Local Operators and Quantum Chaos

by

Daniel Eric Parker

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Joel Moore, Chair
Professor Ehud Altman
Professor David Limmer

Summer 2020

The dissertation of Daniel Eric Parker, titled Local Operators and Quantum Chaos, is approved:

Chair ______________________________ Date ____________

______________________________ Date ____________

______________________________ Date ____________

University of California, Berkeley

Local Operators and Quantum Chaos

To my parents, Judith Fleishman and Thomas Parker, for all they have done for me.

Contents

Contents **ii**

List of Figures **iv**

List of Tables **viii**

1 Invitation: Tea. Earl Grey. Hot. **1**

2 The Lanczos Algorithm **8**
2.1 Krylov Spaces and Three-Term Recurrances 8
2.2 Infinite Lanczos, Orthogonal Polynomials, and Gauss Quadrature 13

3 A Universal Operator Growth Hypothesis **23**
3.1 Introduction 23
3.2 Synopsis 24
3.3 Preliminaries: The Recursion Method 25
3.4 The Hypothesis 28
3.5 Exponential Growth of Complexities 35
3.6 Growth Rate as a Bound on Chaos 40
3.7 Application to Hydrodynamics 46
3.8 Finite Temperature 50
3.9 Conclusions 55

Appendices **59**
3.A Brief Review of the Recursion Method 59
3.B Moments and Lanczos Coefficients in the SYK Model 61
3.C Numerical Details for 1d Spin Chains 66
3.D A Family of Exact Solution with Linear Growth 68
3.E Derivation of the q-Complexity Bound 70
3.F Geometric Origin of the Upper Bounds 72

4 Local Matrix Product Operators **75**
4.1 iMPOs Introduction 75

4.2 The Idea of Compression 77
4.3 Review of MPOs 79
4.4 Finite MPO Compression 81
4.5 Local Infinite Matrix Product Operators 87
4.6 Canonical forms for Infinite MPOs 93
4.7 Compression of iMPOs 99
4.8 Operator Entanglement and Error Bounds 101
4.9 Relation to Control Theory 105
4.10 iMPO Examples 107
4.11 Unit Cell MPOs 111
4.12 Properties of Compressed Hamiltonians 120
4.13 Conclusions 127

Appendices **129**

4.A Proofs for Local MPOs 129
4.B Proofs for Canonical forms 134
4.C Exact estimates of Schmidt values 139
4.D Elementary operations 141

5 Conclusions **146**

Bibliography **148**

List of Figures

1.1 A cup of tea losing information. 1
1.2 Sketch of a highly-quantum system hooked up to a ammeter to measure electrical transport. 2
1.3 Schematic of the role of chaos in quantum dynamics. 3
1.4 Sketch of the computational resources required to simulate dynamics. 4
1.5 (Top) Sketch of the space of operators, with complexity of the operator increasing from left to right. (Bottom) The equivalent 1D chain. (Middle) The wavefunction on the 1d chain. 5

2.1 Lanczos on a random Hermitian matrix with $N = 40$. The initial vector is also chosen randomly. (Left) The true spectrum λ_i is represented by vertical lines, while the eigenvalues $\theta_i^{(n)}$ of T_n are represented by red crosses. One can see that eigenvalues with the largest magnitude converge first. (Right) Error in the eigenvalues versus the number of Lanczos iterations n. The error falls off roughly exponentially. 12
2.2 (Left) Comparison of the full spectral density of the Chaotic Ising model with its approximation as a sum of δ-functions via Eq. (2.22) in the Krylov space $K_N(\widehat{H}; \psi)$. (Right) Time-evolution of a state $|\psi(t)\rangle = e^{-i\widehat{H}t} |\psi\rangle$ in the full Hilbert space and the Krylov spaces $K_N(\widehat{H}; \psi)$. Throughout we take $L = 12$ (dimension 4096) with open boundary conditions, and $|\psi\rangle$ chosen randomly. 19

3.1 Artist's impression of the space of operators and its relation to the 1d chain defined by the Lanczos algorithm starting from a simple operator $\mathcal{O}$. The region of complex operators corresponds to that of large n on the 1d chain. Under our hypothesis, the hopping amplitudes b_n on the chain grow linearly asymptotically in generic thermalizing systems (with a log-correction in one dimension, see Section 3.4). This implies an exponential spreading $(n)_t \sim e^{2\alpha t}$ of the wavefunction φ_n on the 1d chain, which reflects the exponential growth of operator complexity under Heisenberg evolution, in a sense we make precise in Section 3.5. The form of the wavefunction φ_n is only a sketch; see Fig. 3.5 for a realistic picture. . . . 26

3.2 Lanczos coefficients in a variety of models demonstrating common asymptotic behaviors. “Ising” is $H = \sum_i X_i X_{i+1} + Z_i$ with $\widehat{\mathcal{O}} = \sum_j e^{iq_j} Z_j$ ($q = 1/128$ here and below) and has $b_n \sim O(1)$. “X in XX” is $H = \sum_i X_i X_{i+1} + Y_i Y_{i+1}$ with $\widehat{\mathcal{O}} = \sum_j X_j$, which is a string rather than a bilinear in the Majorana fermion representation, so this is effectively an interacting integrable model that has $b_n \sim \sqrt{n}$. XXX is $H = \sum_i X_i X_{i+1} + Y_i Y_{i+1} + Z_i Z_{i+1}$ with $\widehat{\mathcal{O}} = \sum_j e^{iq_j}(X_j Y_{j+1} - Y_j X_{j+1})$ that appears to obey $b_n \sim \sqrt{n}$. Finally, SYK is (3.18) where $q = 4$ and $J = 1$ and $\widehat{\mathcal{O}} = \sqrt{2}\gamma_1$ with $b_n \sim n$. The Lanczos coefficients have been rescaled vertically for display purposes. Numerical details are given in Appendices 3.B and 3.C. . . 29

3.3 Illustration of the spectral function and the analytical structure of $C(t), t \in \mathbb{C}$. When the Lanczos coefficients have linear growth rate α, $\Phi(\omega)$ has exponential tails $\sim e^{-|\omega|/\omega_0}$ with $\omega_0 = 2\alpha/\pi$; $C(t)$ is analytical in a strip of half-width $1/\omega_0$ and the singularities closest to the origin are at $t = \pm i/\omega_0$. See Appendix 3.A for further discussion. 32

3.4 *(a)* Lanczos coefficients in a variety of strongly interacting spin-half chains: $H_1 = \sum_i X_i X_{i+1} + 0.709 Z_i + 0.9045 X_i$, $H_2 = H_1 + \sum_i 0.2 Y_i$, $H_3 = H_1 + \sum_i 0.2 Z_i Z_{i+1}$. The initial operator $\widehat{\mathcal{O}}$ is energy density wave with momentum $q = 0.1$. *(b)* Crossover to apparently linear growth as interactions are added to a free model. Here $H = \sum_i X_i X_{i+1} - 1.05 Z_i + h_X X_i$, and $\widehat{\mathcal{O}} \propto \sum_i 1.05 X_i X_{i+1} + Z_i$. The b_n's are bounded when $h_X = 0$ but appears to have asymptotically linear growth for any $h_X \neq 0$. Logarithmic corrections are not clearly visible in the numerical data. Numerical details are given in Appendix 3.C. 34

3.5 The exact solution wavefunction (3.25) in the semi-infinite chain at various times. The wavefunction is defined only at $n = 0, 1, 2 \ldots$, but has been extrapolated to intermediate values for display. 36

3.6 (a) The growth rate α versus the classical Lyapunov exponent $\lambda_L/2$ in the classical Feingold-Peres model of coupled tops, (3.46). $\alpha \geq \lambda_L/2$ in general, with equality around the $c = 0$ where the model is the most chaotic. The growth rate appears to be discontinuous at the non-interacting points $c = \pm 1$, similarly to Fig. 3.4-(b). (b) The first 40 Lanczos coefficients of quantum $s = 2, \ldots, 32$ and classical ($s = \infty$) FP model, with $c = 0$. 46

3.7 Numerical computation of the diffusion coefficient for the energy density operator $\widehat{\mathcal{O}} = \mathcal{E}_q$ in $H = \sum_i X_i X_{i+1} - 1.05 Z_i + 0.5 X_i$. (a) The Lanczos coefficients for $q = 0.15$ are fit to (3.50) with $\alpha = 0.35$ and $\eta = 1.74$. We found it actually better not to approximate $G^{(N)}(z)$ by $\widetilde{G}^{(N)}(z)$, but instead by $\widetilde{G}^{(N+\delta)}(z)$ for some integer offset δ so that $\eta \approx 1$ (in the example shown, $\delta = 12$). Large η or negative values lead to numerical pathologies. (b) The approximate Green's function (3.52) at $q = 0.15$. The arrow shows the “leading” pole that governs diffusion. (c) The locations of the leading poles for a range of q. One can clearly see the diffusive dispersion relation $z = iDq^2/2 + \mathrm{O}(q^4)$. Fitting yields a diffusion coefficient $D = 3.3(5)$. 51

3.8 Exact Lyapunov exponent $\lambda_L(T)$ (3.97) and growth rate $\alpha(T)$ with the Wightman inner product (3.99) of the SYK model in the large-q limit as a function of temperature (in units of coupling constant $\mathcal{J}$). The conjectured bound $\lambda_L(T) \leq 2\alpha(T)_W$ is exactly saturated at all temperatures, while the universal bound $\lambda_L(T) \leq 2\pi T$ only saturates in the zero temperature limit. 55

3.B.1Change in the growth rate near integrability for the SYK model with $q = 2$ and $q = 4$ (3.85). The ratio of the $q = 4$ to $q = 2$ term is given by J, and the model becomes free at $J = 0$. 64

3.C.1The size distribution of the Pauli strings in the Krylov vectors $\widehat{\mathcal{O}}_n$ for the Hamiltonian H_1 with parameters and initial operator as in Fig. 3.4. Though the distribution drops quickly after its peak, $P_n(s)$ is supported on $[0, \lfloor n/2 \rfloor + 2]$. . . 68

4.1 A finite-state machine that generates Eq. 4.3. 79

4.2 Compression of a finite MPO representing the Hamiltonian (4.34). (a) The bond dimensions for: $\widehat{W}$, the naive MPO representation of H_1; $\widehat{W}_L$, the left-canonical representation by Alg. 1; $\widehat{W}_C$, the compressed MPO by Alg. 2, and $\widehat{W}'_C$, the result of the standard MPS compression. (b,c) The Schmidt spectra of $\widehat{W}'_C$ and almost-Schmidt spectra of $\widehat{W}_C$ at the sites denoted by the triangle and square, respectively. The numerical precision was taken to be $\varepsilon_{\text{can}} = 10^{-12}$ for canonicalization and $\varepsilon_C = 10^{-4}$ for compression. 86

4.3 Compression of the iMPO representing (4.95). Main: The almost-Schmidt spectra of iMPOs representing $\widehat{H}_2$ with spatial cutoff R ranging from 32 to 512. As $R \to \infty$, the largest s_a converge to a point-wise limit, while the long tails rapidly decays (so the latter are finite-R artifacts). Inset: the bond dimensions of the iMPO before and afterwards with a cutoff of $\varepsilon = 10^{-4}$. Other numerical thresholds are the same as Fig. 4.2. 108

4.4 Results of an iMPO implementation of the Lanczos algorithm, applied to $\widehat{H} = \frac{1}{2}\sum_n Z_n Z_{n+1} - 1.05 Z_n + 0.5 X_n$ and $\widehat{O} = \sum_n Z_n$. (a) The Lanczos coefficients b_n computed by the iMPO implementation with SVD truncation threshold ε, compared to the exact method ("ED") of Ref. [1]. (b) Error in the b_n's at precision ε (compared to "ED" values). (c) Bond dimension of the operators $\widehat{O}_n$. The growth rate is roughly $\mathrm{O}(n^2)$. (d) The almost Schmidt spectra of $\widehat{O}_{10}$. A large gap is visible at $a \sim 60$ where s_a drops by $\sim 10^{-6}$. (e) The almost Schmidt spectra of $\widehat{O}_{20}$. The gap is barely visible even with the smallest ε; the error starts to grow rapidly around the same n. 110

4.5 An example of the finite state machine for an MPO with unit cell size 4. One on-site operator is placed for each arrow, and the arrows wrap around from right to left. Each gray box represents the data stored in one tensor. The UCMPO has bond dimension $(\chi_4, \chi_1, \chi_2, \chi_3) = (5, 4, 4, 5)$ and is loop free. The blue numbers are a (non-unique) topological ordering for the nodes. 117

4.6 Ground state fidelity as a function of MPO singular value cutoff for the Fraction Quantum Hall effect Eq. (4.146). 126

List of Tables

2.1 Classical orthogonal polynomial systems and their corresponding weight functions and Lanczos coefficients from [10] or [9]. These differ between sources due to normalization conventions — the polynomials may be monic, orthonormal, or respect some other choice $\langle \pi_n, \pi_m \rangle = z(n)\delta_{nm}$. Here C_n are the Catalan numbers and $\chi_e(n)$ is 1 for n even and 0 for n odd. 16

3.1 Asymptotic behavior of Lanczos coefficients. The first and last rows are known rigorously, while the middle two are conjectures supported by many specific examples. We will see below that the conjecture for chaotic models is slightly modified in $1 + 1d$. 29

3.2 Examples of free/integrable/chaotic models where the Lanczos coefficients have constant/square-root/linear growth. The dynamics of the model are determined by level statistics and "Free*" means the model is free, but the operator in question does not have a simple or local representation in the free particle description. Analytical evidence means either an exact formula for the coefficients is available or analytical bounds are used to deduced the asymptotics. Numerical evidence means that the Lanczos coefficients were computed numerically and appear to have a clear trend. 30

3.3 The growth rate α versus half the OTOC-Lyapunov exponent $\lambda_L/2$ in the q-SYK model (3.18) in units of $\mathcal{J} = \sqrt{q}2^{(1-q)/2}J$. Here α is obtained by exact numerical methods discussed in Appendix 3.B, while λ_L is taken from the Appendix of [41]. The q-SYK is physical only for even integers q, but large-N methods allow an extrapolation to any $q \geq 2$. 42

3.4 Correspondence between finite and infinite temperature definitions and results. 50

4.1 Properties of different set of iMPOs: strictly local (SL), first degree (FD), and the set of general (Gen) iMPOs without restriction. 90

Acknowledgments

It is a truth insufficiently acknowledged that a successful scientist does not work in a vacuum, but must be immersed in a community. I have been incredibly privileged and fortunate to be surrounded by a first-rate community that has supported me at every turn, and any measure of success I have achieved would have been impossible without the constant and unfailing support I have received.

I extend particular thanks to my family: to my parents, Judith Fleishman and Thomas Parker, for fostering and encouraging my curiosity as long as I can remember, for teaching me that any question can be answered if you look in the right place and know enough math, and for their unwavering support every step of the way. I thank my brother, Gregory Parker, for encouraging me by example to greater diligence and reminding me that every claim must be met with a dose of healthy skepticism.

I also extend my sincere thanks to all the incredible mentors I have had in Berkeley. To Ehud Altman, whose incredible curiosity and love of knowledge have been a constant inspiration. To Joe Orenstein, who taught me to appreciate the power and beauty of symmetry in physics, and kept me grounded in reality. To Romain Vasseur, who inspired my interest in quantum dynamics and taught me all the unspoken practical parts of being a modern physicist. To Thomas Scaffidi, for always humoring even my most basic questions and for showing me how to clearly understand even the most confusing of topics. To Takahiro Morimoto, who always knew the answer to any question I could think of and generously shared his grand supply of knowledge with me. To Xiangyu Cao, whose mathematical approach to physics and incredible conversational style of working has turned every days of physics into a source of joy. And finally, to my advisor extrordinaire, Joel Moore, whose kindness and wisdom have made my journey possible.

Lastly, I must thank all the wonderful friends whose support made my PhD possible: Shashwat Silas, Richard Nally, Layne Frechette, Adam Scherlis, Campbell Hewett, Miranda Rutherford, Will Berdanier, Vir Bulchandani, Dan Borgnia, Tomo Soejima, Aaron Szasz, Roger Curley, Maia Werbos, Morgan Presley, Hannah Klion, Illan Halpern, and Benjamin Burdick. Without any of you, I would not be where I am now, and I would not be who I am today. Thank you all.

Chapter 1

Invitation: Tea. Earl Grey. Hot.

Quantum chaos is responsible for transforming the microscopic, quantum mechanical, formulation of our world into the macroscopic reality we are all familiar with. However, despite the fact that quantum mechanics is now nearly a century old, there is no agreed-upon definition for quantum chaos. Understanding quantum chaos is thus a primary goal in the field of quantum dynamics, but a highly difficult one. As an invitation to the topic, and to the ideas in the rest of this thesis, let us start with a relatively non-technical introduction to quantum chaos through the idea of irreversibility.

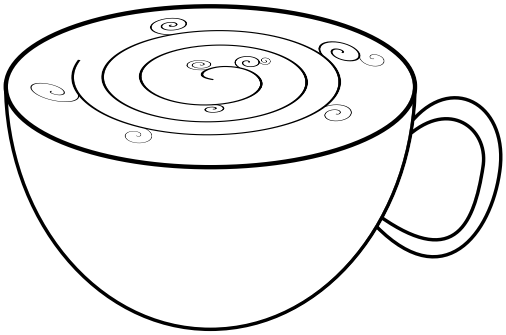


Figure 1.1: A cup of tea losing information.

Irreversibility is a common phenomenon in everyday life.[1] As an analogy, suppose you have a nice cup of tea (say, Earl Grey. Hot.). If you add a splash of milk, it will soon flow into whorls and eddies, then smaller eddies will branch off the eddies, and yet finer eddies around those in an ever-more-intricate pattern. If one waits long enough, the tea will become a uniform tan color and we say it has reached *equilibrium*. This process can be hurried along by stirring the tea (say, clockwise). But if one stirs it back the other way (say, counterclockwise), it will remain tan and never separate out into tea and milk; this process is *irreversible*.

[1] This discussion is based on my presentation at the Pappalardo Fellowship interview on 13 December 2019.

Irreversibility is a classical phenomenon and is often called the "arrow of time," since it only proceeds in one direction. A modern perspective of irreversibility is that one should focus on information itself. In this case, how much information is needed to describe the tea/milk system? At first, the milk is isolated to a stream or a few droplets and therefore its location is easy to describe. As mixing proceeds, the boundary between the milk and tea becomes more and more complex — more "chaotic" — and more and more information is required to keep track of the boundary. Once the tea comes to its tan equilibrium, the milk is a countless myriad of microscopic droplets in ever-shifting positions. At this point, the boundary between the liquids is so complex that it is not practical or even possible to describe it exactly. The information has been lost. However, this does not mean we cannot understand the system. In fact, a single number, the concentration of milk in the tea, suffices to describe all important (i.e. macroscopic) aspects of the system. We have thus arrived at a low-information *thermodynamic* description of the system.

Chaos, roughly, is the mechanism by which information is lost in this process, and a system is more chaotic if the rate of information loss is higher. A highly-chaotic system takes a vast amount of information to describe exactly. But conveniently, the more chaotic a system, the sooner a thermodynamic description will be applicable. To understand quantum chaos, we must understand the quantum version of this analogy.

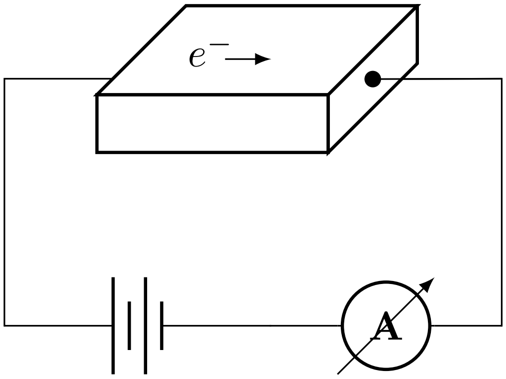


Figure 1.2: Sketch of a highly-quantum system hooked up to a ammeter to measure electrical transport.

Let us now suppose that we have a highly quantum system, such as a high-temperature superconductor or a sample of twisted bilayer graphene. The first and simplest experiments performed on a new material are usually transport measurements: how is heat or charge transported through the system (Fig 1.2). Even in quantum systems, transport is governed by the classical partial differential equations (PDEs) of *hydrodynamics*. For instance, the energy density $\epsilon(x,t)$ is governed by the heat equation

$$\partial_t \epsilon = D\nabla^2 \epsilon, \tag{1.1}$$

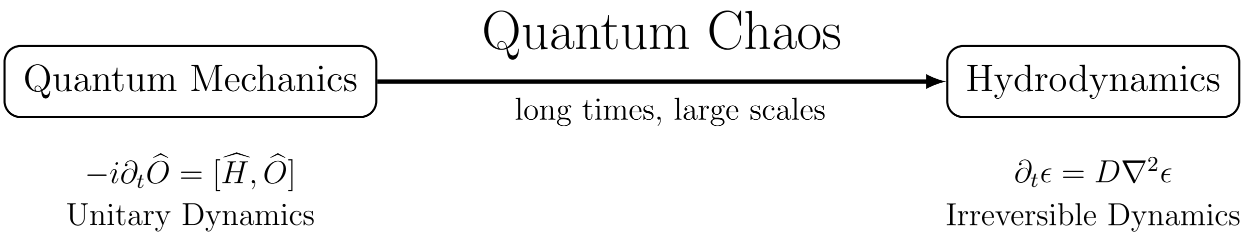


Figure 1.3: Schematic of the role of chaos in quantum dynamics.

where the constant D is called the thermal diffusivity. Once the hydrodynamic datum D is measured by experiment, this single number suffices to describe the macroscopic flow of heat in the system.

On the other hand, the microscopic description of the system must be quantum mechanical. Suppose one has a quantum model for the system: a Hamiltonian $\widehat{H}$ and, for each point x in space, an operator $\widehat{\mathcal{O}}_x$ corresponding to the energy density at x. The dynamics of $\widehat{\mathcal{O}}_x$ are governed by the by the Heisenberg equations of motion

$$-i\partial_t\widehat{\mathcal{O}}_x = [\widehat{H}, \widehat{\mathcal{O}}_x], \tag{1.2}$$

and the energy density is an observable $\epsilon(x,t) = \langle\widehat{\mathcal{O}}_x(t)\rangle$.

To evaluate the accuracy of a quantum model, one must find the hydrodynamic description and compare to experiments. Specifically, one must analyze the model at long times and large scales to determine

(H1). the hydrodynamic equations of motion of the system, such as the wave equation, the heat equation, or even the Einstein field equations, and

(H2). the coefficients ("hydrodynamic data") needed to fully specify the PDE, such as the thermal diffusivity.

Unfortunately, this analysis is often extremely difficult or even impossible. For instance, many reasonable-seeming models of high-temperature superconductors have been proposed which cannot be confirmed nor falsified because it is not known how to compute the hydrodynamic data from the theory.

At a deeper level, this is also a structural puzzle. Quantum dynamics is unitary, so no information is ever lost. But PDEs like the heat equation are irreversible, and information is being lost at every moment. How, then, can we go from the quantum mechanical description to hydrodynamics (Fig. 1.3)? This is the role of quantum chaos. Although information is never destroyed in quantum mechanics, it can become *hidden* or *inaccessible* and thus virtually impossible to recover.

Let us examine the problem from an information-centric perspective. If we could simulate Eq. (1.2) to long times, we could immediately find H1 and H2. How many computational resources (i.e. how much RAM or how many CPUs) would it take to do this? Unfortunately, the amount of information one must keep track of to reach time t scales as e^t, making it impossible to reach large t even on (classical) supercomputers.[2] By contrast, the hydrodynamical description requires only a polynomially large number of resources to simulate the system at large times (Fig. 1.4).[3] So much of the exponentially-large exact description of the system must be unnecessary; only a small subset of the information is needed to specify the hydrodynamics. The fundamental mathematical question involved in quantum chaos, from this perspective, is identifying exactly which parts of the description are unnecessary.

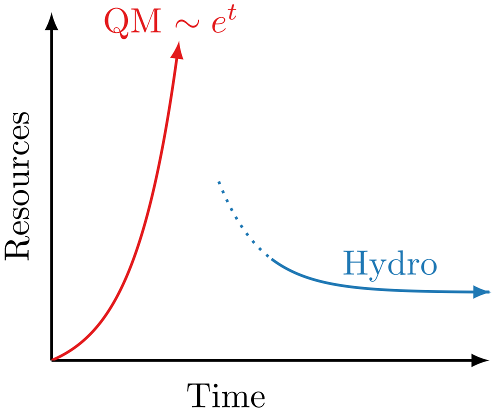


Figure 1.4: Sketch of the computational resources required to simulate dynamics.

