007. doi: 10.1103/PhysRevLett.93.227205. URL <http://link.aps.org/doi/10.1103/PhysRevLett.93.227205>.
 - [58] G. Vidal. Entanglement renormalization. *Physical Review Letters*, 99(22):220405, 2007. ISSN 00319007. doi: 10.1103/PhysRevLett.99.220405. 21, 37, 55, 92, 94
 - [59] Thomas J Elliott, Chengran Yang, Felix C Binder, Andrew JP Garner, Jayne Thompson, and Mile Gu. Extreme dimensional compression with quantum modelling. *arXiv preprint arXiv:1909.02817*, 2019. 22, 36
 - [60] Chengran Yang, Felix C Binder, Mile Gu, and Thomas J Elliott. Measures of distinguishability between stochastic processes. *arXiv preprint arXiv:1909.08366*, 2019. 22, 40, 98

- [61] William Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 2008. 25
- [62] Radu Balescu. Equilibrium and nonequilibrium statistical mechanics. *STIA*, 76:32809, 1975. 25
- [63] Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011. 25
- [64] Claude E Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948. 25
- [65] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 26, 71
- [66] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE, 2007. 26, 85
- [67] R. W. Clarke, M. P. Freeman, and N. W. Watkins. Application of computational mechanics to the analysis of natural data: An example in geomagnetism. *Physical Review E*, 67(1):016203, 2003. ISSN 1063-651X. doi: 10.1103/PhysRevE.67.016203. URL <http://link.aps.org/doi/10.1103/PhysRevE.67.016203>. 26
- [68] D. P. Varn, G. S. Canright, and J. P. Crutchfield. Discovering planar disorder in close-packed structures from x-ray diffraction: Beyond the fault model. *Physical Review B - Condensed Matter and Materials Physics*, 66(17):174110, 2002. ISSN 1550235X. doi: 10.1103/PhysRevB.66.174110.
- [69] D. P. Varn, G. S. Canright, and J. P. Crutchfield. Machine spectral reconstruction theory: A direct method for inferring planar disorder and structure from X-ray diffraction studies. *Acta Crystallographica Section A: Foundations of Crystallography*, 69(2):197–206, 2013. ISSN 01087673. doi: 10.1107/S0108767312046582.
- [70] D. P. Varn and J. P. Crutchfield. Chaotic crystallography: How the physics of information reveals structural order in materials. *Current Opinion in Chemical Engineering*, 7:47–56, 2015. ISSN 22113398. doi: 10.1016/j.coche.2014.11.002. URL <http://dx.doi.org/10.1016/j.coche.2014.11.002>. 26
- [71] James P. Crutchfield. The calculi of emergence: computation, dynamics and induction. *Physica D: Nonlinear Phenomena*, 75(1-3):11–54, 1994. ISSN 01672789. doi: 10.1016/0167-2789(94)90273-9. URL <http://linkinghub.elsevier.com/retrieve/pii/0167278994902739>. 27
- [72] Nicolás Perry and P. M. Binder. Finite statistical complexity for sofic systems. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 60(1):459–463, 1999. ISSN 1063651X. doi:

- 10.1103/PhysRevE.60.459. URL <http://www.ncbi.nlm.nih.gov/pubmed/11969783>. 28
- [73] James P Crutchfield. Observing Complexity and The Complexity of Observation. *Inside versus Outside*, 63(April):235–272, 1994. doi: 10.1007/978-3-642-48647-0_14. URL http://link.springer.com/chapter/10.1007/978-3-642-48647-0_{_}14. 28
- [74] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014. 29
- [75] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012. 29
- [76] Dan Klein and Christopher D Manning. Fast exact inference with a factored model for natural language parsing. In *Advances in neural information processing systems*, pages 3–10, 2003. 29
- [77] Stephen E Levinson, Lawrence R Rabiner, and Man Mohan Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *Bell System Technical Journal*, 62(4):1035–1074, 1983. 29, 86, 110
- [78] John Bechhoefer. Hidden markov models for stochastic thermodynamics. *New Journal of Physics*, 17(7):075003, 2015. 29
- [79] P Baldi, Y Chauvin, T Hunkapiller, and M A McClure. Hidden markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences*, 91(3):1059–1063, 1994. ISSN 0027-8424. doi: 10.1073/pnas.91.3.1059. 29
- [80] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006. 29
- [81] Jyri Kivinen and Christopher Williams. Multiple texture boltzmann machines. In *Artificial Intelligence and Statistics*, pages 638–646, 2012. 29
- [82] Ilya Sutskever, Geoffrey E Hinton, and Graham W Taylor. The recurrent temporal restricted boltzmann machine. In *Advances in neural information processing systems*, pages 1601–1608, 2009. 29
- [83] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 29