Let us therefore examine the quantum dynamics and find the redundant information at a conceptual level. (Making these ideas precise and more rigorous is the subject of the first few chapters.) The formal solution to Eq. (1.2) is an infinite sum of nested commutators

$$\widehat{\mathcal{O}}(t) = e^{i\widehat{H}t}\widehat{\mathcal{O}}e^{-i\widehat{H}t} = \widehat{\mathcal{O}} + it[\widehat{H},\widehat{\mathcal{O}}] + \frac{(it)^2}{2}[[\widehat{H},\widehat{\mathcal{O}}],\widehat{\mathcal{O}}] + \cdots \tag{1.3}$$

To visualize this, let us picture the space of operators, arranged with simple, local operators on the left and more complicated, non-local operators towards the right, as shown in Fig. 1.5. The simple local operator $\widehat{\mathcal{O}}(t)$ is the black dot on the left. The commutator $[\widehat{H},\widehat{\mathcal{O}}]$ (represented by yellow dots) is slightly more complicated and less local, so it can be written as the sum of a few terms in any local basis . The second commutator $[[\widehat{H},\widehat{\mathcal{O}}],\widehat{\mathcal{O}}]$ (blue dots) is again more complicated and less local, so it will be the sum of a larger number of local terms, and so on and so forth. As usual with a Taylor series, the few terms are sufficient to describe the time-evolved operator at early times, but more and more terms are needed as time goes on. At long times, $\widehat{\mathcal{O}}(t)$ becomes ever more complex and less local as it moves out into the space of operators, as indicated by the arrows.

As time goes on, the further out one goes into the space of operators, and the harder it is to observe the operators. Each operator involves increasingly non-local correlations, which

[2] At present, supercomputer computations cannot get much past $t = O(10^1)$ even for the simplest 1d

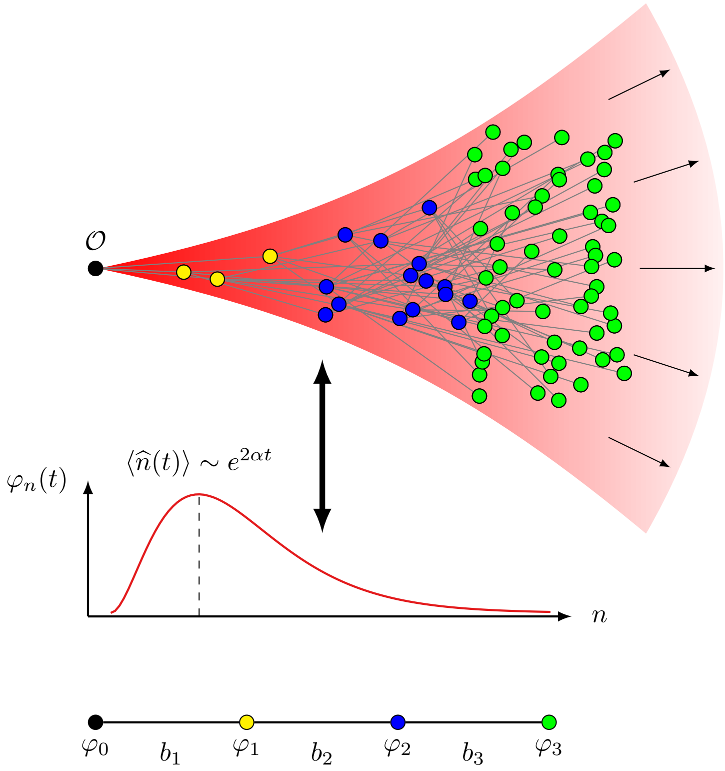


Figure 1.5: (Top) Sketch of the space of operators, with complexity of the operator increasing from left to right. (Bottom) The equivalent 1D chain. (Middle) The wavefunction on the 1d chain.

are exponentially hard to measure experimentally or require exponentially many resources to simulate. So although the operator moves out into operator space under unitary dynamics, and no information is ever lost, the difficulty of retrieving that information grows incredibly fast. For any finite amount of resources, there will be a time at which reconstructing the full operator is too costly, and the information is effectively lost.

To quantify this process, we use a basis specially adapted to these dynamics. As we shall see in Chapter 3, one may always make a unitary transformation so that the dynamics (1.3) in space of operators is mapped onto an equivalent 1d quantum mechanics problem. Equation (1.2) is then equivalent to

$$-i\partial_t\varphi_n = b_{n-1}\varphi_{n-1} + b_n\varphi_{n+1}; \qquad \varphi_n(t=0) = \delta_{n_0}, \quad b_n > 0, \tag{1.4}$$

where $\varphi_n(t)$ is the component of the "operator space wavefunction" on the nth site and $\varphi_0(t) = \epsilon(t)$ is the observable of interest. The b_n's are called the *Lanczos coefficients* and

chaotic systems.

[3] Of course, there is only a hydrodynamic description for a few observables. Generic observables are probably "irreducibly difficult" to compute.

will play a prominent technical role in this thesis. Although (1.4) is completely equivalent to (1.2), it is vastly easier to understand. Crucially, we know that operators further to the right are more "complex".

We can quantify the complexity of the operator space wavefunction via the expectation of the position operator

$$\langle \widehat{n}(t) \rangle = \sum_{n \geq 0} n |\varphi_n(t)|^2. \tag{1.5}$$

One can show that $\langle \widehat{n}(t) \rangle$ increases at most exponentially: for any Hamiltonian, there are constants $A, M > 0$ such that

$$\langle \widehat{n}(t) \rangle \leq A e^{Mt}. \tag{1.6}$$

For non-chaotic systems, the position can grow much slower than this bound. For example, in a free system, $\langle \widehat{n}(t) \rangle \propto t$. But in chaotic quantum systems, exponential growth is achieved and there is a constant $\alpha > 0$, characteristic of the system, so that

$$\langle \widehat{n}(t) \rangle \propto e^{2\alpha t}. \tag{1.7}$$

In a chaotic system, therefore, the wavefunction "runs away" exponentially fast into the space of operators of increasing complexity. Or, in other words, chaotic systems are ones where information is effectively lost as quickly as possible. From our perspective, Eq. (1.7) can be taken as a definition of which quantum systems are chaotic.

Not only is this characterization of quantum chaos conceptually satisfying, but also practically applicable. Once one knows the rate α at which information is lost, there is an algorithm to compute both the hydrodynamic equations (H1) *and* the hydrodynamic data (H2) of the model. This algorithm is remarkably efficient. For the simplest one-dimensional chaotic systems, simulating the dynamics to long times ($t \approx 20$) and extracting the hydrodynamic data takes several hundred thousand CPU hours on a supercomputer. This algorithm gives the same answer in a few seconds on a laptop. So, as often occurs, a different conceptual viewpoint suggests a more efficient computational method.

Now that we have sketched some of the ideas of quantum chaos, we may summarize the aim of this work. This thesis will offer a new definition of quantum chaos:

A system is chaotic if the expectation $\langle \widehat{n}(t) \rangle$ grows exponentially.

We explore two main consequences of this definition:

1. Conceptually, information is lost as soon as possible in a chaotic system.

2. Computationally, quantum chaos provides an efficient algorithm to compute emergent hydrodynamics (H1 and H2).

In the following chapters, we will state these results more carefully, and describe the necessary technical assumptions and caveats.

The rest of this work is organized as follows.

- Chapter 2 introduces the Lanczos algorithm. The Lanczos algorithm for tridiagonalizing matrices is a main technical tool in this work. Its origin is in numerical linear algebra where it is used to find the extremal eigenvalues of a matrix. In the infinite dimensional setting, however, it becomes an analytical tool. We will discuss how Lanczos gives rise to the classical orthogonal polynomials, how it can be used to approximate distributions, and its connection to the classical Moment Problem. Several numerical and physical applications are given as examples.

- Chapter 3 presents a universal operator growth hypothesis, which provides a formulation of quantum chaos in terms of the Lanczos coefficients described above. After making these ideas more precise, the chapter will introduce the idea of a “Q-complexity” and show a relation between the Lanczos coefficients and out-of-time-order correlators and provide a new bound on measures of chaos. We will then give the algorithm for computing the hydrodynamic coefficients and conclude with speculations about how these results may be extended to finite temperature. The material from this chapter is mainly drawn from [1].

- Chapter 4 will switch focus from quantum chaos to the operators themselves and answer the question: what is the most efficient way to represent or approximate a local operator in the thermodynamic limit? Working in the framework of matrix product operators (MPOs), we shall give explicit and efficient algorithms for compressing operators, i.e. finding the most accurate representation of an operator with a fixed number of resources. As practical applications, we show how compression can be used to run the Lanczos algorithm with MPOs directly in the thermodynamic limit, and also how to accurately compress long-range 2D Hamiltonians to small enough sizes to find their ground states. The material in this chapter is mainly drawn from [2].

Chapter 2

The Lanczos Algorithm

This chapter will explore the Lanczos algorithm, a necessary predicate to our study of operator growth, in both the finite and infinite dimensional cases.

The first part of this chapter, Section 2.1, will present the finite dimensional Lanczos algorithm. The finite Lanczos algorithm is a tool for tridiagonalizing a matrix; just as one can compute an eigendecompostion of a matrix $A = U^\dagger DU$, one can find a change of basis $A = V^\dagger TV$ so that T is tridiagonal. In many ways, the fact that T is nearly diagonal means it is nearly as useful as a full eigendecomposition — but much more efficient to compute. In fact, modern algorithms for computing eigenvalues of a matrix proceed by first reducing the matrix to tridiagonal form, then finding the eigenvalues from there [3]. Moreover, just a few iterations of the algorithm are sufficient to approximate the extremal eigenvalues to high precision. The successes and applications of the Lanczos algorithm are so numerous that it was selected as one of SIAM's top 10 algorithms of the 20th century [1] [4]. This section will therefore hew closely to the numerical analysis literature [3, 5].

In the second half of this chapter, Section 2.2, we upgrade to the infinite-dimensional setting. The infinite Lanczos algorithm takes on an analytical character. We will show how the infinite Lanczos algorithm underlies the theory of orthogonal polynomials, and briefly discuss how it solves the classical moment problem and leads to the method of Gauss quadrature for integration against distributions [6]. We shall close with a physical application: how quadrature can be used to approximately exponentiate a Hamiltonian.

2.1 Krylov Spaces and Three-Term Recurrances

As a warm-up and motivation, let us start with the following question: given a matrix A, what is its largest eigenvector? Computationally, perhaps the simplest way to answer this question is with the method of power iteration.

[1] For comparison, the other 9 are: Monte Carlo methods, simplex method for linear programming, (Householder) matrix decompisitions, the Fortran compiler, the QR algorithm, Quicksort, the fast-Fourier transform, integer relation detection, and the fast multipole algorithm.

Power Iteration

Suppose that A is a Hermitian matrix of size N, so it has eigenvalues

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|, \tag{2.1}$$

and corresponding (normalized) eigenvectors $\boldsymbol{z}_k$. (We keep these conditions through the rest of this section, unless otherwise stated.)

If we start with a random vector $\boldsymbol{v} = \sum_{k=1}^N v_k \boldsymbol{z}_k$, then applying A will enhance the component in the largest subspace:

$$A\boldsymbol{v} = \sum_{k=1}^N \lambda_k v_k \boldsymbol{z}_k,$$

so $A\boldsymbol{v}$ is closer to $\boldsymbol{z}_1$ than $\boldsymbol{v}$. Iterating this,

$$A^n \boldsymbol{v} = \sum_{k=1}^N \lambda_k^n v_k \boldsymbol{z}_k = \lambda_1^n (v_1 \boldsymbol{z}_1 + \boldsymbol{\epsilon}_n); \qquad \boldsymbol{\epsilon}_n = \sum_{k=2}^N \left(\frac{\lambda_k}{\lambda_1}\right)^n v_k \boldsymbol{z}_k,$$

and the error term decays exponentially:

$$||\boldsymbol{\epsilon}_n|| \leq \max_{2\leq k\leq n} |v_k| \cdot \left|\frac{\lambda_2}{\lambda_1}\right|^k \xrightarrow{n\to\infty} 0,$$

where $||\cdot||$ is a norm on $\mathbb{C}^N$. So the largest vector is

$$\boldsymbol{z}_1 = \lim_{n\to\infty} \boldsymbol{v}_n; \qquad \boldsymbol{v}_{n+1} = \frac{A\boldsymbol{v}_n}{|A\boldsymbol{v}_n|}, \tag{2.2}$$

for any initial vector $\boldsymbol{v}_0 = \boldsymbol{v}$ whose overlap with $\boldsymbol{z}_1$ is non-zero. This method is known as *power iteration.* Matrix-vector multiplication is an $O(N^2)$ operation, so the power method is quite computationally efficient.

From another perspective, though, power iteration is rather wasteful. Since $\boldsymbol{v}_k$ converges quite quickly to the largest eigenvector, each new vector $\boldsymbol{v}_{k+1}$ is almost linearly dependent with the one before it. Furthermore, the error term

$$\boldsymbol{\epsilon}_n = \sum_{k=2}^N \left(\frac{\lambda_k}{\lambda_1}\right)^n v_k \boldsymbol{z}_k = \left(\frac{\lambda_2}{\lambda_1}\right)^n \left[v_2 \boldsymbol{z}_2 + \boldsymbol{\epsilon}_n^{(2)}\right]; \qquad \boldsymbol{\epsilon}_n^{(2)} = \sum_{k=3}^N \left(\frac{\lambda_k}{\lambda_2}\right)^n v_k \boldsymbol{z}_k,$$

is converging to $\boldsymbol{z}_2$, and the error to *that* converging to $\boldsymbol{z}_3$, and so on. So if we were to only orthogonalize each $\boldsymbol{v}_{k+1}$ with the ones before it, we could find not only the largest eigenvector, but *all* of the largest eigenvectors with the same amount of work.

Krylov Spaces

We now formalize this idea. Given a Hermitian matrix A and a starting vector $\boldsymbol{v}$, define the **Krylov spaces**

$$\mathcal{K}_n(A;\boldsymbol{v}) = \operatorname{span}\{\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots, A^{n-1}\boldsymbol{v}\}. \tag{2.3}$$

Not only is $\boldsymbol{z}_1 \approx \boldsymbol{v}_n \in \mathcal{K}_n(A;\boldsymbol{v})$, but we shall see that $\mathcal{K}_n$ contains good estimates for the n largest eigenvectors of A, each converging exponentially fast.

We can now introduce the Lanczos algorithm, which computes a natural basis for the Krylov space. Quite explicitly, it is the Gram-Schmidt process specialized to the case where the vectors are chosen to be the generators $\{\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots, A^{n-1}\boldsymbol{v}\}$ of a Krylov space. As remarked above, $A^n\boldsymbol{v}$ is nearly linearly dependent to the previous vector $A^{n-1}\boldsymbol{v}$, so it makes sense to perform the Gram-Schmidt process: start with $\boldsymbol{v}$, then iteratively apply A and orthogonalize against previous basis vectors. We will see that — almost miraculously — it is only necessary to orthogonalize against a *single* previous vector.

Explicitly, let $\boldsymbol{v}_1 := \boldsymbol{v}$ and, for $n > 1$, iteratively define

$$\boldsymbol{u}_{n+1} := A\boldsymbol{v}_n - \sum_{k=1}^{n} \boldsymbol{v}_k T_{kn}, \tag{2.4a}$$

$$T_{kn} := \langle A\boldsymbol{v}_n, \boldsymbol{v}_k\rangle, \tag{2.4b}$$

$$T_{n+1,n} := ||\boldsymbol{u}_{n+1}||, \tag{2.4c}$$

$$\boldsymbol{v}_{n+1} := \boldsymbol{u}_{n+1}/T_{n+1,n} \tag{2.4d}$$

Assuming for the moment that $T_{n+1,n} \neq 0$ for $n < N$, then $A\boldsymbol{v}_N$ must be linearly dependent with the previous $\boldsymbol{v}_k$'s, so $\boldsymbol{y}_{N+1} = \boldsymbol{0}$ and the process terminates. Therefore $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}$ form a complete basis. Using $\boldsymbol{u}_{n+1} = T_{n+1,n}\boldsymbol{v}_{n+1}$, we have the relation

$$A\boldsymbol{v}_n = \sum_{k=1}^{n+1} \boldsymbol{v}_k T_{kn}, \tag{2.5}$$

or, in matrix form,

$$AV = VT, \tag{2.6}$$

where V is the unitary matrix whose columns are the $\boldsymbol{v}_n$'s and

$$T = \begin{bmatrix} T_{11} & T_{12} & T_{13} & \cdots & T_{1N} \\ T_{21} & T_{22} & T_{23} & \ddots & \vdots \\ 0 & T_{32} & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & T_{N-1,N} \\ 0 & \cdots & 0 & T_{N,N-1} & T_{NN} \end{bmatrix}, \tag{2.7}$$

is an upper-Hessenberg matrix (i.e. upper-triangular with one additional diagonal below the middle).

Since V is unitary and A is Hermitian, we have the decomposition $T = V^\dagger A V$, so

$$T^\dagger = \left(X^\dagger A X\right)^\dagger = X^\dagger A^\dagger \left(X^\dagger\right)^\dagger = X^\dagger A X = T.$$

Therefore T is not just upper-Hessenberg, but actually a symmetric, tridiagonal matrix:

$$T := \begin{bmatrix} a_1 & b_1 & & \cdots & 0 \\ b_1 & a_2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & b_{N-1} \\ 0 & \cdots & & b_{N-1} & a_N \end{bmatrix} \tag{2.8}$$

where we have defined the **Lanczos coefficients** $a_n := T_{n,n}$ and $b_n := T_{n+1,n} = T_{n,n+1}$. In other words, $A\boldsymbol{v}_n$ is automatically orthogonal to $\boldsymbol{v}_k$ for $k < n-1$ and so we need only orthogonalize $A\boldsymbol{v}_n$ against $\boldsymbol{v}_n$ and $\boldsymbol{v}_{n-1}$ to compute an orthonormal basis. Let us state this result as a theorem.

Theorem 1 (Lanczos Algorithm)**.** *Suppose* $(\mathbb{V}, ||\cdot||)$ *is an inner-product space of dimension* N*. Suppose* A *is a Hermitian operator on* $\mathbb{V}$ *and suppose* $\boldsymbol{v} \in \mathbb{V}$ *is a unit vector. Define* $b_0 := 1$*,* $\boldsymbol{v}_1 := \boldsymbol{v}$*,* $\boldsymbol{v}_0 := \boldsymbol{0}$*, and for* $1 \le n \le N-1$ *iteratively define*[2]

$$a_n := \langle \boldsymbol{v}_n, A\boldsymbol{v}_n\rangle \tag{2.9a}$$

$$\boldsymbol{u}_{n+1} := A\boldsymbol{v}_n - a_n\boldsymbol{v}_n - b_n\boldsymbol{v}_{n-1} \tag{2.9b}$$

$$b_{n+1} := ||\boldsymbol{u}_{n+1}|| \tag{2.9c}$$

$$\boldsymbol{v}_{n+1} := \boldsymbol{u}_{n+1}/b_{n+1}. \tag{2.9d}$$

Then the ***Krylov vectors*** $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ *are an orthonormal basis for* $\mathcal{K}_n(A; \boldsymbol{v})$*, satisfy the* ***three-term recurrance relation***

$$b_{n+1}\boldsymbol{v}_{n+1} = (A - a_{n+1})\boldsymbol{v}_n - b_n\boldsymbol{v}_{n-1}, \tag{2.10}$$

and, finally,

$$T = V^\dagger A V, \tag{2.11}$$

is a tridiagonal decomposition of A*.*

A few comments on this result are in order. First, Eqns. (2.9) are called the **Lanczos Algorithm**. Computationally, finding n Krylov vectors requires n matrix-vector multiplications — a relatively cheap operation, particularly if A can be expressed as a sparse matrix. Second, even a few iterations of the Lanczos algorithm is sufficient to form good estimates for

[2] In principle, this process can terminate early if $b_M = 0$ for some $M < N$. However, this occurs if and only if $\boldsymbol{v}$ has overlap with only M of the eigenvectors of A. In this case, one should pick a new starting vector $\boldsymbol{v}_M$ orthogonal to all the previous ones and continue. If $\boldsymbol{v}$ is chosen randomly, this occurs almost never.

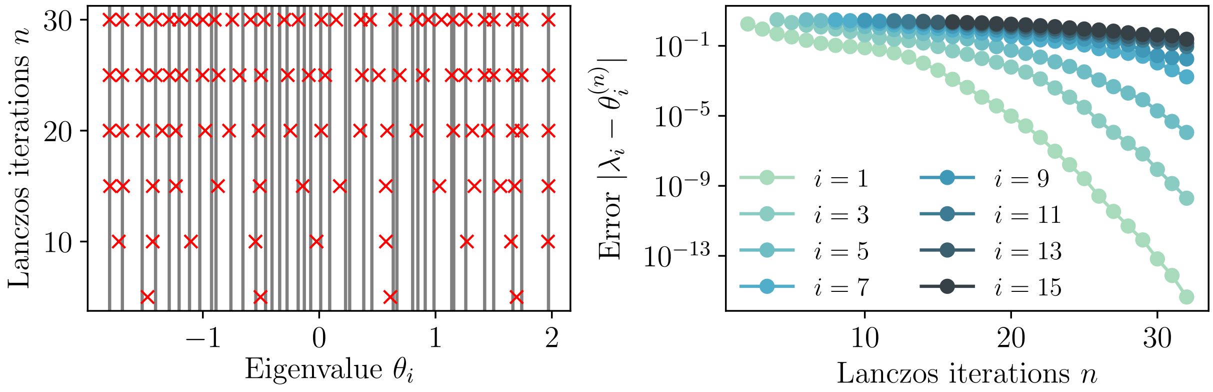


Figure 2.1: Lanczos on a random Hermitian matrix with $N = 40$. The initial vector is also chosen randomly. (Left) The true spectrum λ_i is represented by vertical lines, while the eigenvalues $\theta_i^{(n)}$ of T_n are represented by red crosses. One can see that eigenvalues with the largest magnitude converge first. (Right) Error in the eigenvalues versus the number of Lanczos iterations n. The error falls off roughly exponentially.

the eigenvalues with largest magnitude. As a rule of thumb, n iterations will lead to numerically precise estimates for the $\frac{2}{3}n$ largest eigenvalues. This is illustrated in Fig. 2.1. Third, the condition that A is Hermitian can be relaxed. If A was simply normal, then $T \neq T^\dagger$ in general and will be a full upper-Hessenberg matrix, so one must orthogonalize against all previous vectors. This generalization is called the Arnoldi algorithm, and is by Eqs. (2.4). Another variant is bi-Lanczos, which instead constructs *two* bases, one for $\mathcal{K}_n(A)$ and one for $\mathcal{K}_n(A^\dagger)$. Four, the algorithm as presented here is susceptible to numerical instabilities. Once n exceeds about 40, the accumulation of numerical error means the basis vectors lose orthogonality. There are a large number of variants on the basic algorithm to avoid these numerical issues, such as the Implicitly Restarted Arnoldi Method [7].

Let us examine in more detail how the eigenvalues converge under Lanczos iteration. After n Lanczos steps, we can construct the N-by-n projector matrix V_n whose columns are $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ and the tridiagonal matrix $T_n = V_n^\dagger A V_n$, which is the upper n-by-n block of T. Suppose we have an eigendecomposition of T_n in the Krylov space $\mathcal{K}_n(A; \boldsymbol{v})$:

$$T_n \boldsymbol{s}_i = \theta_i \boldsymbol{s}_i.$$

One can show[3] that $\theta_i \to \lambda_i$ and, projecting back to the full vector space, $V\boldsymbol{s}_i \to \boldsymbol{z}_i$, the ith eigenvector of A. We quote a theorem originally due to Saad that show this convergence is essentially exponentially fast in n.

Theorem 2 (Saad, or Thm. 10.1.4 of [3])**.** *After n steps of the Lanczos algorithm, for*

$1 \le i \le k$,

$$\lambda_i \ge \theta_i \ge \lambda_i - (\lambda_1 - \lambda_N)\left(\frac{\kappa_i \tan\phi_i}{c_{n-i}(1+2\rho_i)}\right)^2, \tag{2.12}$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N$ and

$$\rho_i = \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_n} > 0, \quad \kappa_i = \prod_{j=1}^{i-1} \frac{\theta_j - \lambda_n}{\theta_j - \lambda_i}, \cos\phi_i = |\langle \boldsymbol{v}_1, \boldsymbol{z}_i \rangle|,$$

and $c_n(x)$ is the Chebyshev polynomial of the first kind.