- [84] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. [29](#)
- [85] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002. [31](#)
- [86] Farzad Ghafari, Nora Tischler, Jayne Thompson, Mile Gu, Lynden K. Shalm, Varun B. Verma, Sae Woo Nam, Raj B. Patel, Howard M. Wiseman, and Geoff J. Pryde. Dimensional quantum memory advantage in the simulation of stochastic processes. *Phys. Rev. X*, 9:041013, Oct 2019. doi: 10.1103/PhysRevX.9.041013. URL <https://link.aps.org/doi/10.1103/PhysRevX.9.041013>. [36](#), [100](#)
- [87] Farzad Ghafari, Nora Tischler, Carlo Di Franco, Jayne Thompson, Mile Gu, and Geoff J Pryde. Interfering trajectories in experimental quantum-enhanced stochastic simulation. *Nature communications*, 10(1):1–8, 2019.
- [88] Matthew S Palsson, Mile Gu, Joseph Ho, Howard M Wiseman, and Geoff J Pryde. Experimentally modeling stochastic processes with less memory by the use of a quantum processor. *Science Advances*, 3(2):e1601302, 2017. [36](#)
- [89] C. Schön, E. Solano, F. Verstraete, J. I. Cirac, and M. M. Wolf. Sequential generation of entangled multiqubit states. *Physical Review Letters*, 95(11):110503, 2005. ISSN 00319007. doi: 10.1103/PhysRevLett.95.110503. [37](#), [43](#), [55](#), [92](#), [94](#)
- [90] Norbert Schuch, Ignacio Cirac, and David Pérez-García. Peps as ground states: Degeneracy and topology. *Annals of Physics*, 325(10):2153–2192, 2010. [37](#)
- [91] G. Vidal. A class of quantum many-body states that can be efficiently simulated. 110501(September):110501, 2008. ISSN 0031-9007. doi: 10.1103/PhysRevLett.101.110501. URL <http://arxiv.org/abs/quant-ph/0610099>{%}0Ahttp://dx.doi.org/10.1103/PhysRevLett.101.110501. [37](#), [110](#)
- [92] F. Verstraete, V. Murg, and J. I. Cirac. Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems, 2008. ISSN 00018732. [37](#)
- [93] Frank Verstraete, Michael M Wolf, David Perez-Garcia, and J Ignacio Cirac. Criticality, the area law, and the computational power of projected entangled pair states. *Physical review letters*, 96(22):220601, 2006. [39](#)
- [94] T. H. Johnson, T. J. Elliott, S. R. Clark, and D. Jaksch. Capturing exponential variance using polynomial resources: Applying tensor networks to nonequilibrium stochastic processes. *Physical Review Letters*, 114(9):090602, 2015. ISSN 10797114. doi: 10.1103/PhysRevLett.114.090602. [40](#)

- [95] Dorit Aharonov and Amnon Ta-Shma. Adiabatic Quantum State Generation and Statistical Zero Knowledge. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, page 20, 2003. ISBN 1581136749. doi: 10.1145/780542.780546. URL <http://arxiv.org/abs/quant-ph/0301023>. 46
- [96] Srinivasan Arunachalam and Ronald de Wolf. Optimal Quantum Sample Complexity of Learning Algorithms. *arxiv preprint*, 2016. URL <http://arxiv.org/abs/1607.00932>. 46
- [97] Justin B. Spring, Benjamin J. Metcalf, Peter C. Humphreys, W. Steven Kolthammer, Xian Min Jin, Marco Barbieri, Animesh Datta, Nicholas Thomas-Peter, Nathan K. Langford, Dmytro Kundys, James C. Gates, Brian J. Smith, Peter G.R. Smith, and Ian A. Walmsley. Boson sampling on a photonic chip. *Science*, 339(6121):798–801, 2013. ISSN 10959203. doi: 10.1126/science.1231692. URL <http://www.sciencemag.org/content/339/6121/798.abstract>. 46
- [98] Andrea Crespi, Roberto Osellame, Roberta Ramponi, Daniel J. Brod, Ernesto F. Galvão, Nicolò Spagnolo, Chiara Vitelli, Enrico Maiorino, Paolo Mataloni, and Fabio Sciarrino. Integrated multimode interferometers with arbitrary designs for photonic boson sampling. *Nature Photonics*, 7(7):545–549, 2013. ISSN 17494885. doi: 10.1038/nphoton.2013.112.
- [99] A. P. Lund, A. Laing, S. Rahimi-Keshari, T. Rudolph, J. L. O’Brien, and T. C. Ralph. Boson sampling from a Gaussian state. *Physical Review Letters*, 113(10):100502, 2014. ISSN 10797114. doi: 10.1103/PhysRevLett.113.100502.
- [100] Scott Aaronson and Lijie Chen. Complexity-Theoretic Foundations of Quantum Supremacy Experiments. *arXiv preprint*, 2016. ISSN 18688969. URL <http://arxiv.org/abs/1612.05903>. 46
- [101] Norbert Schuch, David Pérez-García, and Ignacio Cirac. Classifying quantum phases using matrix product states and projected entangled pair states. *Physical Review B - Condensed Matter and Materials Physics*, 84(16):165139, 2011. ISSN 10980121. doi: 10.1103/PhysRevB.84.165139. 51
- [102] Sarah Marzen and James P Crutchfield. Informational and Causal Architecture of Continuous-time Renewal Processes. *Journal of Statistical Physics*, 168(1):109–127, jul 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1793-z. URL <https://doi.org/10.1007/s10955-017-1793-z>. 57, 106
- [103] Matthew S Palsson, Mile Gu, Joseph Ho, Howard M Wiseman, and Geoff J Pryde. Experimentally modeling stochastic processes with less memory by the use of a quantum processor. *Science Advances*, 3(2):e1601302, 2017. doi: 10.1126/sciadv.1601302. URL <http://advances.sciencemag.org/content/3/2/e1601302>. 61

- [104] Andrew J P Garner, Qing Liu, Jayne Thompson, Vlatko Vedral, et al. Provably unbounded memory advantage in stochastic simulation using quantum mechanics. *New Journal of Physics*, 19(10):103009, 2017. [110](#)
- [105] Thomas J Elliott, Andrew J P Garner, and Mile Gu. Memory-efficient tracking of complex temporal and symbolic dynamics with quantum simulators. *New Journal of Physics*, 21:013021, 2019. [61](#), [110](#)
- [106] Alex Monràs and Andreas Winter. Quantum learning of classical stochastic processes: The completely positive realization problem. *Journal of Mathematical Physics*, 57(1):015219, 2016. ISSN 00222488. doi: 10.1063/1.4936935. [61](#)
- [107] Adán Cabello, Mile Gu, Otfried Gühne, Jan-Åke Larsson, and Karoline Wiesner. Thermodynamical cost of some interpretations of quantum theory. *Physical Review A*, 94(5):052127, 2016.
- [108] Adán Cabello, Mile Gu, Otfried Gühne, and Zhen-Peng Xu. Optimal classical simulation of state-independent quantum contextuality. *Physical Review Letters*, 120(13):130401, 2018.
- [109] Prahlad Warszawski and Howard M Wiseman. Open quantum systems are harder to track than open classical systems. *arXiv:1905.10935*, 2019. [61](#)
- [110] Walter L Smith. Renewal theory and its ramifications. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 243–302, 1958. [62](#)
- [111] Sarah E Marzen and James P Crutchfield. Informational and causal architecture of discrete-time renewal processes. *Entropy*, 17(7):4891–4917, 2015. [67](#)
- [112] Johannes Söding. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, 21(7):951–960, 2004. [71](#)
- [113] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. [71](#)
- [114] Jorge Silva and Shrikanth Narayanan. Upper bound kullback–leibler divergence for transient hidden markov models. *IEEE Transactions on Signal Processing*, 56(9):4176–4188, 2008. [71](#)
- [115] Werner Horsthemke. Noise induced transitions. In *Non-Equilibrium Dynamics in Chemical Systems*, pages 150–160. Springer, 1984. [71](#)
- [116] Robert A Kennewick, David Locke, Michael R Kennewick, Richard Kennewick, Tom Freeman, et al. System and method for filtering and eliminating noise from natural language utterances to improve speech recognition and parsing, March 20 2012. US Patent 8,140,327. [71](#)