In other words, the convergence of the ith eigenvalue is controlled by the gap between the ith and $(i+1)$th eigenvalues, normalized to the diameter of the spectrum. In practice, we are interested in a particular eigenvalue (say, the third largest), as the number of Lanczos iterations is increased. In that case,

$$|\lambda_i - \theta_i^{(n)}| \approx [c_{n-i}(1+2\rho_i)]^{-2} = O(e^{-n}),$$

since $c_n(x) = 2^{n-1}x^n + o(x)$ blows up exponentially with n for x outside $[-1, 1]$. The finite Lanczos algorithm therefore provides a quickly converging estimate for all the extremal eigenvalues of a matrix.

2.2 Infinite Lanczos, Orthogonal Polynomials, and Gauss Quadrature

This section will present the Lanczos algorithm in the infinite-dimensional setting, where it becomes an analytic tool that underlies the theory of orthogonal polynomials, gives a practical way to integrate against distributions, and is closely related the so-called ‘moment problem’.

Orthogonal Polynomials

Let us start in the realm of orthogonal polynomials. We shall see that all the classical families of orthogonal polynomials — Chebychev, Hermite, Laguerre, etc — all come from the Lanczos algorithm. Our treatment will closely follow the references [8] and [9].

Definition 3. An **orthogonal polynomial sequence** (OPS) $\{p_n(x)\}_{n\ge 0}$ with respect to a linear form $L : \mathbb{C}[x] \to \mathbb{C}$ if

1. for all n, $p_n(x)$ is a polynomial of degree n,
2. $L[p_m(x)p_n(x)] = 0$ if $m \ne n$,
3. for all n, $L[p_n(x)^2] \ne 0$.

The polynomials p_n can be normalized arbitrarily. Two standard choices are monic polynomials, with $p_n(x) = x^n + \cdots$ or orthonormal polynomials, so $L[p_n p_m] = \delta_{nm}$.

The linear form is commonly given as an integral

$$L[p(x)] = \int_a^b p(x)w(x)\,dx \tag{2.13}$$

where $w(x) \geq 0$ is called a **weight distribution** [3]

A choice of weight w makes L into an inner product:

$$\langle p, q\rangle_w := L[pq^*] = \int p(x)q(x)^* w(x)\,dx. \tag{2.14}$$

We will also need the **moments** of w, defined as

$$\mu_n := L[x^n] = \int x^n w(x)\,dx. \tag{2.15}$$

These are strictly positive since w is positive.

One should think of orthogonal polynomials as an infinite-dimensional linear algebra problem, where $w(x)$ is the spectral distribution or density of states for an operator. Multiplication by the finite matrix A is therefore replaced with multiplication by x. The analogue of the Krylov spaces $K_n(A; \boldsymbol{v})$ are now the span of $\{1, x, x^2, \ldots, x^n\}$. Just as before, we can apply Gram-Schmidt to iteratively orthogonalize this sequence, producing a natural basis.

Theorem 4 (Infinite Lanczos; Thm. 3.3 of [8]). *Suppose L is a linear form given by a weight function $w(x)$. Then there exists constants $\{a_n\}_{n\geq 1}$, $\{b_n\}_{n\geq 1}$ such that*

$$\begin{aligned}
\pi_0(x) &= 1 &(2.16a)\\
\pi_1(x) &= x &(2.16b)\\
b_{n+1}\pi_{n+1}(x) &= (x - a_n)\pi_n(x) - b_n\pi_{n-1}(x), \quad n > 1 &(2.16c)\\
a_n &= \langle \pi_n, x\pi_n\rangle_w &(2.16d)\\
b_n &= \langle \pi_n, x\pi_{n-1}\rangle_w = ||\pi_n||_w &(2.16e)
\end{aligned}$$

gives a OPS of orthonormal polynomials with respect to w (assuming that none of the b_n's vanish).

One can see this is exactly Eq. (2.10) with $\boldsymbol{v}_n$ replaced by π_n and A replaced by x.

Proof. The proof is essentially the same as the finite case, but with different notation. Suppose for induction that $\{\pi_k\}_{k=0}^n$ is an OPS with respect to w, and thus forms a basis for

[3] There are several technical complexities associated to $w(x)$ which we shall ignore, particularly that $w(x)$ must form a positive definite measure. We assume that all integrals are well-defined and $\int p(x)w(x)dx > 0$ for any polynomial p. See Chapter 2 of [8] for a more careful discussion.

polynomials up to degree n (The base case holds by definition.) Then multiplying by x gives a degree $n+1$ polynomial, which will define π_{n+1}:

$$T_{n+1,n}\pi_{n+1}(x) = x\pi_n(x) - \sum_{k=0}^{n} \pi_k(x) T_{kn}; \quad T_{kn} = \langle x\pi_n, \pi_k \rangle \text{ for } k \leq n.$$

Here $T_{n,n+1}$ is an as-yet-unknown constant so that $\pi_{n+1}(x)$ is normalized. For $k < n$, $x\pi_k$ is a polynomial of degree $k+1$, and we can expand it as $x\pi_k(x) = \sum_{j=0}^{k+1} c_j \pi_j(x)$. Then

$$T_{kn} = \langle x\pi_n, \pi_k \rangle = \langle \pi_n, x\pi_k \rangle = \sum_{j=0}^{k+1} c_j \langle \pi_n, \pi_j \rangle = c_n \delta_{k+1,n},$$

where we used the fact that x is self-adjoint. So $T_{nk} = 0$ for $k+1 < n$ (i.e. T is upper-Hessenberg as an infinite matrix) and we have the relation

$$T_{n+1,n}\pi_{n+1}(x) = (x - T_{nn})\pi_n(x) - T_{n-1,n}\pi_{n-1}(x), \tag{2.17}$$

where we have defined $a_n = \langle \pi_n, x\pi_n \rangle$ and $b_n = \langle \pi_n, x\pi_{n-1} \rangle$ for convenience. Put

$$T_{n+1,n} := ||(x - T_{nn})\pi_n(x) - T_{n-1,n}\pi_{n-1}(x)||$$

then $\pi_{n+1}(x)$ is completely defined and, by construction, orthonormal to the previous polynomials. Finally, note that Eq. (2.17) implies

$$\langle x\pi_n, \pi_{n+1} \rangle = T_{n+1,n} = \langle \pi_n, x\pi_{n+1} \rangle = \langle x\pi_{n+1}, \pi_n \rangle^* = T^*_{n,n+1}.$$

But $T_{n+1,n} \geq 0$ by definition, so $T_{n+1,n} = T_{n,n+1}$ (i.e. T is symmetric and tridiagonal as a matrix). For convenience, we put $a_n := T_{nn}$ and $b_n := T_{n-1,n}$, which gives Eq. (2.16) above. □

We can package Eq. (2.16) in terms of an infinite tridiagonal matrix, often called a Jacobi matrix:

$$J := \begin{bmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & & \\ & b_2 & a_3 & b_3 & \\ & & b_3 & a_4 & \ddots \\ & & & \ddots & \ddots \end{bmatrix}. \tag{2.18}$$

If we write $\boldsymbol{P}_n = (\pi_0, \pi_1, \pi_2, \ldots, \pi_{n-1})^T$, then

$$x\boldsymbol{P}_n(x) = J_n \boldsymbol{P}_n(x) + b_n \pi_n(x) \boldsymbol{e}_n \tag{2.19}$$

where J_n is the upper-left n-by-n block of J and $(\boldsymbol{e}_n)_j = \delta_{nj}$ is the standard (column) basis vector. This makes it easy to compute the moments. We have $x\pi_n = J_{nk}\pi_k$ and

$$x\pi_0(x) = \sum_k J_{0k}\pi_k, \quad x^2\pi_0(x) = \sum_{k,\ell} J_{0k}J_{k\ell}\pi_\ell(x), \ldots, x^n\pi_0(x) = \sum_k [J^n]_{0k}\pi_k(x)$$

Polynomials	**Domain**	**Weight Fcn.**	a_n	b_n	**Moments** μ_n
Legendre	$[-1,1]$	1	0	$\sqrt{\frac{n^2}{(2n-1)(2n+1}}$	$\frac{1}{n/2}\chi_e(n)$
Chebyshev (1st kind)	$[-1,1]$	$2[1-x^2]^{-1/2}$	0	$\frac{1}{2}+\delta_{n1}/\sqrt{2}$	$\frac{1}{4^{(n/2)}}C_{n/2}\chi_e(n)$
Chebyshev (2nd kind)	$[-1,1]$	$4[1-x^2]^{1/2}$	0	$\frac{1}{2}$	$\frac{1}{4^{(n/2)}}\binom{n}{n/2}\chi_e(n)$
Laguerre	$[0,\infty)$	$2\pi e^{-x}$	$2n+1$	k	$(n+1)!$
Hermite	$(-\infty,\infty)$	$\sqrt{4\pi}e^{-x^2}$	0	$\sqrt{\frac{k}{2}}$	$\prod_{k=1}^{n}(2n-1)\chi_e(n)$

Table 2.1: Classical orthogonal polynomial systems and their corresponding weight functions and Lanczos coefficients from [10] or [9]. These differ between sources due to normalization conventions — the polynomials may be monic, orthonormal, or respect some other choice $\langle \pi_n, \pi_m \rangle = z(n)\delta_{nm}$. Here C_n are the Catalan numbers and $\chi_e(n)$ is 1 for n even and 0 for n odd.

so

$$\mu_n = \int x^n w(x)\,dx = \int \sum_k [J^n]_{0k}\pi_k(x)w(x)\,dx = [J^n]_{00} \tag{2.20}$$

since $\int \pi_k(x)w(x)\,dx = \langle 1, \pi_k \rangle = \delta_{0k}$, so long as the weight is normalized as $\int w(x) = 1$.

We have now shown one form that the Lanczos algorithm takes in the infinite-dimensional case, namely that it takes the infinite dimensional initial data of a weight function or distribution and translates it into discrete polynomials, which are often much easier to compute and work with. Given a weight function, Gram-Schmidt on the sequence $\{x^n\}$ produces a family of orthogonal polynomials. All classical families of orthogonal polynomials can be expressed in this way, and some of the most common are given in Table 2.1.

To give a taste of the utility of the infinite dimensional Lanczos algorithm, we will now discuss two classic topics: Gauss quadrature, and the Moment Problem. Ironically, each of these was studied as a subject in its own right before the finite-dimensional Lanczos algorithm was conceived of.

Gauss Quadrature

Quadrature is an efficient method of integrating functions against distributions [6]. The technique dates back to the 1800s, when to evaluate an integral which could not be solved analytically, one would have to evaluate the integrand by hand at many points. Gauss quadrature provides an optimal set of points at which to evaluate the integrand to get the quickest-converging estimation. Today, Gauss quadrature and its many variations are the standard way to numerically evaluate integrals. We will give a brief taste of this wide subject following [6], emphasizing the connection to the Lanczos algorithm and ending with an application in physics.

If $w(x)$ is a weight function as before, a **quadrature rule** of order N for a function f is a relation

$$\int_a^b f(x)w(x)\,dx = \sum_{j=1}^{N} w_j f(\theta_j) + R[f] \tag{2.21}$$

where the w_j are called the *weights*, the points θ_j are the *nodes* and $R[f]$ is a remainder function. One quadrature rule is Riemann integration, with evenly spaced nodes $\theta_j = a + j(b-a)/N$ and equal weights $w_j = (b-a)/N$. However, the convergence may be very slow, especially if the weight function (distribution) has many singularities. By using irregularly spaced nodes and unequal weights, one may speed up convergence significantly. The fundamental idea is to use Lanczos to find the 'best' approximation of $w(x)$ by N δ-functions:

$$w(x) \approx \sum_{j=1}^{N} w_j \delta(x-\theta_j). \tag{2.22}$$

(We make the word 'best' precise below.)

Suppose that we perform the Lanczos algorithm on $w(x)$ and find the first N polynomials $\boldsymbol{P}_N = (p_0, p_1, \ldots, p_{N-1})^T$ and the Jacobi matrix J_N. There is a close connection between the polynomial p_{n-1} and the eigenvalues of J_N. Suppose $\{\theta_j\}_{j=1}^N$ are the zeros of $p_{n-1}(x)$. Putting this into Eq. (2.19)

$$\theta_j \boldsymbol{P}_N(\theta_j) = J_N \boldsymbol{P}_N(\theta_j) + b_N p_N(\theta_j)\boldsymbol{e}_N = J_N \boldsymbol{P}_N(\theta_j),$$

so $\boldsymbol{P}_N(\theta_j)$ is an eigenvector of J_N with eigenvalue θ_j. Define Z_N as the unitary matrix whose columns are $\boldsymbol{z}_j := \boldsymbol{P}_N(\theta_j)/||\boldsymbol{P}_N(\theta_j)||$, so that $Z_N J_N Z_N^\dagger = \Theta_N = \mathrm{diag}(\theta_1, \ldots, \theta_N)$. As our notation suggests, the θ_j's become the nodes, and the eigenvectors will become the weights.

To see this, let's first consider a polynomial $f(x) = \sum_{n=0}^{N-1} f_n x^n$ of degree at most N and suppose $\int w(x)\,dx = 1$ is normalized. Then we change basis to the π_n's,

$$\int f(x)w(x)\,dx = \int \sum_n f_n x^n w(x)\,dx = \sum_n f_n \mu^n = \sum_n f_n [J^n]_{00} = [f(J)]_{00}, \tag{2.23}$$

where we have used Eq. (2.20). Then we go to the eigenbasis

$$\int f(x)w(x)\,dx = [f(J)]_{00} = [Z_N f(\Theta_N) Z_N^\dagger]_{00} = \sum_{j=1}^{n} |\boldsymbol{e}_0 \cdot \boldsymbol{z}_j|^2 f(\theta_j).$$

which is the formula for Gauss quadrature, provided we identify the eigenvalues θ_j of J_N as the nodes with weights

$$w_j := |\boldsymbol{z}^j \cdot \boldsymbol{e}_0|^2. \tag{2.24}$$

We have hence found the nodes and weights for Gauss quadrature and shown it is exact for polynomials up to order N. However, even for arbitrary functions, the convergence is very good and the error is controlled by the following theorem.

Theorem 5 (Theorem 6.3 of [6])**.** *Gauss Quadrature of order N is exact for polynomials of degree $2N-1$ and, if $f^{(2n)}(x).0$ for $a \leq x \leq b$, then the remainder is bounded by*

$$|R[f]| \leq (b_1^2 \cdots b_{N-1})^2 \sup_{x\in[a,b]} \frac{f^{(2N)}(x)}{(2N)!} \xrightarrow{N\to\infty} 0. \tag{2.25}$$

These conditions can be relaxed, but the essential point is that the quadrature converges quite quickly with N for arbitrary and even highly singular distributions $w(x)$. (There are also common variations on the Gauss quadrature rules so that the approximation to the integral is always above or below the true integral, so a tight error bound can be found on the true value [6].) It is somewhat unsurprising that polynomials up to degree $2N-1$ are correctly captured, since Eq.(2.19) tells us

$$\mu_{2n} = \int x^n w(x)\, dx = [J^{2n}]_{00} = ||J^n \boldsymbol{e}_0||^2,$$

so as long as $n < N$, the matrix multiplication doesn't spill outside of the first N dimensions and we can replace $J \to J_N$ exactly.

As a numerical technique, the procedure for Gauss quadrature is as follows. For a distribution $w(x)$, run Lanczos to step N and find the corresponding Jacobi matrix J_N and polynomials $\{p_n\}$, then compute the nodes θ_j and the weights w_j and use them in Eq. (2.21).

Analytically, however, perhaps the most interesting part of Gauss quadrature is Eq. (2.22); Gauss quadrature approximates a distribution as a sum of N δ-functions such that the first $2N$ moments are reproduced exactly. This deepens the results of Section 2.1. If we take our weight function to be the spectral density of an operator or matrix A, then N iterations of Lanczos *not only gives good estimates for the extremal eigenvalues (as we saw before), but also accurately captures the middle of the spectrum.* We will use this fact to great advantage in Chapter 3 to understand the off-diagonal distribution from ETH.

For now, though, let us give a quick application. Suppose $\widehat{H}$ is a Hamiltonian and we want to compute the time-evolution of a state

$$e^{i\widehat{H}t} \left|\psi\right\rangle .$$

Even if $\widehat{H}$ is too large to perform exact diagonalization and find all eigenstates (say, a spin chain with $L = 30$ spins), we can still perform a few dozen steps of the Lanczos algorithm, as that only requires cheap matrix-vector multiplication.

If we perform N Lanczos steps to find $K_N(\widehat{H}, |\psi\rangle)$, then we have the relation $\widehat{H}V_N = V_N T_N$, where T_N is an N-by-N matrix. Just as Eq. (2.22) is the best approximation of $w(x)$ by n δ-functions, this is the best approximation of H by N eigenvectors

$$\widehat{H} \approx V_N T_N V_N^\dagger.$$

So

$$e^{i\widehat{H}t} \left|\psi\right\rangle \approx e^{iV_N T_N V_N^\dagger t} \left|\psi\right\rangle = V_N e^{iT_N t} V_N^\dagger \left|\psi\right\rangle = V_N e^{iT_N t} \cdot \boldsymbol{e}_0. \tag{2.26}$$

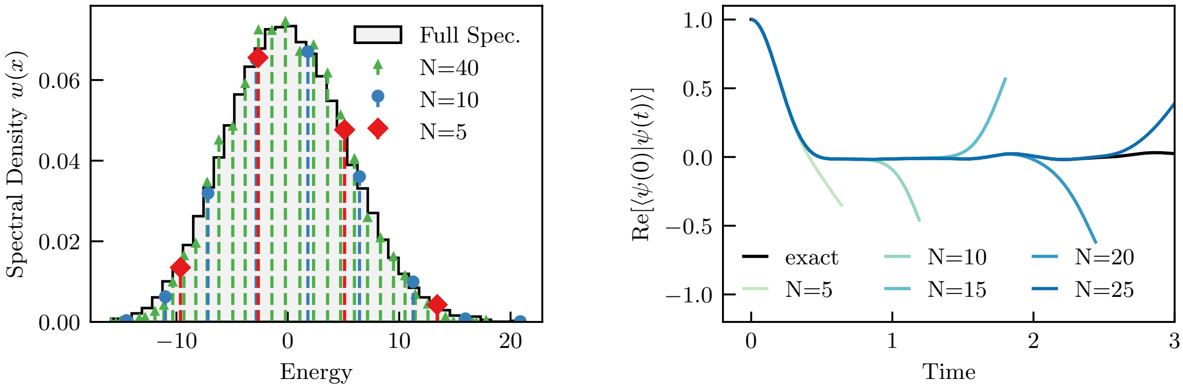


Figure 2.2: (Left) Comparison of the full spectral density of the Chaotic Ising model with its approximation as a sum of δ-functions via Eq. (2.22) in the Krylov space $K_N(\widehat{H};\psi)$. (Right) Time-evolution of a state $|\psi(t)\rangle = e^{-i\widehat{H}t}|\psi\rangle$ in the full Hilbert space and the Krylov spaces $K_N(\widehat{H};\psi)$. Throughout we take $L = 12$ (dimension 4096) with open boundary conditions, and $|\psi\rangle$ chosen randomly.

The convergence of this approximation is subtle, and beyond our scope. However, this will always work at least as well as expanding $e^{i\widehat{H}t}$ in a Taylor series and often far better.

As a concrete example, consider the Chaotic Ising model

$$\widehat{H} = \sum_{n=1}^{L} \sigma_n^z \sigma_{n+1}^z + h_x \sigma_n^x + h_z \sigma^z. \tag{2.27}$$

Fig. 2.2 shows how the distribution Eq. (2.22) approximates the full distribution through the whole spectrum and how Eq. (2.26) approximates the time-evolution for longer and longer times as the number of Lanczos steps N grows.

The Moment Problem

In this section, we touch upon the connection between Lanczos and the classical moment problem. The classical moment problem (due to Stieltjes in the 1890s and Hamburger in the 1920s [8]) is as follows: suppose $w(x)$ is a distribution on $\mathbb{R}$ whose moments $\{\mu_0, \mu_1, \mu_2, \ldots\}$ are all known. What is $w(x)$?

We will briefly sketch out a constructive solution. A complete treatment of the problem may be found in [11]. The idea is to use the moments to find the Lanczos coefficients, then use the Lanczos coefficients to find the nodes and weights to approximate the distribution in a convergent way.

First, given the first $2n$ moments $\{\mu_0, \ldots, \mu_{2n-1}\}$, one may compute the first n Lanczos coefficients $\{a_1, \ldots, a_n\}$ and $\{b_1, \ldots, b_n\}$. This is because running the Lanczos algorithm does not actually require the distribution w, but only the moments. To see this, suppose for induction that $\{a_k\}_{k=0}^n$ and $\{b_k\}_{k=0}^n$ are both known as we as corresponding polynomials $\pi_k(x) = \sum_{j=0}^{k-1} p_j^k x^j$ for $\{\pi_k\}_{k=0}^n$. So

$$a_n = \langle \pi_n, x\pi_n \rangle = \sum_{i,j=0}^{n-1} p_i^n p_j^n \langle x^i, x^{j+1} \rangle = \sum_{i,j=0}^{n-1} p_i^n p_j^n \mu_{i+j+1},$$

which is then entirely determined by the moments. Similarly, one may compute b_n in terms of the moments and, using Eq. (2.16), find π_{n+1}. This allows us to translate the first $2N$ moments into the first N of the a's and b's. We note that, computationally, this process is incredibly ill-conditioned; to compute b_n to N digits of precision, one needs $O(e^N)$ digits of precision on the μ_n's.

Equipped with the first n Lanczos coefficients, we can estimate the distribution as a sum of δ-functions as in Eq. (2.22). Let us express this distribution in terms of the Jacobi matrix. If $J_n = \sum_j \theta_j \boldsymbol{z}_j \boldsymbol{z}_j^\dagger$ is an eigendecomposition, then

$$w_n(x) := \sum_{j=1}^n |\boldsymbol{z}_j \cdot \boldsymbol{e}_0|^2 \delta(x - \theta_i) = \operatorname{Im} \boldsymbol{e}_0^T \cdot \sum_j \frac{\boldsymbol{z}_j \boldsymbol{z}_j^\dagger}{x - i\epsilon - \theta_j} \cdot \boldsymbol{e}_0 = \operatorname{Im} \left[\frac{1}{(x - i\epsilon)I - J_n} \right]_{00},$$

where we have used the Cauchy principle value formula $\frac{1}{x-\lambda} = \mathcal{P}\frac{1}{x-\lambda} + i\delta(x-\lambda)$ (i.e. this relation holds as a distribution). One may then write a formal solution

$$w(x) = \lim_{n\to\infty} w_n(x). \tag{2.28}$$

The question, then, is when this converges. This is a somewhat tricky question and requires some amount of careful analysis, which is beyond our scope. Historically, the problem was treated by expressing $w_n(x)$ as a continued fraction (which follow from expanding $[(x - J_n)^{-1}]_{00}$ via Kramer's rule, an idea we will return to in Chap. 3 below). Studying the convergence of the continued fractions leads to the following criterion

Theorem 6 (Hamburger; Thm. 6.1 of [8])**.** *The Hamburger moment problem has at least one solution if and only if the determinants*

$$\det[\mu_{i+j}]_{i,j=0}^n \tag{2.29}$$

are strictly positive for all n.

One can check that this criterion is exactly what is needed when computing the Lanczos coefficients from the moments for the denominators to be non-zero. However, even with this criterion, the solution to the Hamburger problem need not be unique. If there is a unique distribution, then the moments $\{\mu_n\}$ are called *determinate*, as they entirely determine the distribution. Otherwise, the moment problem is called *indeterminate* and there are in fact infinitely continuous and infinitely many discrete solutions [12]. One sufficient condition for the moment problem to be determinate is as follows.

Theorem 7 (Carleman's Condition [12]). *Suppose*

$$\lim_{N\to\infty}\sum_{k=1}^{N}\mu_{2k}^{-1/2k}=\infty \textit{ or equivalently } \lim_{N\to\infty}\sum_{k=1}^{N}\frac{1}{b_n}=\infty, \tag{2.30}$$

then the moment problem is determinate.