- [117] J. Huang and B. Kingsbury. Audio-visual deep learning for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7596–7599, May 2013. doi: 10.1109/ICASSP.2013.6639140. [71](#)
- [118] Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909. [71](#)
- [119] Donald Bures. An extension of kakutani’s theorem on infinite product measures to the tensor product of semifinite w^* -algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969. [71](#), [73](#), [79](#)
- [120] B-H Juang and Lawrence R Rabiner. A probabilistic distance measure for hidden markov models. *AT&T technical journal*, 64(2):391–408, 1985. [71](#)
- [121] E. F. Krause. *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*. Dover Books on Mathematics Series. Dover Publications, 1986. [71](#)
- [122] Rune B Lyngso, Christian NS Pedersen, and Henrik Nielsen. Metrics and similarity measures for hidden markov models. In *Proceedings of ISMB*, volume 99, pages 178–186, 1999. [75](#)
- [123] B. Fuglede and F. Topsøe. Jensen-shannon divergence and hilbert space embedding. In *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, pages 31–, June 2004. doi: 10.1109/ISIT.2004.1365067. [71](#), [73](#)
- [124] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007.
- [125] Sayed Mohammad Ebrahim Sahraeian and Byung-Jun Yoon. A novel low-complexity hmm similarity measure. *IEEE Signal Processing Letters*, 18(2): 87–90, 2011. [71](#), [72](#), [73](#)
- [126] Stephen E Levinson, Lawrence R Rabiner, and Man Mohan Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *Bell System Technical Journal*, 62(4):1035–1074, 1983. [72](#), [73](#)
- [127] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012. [74](#), [97](#)
- [128] Oskar Perron. Zur theorie der matrices. *Mathematische Annalen*, 64(2): 248–263, 1907. [77](#), [81](#)
- [129] F. G. Frobenius. *Über Matrizen aus nicht negativen Elementen*. Königliche Akademie der Wissenschaften, 1912. [77](#), [81](#)

- [130] RV Mises and Hilda Pollaczek-Geiringer. Praktische verfahren der gleichungsauflösung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 9(2):152–164, 1929. [78](#)
- [131] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017. [92](#)
- [132] Kwok Ho Wan, Oscar Dahlsten, Hlér Kristjánsson, Robert Gardner, and MS Kim. Quantum generalisation of feedforward neural networks. *npj Quantum Information*, 3(1):36, 2017. doi: 10.1038/s41534-017-0032-4. [92](#)
- [133] Zhao-Yu Han, Jun Wang, Heng Fan, Lei Wang, and Pan Zhang. Unsupervised generative modeling using matrix product states. *Phys. Rev. X*, 8:031012, Jul 2018. doi: 10.1103/PhysRevX.8.031012. URL <https://link.aps.org/doi/10.1103/PhysRevX.8.031012>. [92](#)
- [134] Ivan Glasser, Nicola Pancotti, and J. Ignacio Cirac. Supervised learning with generalized tensor networks. pages 4799–4807, 2018. ISSN 10495258. URL <http://arxiv.org/abs/1806.05964>.
- [135] E. Miles Stoudenmire and David J. Schwab. Supervised Learning with Quantum-Inspired Tensor Networks. In *Advances in Neural Information Processing Systems*, pages 4799—4807. 2016. [111](#)
- [136] E Miles Stoudenmire. Learning relevant features of data with multi-scale tensor networks. *Quantum Science and Technology*, 3(3):034003, 2018. URL <http://stacks.iop.org/2058-9565/3/i=3/a=034003>. [111](#)
- [137] Siddarth Srinivasan, Geoff Gordon, and Byron Boots. Learning hidden quantum markov models. *arXiv preprint arXiv:1710.09016*, 2017.
- [138] Marcello Benedetti, Delfina Garcia-Pintos, Oscar Perdomo, Vicente Leyton-Ortega, Yunseong Nam, and Alejandro Perdomo-Ortiz. A generative modeling approach for benchmarking and training shallow quantum circuits. *npj Quantum Information*, 5(1):45, 2019. [108](#)
- [139] X Gao, Z-Y Zhang, and L-M Duan. A quantum machine learning algorithm based on generative models. *Science advances*, 4(12):eaat9004, 2018.
- [140] Jin-Guo Liu and Lei Wang. Differentiable learning of quantum circuit born machines. *Physical Review A*, 98(6):062324, 2018. [92](#), [108](#)
- [141] Richard J Rossi. *Mathematical statistics: an introduction to likelihood based inference*. John Wiley & Sons, 2018. [92](#)
- [142] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [92](#), [93](#), [94](#), [98](#)

- [143] F. Verstraete, J. J. García-Ripoll, and J. I. Cirac. Matrix product density operators: Simulation of finite-temperature and dissipative systems. *Physical Review Letters*, 93(20):1–5, 2004. ISSN 00319007. doi: 10.1103/PhysRevLett.93.207204. [92](#), [94](#)
- [144] Matthew Ho, Mile Gu, and Thomas J Elliott. Robust inference of memory structure for efficient quantum modeling of stochastic processes. *Physical Review A*, 101(3):032327, 2020. [92](#)
- [145] Ted Petrie. Probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 40(1):97–115, 1969. [93](#)
- [146] B. . Juang and L. R. Rabiner. A probabilistic distance measure for hidden markov models. *AT T Technical Journal*, 64(2):391–408, Feb 1985. ISSN 8756-2324. doi: 10.1002/j.1538-7305.1985.tb00439.x. [93](#)
- [147] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988. [93](#)
- [148] T Tieleman and G Hinton. Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *Technical Report.*, 2017. [93](#)
- [149] N. F. Travers and J. P. Crutchfield. Exact Synchronization for Finite-State Sources. *Journal of Statistical Physics*, 145(5):1181–1201, dec 2011. ISSN 00224715. doi: 10.1007/s10955-011-0342-4. [96](#)
- [150] N. F. Travers and J. P. Crutchfield. Asymptotic Synchronization for Finite-State Sources. *Journal of Statistical Physics*, 145(5):1202–1223, dec 2011. ISSN 00224715. doi: 10.1007/s10955-011-0349-x. [96](#)
- [151] K Burnham and D Anderson. Model selection and multi-model inference . 2nd edn.(springer: New york.). 2002. [97](#)
- [152] Art Hobson. *Concepts in statistical mechanics*. CRC Press, 1987. [97](#)
- [153] A Dembo O Zeitouni, A Dembo, and O Zeitouni. Large deviations techniques and applications, 1998. [97](#)
- [154] James P Crutchfield, Christopher J Ellison, and John R Mahoney. Times barbed arrow: Irreversibility, crypticity, and stored information. *Physical review letters*, 103(9):094101, 2009. [100](#), [109](#)
- [155] M. Horodecki and J. Oppenheim. Fundamental limitations for quantum and nanoscale thermodynamics. *Nat Commun*, (4):2059, 2013. doi: 10.1038/ncomms3059. [102](#)
- [156] Brian Coyle, Daniel Mills, Vincent Danos, and Elham Kashefi. The born supremacy: Quantum advantage and training of an ising born machine. *npj Quantum Information*, 6(1):1–11, 2020. [108](#)

- [157] James E. Hanson and James P. Crutchfield. Computational mechanics of cellular automata: An example. *Physica D: Nonlinear Phenomena*, 103(1-4): 169–189, 1997. ISSN 01672789. doi: 10.1016/S0167-2789(96)00259-X. [110](#)
- [158] N. Schuch, Ignacio Cirac, and David Pérez-García. PEPS as ground states: Degeneracy and topology. *Annals of Physics*, 325(10):2153–2192, 2010. ISSN 00034916. doi: 10.1016/j.aop.2010.05.008. [110](#)
- [159] Anthony R Cassandra, Leslie Pack Kaelbling, and Michael L Littman. Acting optimally in partially observable stochastic domains. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, pages 1023–1028. AAAI Press, 1994. [110](#)
- [160] Cina Aghamohammadi, Samuel P Loomis, John R Mahoney, and James P Crutchfield. Extreme quantum memory advantage for rare-event sampling. *Physical Review X*, 8(1):011025, 2018. [110](#)
- [161] William Huggins, Piyush Patel, K. Birgitta Whaley, and E. Miles Stoudenmire. Towards Quantum Machine Learning with Tensor Networks. 2018. URL <http://arxiv.org/abs/1803.11537>. [111](#)
- [162] Chu Guo, Zhanming Jie, Wei Lu, and Dario Poletti. Matrix Product Operators for Sequence to Sequence Learning. pages 1–13, 2018. doi: arXiv:1803.10908v2. URL <http://arxiv.org/abs/1803.10908>. [111](#)