We can use this to give a physically useful result, which will be used many times below: the spectral function (closely related to the off-diagonal distribution in the Eigenstate Thermalization Hypothesis) is well-defined in the thermodynamic limit. Suppose that we have a spin chain with on-site dimension d, and $\widehat{H}$ is a k-body translationally-invariant Hamiltonian (i.e. each term acts on at most k sites at a time). Suppose $\widehat{\mathcal{O}}$ is a local, translationally-invariant operator. We can consider this as a vector in the space of operators, equipped with the Frobenius norm, and consider the superoperator $\mathcal{L}=[\widehat{H},\cdot]$. If $\widehat{H}\left|n\right\rangle=\epsilon_n\left|n\right\rangle$ are the eigenvectors in the thermodynamic limit, the spectral distribution is formally defined as

$$\Phi(\omega):=\sum_{n,m}|O_{nm}|^2\delta(\omega-\epsilon_{nm})$$

where $\epsilon_{nm}=\epsilon_n-\epsilon_m$ and $O_{nm}=\langle n|\widehat{\mathcal{O}}|m\rangle$. However, computationally, this is not a useful form to integrate against, because it would require exact diagonalization of the Hamiltonian in the thermodynamic limit.

We show in Chap. 3 that the moments

$$\mu_{2n}=(\widehat{\mathcal{O}}|\mathcal{L}^{2n}|\widehat{\mathcal{O}})$$

satisfy the bound

$$\mu_{2n}\le C(k)(2n)! \text{ or } b_n\le c(k)n$$

where $C(k)$ and $c(k)$ are positive constants that depend on k. Then

$$\sum_{k=1}^{N}\frac{1}{b_n}\ge c(k)\sum_{k=1}^{N}\frac{1}{n}\xrightarrow{N\to\infty}\infty,$$

so Carleman's condition is satisfied. The spectral function is then given by (2.28) and we can each integrate low-degree polynomials against it by working with the finite δ-function estimates w_n. In principle one can then compute any integral against the spectral distribution, although in practice the convergence may be too slow for this to be workable. Nevertheless, this allows the spectral distribution to be upgraded from a formal object to one which may be used computationally and numerically, directly in the thermodynamic limit.

In summary, we have presented the Lanczos algorithm in several different guises. In the finite-dimensional case, we have shown it produces an orthonormal basis for the Krylov spaces, and gives good estimates for the extremal eigenvectors and eigenvalues of a matrix. In the infinite dimensional case, we have shown that the Lanczos algorithm generates the family of orthogonal polynomials associated to a distribution. We have also demonstrated the close connection between the Lanczos algorithm and Gauss curvature and, finally, demonstrated how Lanczos solves the classical moment problem. In the next chapter we shall see the utility of these ideas in the context of quantum dynamics. We shall see how the asymptotic behavior of the Lanczos coefficients is closely connected to quantum chaos.

Chapter 3

A Universal Operator Growth Hypothesis

3.1 Introduction

The emergence of ergodic behavior in quantum systems is an old puzzle[1]. Quantum mechanical time-evolution is local and unitary, but many quantum systems are effectively described by irreversible hydrodynamics, involving familiar quantities such as electrical conductivity. Understanding this emergent thermal behavior at both a conceptual and computational level is a central goal of theoretical research on quantum dynamics, of which a cornerstone is the Eigenstate Thermalization Hypothesis [13, 14, 15, 16, 17].

Recent work has shifted focus from states to *operator growth* in many-body systems [18, 19, 20, 21, 22, 23]. Under Heisenberg-picture evolution, simple operators generically decay into an infinite "bath" of increasingly non-local operators. The emergence of this dissipative behavior from unitary dynamics is believed to be at the origin of thermalization, the decay of dynamical correlation functions, and the accuracy of hydrodynamics at large scales. This picture was recently confirmed in random unitary models of quantum dynamics [18, 19], and extended to increasingly realistic systems involving conservation laws [20, 21], Floquet dynamics [23], and even interacting integrable models [22].

While random unitary models are valuable proxies for studying operator growth, one would like to confirm this picture in genuine Hamiltonian systems. In semiclassical systems, a quantitative measure is provided by the out-of-time-order correlation function (OTOC). The classical butterfly effect gives rise to an exponential growth of the OTOC, characterized by the Lyapunov exponent, which may be computed in a variety of models. It is conjectured that the Lyapunov exponent is bounded [24] and this bound is achieved in certain large-N strongly interacting models with a classical gravity dual, such as the Sachdev-Ye-Kitaev (SYK) model [25, 26, 27]. Unfortunately, the OTOC does not necessarily exhibit expo-

[1] This chapter is mainly drawn from [1], which is joint work with Xiangyu Cao, Alexander Avdoshkin, Thomas Scaffidi, and Ehud Altman

nential growth outside of semiclassical or large-N limits, rendering the Lyapunov exponent ill-defined [28, 29, 21, 30]. A general theory of operator growth under generic, non-integrable Hamiltonian dynamics is, therefore, still lacking.

The amount of information required to describe a growing operator increases exponentially in time. Computationally, this bars the exact calculation of operators at long times. Yet, the exponential size of the problem has a positive aspect: it acts as a thermodynamic bath, so a statistical description should emerge and become nigh-exact. This idea indicates operator growth should be governed by some form of universality. In this work we present a hypothesis specifying universal properties of growing operators in non-integrable quantum systems in any dimension.

3.2 Synopsis

Our hypothesis has a simple formulation in the framework of the continued fraction expansion or *recursion method*, which we review in Section 3.3. This is a well-understood technique, dating back to the 1980s [31], and has recently been used to compute conductivities in strongly-interacting systems [32, 33, 34]. It is surveyed in great detail in Ref. [10]. Essentially, it converts any linear-response calculation to the problem of a quantum particle on a half chain, with the hopping matrix elements given by the Lanczos coefficients b_n. Section 3.4 presents our hypothesis: operators in generic, non-integrable systems have Lanczos coefficients with asymptotically linear growth with n, suppressed by a logarithmic correction in one dimension. The linear growth rate, denoted α, is the central quantity of this work. It has dimensions of energy and can be bounded from above by the local bandwidth [see (3.16) and (3.23)]. In light of this, the hypothesis essentially asserts that the Lanczos coefficients grow as fast as possible in non-integrable systems. Although we are unable to prove the hypothesis rigorously, we shall support it with extensive numerical evidence, calculations in SYK models, and general physical arguments in Section 3.4. In particular, the hypothesis is equivalent to the exponential decay of the spectral function at high frequency, which can be (and has been) observed experimentally [35, 36, 37].

We explore several consequences of the hypothesis. In Section 3.5, we develop a precise picture of the universal growth of operators. We show that under the hypothesis, the 1d quantum mechanics, governed by the Lanczos coefficients $b_n \sim \alpha n$, captures the irreversible process of simple operators evolving into complex ones. Furthermore, the 1d wavefunction delocalizes exponentially fast on the n axis, at a rate exactly given by α. Asymptotically, the expected position of the 1d wavefunction satisfies

$$(n)_t \sim e^{2\alpha t} . \tag{3.1}$$

The expectation value $(n)_t$ has a succinct interpretation as an upper bound for a large class of operator complexity measures called "q-complexities", which we define in section 3.5. Crucially, this class includes out-of-time-order correlators. This allows us to establish a quantitative connection between α and the Lyapunov exponent, which will be the subject of Section 3.6. We show for quantum systems at infinite temperature that the growth rate gives an upper bound for the Lyapunov exponent whenever the latter is well-defined:

$$\lambda_L \leq 2\alpha\,. \tag{3.2}$$

For classical systems, this statement is a conjecture but it is posible to prove a somewhat weaker bound. We check (3.2) in the SYK model and a classical tops model, and find it to be tight in both cases.

A further application of the hypothesis, discussed in Section 3.7, is a semi-analytical technique to compute diffusion coefficients of conserved quantities. We leverage the hypothesis to extend classical methods of the continued fraction expansion to directly compute the pole structure of the Green's function, thus revealing the dispersion relation of the dynamics.

Section 3.8 discusses the generalization to finite temperatures, which involves many open questions. Nevertheless, we show that the universal bound on chaos $\lambda_L \leq 2\pi k_B T/\hbar$ [24] can be implied and improved by a proper finite-temperature extension of the bound (3.2), and provide evidence supporting this conjecture. We conclude in Section 3.9 by discussing conceptual implications of our results and perspectives for future work.

3.3 Preliminaries: The Recursion Method

We briefly review the recursion method in order to state the hypothesis. A comprehensive treatment may be found in [10]. Consider a local Hamiltonian H and fix a Hermitian operator $\widehat{\mathcal{O}}$. We regard the operator as a state $|\widehat{\mathcal{O}})$ in the Hilbert space of operators, endowed with the infinite-temperature inner product $(\widehat{\mathcal{O}}_1|\widehat{\mathcal{O}}_2) := \mathrm{Tr}[\widehat{\mathcal{O}}_1^\dagger \widehat{\mathcal{O}}_2]/\,\mathrm{Tr}[1]$. We write $||\widehat{\mathcal{O}}|| := (\widehat{\mathcal{O}}|\widehat{\mathcal{O}})^{1/2}$ for the norm. We will focus on systems in the thermodynamic limit.

Just as states evolve under the Hamiltonian operator, operators evolve under the Liouvillian superoperator $\mathcal{L} := [H, \cdot]$. Our central object is the autocorrelation function

$$C(t) = \mathrm{Tr}[\widehat{\mathcal{O}}(0)\widehat{\mathcal{O}}(t)]/\,\mathrm{Tr}[1] = (\widehat{\mathcal{O}}|\exp{(i\mathcal{L}t)}\,|\widehat{\mathcal{O}}), \tag{3.3}$$

where the second equality follows from Baker-Campbell-Hausdorff.

Computing $C(t)$ is inherently difficult. Suppose $\widehat{\mathcal{O}}(t=0)$ is a relatively simple operator that can be written as the sum of a few basis vectors in any local basis [2]. As the spatial support of $\widehat{\mathcal{O}}(t)$ grows, the number of non-zero coefficients of $\widehat{\mathcal{O}}(t)$ in any local basis can blow up exponentially. To make progress, one must compress this information. Intuitively, there are so many basis vectors at a given spatial size or "complexity" that we can think of them as a thermodynamic bath; no single basis vector has much individual relevance, only their statistical properties are important. In this interpretation, the operator flows though a series of "operator baths" of increasing size. The dynamics of an operator is then reduced to

[2] A local basis in, say, a spin chain is any basis related to the basis of Pauli strings by a finite-depth local unitary circuit.

how the baths are connected — a much simpler problem. In particular, the second law then dictates that an operator eventually flows to the largest possible baths, running irreversibly away from small operators. This is shown schematically in Fig. 3.1.

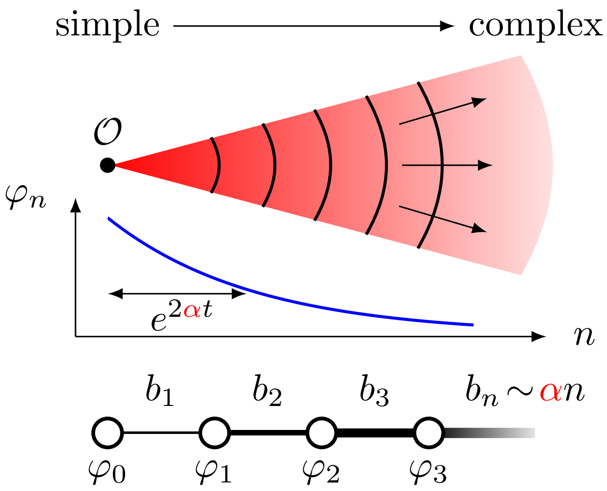


Figure 3.1: Artist's impression of the space of operators and its relation to the 1d chain defined by the Lanczos algorithm starting from a simple operator $\mathcal{O}$. The region of complex operators corresponds to that of large n on the 1d chain. Under our hypothesis, the hopping amplitudes b_n on the chain grow linearly asymptotically in generic thermalizing systems (with a log-correction in one dimension, see Section 3.4). This implies an exponential spreading $(n)_t \sim e^{2\alpha t}$ of the wavefunction φ_n on the 1d chain, which reflects the exponential growth of operator complexity under Heisenberg evolution, in a sense we make precise in Section 3.5. The form of the wavefunction φ_n is only a sketch; see Fig. 3.5 for a realistic picture.

We now quantify this idea precisely. This is done by applying the Lanczos algorithm, which iteratively computes a tridiagonal representation of a matrix. The idea is to find the sequence $\{\mathcal{L}^n |\widehat{\mathcal{O}})\}$, and then apply Gram-Schmidt to orthogonalize. Explicitly, start with a normalized vector $|\widehat{\mathcal{O}}_0) := |\widehat{\mathcal{O}})$. As a base case, let $|\widehat{\mathcal{O}}_1) := b_1^{-1}\mathcal{L}\,|O_0)$ where $b_1 := (\widehat{\mathcal{O}}_0\mathcal{L}|\mathcal{L}\widehat{\mathcal{O}}_0)^{1/2}$. Then inductively define

$$\begin{aligned} |A_n) &:= \mathcal{L}\,|\widehat{\mathcal{O}}_{n-1}) - b_{n-1}\,|\widehat{\mathcal{O}}_{n-2})\ , \\ b_n &:= (A_n|A_n)^{1/2}\,, \\ |\widehat{\mathcal{O}}_n)\ &:=\ b_n^{-1}\,|A_n)\ . \end{aligned} \tag{3.4}$$

The output of the algorithm is a sequence of positive numbers, $\{b_n\}$, called the *Lanczos coefficients*, and an orthonormal sequence of operators, $\{|\widehat{\mathcal{O}}_n)\}$, called the *Krylov basis*. (This is a bit of a misnomer, as the Krylov basis spans *an* operator space containing $\widehat{\mathcal{O}}(t)$ for any t, but does not usually span the full space of operators). The Liouvillian is tridiagonal in this basis:

$$L_{nm} := (\widehat{\mathcal{O}}_n|\mathcal{L}|\widehat{\mathcal{O}}_m) = \begin{pmatrix} 0 & b_1 & 0 & 0 & \cdots \\ b_1 & 0 & b_2 & 0 & \cdots \\ 0 & b_2 & 0 & b_3 & \cdots \\ 0 & 0 & b_3 & 0 & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}. \tag{3.5}$$

We make four remarks. First, if the operator Hilbert space is d-dimensional with d finite (or if the subspace spanned by $|\widehat{\mathcal{O}}_0), |\widehat{\mathcal{O}}_1), |\widehat{\mathcal{O}}_2), \dots$ is so), the algorithm will halt at $n = d+1$: in this work, we work always in the thermodynamic limit and discard this non-generic situation. Second, the Lanczos algorithm presented here is adapted to operator dynamics. Generally, a tridiagonal matrix will have non-zero diagonal entries, but they vanish in (3.5). This is because one can inductively show that $i^n \widehat{\mathcal{O}}_n$ is Hermitian for all n, hence $(\widehat{\mathcal{O}}_n|\mathcal{L}|\widehat{\mathcal{O}}_n) = 0$. Third, the knowledge of the Lanczos coefficients $b_1, \dots, b_n$ is equivalent to that of the *moments* $\mu_2, \mu_4, \dots, \mu_{2n}$, defined as the Taylor series coefficients of the correlation function

$$\mu_{2n} := (\widehat{\mathcal{O}}|\mathcal{L}^{2n}|\widehat{\mathcal{O}}) = \frac{d^{2n}}{dt^{2n}} C(t)\big|_{t=0} \tag{3.6}$$

The non-trivial transformation between the Lanczos coefficients and the moments is reviewed in Appendix 3.A. Fourth, the Lanczos coefficients have units of energy.

In the Krylov basis, the correlation function $C(t)$ is:

$$C(t) = \left(e^{iLt}\right)_{00}. \tag{3.7}$$

Hence the autocorrelation depends only on the Lanczos coefficients, and not on the Krylov basis. One way to interpret the Lanczos coefficients, which we will employ extensively below, is as the hopping amplitudes of a semi-infinite tight-binding model — see Fig. 3.1. The wavefunction on the semi-infinite chain is defined as $\varphi_n(t) := i^{-n}(\widehat{\mathcal{O}}_n|\widehat{\mathcal{O}}(t))$. Heisenberg evolution of $\widehat{\mathcal{O}}(t)$ becomes a discrete Schrödinger equation:

$$\partial_t \varphi_n = -b_{n+1}\varphi_{n+1} + b_n \varphi_{n-1}, \quad \varphi_n(0) = \delta_{n0}. \tag{3.8}$$

where $b_0 = \varphi_{-1} = 0$ by convention. The autocorrelation is simply $C(t) = \varphi_0(t)$, so the Lanczos coefficients are completely equivalent to the autocorrelation function.

Just as different bases are well-suited for particular computations, a number of equivalent represention of the autocorrelation function appear in this work, namely the Green's function

$$G(z) = \left(\widehat{\mathcal{O}}\left|\frac{1}{z-\mathcal{L}}\right|\widehat{\mathcal{O}}\right) = i\int_0^\infty e^{-izt} C(t)\, dt. \tag{3.9}$$

and the spectral function

$$\Phi(\omega) \ = \ \int_{-\infty}^{\infty} C(t) e^{-i\omega t}\, dt\,. \tag{3.10}$$

In summary, we have reviewed five equivalent ways to describe the dynamics

$$C(t) \leftrightarrow G(z) \leftrightarrow \Phi(\omega) \leftrightarrow \{\mu_{2n}\} \leftrightarrow \{b_n\} \tag{3.11}$$

Just as with a choice of basis, we shall use the most convenient representation for the task at hand and translate freely between them. We note that $\{b_n\}$ is special in the sense that it is a non-linear representation of the autocorrelation while all other representations are linearly related. We provide the details on the mapping to b_n in Appendix 3.A, with a particular focus on asymptotic properties.

The nonlinearity involved in $\{b_n\}$ also makes them more abstract. Intuitively, we can think of the Krylov basis $\{\widehat{\mathcal{O}}_n\}$ as stratifying operators by their 'complexity' (with respect to the initial operator $\widehat{\mathcal{O}}$), and b_n's describe how operators of different complexities transform into one another. The goal of this work is to study aspects of operator growth that can be reduced to the quantum mechanics on this semi-infinite chain.

3.4 The Hypothesis

We now state the hypothesis. Informally, *in a chaotic quantum system, the Lanczos coefficients $\{b_n\}$ should grow as fast as possible.* The maximal possible growth rate turns out to be linear (with logarithm corrections in 1D). Our precise statement is therefore as follows. Suppose that H describes an infinite, non-integrable [3], many-body system in dimension $d > 1$ and $\widehat{\mathcal{O}}$ is a local operator having zero overlap with any conserved quantity (in particular, $(\widehat{\mathcal{O}}|H) = 0$). Then the Lanczos coefficients are asymptotically linear:

$$b_n = \alpha n + \gamma + \mathrm{o}(1), \tag{3.12}$$

for some real constants $\alpha > 0$ and γ. This linear growth is an example of universality. We will refer to α as the *growth rate*, and it will play a multitude of roles. In fact, it quantitatively captures the growth of "operator complexity" in a precise sense (Section 3.5). On the other hand, it is observable by standard linear response measures (Section 3.4). This section first describes why linear growth is maximal, amasses a weight of evidence in favor of the hypothesis, and finally discusses the special case of one dimension.

We note that the idea of classifying operator dynamics by Lanczos coefficients asymptotics is as old as the recursion method itself. Many examples have been explored, resulting in a broad zoology, as surveyed in [10]. In particular, it is known that non-interacting models (such as lattice free fermions) give rise to a *bounded* sequence $b_n \sim \mathrm{O}(1)$. If we start with a two-body operator $\widehat{\mathcal{O}}$ in such free models, all $\widehat{\mathcal{O}}_n$'s will remain two-body. In this sense, the operator dynamics is simple. In this work, we focus on the opposite extreme of generic chaotic dynamics. To our knowledge, the ubiquity of asymptotically linear growth in these

[3] As a working definition, we say that a system is integrable if it has an extensive number of quasi-local conserved quantities.

Asymptotic	**Growth Rate**	**System Type**
$b_n \sim O(1)$	constant	Free models
$b_n \sim O(\sqrt{n})$	square-root	Integrable models
$b_n \sim \alpha n$	linear	Chaotic models
$b_n \gtrsim O(n)$	superlinear	Disallowed

Table 3.1: Asymptotic behavior of Lanczos coefficients. The first and last rows are known rigorously, while the middle two are conjectures supported by many specific examples. We will see below that the conjecture for chaotic models is slightly modified in $1+1d$.

systems and its consequences have not been systematically studied in quantum systems. Interacting models with obstructions to thermalization (e.g., integrable systems) lead to more involved behaviors, which have not been thoroughly explored. Nevertheless, a square root behavior $b_n \sim \sqrt{n}$ has been observed in a few examples ([10, 38], see also Fig. 3.2). The situation is summarized in Table 3.1 and a number of examples either from our work or previous literature are given in Table 3.2.

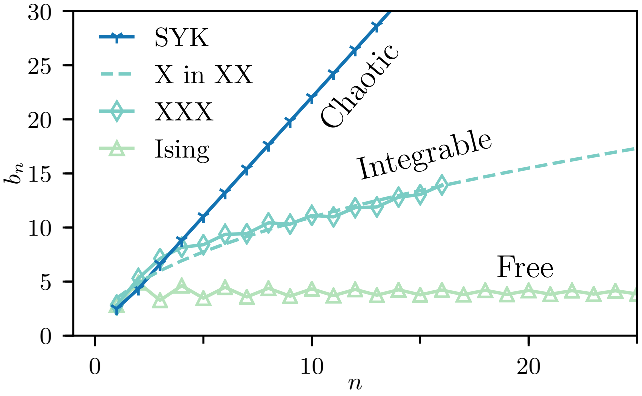


Figure 3.2: Lanczos coefficients in a variety of models demonstrating common asymptotic behaviors. “Ising” is $H = \sum_i X_i X_{i+1} + Z_i$ with $\widehat{\mathcal{O}} = \sum_j e^{iq_j} Z_j$ ($q = 1/128$ here and below) and has $b_n \sim O(1)$. “X in XX” is $H = \sum_i X_i X_{i+1} + Y_i Y_{i+1}$ with $\widehat{\mathcal{O}} = \sum_j X_j$, which is a string rather than a bilinear in the Majorana fermion representation, so this is effectively an interacting integrable model that has $b_n \sim \sqrt{n}$. XXX is $H = \sum_i X_i X_{i+1} + Y_i Y_{i+1} + Z_i Z_{i+1}$ with $\widehat{\mathcal{O}} = \sum_j e^{iq_j}(X_j Y_{j+1} - Y_j X_{j+1})$ that appears to obey $b_n \sim \sqrt{n}$. Finally, SYK is (3.18) where $q = 4$ and $J = 1$ and $\widehat{\mathcal{O}} = \sqrt{2}\gamma_1$ with $b_n \sim n$. The Lanczos coefficients have been rescaled vertically for display purposes. Numerical details are given in Appendices 3.B and 3.C.

Model	**Op.**	**Dynamics**	**Lanczos**	**Evidence**	**Ref.**
Ising	$\widehat{Z}$	Free	$O(1)$	Analytic	[10]
XX	$\widehat{Z}$	Free	$O(1)$	Analytic	[10]
SYK$^{(2)}$	γ	Free	$O(1)$	Analytic	[24]
XX	$\widehat{X}$	Free*	$O(\sqrt{n})$	Analytic	[10]
Free Fermions in Disguise	$\widehat{Z}$	Free*	$O(\sqrt{n})$	Numerical	[39]
MBL	$\widehat{Z}$	Int.	$O(\sqrt{n})$	Numerical	
XXZ	$\widehat{Z}$	Int.	$O(\sqrt{n})$	Numerical	Fig. 3.2, [40]
Chaotic Ising	$\widehat{Z}$	Chaotic	$O(n)$	Numerical	Fig. 3.2
XXZ + NNN	$\widehat{Z}\widehat{Z}$	Chaotic	$O(n)$	Numerical	[40]
SYK$^{(4)}$	γ	Chaotic	$O(n)$	Numerical	Fig. 3.2
SYK$^{(\infty)}$	γ	Chaotic	$O(n)$	Analytic	[41]
SYK Hopping	γ	Chaotic	$O(n)$	Analytic	
2D Fermi Hubbard	$\widehat{J}$	Chaotic	$O(n)$	Numerical	[42]
Bouch Model	$\widehat{X}$	Chaotic	$O(n)$	Analytic	[43]

Table 3.2: Examples of free/integrable/chaotic models where the Lanczos coefficients have constant/square-root/linear growth. The dynamics of the model are determined by level statistics and "Free*" means the model is free, but the operator in question does not have a simple or local representation in the free particle description. Analytical evidence means either an exact formula for the coefficients is available or analytical bounds are used to deduced the asymptotics. Numerical evidence means that the Lanczos coefficients were computed numerically and appear to have a clear trend.

Upper Bounds

We start by showing that linear growth is the maximal possible growth of the Lanczos coefficients. This is most easily done starting with the spectral function. In interacting many-body systems, the spectral function has a tail extending to arbitrarily high frequencies. The asymptotic behavior of the tail is directly related to the Lanczos coefficients, with faster growth of Lanczos coefficients corresponding to slower decay of $\Phi(\omega)$. The precise asymptotic behavior is [44, 45]

$$b_n \sim n^\delta \Longleftrightarrow \Phi(\omega) \sim \exp(-|\omega/\omega_0|^{1/\delta}) \tag{3.13}$$

for any $\delta > 0$ and some constant ω_0. In particular, $\delta = 1$ corresponds to asymptotically linear Lanczos coefficients and an exponentially decaying spectral function.

The decay of the spectral function is constrained by a powerful bound. A rigorous and general result of Refs [46] (see also [47, 48, 49], and Appendix 3.F for a self-contained proof) is that, given an r-local lattice Hamiltonian $H = \sum_i h_i$ in any dimension,

$$\Phi(\omega) \leq Ce^{-\kappa|\omega|}, \ \kappa = \frac{1}{2eG_r||h_i||} \tag{3.14}$$

for some $C > 0$ and a known O(1) geometrical factor G_r. We may conclude $\delta \leq 1$ in (3.13), so the Lanczos coefficients grow at most linearly.

When linear growth of the b_n's is achieved, the growth rate α is quantitatively related to the exponential decay rate in the spectral function. In fact, Appendix 3.A shows the following asymptotics are equivalent (see Fig. 3.3):

$$b_n = \alpha\, n + \mathrm{O}(1)\,, \tag{3.15a}$$

$$\Phi(\omega) = e^{-\frac{|\omega|}{\omega_0} + \mathrm{o}(\omega)}, \ \omega_0 = \frac{2}{\pi}\alpha, \tag{3.15b}$$

We stress that this is a purely mathematical equivalence, which holds independently of physical considerations such as the dimension, the temperature, or even if the system is quantum or classical. However, this equivalence has a key physical consequence: it implies that α is observable in linear response measurements. In fact, high-frequency power spectra for quantum spin systems can be measured with nuclear magnetic resonance, and exponential decays were reported for CaF_2 [35, 36, 37]. This experimental technique therefore provides a practical way of measuring α. On a theoretical note, the spectral function also appears in the off-diagonal Eigenstate Thermalization Hypothesis, which is therefore related to our hypothesis.

Additionally, comparing (3.14) and (3.15) shows that $\alpha \leq \pi/2\kappa$, so the growth rate is limited by the local bandwidth of the model and the geometry:

$$\alpha \leq \pi e G_r ||h_i||\,, \tag{3.16}$$

c.f. (3.14). This inequality is the consequence of the natural energy scale for the Lanczos coefficients being set by the local bandwidth. However, we shall see that α itself is not merely the bandwidth, but contains a great deal of physical information about the system.

We find it useful to dispel a possible misconception related to the high-frequency tail of the spectral function $\Phi(\omega)$. On dimensional grounds it is tempting — though ultimately erroneous — to interpret (3.15) as a statement about the short-time behavior of $C(t)$. To see why this is wrong, notice that the short-time behavior is captured by the first moment alone, as $C(t) = 1 - \mu_2 t^2/2 + O(t^4)$. The high-frequency information instead governs the asymptotics of moments μ_{2n} as $n \to \infty$ (which involve increasingly large operators) and the analytical structure of $C(t)$ on the imaginary-t axis, as shown in Fig. 3.3. In particular, the exponential decay rate sets the location of the closest pole to the origin on the imaginary axis. The high-frequency information also does not control the large time limit $t \to +\infty$; we will come back to this point in Section 3.7 below. In brief, the hypothesis governs large ω behavior of $\Phi(\omega)$ and, correspondingly, the behavior of $C(t)$ on the *imaginary* axis. Explicitly, a growth rate of α gives rise to a singularity at

$$t = \pm\frac{i\pi}{2\alpha}\,. \tag{3.17}$$

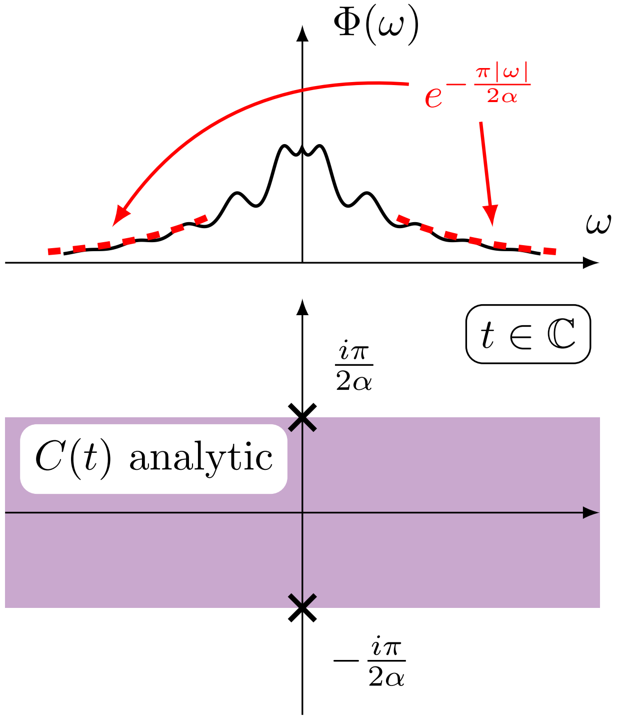


Figure 3.3: Illustration of the spectral function and the analytical structure of $C(t), t \in \mathbb{C}$. When the Lanczos coefficients have linear growth rate α, $\Phi(\omega)$ has exponential tails $\sim e^{-|\omega|/\omega_0}$ with $\omega_0 = 2\alpha/\pi$; $C(t)$ is analytical in a strip of half-width $1/\omega_0$ and the singularities closest to the origin are at $t = \pm i/\omega_0$. See Appendix 3.A for further discussion.

Analytical Evidence

The upper bounds of the previous section show that the Lanczos coefficients cannot grow faster than linearly. We now show that this bound is tight through two analytic examples.

It is an ironic point that the assumptions for the hypothesis (3.12) fail in virtually all known solvable models, as those are often integrable, or even non-interacting. This explains why, to the best of our knowledge, linear growth was not recognized in any of the extensive literature on the recursion method as a universal behavior (except for certain classical systems [50]). However, there is one solvable model where we can compute the linear behavior analytically: the SYK model (see, e.g. [26, 27, 25]). Its Hamiltonian is

$$H_{\mathrm{SYK}}^{(q)} = i^{q/2} \sum_{1\le i_1<i_2<\cdots<i_q\le N} J_{i_1\ldots i_q}\gamma_{i_1}\gamma_{i_2}\cdots\gamma_{i_q} \tag{3.18}$$

where the γ_i's, with $1 \le i \le N$, are Majorana fermions with anti-commutators $\{\gamma_i, \gamma_j\} = \delta_{ij}$, and the $J_{i_1\ldots i_q}$'s are disordered couplings drawn from a Gaussian distribution with mean zero and variance $(q-1)!J^2/N^{q-1}$. We study the dynamics of a single Majorana $\widehat{\mathcal{O}} =$

$\sqrt{2}\gamma_1$ [41]. To leverage the SYK solvability, we shall compute the moments $\mu_{2n} = (\widehat{\mathcal{O}}|\mathcal{L}^{2n}|\widehat{\mathcal{O}})$, averaged over disorder in the large-N limit. For any finite q, the moments can be computed efficiently, thanks to the well-known large-N Schwinger-Dyson type equations satisfied by the correlation functions. The self-averaging properties of the SYK model allows the typical Lanczos coefficients to be computed from the averaged moments via a general numerical procedure [10]. This is described in detail in Appendix 3.B.

We find that the Lanczos coefficients follow the universal form (3.12) quite closely, as shown in Fig. 3.4(a). In the large-q limit, there is a closed form expression for the coefficients, computed in Appendix 3.B:

$$b_n^{\text{SYK}} = \begin{cases} \mathcal{J}\sqrt{2/q} + \mathrm{O}(1/q) & n = 1 \\ \mathcal{J}\sqrt{n(n-1)} + \mathrm{O}(1/q) & n > 1\,, \end{cases} \tag{3.19}$$

where $\mathcal{J} = \sqrt{q}\,2^{(1-q)/2}J$. Therefore in the large-q limit, the SYK model follows the universal form (3.12) with $\alpha = \mathcal{J}$. We may conclude that our hypothesis is obeyed in a canonical model of quantum chaos and that the upper bound of linear growth of the Lanczos coefficients is tight.

The SYK model is quite unusual in several respects: it is a disordered, large-N model in zero dimensions. However, none of these special features are required to achieve linear growth. To demonstrate this we turn to a model studied in the mathematical literature, defined on the 2d square lattice [43]:

$$H = \sum_{x,y} X_{x,y}Z_{x+1,y} + Z_{x,y}X_{x,y+1} \tag{3.20}$$

where X and Z are the normal Pauli matrices. A theorem [43] states that the moments of the operator $X_{0,0}$ grow as

$$\mu_{2n} = n^{2n}e^{O(n)} \tag{3.21}$$

which implies that the Lanczos coefficients grow linearly (see Appendix 3.A for translation between asymptotics). Thus linear growth (3.12) is a tight-upper bound for the growth of the Lanczos coefficients in dimensions greater than one for "realistic" spin models. The content of our hypothesis is that achieving this upper bound is generic in chaotic systems.

The Special Case $d = 1$

We now turn to the special case of one dimensional systems. Let us first present some numerical evidence. Fig. 3.4(a) shows the Lanczos coefficients for a variety of spin models in the thermodynamic limit. (Numerical details are given in Appendix 3.C.) One can clearly see that the asymptotic behavior still *appears* linear whenever the model is non-integrable. There is often an onset period before the universal behavior sets in; the first few Lanczos coefficients are highly model-dependent. We have observed that the more strongly-interacting

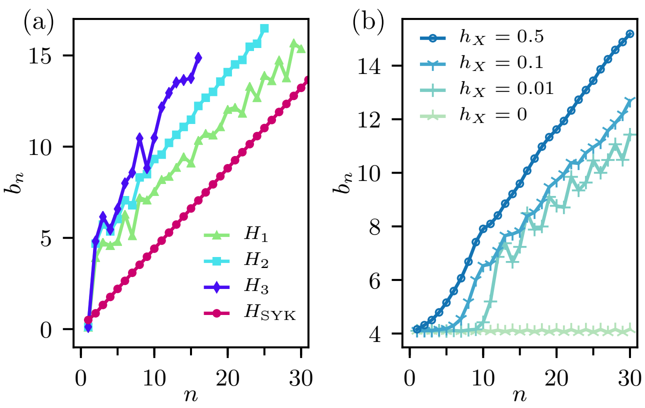


Figure 3.4: *(a)* Lanczos coefficients in a variety of strongly interacting spin-half chains: $H_1 = \sum_i X_i X_{i+1} + 0.709 Z_i + 0.9045 X_i$, $H_2 = H_1 + \sum_i 0.2 Y_i$, $H_3 = H_1 + \sum_i 0.2 Z_i Z_{i+1}$. The initial operator $\widehat{\mathcal{O}}$ is energy density wave with momentum $q = 0.1$. *(b)* Cross-over to apparently linear growth as interactions are added to a free model. Here $H = \sum_i X_i X_{i+1} - 1.05 Z_i + h_X X_i$, and $\widehat{\mathcal{O}} \propto \sum_i 1.05 X_i X_{i+1} + Z_i$. The b_n's are bounded when $h_X = 0$ but appears to have asymptotically linear growth for any $h_X \neq 0$. Logarithmic corrections are not clearly visible in the numerical data. Numerical details are given in Appendix 3.C.

the system, the sooner universal behavior appears [4]. Fig. 3.4(b) shows the robustness of this asymptotic behavior. The pure transverse field Ising model may be mapped to free fermions so, as expected, the Lanczos coefficients are bounded. But as soon as a small integrability-breaking interaction is added, the coefficients appear to become asymptotically linear, and the asymptotic behavior sets in at smaller n as the strength of the interaction increases. This is reminiscent of the crossover from Poisson to GOE distributed level statistics as integrability is broken [51, 52]. Observe also that the slope of the asymptotic growth depends only weakly on the (integrability breaking) interaction strength. This seems to be a general phenomenon, as it occurs also in the SYK model plus two body interactions, see Fig. 3.B.1 for details.

The numerical evidence is apparently compatible with linear growth of the Lanczos coefficients in 1d — but only apparently. We can see this by considering the singularity structure of the correlation function. When the Lanczos coefficients achieve linear growth, there is a singularity in $C(t)$ on the imaginary axis, given by Eq. (3.3). However, there is a classical theorem [53] which says, roughly, that $C(t)$, $t \in \mathbb{C}$, is *entire* for any local system in one dimension. Lanczos coefficients, therefore, must have strictly sublinear growth in one dimension. We note that this is an entirely geometric constraint, and has been previously noted by several works in a variety of contexts [46, 49], and derive it from first principles in Appendix 3.F.

[4] This is quite fortuitous, computationally: as a general rule, in more strongly interacting systems, exponentially more parameters are required to compute a given b_n, so fewer b_n's may be computed overall.

To formulate the hypothesis in one-dimension, we return to the informal version: *the Lanczos coefficients should grow as fast as possible.* More concretely, the Lanczos coefficients should achieve the upper-bound imposed by the geometry. Following [43], we compute this bound in Appendix 3.F and can therefore formulate the hypothesis as follows. Suppose H describes an infinite, non-integrable, many-body system in dimension d and $\widehat{\mathcal{O}}$ is a local operator having zero overlap with any conserved quantity. Then the asymptotic behavior of the Lanczos coefficients is

$$b_n = \begin{cases} A\frac{n}{W(n)} + \mathrm{o}(n/\ln n) \sim A\frac{n}{\ln n} + \mathrm{o}(n/\ln n) & d = 1 \\ \alpha n + \gamma + \mathrm{o}(1) & d > 1 \end{cases} \tag{3.22}$$

for some constants α, γ, A and W is the Lambert W-function which is defined by the implicit equation $z = W(ze^z)$ and has the asymptotic $W(n) = \ln n - \ln\ln n + o(1)$. In other words, the hypothesis acquires a logarithmic correction in one dimension. The coefficient A, like the growth rate α, has dimensions of energy and can be bounded above by the local bandwidth; for Hamiltonians with nearest-neighbor local term h_x, we have (see Appendix 3.F)

$$A \leq \frac{4}{e}||h_x|| \,. \tag{3.23}$$

We note that, unlike in higher dimensions, we are not aware of any analytic examples which achieve the maximal growth rate in 1D, leaving open the possibility that the first line of (3.22) is an over-estimate.

In some sense, the linear growth "barely breaks" in one dimension; the Lanczos coefficients can still grow faster than $b_n \sim n^\delta$ for any $\delta < 1$. The phenomenological difference between linear growth in all dimensions and (3.22) is often slight — such as in Fig. 3.4. Indeed, resolving logarithmic corrections in numerical data is a hard problem that often requires several decades of scaling. Altogether, we see that there is a subtle logarithmic correction to the operator growth hypothesis in one dimension.

3.5 Exponential Growth of Complexities

Now that we have presented evidence in favor of the hypothesis, we shall turn to the analysis of its consequences. In this section we study the universal behavior of operators which have linear growth of Lanczos coefficients with rate α. This is done in two steps. First, by studying the quantum mechanics problem (3.8) on the semi-infinite chain, we show that α measures the rate of exponential growth in operator complexity, in a sense we shall make precise below. Second, we prove that α gives an upper bound on a large class of operator complexity measures. Finally we shall remark on the case of linear growth with log-corrections.

We remark that our notion of complexity is *prima facie* distinct from other notions bearing the same name, such as circuit complexity (see the reviews [54, 55] and references

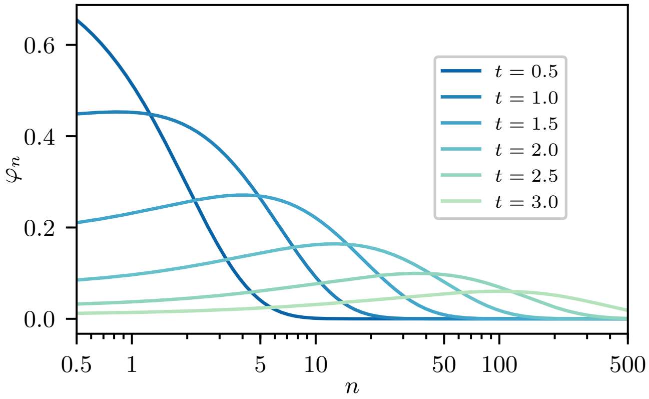


Figure 3.5: The exact solution wavefunction (3.25) in the semi-infinite chain at various times. The wavefunction is defined only at $n = 0, 1, 2 \ldots$, but has been extrapolated to intermediate values for display.

therein). Indeed, a satisfactory definition of operator complexity of any sort is an unresolved problem, and may not have a unique answer.

Exponential Growth in the Semi-infinite Chain

Recall that the Lanczos algorithm reduces the operator dynamics to a discrete Schrödinger equation (3.8),

$$\partial_t \varphi_n = -b_{n+1}\varphi_{n+1} + b_n \varphi_{n-1}, \quad \varphi_n(0) = \delta_{n0}\,.$$

We shall analyze this quantum mechanics problem when the hypothesis is satisfied in $d > 1$, i.e. $b_n = \alpha n + \gamma + o(1)$.

As a first step, we take the continuum limit, by linearizing around momenta 0 and π. This yields a Dirac equation $\partial_t \varphi = \pm 2\alpha x \partial_x \varphi$, whose characteristic curves $x \propto e^{\pm 2\alpha t}$ show the wavefunction spreads exponentially fast to the right in the semi-infinite chain with rate 2α. We remark that among all power-law Lanczos coefficient asymptotics $b_n \sim n^\delta$, the linear growth $\delta = 1$ is the only one which results in exponential spreading. When $\delta > 1$, the characteristic curves reach $x = \infty$ at finite time [5]. When $\delta < 1$, the spreading follows a power law $x \sim t^{1/(1-\delta)}$. In the case of $d = 1$, with the logarithmic correction, the wavefunction spreads as a stretched exponential — faster than any power law, but still slower than exponential.

To undertake a more careful analysis of the wavefunction on the semi-infinite chain, we employ a family of exact solutions. Suppose

$$\widetilde{b}_n := \alpha\sqrt{n(n-1+\eta)} \xrightarrow{n \gg 1} \alpha n + \gamma, \tag{3.24}$$

[5] This seems non-physical and indeed, has only been observed in exotic classical systems [50]. It is ruled out whenever the dynamics are local by Eq. (3.14).

where $\eta = 2\gamma/\alpha + 1$. For any system when the hypothesis is satisfied, the b_n's will approach the $\widetilde{b}_n$'s asymptotically, so the properties of the exact solution using the $\widetilde{b}_n$'s are universal properties at large n. It is shown in Appendix 3.D that the full wavefunction for the operator evolving under the $\widetilde{b}_n$'s is

$$|\widehat{\mathcal{O}}(t)) = \sum_{n=0}^{\infty} \sqrt{\frac{(\eta)_n}{n!}} \tanh(\alpha t)^n \operatorname{sech}(\alpha t)^\eta i^n \, |\widehat{\mathcal{O}}_n) \tag{3.25}$$

where $(\eta)_n = \eta(\eta+1)\cdots(\eta+n-1)$ is the Pochhammer symbol and $|\widehat{\mathcal{O}}_n)$ is the nth Krylov basis vector. Note that this example is not artificial but arises naturally in the SYK model, studied in Section 3.6 below.

The exact solution (3.25) benefits from a detailed analysis. Recall that the component of the wavefunction at some fixed site n is $\varphi_n(t) = (-i)^n(\widehat{\mathcal{O}}_n|\widehat{\mathcal{O}}_0(t))$. For each n, $\varphi_n(t)$ is a purely real function which starts at 0 (for $n > 1$), increases monotonically until reaching a maximum at $t \sim \ln n$, then decreases as $\sim e^{-\alpha\eta t}$. The fact that exponential decay, reminiscent of dissipative dynamics, emerges under unitary evolution is quite remarkable, and is only possible in an infinite chain [6]. Physically, the wavefunction is decaying by "escaping" off to $n \to \infty$, which serves as a bath. Note, however, that the hypothesis is not sufficient to show that $\varphi_n(t)$ decays exponentially with time for *small* n, a fact whose consequences are studied in 3.7 below.

We now come to a central consequence of the linear growth hypothesis: the exponential spreading of the wavefunction. At any fixed time and large n, the wavefunction (3.25) has the form $|\varphi_n(t)|^2 \sim e^{-n/\xi(t)}$, where $\xi(t)$ is a "delocalization length" that grows exponentially in time: $\xi(t) \sim e^{2\alpha t}$ for $\alpha t \gg 1$. This exponential spreading is reflected in the expected position of the operator wavefunction (3.25) on the semi-infinite chain

$$(n)_t := (\widehat{\mathcal{O}}(t)|n|\widehat{\mathcal{O}}(t)) = \eta \sinh(\alpha t)^2 \sim e^{2\alpha t}\,, \tag{3.26}$$

More generally, $(n^k)_t \sim e^{2k\alpha t}$ for $k \geq 1$. This result agrees, of course, with the one obtained in the simple continuum-limit above. We believe that the asymptotic growth in (3.26) holds whenever the Lanczos coefficients grow linearly. Although we have not proven this assertion, we have checked that it holds for many cases, such as artificially generated sequences of Lanczos coefficients $b_n = \alpha n + \gamma_n$ with various kinds of bounded "impurity" terms $\gamma_n \sim \mathrm{O}(1)$. We will consider (3.26) as a fact that follows directly from the hypothesis: the position of an operator in the abstract Krylov space grows exponentially in time.

We may interpret this exponential growth as a quantitative measure of the *irreversible* tendency of quantum operators to run away towards higher "complexity" [56]. Indeed, we identify the position on the semi-infinite chain $(n)_t$ as a notion of operator complexity. We refer to $(n)_t$ as the "Krylov-complexity" (or "K-complexity" for short) of an operator. After

[6] This shows the importance of the thermodynamic limit. With any finite-dimensional Hilbert space, the chain would be finite, and the results in this section would be affected.

all, as n increases, the operators $\widehat{\mathcal{O}}_n$ becomes more "complex", in the following sense: in the Heisenberg-picture, the equations of motions for $\widehat{\mathcal{O}}_n$'s form a hierarchy:

$$\begin{aligned} -i\dot{\widehat{\mathcal{O}}}_0(t) &= b_1\widehat{\mathcal{O}}_1(t)\,, \\ -i\dot{\widehat{\mathcal{O}}}_1(t) &= b_1\widehat{\mathcal{O}}_0(t) + b_2\widehat{\mathcal{O}}_2(t)\,, \\ -i\dot{\widehat{\mathcal{O}}}_2(t) &= b_2\widehat{\mathcal{O}}_1(t) + b_3\widehat{\mathcal{O}}_3(t)\,, \\ &\vdots \end{aligned} \tag{3.27}$$

that is, the dynamics of $\widehat{\mathcal{O}}_n(t)$ depends on $\widehat{\mathcal{O}}_{n+1}(t)$. This is analogous to the BBGKY hierarchy in statistical mechanics, in which the evolution of the n-particle distribution depends on the $(n+1)$-particle one. Similarly, as n increases, the $\widehat{\mathcal{O}}_n$'s becomes less local in real space, involve more basis vectors in any local basis, and are more difficult to compute. We remark that K-complexity is a distinct notion from precise terms such as circuit complexity and no relation should be inferred between the two. Closer precedents are the ideas of f-complexity and s-complexity [57].

We know from Section 3.4 that linearly growing Lanczos coefficients are the maximal rate so, in turn, the wavefunction may not spread faster than exponentially. Thus the hypothesis in $d > 1$ implies that non-integrable systems have maximal growth of K-complexity: exponential, with rate 2α.

A Bound on Complexity Growth

The physical meaning of K-complexity is far from transparent. After all, it depends on the rather abstract Krylov basis, the initial operator, and the choice of dynamics. To help pin down the idea of K-complexity, we study its relation to more familiar quantities. We shall consider a class of observables, "q-complexities" (q stands for *quelconque*), that includes familiar notions like out-of-time-order correlators and operator size. We will show that the growth of any q-complexity is bounded above by K-complexity.

We now define the q-complexity. Suppose $\mathcal{Q}$ is a superoperator that satisfies two properties:

1. $\mathcal{Q}$ is positive semidefinite. We denote its eigenbasis as $|q_a)$, indexed by a, so that

$$\mathcal{Q} = \sum_a q_a\,|q_a)\,(q_a|\,,\; q_a \geq 0\,. \tag{3.28a}$$

2. There is a constant $M > 0$ such that

$$(q_b|\mathcal{L}|q_a) = 0 \text{ if } |q_a - q_b| > M\,, \tag{3.28b}$$

$$(q_a|\widehat{\mathcal{O}}) = 0 \text{ if } |q_a| > M\,. \tag{3.28c}$$

Then q-complexity is defined to be the expectation value

$$(\mathcal{Q})_t := (\widehat{\mathcal{O}}(t)|\mathcal{Q}|\widehat{\mathcal{O}}(t)), \tag{3.29}$$

where $\widehat{\mathcal{O}}(t)$ is evolved under the Liouvillian dynamics of $\mathcal{L}$. A q-complexity is, in principle, an observable, and requires Hamiltonian (or Liouvillian) dynamics. The rationale for the conditions is as follows: (3.28a) ensures the q-complexity is always non-negative, (3.28b) guarantees it cannot change too much under one application of the Liouvillian, and (3.28c) assigns a low complexity to the initial operator. To illustrate this concept, we now consider three examples: K-complexity, operator size, and out-of-time-order correlators.

Example 1: K-complexity.The K-complexity is always a q-complexity, with

$$\mathcal{Q} = \sum_n n\,|\widehat{\mathcal{O}}_n)\,(\widehat{\mathcal{O}}_n|\,.$$

The basis $|q_a)$ is just the Krylov basis $|\widehat{\mathcal{O}}_n)$ and the conditions (3.28b) and (3.28c) are satisfied by construction of the Krylov basis with $M = 1$.

Example 2: operator size. A second example of a q-complexity is provided by *operator size* [41]. For concreteness, we work in the framework of a spin-1/2 model (though Majorana fermions or higher spins work equally well). Consider the basis of *Pauli strings*, e.g. strings $IXYZII\cdots$ with finitely many non-identity operators. Define $\mathcal{Q}$ to be diagonal in this basis, where the action of $\mathcal{Q}$ on a Pauli string is the number of non-identity Pauli's. So, for instance, $\mathcal{Q}\,|IXYZI\cdots) = 3\,|IXYZI\cdots)$. The eigenvectors of $\mathcal{Q}$ have non-negative eigenvalues, so $\mathcal{Q}$ is positive semi-definite.

Any choice of dynamics with at most M-body interactions (even long-ranged ones) will satisfy (3.28b), while (3.28c) requires simple that $\widehat{\mathcal{O}}$ is d-local. So, under these conditions, the q-complexity $(\mathcal{Q})_t$ becomes the average size of Pauli strings contained in $\widehat{\mathcal{O}}(t)$:

$$(\mathcal{Q})_t = \sum_{\pi \in \text{Pauli strings}} \text{size}(\pi)\,|(\pi|\widehat{\mathcal{O}}(t))|^2\,. \tag{3.30}$$

Example 3: OTOCs. Our third — and most interesting — example of q-complexity is out-of-time-order commutators (OTOCs). Given $\widehat{\mathcal{O}}(t)$, each choice of local operator V defines an OTOC $([V, \widehat{\mathcal{O}}(t)] \mid [V, \widehat{\mathcal{O}}(t)])$. For simplicity, we work with a many-body lattice model, and consider an on-site operator V_i. We then define the OTOC superoperator by

$$\mathcal{Q} := \sum_i \mathcal{Q}_i, \quad (A|\mathcal{Q}_i|B) := ([V_i, A] \mid [V_i, B]), \tag{3.31}$$

where the sum runs over all lattice sites i. Provided the Hamiltonian and initial operator are r-local, and that the dimension D of the on-site Hilbert space is finite, (3.31) is a q-complexity.

To see this, let us work in the eigenbasis of $\mathcal{Q}$. For each site i, there is a basis $\mathcal{Q}_i\,|q_{i,a}) = q_{i,a}\,|q_{i,a})$ with $1 \leq a \leq D^2$. We take $|q_{i,0})$ to be the identity operator with eigenvalue 0, and note that $0 \leq q_{i,a} \leq Q$ for some finite Q. Since $[\mathcal{Q}_i, \mathcal{Q}_j] = \delta_{ij}$, the eigenbasis for the full operator space is the tensor product of the on-site bases. So for any sequence $\boldsymbol{a} = \{a_i\}$, $|q_{\boldsymbol{a}}) := \otimes_i\,|q_{i,a_i})$ is an eigenvector satisfying

$$\mathcal{Q}\,|q_{\boldsymbol{a}}) = q_{\boldsymbol{a}}\,|q_{\boldsymbol{a}})\,, \quad q_{\boldsymbol{a}} = \sum_i q_{i,a_i} \geq 0. \tag{3.32}$$

For the eigenvalue to be finite, a_i must be zero for all but a finite number of i's and all eigenvalues are non-negative, so $\mathcal{Q}$ is positive semidefinite. Since the Hamiltonian is r-local, the matrix element $(q_{\boldsymbol{a}}|\mathcal{L}|q_{\boldsymbol{b}}) \neq 0$ only if $\boldsymbol{a}$ and $\boldsymbol{b}$ differ on at most r sites. So by (3.32), we may bound the difference $|q_{\boldsymbol{a}} - q_{\boldsymbol{b}}| \leq M = rQ$. Similarly, any r-local operator satisfies (3.28c). Having verified all the properties (3.28), we may conclude that OTOCs of this form are a q-complexity.

OTOCs are known to be closely related to the operator size [41, 24]. It is usually possible to bound either quantity from the other, and to choose V_i such that the OTOC reduces to the operator size.

We have now seen three examples of q-complexities, two of which are quantities that have been studied in recent times to understand the complexity of operators. We remark that q-complexities (including K-complexity) are quadratic in $O(t)$ and *not* linear response quantities, although the growth rate α is, via the spectral function. We will see in Section 3.6 that q-complexities may also be applied to classical systems, though they work somewhat differently there.

A rigorous argument in Appendix 3.E proves that, for any q-complexity,

$$(\mathcal{Q})_t \leq C(n)_t,\; C = 2M\,. \tag{3.33}$$

The following section will focus on the application of this general bound in the specific case of OTOCs.

To close this section, we show how the above results are affected by the log-correction to linear growth in 1d from Eq. (3.22): $b_n \sim An/\ln n$. The continuum Dirac equation analysis yields a stretched exponential growth of K-complexity:

$$(n)_t \sim e^{\sqrt{At}}\,, \tag{3.34}$$

which is slower than any exponential growth but faster than any power law. Combined with (3.33), we conclude that all q-complexities have at most stretched exponential growth in 1d.

3.6 Growth Rate as a Bound on Chaos

We showed in the preceding section that K-complexity provides an upper bound for any q-complexity whatsoever, which includes certain types of OTOCs. Combining (3.33) and

(3.26), we see that q-complexities grow at most exponentially in time, at least when the hypothesis holds for $d > 1$. If that is the case, with $(Q)_t \sim e^{\lambda_Q t}$, then the exponent is bounded above by 2α:

$$\lambda_Q \leq 2\alpha\,. \tag{3.35}$$

In the rest of this section we focus on the case where the q-complexity is an OTOC. When the OTOC grows exponentially at late times,

$$(\mathcal{Q}^{\mathrm{OTOC}})_t \sim e^{\lambda_L t}\,, \tag{3.36}$$

its growth rate λ_L is called the Lyapunov exponent, since in the classical limit it reduces to the Lyapunov exponent characterizing the butterfly effect in classical deterministic chaos [7]. We can then state following bound on Lyapunov exponents: *for any system at infinite temperature where the operator growth hypothesis holds, then*

$$\lambda_L \leq 2\alpha, \tag{3.37}$$

where we put $\lambda_L = 0$ whenever the OTOC grows slower than exponentially, and similarly for α. This follows directly from (3.33) and (3.26), so we have essentially proven (3.37) as a mathematical proposition.

It is interesting to note that, as λ_L is defined via a four-point correlation function (the OTOC), while α depends on a two-point correlation function ($C(t)$), the bound (3.37) can be interpreted as a relation between correlation functions of distinct nature. Such a relation is, to our knowledge, rather unusual (see [59] for a recent result). However, this point of view is not how we derived (3.37); an alternative proof working directly with the correlation functions would be illuminating.

Remarkably, the bound (3.37) appears to be valid under much less restrictive assumptions — at any temperature and in either classical or quantum systems. In this section, we examine the cases of quantum and classical systems at infinite temperature, and leave that of finite temperatures to Section 3.8 below.

SYK Model

We illustrate the bound (3.37) for the SYK model (3.18). At infinite temperature, no analytic formula for the Lyapunov exponent is available, but it has been computed numerically in, e.g. [41, 25]. Table 3.3 shows that not only does (3.37) hold for the whole range of q-SYK models, but α is almost equal to $\lambda_L/2$, with exact agreement in the limit $q \to \infty$ [8]. These results show that the bound $\lambda_L \leq 2\alpha$ is tight: the prefactor cannot be improved in general. Moreover, in the large q limit, the probability distribution $|\varphi_n(t)|^2$ on the semi-infinite line is identical to the operator size distribution of $\gamma_1(t)$ [41]. (See (3.95) in Appendix 3.B for the

[7] To be precise, the OTOC measures a *generalized* Lyapunov exponent with $q = 2$, which is greater or equal to the typical one [58]

[8] Indeed, the difference may well be a numerical effect, see [41].

precise statement.) So the large-q SYK model is an instance where the quantum mechanics problem on the semi-infinite chain can be concretely interpreted.

We remark that in models with all-to-all interactions like SYK and its variants may be the only circumstances where the bound (3.37) can be nearly saturated. For spatially extended quantum systems with finitely many local degrees of freedom, Lieb-Robinson bounds [60] and its long-range generalizations [61] guarantee that the OTOC has slower-than-exponential growth in most physical systems at infinite temperature [9].

Such a difference can be understood as follows. Due to the lack of spatial structure in the SYK model, we expect operator complexity (by any reasonable definition) is almost completely captured by operator size which, in turn, is directly probed by OTOCs. In finite-dimensional systems, complexity should be a distinct concept from operator size. For instance, long Pauli strings generated in the non-interacting Ising models have nonetheless low complexity, since they can be transformed to simple few-body operators under the Jordan-Wigner transform. In non-integrable systems, by contrast, operator size growth is limited by Lieb-Robinson, while complexity can grow exponentially in the *bulk* of an operator's support.

q	2	3	4	7	10	∞
$\alpha/\mathcal{J}$	0	0.461	0.623	0.800	0.863	1
$\lambda_L/(2\mathcal{J})$	0	0.454	0.620	0.799	0.863	1

Table 3.3: The growth rate α versus half the OTOC-Lyapunov exponent $\lambda_L/2$ in the q-SYK model (3.18) in units of $\mathcal{J} = \sqrt{q}2^{(1-q)/2}J$. Here α is obtained by exact numerical methods discussed in Appendix 3.B, while λ_L is taken from the Appendix of [41]. The q-SYK is physical only for even integers q, but large-N methods allow an extrapolation to any $q \geq 2$.

Classical Chaos

We now transition to the classical setting. After briefly explaining how the recursion method carries over almost verbatim to classical systems, we shall examine the classical form of the bound (3.37). However, the arguments of Section 3.5 do not carry over in full, and we are only able to prove a *weaker* bound. We close with a numerical case-study that suggests the stronger conjectural bound may well be true (and tight).

A (Weaker) Bound on Classical Chaos

The recursion method applies to classical and quantum systems in exactly the same manner [10]. Classically, operator space is the space of functions on classical phase space and

[9] Indeed, generalized Lieb-Robinson bounds state that the OTOC between $\widehat{\mathcal{O}}(t)$ and V_i is exponentially small if the site i lies out of some volume which grows sub-exponentially. Then, a sum like (3.31) is essentially that volume.

the Liouvillian $\mathscr{L} = i\{\mathscr{H}, \cdot\}$ is defined by the Poisson bracket against the classical Hamiltonian $\mathscr{H}$ (we take $\hbar = 1$). The appropriate classical inner product at infinite temperature is $(f|g) = \int f^* g \, d\Omega$, where $d\Omega$ is the symplectic volume form on the phase space [10]. The Liouvillian $\mathscr{L}$ is a self-adjoint operator, and the entire framework of the Lanczos coefficients carries over wholesale.

Indeed, the Lanczos coefficients have been studied *more* in the classical context. It is known [50, 10] that linear growth of the Lanczos coefficients appears in general finite-dimensional, non-linear systems, to which we restrict ourselves [11]. The growth rate α is well-defined in such systems, as is the (classical) Lyapunov exponent λ_L, and the bound (3.37) takes on the same form as before: $\lambda_L \leq 2\alpha$. In short, the similarity of classical and quantum Liouvillian evolution means that the recursion method — and its consequences — carry over unchanged.

There is, however, one important caveat: a classical OTOC does *not* generally qualify as a q-complexity. We will demonstrate this through an explicit, and instructive, example. Let us consider a single classical $SU(2)$ spin. Its classical phase space is the two-sphere, and classical operator space is spanned by the basis of spherical harmonics $|Y_\ell^m)$, $\ell = 0, 1, 2\ldots,$ $m = -\ell, \ldots, \ell$.

A typical Hamiltonian is a polynomial of the classical spin operators $\mathscr{S}^x, \mathscr{S}^y, \mathscr{S}^z$ with Poisson bracket $\{\mathscr{S}^a, \mathscr{S}^b\} = -\varepsilon^{abc}\mathscr{S}^c$. We consider the simple non-linear example

$$\mathscr{H} = J\mathscr{S}^z\mathscr{S}^z + h_x\mathscr{S}^x. \tag{3.38}$$

Using Clebsch-Gordon coefficients one can show that the classical Liouvillian is quite sparse, and only the following matrix elements are non-zero:

$$(Y_m^{\ell\pm1}|\mathscr{L}|Y_m^\ell) \neq 0\,,\ (Y_{m\pm1}^\ell|\mathscr{L}|Y_m^\ell) \neq 0, \tag{3.39}$$

whenever the states in question exist.

We now examine the classical OTOC for the local operator $\mathscr{S}^z$, given by matrix elements of a super-operator $\mathscr{Q}^z$. This operator is diagonal in the basis of spherical harmonics

$$\begin{aligned}(Y_n^k|\mathscr{Q}_z|Y_m^\ell) :=& (\{\mathscr{S}^z, Y_k^n\}|\{\mathscr{S}^z, Y_l^m\}) \\ =& m^2\delta_{nm}\delta_{k\ell},\end{aligned} \tag{3.40}$$

and we may immediately read off the eigenvalues as m^2. When m changes by 1 upon application of the Liouvillian, the eigenvalue m^2 changes by $1 \pm 2m$, which can be arbitrarily large. Hence the condition (3.28b) cannot be satisfied for any finite constant d. It is helpful to recall that Section 3.5 showed the quantum OTOC is a q-complexity whenever the on-site Hilbert space is finite-dimensional. This fails in the case of a spin s, whose on-site dimension $2s + 1$, in the classical limit $s \to \infty$. We have therefore seen that classical OTOCs are not

[10] We therefore require a compact phase space, such as in a classical spin model.

[11] Note that even if the phase space is finite-dimensional, the operator space is infinite-dimensional, allowing an infinite sequence of Lanzcos coefficients.

q-complexities and, hence, the bound (3.37) does not follow from the reasoning of Section 3.5 in the classical case, and remains a conjecture.

Nonetheless, for any Hamiltonian and initial operators that are polynomials of the spin variables $\mathscr{S}^a$, we can show the following general bound

$$\lambda_L \leq 4\alpha\,, \tag{3.41}$$

which is weaker than the conjectured $\lambda_L \leq 2\alpha$.

To show (3.41), observe that by (3.40), the superoperator $\mathcal{R}_z := \mathcal{Q}_z^{\frac{1}{2}}$ satisfies (3.28b), since its has eigenvalue m for Y_m^ℓ, which can change only by δ upon one Liouvillian application, where δ is the polynomial degree of the Hamiltonian. Other conditions in (3.28) are satisfied straightforwardly. We then have

$$e^{\lambda_L t} \sim (\mathcal{Q}_z)_t = (\mathcal{R}_z^2)_t \leq C^2 (n^2)_t \sim e^{4\alpha t}\,, \tag{3.42}$$

which implies (3.41). Here the first $\sim$ is by definition, the the inequality is a straightforward generalization of the bound on q-complexity, Eq. (3.127) of Appendix 3.E, and the last $\sim$ is a generalization of (3.26) (see below that equation).

This argument carries over to the OTOC with spin variables in any direction by spherical symmetry, and applies almost *verbatim* to systems with a few spins, $\mathscr{S}_i^{x,y,z}, i = 1, \ldots, N$. A Lyapunov exponent associated with a finite sum such as

$$\sum_{i=1}^{N} \sum_{a=x,y,z} (\{\mathscr{S}_i^a, \widehat{\mathcal{O}}(t)\} \big| \{\mathscr{S}_i^a, \widehat{\mathcal{O}}(t)\}) \tag{3.43}$$

satisfies the same bound since every term does so. In summary, (3.41) is established in general classical few-spin models. We expect it is possible to show (3.41) rigorously.

An interesting corollary of (3.41) is a relation between chaos and the decay rate of the spectral function. Recall that the linear growth of Lanczos coefficients is equivalent to the exponential decay of the spectral function $\Phi(\omega) \sim \exp(-|\omega|/\omega_0)$ at high frequency, where $\omega_0 = \frac{2}{\pi}\alpha$. Then (3.41) is equivalent to

$$\lambda_L \leq 2\pi\omega_0\,. \tag{3.44}$$

(The conjectured bound would instead imply $\lambda_L \leq \pi\omega_0$.) In numerous classical systems, the power spectrum decay of time series has been used as an empirical probe of deterministic chaos [62, 63, 64, 65, 66, 67, 68]. To the best of our knowledge, the bound (3.44) provides the first quantitative justification for this usage.

We mention that the relation between chaos and *long-time* decay of correlation functions has also been studied: long-time relaxation to equilibrium was shown to be controlled by Ruelle resonances in specific chaotic models [69, 70]. However, the long-time and high-frequency behaviors are *a priori* unrelated, as we discuss further in Section 3.7.

We stress that the growth rate is an upper bound, but *not* a diagnostic of *classical* chaos. Indeed, our bound is correct but not tight for most classical integrable systems which, generically, have non-zero growth rate but no chaos [50].

Unfortunately, we are not able to improve the argument and prove the stronger conjectured bound. Instead, we resort to testing the validity of the conjectured bound (3.37) in a canonical example of classical chaos.

Numerical Case Study

The Feingold-Peres model of coupled tops [71] is a well-studied model of few-body chaos, both classically and at the quantum level [72, 73]. The quantum model is a system of two spin-s particles, 1 and 2, with Hamiltonian

$$H_{\mathrm{FP}} = (1+c)\left[S_1^z + S_2^z\right] + 4s^{-1}(1-c)S_1^x S_2^x \tag{3.45}$$

where $c \in [-1, 1]$ is a parameter and S_i^α satisfy the $SU(2)$ algebra $[S_i^\alpha, S_j^\beta] = i\hbar\delta_{ij}\varepsilon^{\alpha\beta\gamma}S_i^\gamma$ and act on a spin-s Hilbert space. This is non-interacting when $c = \pm 1$ and chaotic in the intermediate region. Correspondingly, the Lanczos coefficients are asymptotic to a constant at $c = \pm 1$ and increase linearly in intermediate regions. However, since the operator space dimension is finite (equal to $(2s+1)^4$), the sequence of Lanczos coefficients is finite; in fact, the Lanczos coefficients saturate. The classical limit is obtained by taking s to infinity. There the Hamiltonian becomes

$$\mathscr{H}_{\mathrm{FP,cl}} = (1+c)\left[\mathscr{S}_1^z + \mathscr{S}_2^z\right] + 4(1-c)\mathscr{S}_1^x \mathscr{S}_2^x \tag{3.46}$$

where $\mathscr{S}_i^\alpha, i = 1, 2$ are two sets of classical $SU(2)$ spins. As an $SU(2)$ representation, the classical operator space contains all integer spins, whereas the quantum operator space has only integer spins up to $2s$.

We compute the classical Lanczos coefficients for the operator $\widehat{\mathcal{O}} \propto S_1^z S_2^z$ ($\mathscr{S}_1^z \mathscr{S}_2^z$ in the classical case). As shown in Fig. 3.6(b), the quantum Lanczos coefficients converge to the classical ones as $s \to \infty$, as expected, and they increase linearly near $c = 0$. We have checked that α does not depend on the choice of initial operator $\widehat{\mathcal{O}}$, so long as $\widehat{\mathcal{O}}$ does not overlap with any conserved quantity.

To test the conjectured bound (3.37), we compare the growth rate α with the classical Lyapunov exponent ($\lambda_L/2$ in our notation), which can be calculated by the standard variational equation method [74]. Remarkably, the data shown in Fig. 3.6(a) corroborates the conjectured bound $\alpha \geq \lambda_L/2$ in the parameter region explored, with equality up to numerical accuracy in the regime $c \approx 0$, where the model is known to be maximally chaotic, with almost no regular orbits [72, 71]. Enlarging the parameter space, for instance by adding terms such as $\mathscr{S}_i^z$ to the Hamiltonian, give further results consistent with the bound. It is thus possible that the conjectured bound is valid in classical systems and becomes tight in highly chaotic ones.

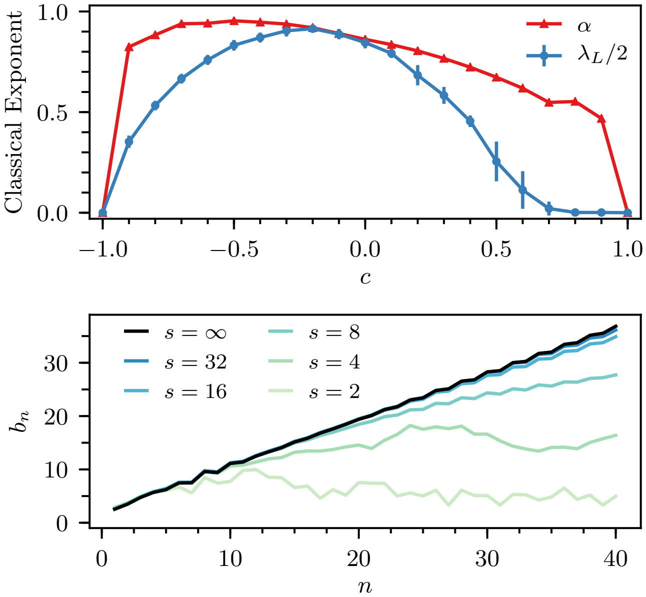


Figure 3.6: (a) The growth rate α versus the classical Lyapunov exponent $\lambda_L/2$ in the classical Feingold-Peres model of coupled tops, (3.46). $\alpha \geq \lambda_L/2$ in general, with equality around the $c = 0$ where the model is the most chaotic. The growth rate appears to be discontinuous at the non-interacting points $c = \pm 1$, similarly to Fig. 3.4-(b). (b) The first 40 Lanczos coefficients of quantum $s = 2, \ldots, 32$ and classical ($s = \infty$) FP model, with $c = 0$.

3.7 Application to Hydrodynamics

Structural information about quantum systems can enable numerical algorithms. As an example, the success of the density matrix renormalization group algorithm is a consequence of the area law of entanglement entropy [75, 76]. We now apply the hypothesis to develop a semi-analytical technique to calculate decay rates and autocorrelation functions of operators and, in particular, compute diffusion coefficients of conserved charges. The key idea is to use the hypothesis to make a meromorphic approximation to the Green's function. This section introduces the continued fraction expansion of the Green's function, describes the zoology of operator decay, and finally presents the semi-analytical method.

Continued Fraction Expansion: Brief Review

We briefly review the continued fraction expansion of the Green's function [10]. The Green's function (3.9) is related to the autocorrelation $C(t)$ by the following transform:

$$G(z) = i\int_0^\infty C(t)e^{-izt}dt\,,\; C(t) = \oint G(z)e^{izt}\frac{dz}{2\pi i}\,, \tag{3.47}$$

where the integration contour is taken to be the shifted real axis shifted down by $-i\epsilon$ for some small $\epsilon > 0$. Since $C(t)$ is bounded on the real axis, $G(z)$ is analytic in the lower half-plane, but may contain singularities on the upper half plane. We shall refer to (3.47) as the Laplace transform, despite the fact that it differs from the usual definition by a factor of i.

In the Krylov basis, $G(z) = [z - L]^{-1}_{00}$ corresponds to all paths that start on the first site, propagate through the chain, and return. We can divide all paths into those that stay on the first site, and those that first hop to the second site, propagate on sites $n \geq 2$, and then return. More formally, for each $n \geq 0$, let $L^{(n)} := L_{p\geq n, q\geq n}$ be the hopping matrix on the semi-infinite chain restricted to sites n and above, and let $G^{(n)}(z) := \left[z - L^{(n)}\right]^{-1}_{nn}$ be the corresponding Green function. (Note that $G^{(0)}(z) = G(z)$.) We then have the following recursion relation — hence the name "recursion method" —

$$G^{(n)}(z) = \frac{1}{z - b_{n+1}^2 G^{(n+1)}(z)}\,,\ n \geq 0\,. \tag{3.48}$$

(For a quick derivation [34], consider the polynomial $P_n(z) := \det(z - L^{(n)})$. By Cramer's rule we have $G^{(n)}(z) = P_{n+1}(z)/P_n(z)$; a cofactor expansion gives $P_n(z) = zP_{n+1}(z) - b_{n+1}^2 P_{n+2}(z)$. Then (3.48) follows from the two preceding equations.)

Applying Eq. (3.48) recursively yields the continued fraction expansion:

$$G(z) = \cfrac{1}{z - \cfrac{b_1^2}{z - \cfrac{b_2^2}{z - \ddots}}}\,. \tag{3.49}$$

To save space, we denote the recursion 3.48 by $G^{(n)} = M_{n+1} \circ G^{(n+1)}$, where M_n is the Möbius transform $w \mapsto 1/(z - b_n^2 w)$ and "$\circ$" denotes function composition. It is crucial that the convergence of the continued fraction expansions is quite subtle and quite different from the convergence of, say, Taylor series. Practically speaking, one can compute only a finite number of the b_n's in most situations. Truncating the expansion by taking the rest of the b_n's to be zero (or any constant) rarely provides a good approximation to the whole function [10].

Hydrodynamical Phenomenology

Long-time and large-wavelength properties of correlation functions are governed by emergent hydrodynamics. For each conserved charge (e.g. energy, spin), the density field should relax to equilibrium in a manner prescribed by a classical partial differential equation. Often this is a diffusion equation, though more exotic possibilities such as anomalous diffusion and ballistic transport (infinite conductivity) can also appear.

A numerical (and sometimes experimental) protocol to probe the emergent hydrodynamics is to study the autocorrelation function of the density wave operator $\widehat{\mathcal{O}}_q = \sum_x e^{iqx} Q_x$

(here Q_x is the operator of the conserved charge at x) at a range of momenta q. The behavior at large time is of especial interest, and can, in turn, be read off from the singularity structure of the Green's function. Let us give a few examples. If the closest pole to the origin is at $z = i\gamma$, then the autocorrelation function will decay exponentially as $e^{-\gamma t}$, while if the location of the closest pole varies quadratically as $z = iDq^2/2$, then the dynamics are diffusive. However, the presence of non-linear terms in addition to the linear diffusive ones can give rise to exotic behavior where the diffusion constant itself becomes a function of frequency. An example of this is $G(z) = [z - iD(z)q^2/2]^{-1}$, where $D(z) = D_0 + D_1\sqrt{z}$. At any fixed q, $G(z)$ has a branch cut in addition the diffusive pole, so although the diffusion constant D_0 is still well-defined, autocorrelation functions decay [77] as a *power law* in time [12]. Regardless, the full singularity structure of the Green's function determines the long-time behavior.

Of course, computing the singularity structure of the Green's function is a demanding task. Even in integrable models, determining if the correct hydrodynamics is, say, diffusion or anomalous diffusion is non-trivial — let alone computing diffusion coefficients (see Refs [78, 79, 80, 81, 82, 83] for recent developments). Indeed, accurately computing diffusion coefficients has been the goal of much recent numerical work [84, 85, 86]. This difficulty is reflected in the continued fraction expansion (3.49): the location of the poles change with each new fraction, so the full analytic structure of $G(z)$ depends on *all* of the b_n's.

Knowing that the coefficients obey the universal form (3.12) is not enough, because even though the wavefunction is spreading out into the semi-infinite chain exponentially fast, we are given no guarantee about the wavefunction at the origin $n = 0$. For instance, the correlation functions $C_1(t) = \operatorname{sech}(\alpha t)$ and $C_2(t) = (1 + t^2)^{-\gamma}$ [10] both correspond to Lanczos coefficients that grow linearly But $C_1(t)$ decays exponentially while $C_2(t)$ decays as a power law, so clearly the asymptotics of b_n alone is insufficient to establish long-time behavior. The power law decay is nonetheless reflected in the Lanczos coefficients for $C_2(t)$, which have an alterating subleading tail. Precisely, they have the form $b_n = \alpha n + \gamma + (-1)^n f_n$ where the f_n's are positive and decay to zero. Therefore determining the long-time tail of $C(t)$ probably requires additional information about the subleading corrections to the hypothesis. In particular, the results in this work are *prima facie* unrelated to a bound on transport [87].

Numerical Diffusion Coefficients

Despite the complex behavior of autocorrelation functions in the time domain, there are situations where the hypothesis alone suffices to compute diffusion coefficients. In the case where the b_n's approach the universal form (3.12) especially quickly and regularly, we are able to make a meromorphic approximation to $G(z)$. The idea is as follows. In the semi-infinite chain picture, we may hope to calculate the first few Lanczos coefficients exactly, so we may describe behavior near the origin $n = 0$ exactly. For large n, on the other hand, the hypothesis gives the coefficients almost exactly, so we can describes the dynamics by some exact solution. By stitching the dynamics at large and small n together, we can hope to find

[12] We thank Achim Rosch for pointing out this possibility.

the dynamics on the whole chain. This allows us to recover a diffusive dispersion relation and numerically extract the diffusion constant in specific models.

We remark that there are a number of existent extrapolation schemes to determine the Green's function from the first few Lanczos coefficients [10, 34]. The new ingredient here is the hypothesis, which controls the approximation.

To make this idea into a precise numerical technique, we need three ingredients: a way to compute the Lanczos coefficients at small n, an exact solution at large n, and a robust way to meld them together. For a 1D spin chain in the thermodynamic limit of large system size, it is straightforward to compute the first few dozen Lanczos coefficients exactly through repeated matrix multiplication. Details are given in Appendix 3.C.

To find the large n-behavior, we employ an exact solution for the quantum mechanics problem on the semi-infinite chain. If the hypothesis is obeyed, then the b_n's also asymptotically approach the form

$$\widetilde{b}_n = \alpha\sqrt{n(n-1+\eta)} \xrightarrow{n\gg 1} \alpha n + \gamma, \tag{3.50}$$

where $\eta = 2\gamma/\alpha + 1$. The agreement is better, of course, at large n. The coefficients $\widetilde{b}_n$ have the virtue that the quantum mechanics problem they describe on the semi-infinite chain is exactly solvable. Appendix 3.D applies the theory of Meixner orthogonal polynomials of the second kind to determine the autocorrelation analytically: $C(t) = \operatorname{sech}(\alpha t)^\eta$. (This is the same exact solution used in Section 3.5 above.) By Laplace transform, the corresponding Green's function is

$$\widetilde{G}_{\alpha,\gamma}(iz) = \frac{1}{\alpha} H(z/\alpha;\eta), \tag{3.51a}$$

$$H(z;\eta) = \frac{2^\eta}{z+\eta} {}_1F_2(\eta, \frac{z+\eta}{2}, \frac{z+\eta}{2}+1; -1), \tag{3.51b}$$

$$\widetilde{G}^{(n)}(z) \;=\; \widetilde{M_n}^{-1} \circ \cdots \circ \widetilde{M_1}^{-1} \circ \widetilde{G}(z) \tag{3.51c}$$

Here ${}_1F_2$ is the hypergeometric function and $\widetilde{M_k}$ depends on $\widetilde{b}_k$. It is crucial that $\widetilde{G}^{(n)}(z)$ is known analytically, so that (3.51) provides the asymptotically exact large n-behavior.

Now we stitch the small and large n information together. The true Green's function $G^{(N)}(z)$ only depends on the coefficients b_n with $n \geq N$. So for sufficiently large N, where the b_n's are approximately the same as the $\widetilde{b}_n$'s, we may approximate

$$\begin{aligned} G(z) &= M_1 \circ \cdots \circ M_N \circ G^{(N)}(z) \\ &\approx M_1 \circ \cdots \circ M_N \circ \widetilde{G}^{(N)}_{\alpha,\gamma}(z), \end{aligned} \tag{3.52}$$

an approximation that becomes better at large N. Equation (3.52) is our semi-analytical approximation to the Green's function. One can check that this is a meromorphic approximation for $G(z)$, whose poles lie only in the upper half plane.

In practice, one must calculate the b_n's until the universal behavior appears and fit α and η. Then the approximate $G(z)$ can be calculated from (3.51) and a sequence of two-by-two matrix multiplications. One can then find the location of the first pole on the imaginary axis for a range of wavevectors q and fit $z = iDq^2/2 + O(q^4)$ to extract the diffusion coefficient D. This procedure is illustrated for the energy diffusion in chaotic Ising model in Fig. 3.7. Almost all the computational effort goes into in computing the first few b_n's exactly. We also note that the extrapolation is carried out with a linear fit to the Lanczos coefficients which is not strictly appropriate to $d = 1$ (the log-correction is missing). Nevertheless, the numerical value of the diffusion coefficient appears to match other methods to within a few percent.[13] Further numerical tests on this example indicate the the exact asymptotics of Lanczos coefficients may not be necessary to compute D to a decent precision.

In short, the hypothesis is sometimes sufficient to describe the emergent hydrodynamic behavior of operators, even if we ignore the log correction in 1d. We reiterate that the hypothesis governs the leading order asymptotics of the Lanzcos coefficients only, while the autocorrelation depends on further corrections, so there is no *a priori* reason it should be computable just from the hypothesis. On the other hand, in the better scenarios, *less* knowledge on the Lanczos coefficients is required to capture the hydrodynamic coefficients. We will provide further examples of this algorithm and discuss its theoretical and practical accuracy in subsequent work.

3.8 Finite Temperature

So far our discussion has been confined to infinite temperature. In this section we generalize to finite temperature. Only a minor modification is required to carry out the Lanczos algorithm at finite temperature so many of our results carry over unaffected. A summary is provided in Table 3.4 for the reader's convenience.

	$T = \infty$	$T < \infty$
Inner Product	$(A\|B) \propto \mathrm{Tr}[A^\dagger B]$	Eq. (3.53)
Lanczos Algorithm	Eq. (3.4)	Eq. (3.55)
$C(t), G(z), \Phi(\omega), \mu_{2n}$	Section 3.3	Eq. (3.56)
$b_n \leftrightarrow C \leftrightarrow G \leftrightarrow \Phi \leftrightarrow \mu$	App. 3.A	App. 3.A
Hypothesis	Eq. (3.22)	Eq. (3.58)
$b_n \sim \alpha n$ for SYK	Eq. (3.93)	Eq. (3.99)
Bound $\lambda_L \leq 2\alpha$	Proven	Conjectured

Table 3.4: Correspondence between finite and infinite temperature definitions and results.

[13] We are greatful to Francisco Machado and Biantian Ye for sharing their density matrix truncation (DMT) results with us.

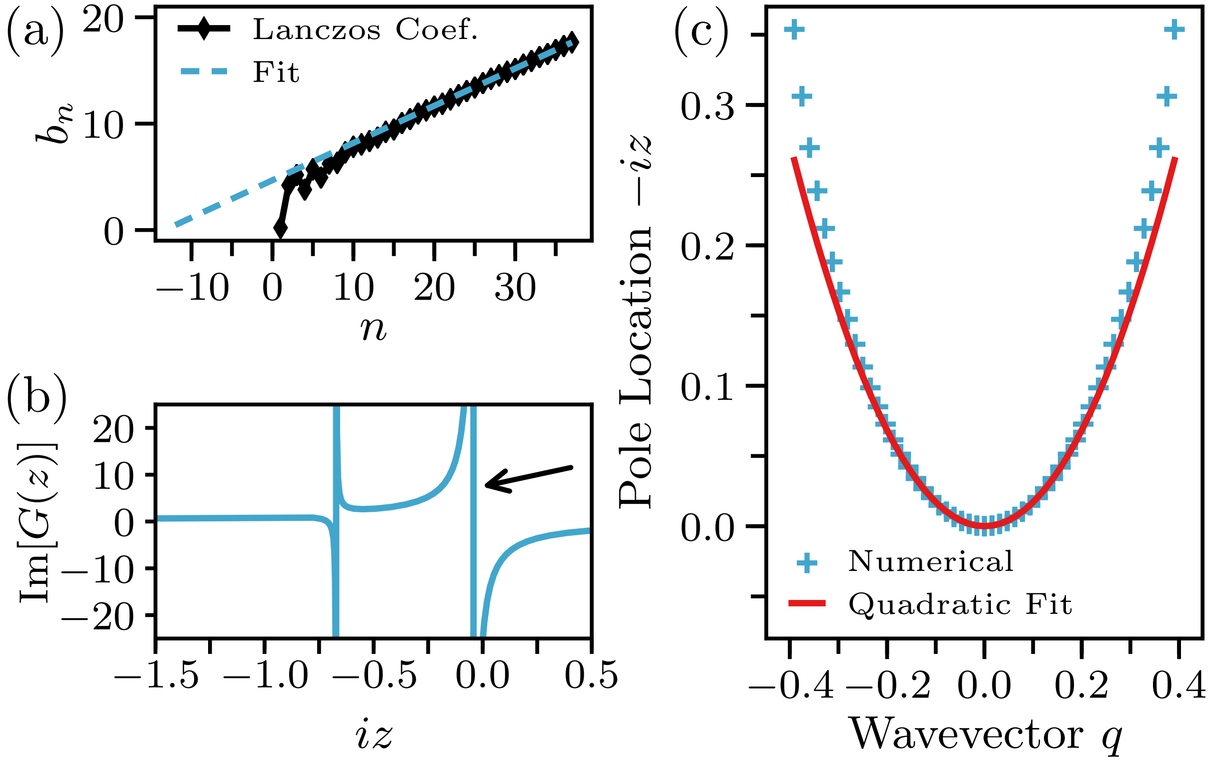


Figure 3.7: Numerical computation of the diffusion coefficient for the energy density operator $\widehat{\mathcal{O}} = \mathcal{E}_q$ in $H = \sum_i X_i X_{i+1} - 1.05 Z_i + 0.5 X_i$. (a) The Lanczos coefficients for $q = 0.15$ are fit to (3.50) with $\alpha = 0.35$ and $\eta = 1.74$. We found it actually better not to approximate $G^{(N)}(z)$ by $\widetilde{G}^{(N)}(z)$, but instead by $\widetilde{G}^{(N+\delta)}(z)$ for some integer offset δ so that $\eta \approx 1$ (in the example shown, $\delta = 12$). Large η or negative values lead to numerical pathologies. (b) The approximate Green's function (3.52) at $q = 0.15$. The arrow shows the "leading" pole that governs diffusion. (c) The locations of the leading poles for a range of q. One can clearly see the diffusive dispersion relation $z = iDq^2/2 + \mathrm{O}(q^4)$. Fitting yields a diffusion coefficient $D = 3.3(5)$.

Choice of Inner Product

A single modification is required to adapt the formalism of recursion method to finite temperature: an operator inner product which incorporates the thermal density matrix. At temperature $T = 1/\beta$ (we set $k_{\mathrm{B}} = 1$), a general operator scalar product is defined by the integral [10]:

$$(A|B)^g_\beta := \frac{1}{Z}\int_0^\beta g(\lambda)\,\mathrm{Tr}[y^{\beta-\lambda} A^\dagger y^\lambda B]\,d\lambda \tag{3.53}$$

where $g(\lambda)$ is some even function on the thermal circle $[0,\beta]$, $y := e^{-H}$, and $Z := \mathrm{Tr}[y^\beta]$ [14]. The choice of the inner product is not arbitrary, but is equivalent to the choice of the correlation function

$$C^g_\beta(t) = (\widehat{\mathcal{O}}|\widehat{\mathcal{O}}(t))^g_\beta = \int_0^\beta g(\lambda)\mathrm{Tr}[\rho_\beta \widehat{\mathcal{O}}^\dagger \widehat{\mathcal{O}}(t+i\lambda)]d\lambda \tag{3.54}$$

(where $\rho_\beta = e^{-\beta H}/Z$), which is in turn determined by the physical context; in fact, only a few choices of g are physically relevant, such as (3.59) and (3.60) below.

Once the inner product is chosen, the Lanczos coefficients are defined by the same Lanczos algorithm with the new norm. Quite explicitly, the recursion is:

$$\begin{aligned} |A_n) &:= \mathcal{L}\,|\widehat{\mathcal{O}}_{n-1})^g_\beta - b^{(g)}_{n-1,T}\,|\widehat{\mathcal{O}}_{n-2})^g_\beta\,, \\ b^{(g)}_{n,T} &:= [(A_n|A_n)^g_\beta]^{1/2}\,, \\ |\widehat{\mathcal{O}}_n)^g_\beta \ &:= \ \left(b^{(g)}_{n,T}\right)^{-1} \ |A_n)\ , \end{aligned} \tag{3.55}$$

for $n = 1,2,3,\ldots$, starting from $|\widehat{\mathcal{O}}_0)^g_\beta := |\widehat{\mathcal{O}})$, $|\widehat{\mathcal{O}}_{-1})^g_\beta := 0$ and $b^{(g)}_{0,T} := 0$. We emphasize that *only* the inner product has been changed compared to the infinite-T version. In fact, the Krylov subspaces $\mathrm{span}\{|\widehat{\mathcal{O}})\,,\mathcal{L}\,|\widehat{\mathcal{O}})\,,\ldots,\mathcal{L}^n\,|\widehat{\mathcal{O}})\}$ are unchanged at finite temperature, and only the notion of orthogonality is different, giving us a new orthogonal basis for those spaces. Also, we have the same *relationships* between the Lanczos coefficients and the correlation function (3.54), as well as its linear transforms, the Green's function and spectral function

$$G^g_\beta(z) := i\int_0^\infty e^{-izt}C^g_\beta(t)dt, \tag{3.56a}$$

$$\Phi^g_\beta(\omega) := \int_{-\infty}^\infty e^{-i\omega t}C^g_\beta(t)dt\,, \tag{3.56b}$$

where the superscript g is not an exponent. For example, the Green function (3.56a) admits the continuous fraction expansion

$$G(z) = \cfrac{1}{z - \cfrac{\Delta^{(g)}_{1,T}}{z - \cfrac{\Delta^{(g)}_{2,T}}{z - \ddots}}}\,,\ \Delta^{(g)}_{n,T} := \left(b^{(g)}_{n,T}\right)^2\,, \tag{3.57}$$

which is identical to (3.49), except that b_n are replaced by the finite-T Lanczos coefficients. Similarly, the results of Appendix 3.A carry over directly.

[14] Precisely, g must satisfy $g(\lambda) \geq 0$, $g(\beta - \lambda) = g(\lambda)$, and $\beta^{-1}\int_0^\beta d\lambda g(\lambda) = 1$. We also restrict to the subspace of operators with zero thermal expectation value, and omit the disconnected term in (3.53).

The statement of the hypothesis at finite temperature is also directly analogous. We hypothesize that a chaotic system should have maximal growth of the Lanczos coefficients,

$$b_{n,T}^{(g)} = \alpha_T^{(g)} n + \gamma + o(1), \tag{3.58}$$

under the same conditions as before. Here $\alpha_T^{(g)} \geq 0$ depends on the inner product. Evidence for the hypothesis at finite T will be provided in Section 3.8.

Though the Lanczos algorithm proceeds in the same way for any choice of inner product, this choice will determine what physical correlation function we end up computing. There are two prominent choices of inner products:

- In linear response theory, we use the "standard" inner product given by $g(\lambda) = [\delta(\lambda) + \delta(\lambda - \beta)]/2$:
  $$(A|B)_\beta^S := \frac{1}{2Z} \operatorname{Tr}[y^\beta A^\dagger B + A^\dagger y^\beta B] \tag{3.59}$$
  that leads to the usual thermal correlation function.

- In quantum field theory, it is often natural to consider the Wightman inner product, which corresponds to $g(\lambda) = \delta(\lambda - \beta/2)$:
  $$(A|B)_\beta^W := \frac{1}{Z} \operatorname{Tr}[y^{\beta/2} A^\dagger y^{\beta/2} B]\,. \tag{3.60}$$
  In particular, this inner product allows us to relate our bound on chaos (3.37) and the finite-temperature bound of Ref. [24].

In equations (3.59) and (3.60) and below, we replace the g by S or W to indicate the choice of standard and Wightman inner product, respectively. At infinite temperature, both inner products reduce to the one $(A|B) = \operatorname{Tr}[A^\dagger B]/\operatorname{Tr}[1]$ considered previously.

The spectral functions of the two choices are related by a well-known identity:

$$\Phi_\beta^W(\omega) = \operatorname{sech}\left(\frac{\omega\beta}{2}\right) \Phi_\beta^S(\omega) \xrightarrow{\omega \gg T} e^{-\beta\omega/2} \Phi_\beta^S(\omega)\,, \tag{3.61}$$

which follows directly from the definition (3.10). The Wightman inner product therefore imposes an extra temperature-dependent exponential decay to the spectral function, due to the suppression of high energy excitation by the two $e^{-\beta H/2}$ factors in (3.60). This observation will be crucial in the following section. On the other hand, it would be very interesting to understand how the high-frequency tail of $\Phi(\omega)_\beta^S$ depends on the temperature.

Bound on Chaos

A key result on quantum chaos at finite temperature is the bound on chaos of Ref. [24]. This universal bound was derived for quantum field theories at finite temperature $T = \beta^{-1}$, and reads as follows

$$\lambda_{L,T} \leq 2\pi T \tag{3.62}$$

in natural units $\hbar = k_{\mathrm{B}} = 1$. It is nontrivial in finite-temperature quantum systems, and is therefore complemented by our bound $\lambda_L \leq 2\alpha$ (3.37) which applies to infinite temperature quantum and classical system. This leads to two natural questions. Can our bound be extended to finite temperature? How does it compare to the universal one?

Since $\alpha_T^{(g)}$ depends on the inner product, and the finite-T OTOC admits various regularizations, it is already a nontrivial task to find the correct formulation of the extension. To make progress we consider the regularization scheme used for four-point OTOCs in [24] to derive the universal bound. This scheme inserts the operators in the thermal circle $[0, \beta)$ with even spacing, as does the Wightman inner product (3.60). This suggests that an extension of the bound $\lambda_L \leq 2\alpha$ to finite temperature can be obtained by comparing the finite-T Lyapunov exponent (as defined in [24]) and the finite-T growth rate defined with the Wightman inner product:

$$\lambda_{L,T} \leq 2\alpha_T^{(W)} \quad \text{(conjecture)}\,. \tag{3.63}$$

We stress that this is a conjecture below infinite temperature. Nevertheless, as we show in Section 3.8 below, exact results in the q-SYK model suggest that (3.63) is plausible and tight.

We now turn to the relation between the conjecture (3.63) and the universal bound, and show that the former infers the latter. By (3.61), the Wightman spectral function decays at least as fast as $e^{-\beta\omega/2}$ at high frequency (because $\Phi_\beta^S(\omega) \leq 1$). By (3.15), this is equivalent to the following upper bound on the Lanczos coefficients growth rate:

$$\alpha_T^{(W)} \leq \pi T\,, \tag{3.64}$$

where $\alpha_T^{(W)}$ denotes the growth rate with Wightman inner product. Therefore, the conjecture (3.63), if true, would be tighter than the universal one $\lambda_{L,T} \leq 2\pi T$ (3.62). At low temperature ($\beta \to \infty$ limit), the decay of $\Phi_\beta^W(\omega)$ is dominated by the factor $e^{-\beta\omega/2}$, so $\alpha_T^{(W)}/(\pi T) \to 1$ and the conjectural bound (3.63) becomes equivalent to the universal one (3.62). This equivalence suggests further the plausibility of the conjecture (3.63).

SYK Model

To illustrate the foregoing discussion, and provide some evidence for the hypothesis at finite-T (3.58) and the conjectural bound on chaos (3.63), let us consider again the example of SYK model.

At low temperatures $T = 1/\beta \ll J$, it is well-known that $\lambda_{L,T} = 2\pi T$ [26] saturates the universal quantum bound (3.62). In this limit, the finite-T autocorrelation function of $\widehat{\mathcal{O}} = \sqrt{2}\gamma_1$ may be computed exactly by conformal invariance [25]. Choosing the Wightman inner product, we have

$$C_\beta^W(t) \propto \operatorname{sech}(t\pi T)^{2/q}\ . \tag{3.65}$$

This is the autocorrelation function of the exact solution (3.25), and corresponds to Lanczos coefficients $b_{n,T}^{(W)} = \pi T\sqrt{n(n-1+\eta)}$. They satisfy the hypothesis (3.58) with $\alpha_T^{(W)} = \pi T$

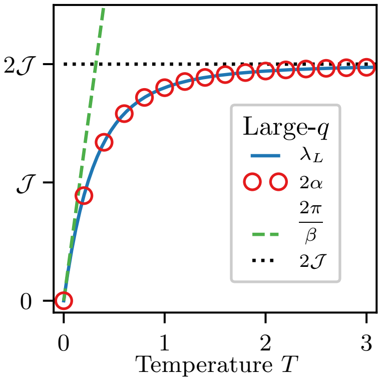


Figure 3.8: Exact Lyapunov exponent $\lambda_L(T)$ (3.97) and growth rate $\alpha(T)$ with the Wightman inner product (3.99) of the SYK model in the large-q limit as a function of temperature (in units of coupling constant $\mathcal{J}$). The conjectured bound $\lambda_L(T) \le 2\alpha(T)_W$ is exactly saturated at all temperatures, while the universal bound $\lambda_L(T) \le 2\pi T$ only saturates in the zero temperature limit.

(3.64). Therefore the low-temperature SYK model saturates also our conjectural bound (3.63).

At finite (but not necessarily low) temperatures, using analytic results in the large-q limit [25], it is not hard to check (see Appendix 3.B) that our conjectured bound (3.63) is saturated, whereas the universal bound (3.62) is not, see Fig. 3.8. This result indicates that an extension of our bound on chaos to finite temperature is at least plausible. The exact agreement between $\alpha_T^{(W)}$ and $\lambda_{L,T}$ is notable given that the former is defined solely from 2-point correlators whereas the latter requires 4-point functions.

We reiterate that the above SYK results depend crucially on the Wightman inner product. If the "standard" inner product (3.59) is chosen instead, the Lanczos coefficients $b_{n,T}^{(S)}$ cannot be extracted from the conformal solution, since that would require the Taylor expansion of $C_\beta^S(t)$ around $t = 0$, at which the conformal solution is non-analytic. A numerical high-temperature expansion (extending the method of Appendix 3.B) and an exact calculation in the large-q limit both indicate that the Lanczos coefficients still grow linearly, but the growth rate *increases* as the temperature decreases.

To summarize, exact calculations in the SYK model support the universal operator growth hypothesis at finite temperature, and the conjectural bound on chaos.

3.9 Conclusions

Discussion

We have presented a hypothesis on the universal growth of operators: the Lanczos coefficients follow the asymptotically linear form $b_n = \alpha n + \gamma + \mathrm{o}(1)$ in non-integrable systems, with a logarithmic correction in 1d. We have seen copious evidence that the hypothesis is satisfied in a wide variety of non-integrable models. Over the course of this work, the growth rate α has emerged as a quantity of prime importance, tying a diverse array of seemingly-disparate ideas together. Let us recount them now:

- $\alpha > 0$ is the slope of asymptotically linear growth of the Lanczos coefficients.
- $\frac{2}{\pi}\alpha = \omega_0$ is the exponential decay rate of the spectral function $\Phi(\omega) \sim e^{-|\omega|/\omega_0}$, which can be (and has been) measured experimentally [35, 36, 37].
- $\pm i\pi/(2\alpha)$ are the locations of the singularities closest to the origin in the (analytic continuation) of the autocorrelation $C(t)$, see Appendix 3.A.
- 2α is the exponential growth rate of Krylov-complexity.
- 2α is an upper bound for the growth of all q-complexities.
- 2α is an upper bound for the Lyapunov exponent (whenever the latter is well-defined), since quantum OTOCs are an example of q-complexities.

We have, of course, put aside the precise conditions and qualifiers of each statement. In light of these results, α plays a central role in operator growth and dynamics of complex systems.

Complexity — especially the Krylov-complexity — arose as a key concept in this work. We would like to highlight its temporal nature which, as we now argue, makes it a more general notion than chaos. Chaos essentially tracks the development of structures at ever-smaller scales in *phase space*. In classical systems, of course, this may proceed indefinitely, while in quantum systems, features smaller than $\hbar$ are ruled out and the process saturates. Chaos therefore cannot carry over straightforwardly to systems deep in the quantum regime, where the phase space volume is comparable to $\hbar$ and saturation occurs almost immediately. The K-complexity, in sharp contrast, measures structures at ever-smaller scales in the *time* domain. We believe this is a fundamental difference; as we have seen, the K-complexity can grow exponentially in quantum systems beyond semiclassical or large-N limits. Operator complexity may well supersede the notion of chaos in quantum dynamics.

Outlook

We would like to understand how our hypothesis can be affected by obstructions to thermalization. Based on evidence available to us, it is tempting to conjecture that they lead to a qualitative slower growth for quantum systems. Confirming this in general would be a remarkable result. However, given the diversity of non-thermalizing situations, it may be more reasonable to explore them on a case by case basis. In free and integrable models, there are an extensive number of conserved local or quasi-local charges. The behavior of the Lanczos coefficients in integrable models is likely non-universal, and depends strongly on the model and operator in question [10]. We wish to gain general analytical insights in this direction (especially for interacting models), by leveraging the knowledge available on the quantum inverse scattering method [88, 89, 90]. Also, it may be desirable to modify the Lanczos algorithm to promote the semi-infinite line to a lattice where the perpendicular direction is generated by commutators against quasi-local conserved charges. Another exceptional case is quantum scar states [91, 92, 93], isolated states that fail to thermalize in otherwise chaotic systems, possibly due to emergent or approximately conserved charges. It would be revealing to see how scars are reflected in the Lanczos coefficients. Finally it would be of great interest to understand the interplay of the hypothesis with many-body localized systems (see [94] and references therein for a review, and [33] for numerical calculations of Lanczos coefficients in disordered spin chains) where thermalization fails.

Our treatment at finite temperatures is far from complete and leaves numerous open questions, especially those concerning the "standard" inner product: How do the Lanczos coefficients grow? If linearly, how does the growth rate depend on the temperature? How can we extend our bound on chaos to finite T? Numerical investigations into these questions are challenging due to the presence of the thermal density matrix [95, 34, 96]. Quantum Monte Carlo seems promising for this problem, as the Lanczos coefficients can be computed without analytic continuation. In low dimensions, DMRG can be also useful: matrix product operators can be used to approximate the thermal state, and the operators in the Lanczos algorithm.

One would like to put the hypothesis on more solid mathematical footing, especially in 1d. Finding analytically tractable models *far from the large-N limit* that achieve the maximal Lanczos coefficient growth seems a formidable problem, which is made even harder by the restriction to time-independent Hamiltonian systems; the only result in this direction is that of [43] in 2d. Many solvable models of quantum chaos (see Refs [23, 97] for notable recent progress) are only defined as unitary maps or Floquet systems. To this respect, a meaningful extension of the hypothesis to such contexts would be a highly rewarding advance.

An alternative route would be to develop an extended (Hermitian) random matrix theory. Standard proofs of the Wigner semicircle law exploit the connections between the moments of a distribution, the combinatorics of Dyck paths, Catalan numbers, and the Stieltjes transform of a distribution [98]. These are directly analogous to the moments μ_{2n}, the combinatorics of Motzkin paths, secant numbers, and the continued fraction expansion for $G(z)$ — all of which arose in the calculation of our exact wavefunction in Appendix 3.D). The non-trivial appearance of the same type of objects in both contexts suggests a strong analogy. We thus speculate that the hypothesis can be derived analytically by introducing a new type of random matrix ensemble that incorporates locality and translation invariance. (This is similar to the framework of [99].) In this case, a Hamiltonian such as $H = \sum_{<x,y>} h_{x,y}$, where $h_{x,y}$ is a random matrix acting on neighboring sites x and y, should obey the hypothesis (3.12) in expectation. Therefore generic, 2-local Hamiltonians would also be expected to obey the hypothesis by concentration of measure. It may well be that showing the hypothesis holds for a specific Hamiltonian is of comparable difficulty to showing the ergodic hypothesis applies to specific classical systems.

Coming back to physics, we argue that there should be a general principle, analogous of the second law of thermodynamics, that governs the operator growth in generic systems. Indeed, the latter is *irreversible*, in the same sense as the dynamics of an isolated gas is so in the thermodynamic limit. We cannot help but wonder what *entropy* is maximized by the operator growth process, and whether any notion of (quantum) dynamical entropy (see e.g. [100, 101, 102, 103] is relevant in describing the process. Elusive as it seems, such a thermodynamic principle might be the ultimate explanation of our empirical observations of ubiquitous maximal operator growth.

To close, we wish to point out that the territory of q-complexities beyond K-complexity and OTOCs is completely unexplored. In generic many-body systems (i.e. not semiclassical) at infinite temperature, these two examples represent two extremes, showing maximal and non-existent exponential growth rates, respectively. The significant gap between them should be filled with potentially more meaningful measures of complexity. These complexities could be entirely new concepts or disguised forms of existing notions such as circuit complexity and entanglement entropy. Hopefully, charting this *terra incognita* will continue to shed new light on the complex nature of many-body quantum dynamics.

Appendices

3.A Brief Review of the Recursion Method

In this appendix we recall the relations between Lanczos coefficients, correlation function, Green function, spectral function, and moments. These relations are mathematical in nature, and apply to any inner product on the operator space, and thereby to finite as well as infinite temperature. For simplicity, we will omit the sub- and superscripts indicating the inner product.

Let us recall the five equivalent representations of the dynamics of an operator:

$$C(t) \leftrightarrow G(z) \leftrightarrow \Phi(\omega) \leftrightarrow \{\mu_{2n}\} \leftrightarrow \{b_n\}. \tag{3.66}$$

The first four are related by linear transformations given in the text. For instance, the *moments* μ_{2n} are the Taylor expansion coefficients of autocorrelation around $t = 0$:

$$C(-it) := \sum_{n=0}^{\infty} \mu_{2n} \frac{t^{2n}}{(2n)!}, \quad \mu_{2n} := (\widehat{\mathcal{O}}|\mathcal{L}^{2n}|\widehat{\mathcal{O}}), \tag{3.67}$$

where the odd terms vanish provided $\widehat{\mathcal{O}}$ is Hermitian. The moments can also be extracted from the spectral function via

$$\mu_{2n} = \int \omega^{2n} \Phi(\omega)\, d\omega. \tag{3.68}$$

All the transformations between the first four quantities are similarly straightforward.

The Lanczos coefficients, on the other hand, are related to the others via a *non-linear* transformation. The rest of this Appendix discusses how to perform the non-trivial translation between the Lanczos coefficients and the moments both asymptotically and numerically.

From Moments to Lanczos Coefficients

Cumulative products of the first n Lanczos coefficients are given by determinants of the Hankel matrix of moments [10]

$$b_1^2 \dots b_n^2 = \det \left(\mu_{i+j}\right)_{0 \le i,j \le n} . \tag{3.69}$$

If the moments are known, the determinant can be computed efficiently by transforming the Hankel matrix into diagonal form. Doing this iteratively for $k \in [1, n]$ provides a fast algorithm that computes $b_1, \ldots, b_n$ from $\mu_2, \mu_4, \ldots, \mu_{2n}$. The algorithm may be expressed concisely as a recursion relation (see Eq. 3.33 of Ref. [10]) as follows:

$$\begin{aligned} b_n &= \sqrt{M_{2n}^{(n)}}\,, \\ M_{2k}^{(0)} &= \frac{M_{2k}^{(m-1)}}{b_{m-1}^2} - \frac{M_{2k-2}^{(m-2)}}{b_{m-2}^2}\,,\ k = m, \ldots, n\,, \\ M_{2k}^{(0)} &= \mu_{2k}\,,\ b_{-1} = b_0 := 1\,,\ M_{2k}^{(-1)} := 0\,. \end{aligned} \tag{3.70}$$

If an analytic expression for $C(t)$ is known, then an arbitrary number of the Lanczos coefficients may be computed numerically via (3.70). We remark that this algorithm suffers from large numerical instabilities due to repeated floating-point divisions.

From Lanczos Coefficients to Moments

It follows from the tridiagonal form of L that the moments may be expressed in terms of the Lanczos coefficients as

$$\mu_{2n} = (\widehat{\mathcal{O}}|\mathcal{L}^{2n}|\widehat{\mathcal{O}}) = (L^{2n})_{00}. \tag{3.71}$$

If the Lanczos coefficients are known, this is a completely combinatorial object. In particular, the moments are given by a sum over Dyck paths. Formally, a Dyck path of length $2n$ can be defined as a sequence $(h_0, h_1, \ldots, h_2 n)$ such that: $h_0 = h_{2n} = 1/2$; $h_k \geq \frac{1}{2}$ and $|h_k - h_{k+1}| = 1$ for any k. These are often visualized as paths starting at height zero where each segment either increases or decreases the height by one unit, with the constraint that the height is always non-negative and returns to zero at the end. Denoting the set of such paths by $\mathcal{D}_n$, we have

$$\mu_{2n} = \sum_{\{h_k\} \in \mathcal{D}_n} \prod_{k=1}^{2n} b_{(h_k + h_{k-1})/2}\,. \tag{3.72}$$

For example, $\mu_2 = b_1^2$ and $\mu_4 = b_1^4 + b_1^2 b_2^2$. The number of Dyck paths of length $2n$ is given by the Catalan numbers $C_n = \frac{(2n)!}{(n+1)!n!}$. A consequence of (3.72) is the following lower bound:

$$\mu_{2n} \geq b_1^2 \ldots b_n^2\,. \tag{3.73}$$

On the other hand, we have the upper bound $\mu_{2n} \leq \max_{k=1}^{n} (b_k^2)\, C_n$. Applying the upper and lower bounds, linear growth of the Lanczos coefficients b_n corresponds to the following growth rate of moments:

$$\mu_{2n} = \exp(2n \ln n + \mathrm{O}(n))\,. \tag{3.74}$$

This equation is a useful reformulation of the linear growth hypothesis.

If the growth rate is known as well, $b_n = \alpha n + \mathrm{O}(1)$, it is possible to refine the asymptotic by specifying the next order exponential term:

$$\mu_{2n} = \left(\frac{4n\alpha}{e\pi}\right)^{2n} e^{\mathrm{o}(n)} \,. \tag{3.75}$$

Combining this equation with the Stirling formula, the correlation function $C(t) = \sum_n \mu_{2n}(it)^{2n}/(2n)!$ has convergence radius $r = \pi/(2\alpha)$, due to singularities at $t = \pm ir$; in fact, $C(t)$ is analytical in the strip $-r < \mathrm{Im}(t) < r$, see Fig. 3.3. Therefore, the Fourier transform of $C(t)$, which is the spectral density $\Phi(\omega)$, has a exponential decay

$$|\Phi(\omega)| = e^{-|\omega|/\omega_0 + o(\omega)} \,,\, \omega_0 = r^{-1} = 2\alpha/\pi \,. \tag{3.76}$$

We illustrate the above results by a simple example: when $b_n = \alpha\, n$, then $C(t) = \mathrm{sech}(\alpha t)$ and $\Phi(\omega) = \frac{\alpha}{\pi}\,\mathrm{sech}\left(\frac{\pi\omega}{2\alpha}\right)$. The moments $\mu_{2n} = 1, 1, 5, 61, 1385, \dots$ are known as Euler or secant numbers and have the asymptotic behavior $\mu_{2n} = 4\sqrt{\frac{4n}{\pi}}\left(\frac{4n}{\pi e}\right)^{2n}(1 + \mathrm{o}(1))$ [104]. We checked that (3.76) and (3.75) hold in all analytic examples we are aware of in the literature and believe them to hold in general.

3.B Moments and Lanczos Coefficients in the SYK Model

In this section we compute the Lanczos coefficients in the large-N SYK model at infinite temperature with the initial operator $\widehat{\mathcal{O}} = \sqrt{2}\gamma_1$. Most often, this is done by computing the moments and applying the mapping described in Section 3.A.

For convenience, we recall the SYK Hamiltonian and disorder normalization:

$$H^{(q)}_{\mathrm{SYK}} = i^{q/2} \sum_{1 \le i_1 < i_2 < \cdots < i_q \le N} J_{i_1 \dots i_q} \gamma_{i_1} \cdots \gamma_{i_q} \,, \tag{3.77}$$

$$\overline{J^2_{i_1 \dots i_q}} = 0, \tag{3.78}$$

$$\overline{J^2_{i_1 \dots i_q}}^2 = \frac{(q-1)!}{N^{q-1}} J^2, \tag{3.79}$$

where the line denotes disorder averages. We shall extend $J_{i_1 \dots i_q}$ to all $i_1, \dots, i_q$ by antisymmetry. As discussed in the main text, disorder-averaging will be assumed throughout. We first describe the general method, and then discuss the large-q limit.

General Method

Since the moments are closely related to the Green function, they can be calculated by the diagrammatic technique commonly used in the SYK literature. Indeed, μ_{2n} can be represented as a sum over diagrams G diagrams with $2n$ vertices:

$$\mu_{2n} = J^{2n} 2^{(2-q)n} \sum_G C_G \,, \tag{3.80}$$

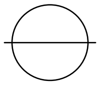
where C_G is the combinatorial factor of the diagram, which counts the number of labellings of the vertices by $1, \ldots, 2n$ such that the labels are increasing from left to right.

Let us illustrate the diagrams with some examples with $q = 4$ and $n = 1, 2$. Direct calculation yields:

$$\begin{aligned}
\mathcal{L}\gamma_1 &= -\sum_{j<k<l} J_{1jkl}\gamma_j\gamma_k\gamma_l \,, \\
\mathcal{L}^2\gamma_1 &= 2^{2-q} \sum_{j<k<l} J_{1jkl}^2 \gamma_1 \\
&\quad + \sum_{j<k<l} J_{1jkl} \sum_{r<s<t} J_{jrst}\gamma_k\gamma_l\gamma_r\gamma_s\gamma_t \\
&\quad + \sum_{j<k<l} J_{1jkl} \sum_{r<s<t} J_{krst}\gamma_j\gamma_l\gamma_r\gamma_s\gamma_t \\
&\quad + \sum_{j<k<l} J_{1jkl} \sum_{r<s<t} J_{lrst}\gamma_j\gamma_k\gamma_r\gamma_s\gamma_t \,.
\end{aligned} \tag{3.81}$$

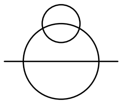
The first two moments μ_2 and μ_4 are (twice) the norm squared of the $\mathcal{L}\gamma_1$ and $\mathcal{L}^2\gamma_1$, respectively. Under disorder averaging, the terms on the right-hand side are orthogonal, and each corresponds to a different diagram:

$$\mu_2 = J^2 2^{(2-q)} \;=$$

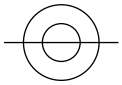


,

$$\mu_4 = J^4 2^{2(2-q)} q \;=$$

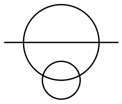


$+$

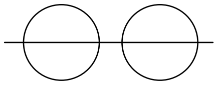


(3.82)

$+$

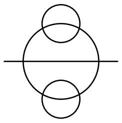


$+$



.

The combinatorial factor is $C_G = 1$ for each of the above graphs. The first non-trivial combinatorial factor is$C_G = 6$ for the diagram



, which contributes to μ_6. The six vertex orderings are $1\genfrac{}{}{0pt}{}{2\;\;3}{4\;\;5}6$, $1\genfrac{}{}{0pt}{}{4\;\;5}{2\;\;3}6$, $1\genfrac{}{}{0pt}{}{2\;\;4}{3\;\;5}6$, $1\genfrac{}{}{0pt}{}{3\;\;4}{2\;\;5}6$, $1\genfrac{}{}{0pt}{}{2\;\;5}{3\;\;4}6$, and $1\genfrac{}{}{0pt}{}{3\;\;5}{2\;\;4}6$.

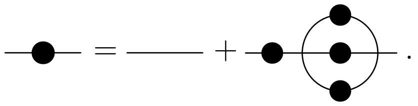
The SYK diagrams encode the Schwinger-Dyson equations governing the autocorrelation and Green's function which are, up to trivial transformations, the exponential and ordinary generating functions of the moments, respectively:

$$zG(z) = 1 + J^2 2^{2-q} G(z)\widetilde{\Sigma}(z), \tag{3.83a}$$

$$\Sigma(t) = C(t)^{q-1}, \tag{3.83b}$$

$$\widetilde{\Sigma}(z) = i\int_0^{\infty} \Sigma(t) e^{-itz} dt, \tag{3.83c}$$

that is, $\widetilde{\Sigma}(z)$ and $\Sigma(t)$ are related by (non-standard) Laplace transform (3.47) just as $G(z)$ and $C(t)$ are. Equation (3.83) can be represented diagrammatically (here for the case $q = 4$) by



(3.84)

The dot represents a general SYK diagram (a fully-dressed Green's function). This is the sum of the bare Green's function, or the time-domain product of $(q-1)$ dressed Green's functions. Note that both exponential and ordinary generating functions are needed to take the combinatorial factors into account: a serial (respectively, parallel) composition of diagrams correspond to product of ordinary (resp. exponential) generating function.

Equation (3.83) has no closed form solution for general q. However, working with the power series representations, it enables the numerical calculation of $\mu_2, \dots, \mu_{2n}$ in polynomial time and space complexity in n. Concretely, the following iteration algorithm can be easily implemented in a computer algebra system:

1. Set $g_0(z) := z^{-1}$, and let $j = 0$.
2. Compute $c_j(t)$ from $g_j(z)$ by replacing z^{-2n-1} with $(it)^{2n}/(2n)!$.
3. Set $\sigma_j(t) := c_j(t)^{q-1}$ up to order t^j.
4. Compute $\widetilde{\sigma}_j(z)$ from $\sigma_j(t)$ by replacing $(it)^{2n}$ with $z^{-2n-1}(2n)!$.
5. Set $g_{j+1}(z) := (1 + J^2 2^{2-q} g(z)\widetilde{\sigma}_j(z))/z$ up to order t^{j+1}.
6. Increment j by 1 and repeat from step 2.

When the above procedure is stopped at $j = n$, the result $g_n(z)$ will be a polynomial truncation of the Green function: $g_n(z) = \sum_{j=0}^{n} \mu_{2j} z^{-2j-1}$, which contains the correct moments up to μ_{2n}. They can be then used to compute Lanczos coefficients $b_1^2, \dots, b_n^2$ by the recipe (3.70). Arbitrary-precision rational number arithmetic is necessary for $n \sim 10^2$, since the moments grow very fast. We calculated b_n for a few different values of q up to $n = 100$, and extracted the linear slope by a linear fit. The results are reported in Table 3.3 and Fig. 3.4 (a).

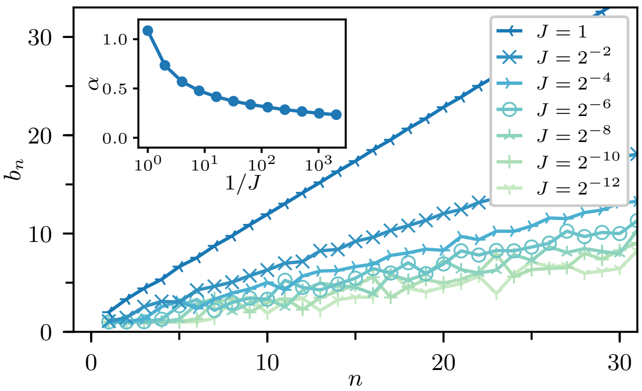


Figure 3.B.1: Change in the growth rate near integrability for the SYK model with $q = 2$ and $q = 4$ (3.85). The ratio of the $q = 4$ to $q = 2$ term is given by J, and the model becomes free at $J = 0$.

The above method can be readily adopted to variants of SYK where two-body and four-body interactions coexist:

$$H = H^{(4)}_{\text{SYK}}(J) + H^{(2)}_{\text{SYK}}(J = 1)\,. \tag{3.85}$$

One only needs to replace the last term in (3.83b) by a sum over $q = 2$ and $q = 4$ with the corresponding coupling constants. Since the $q = 2$ model is non-interacting, eq. (3.85) can be another model to study the effect of weak thermalizing interaction on the Lanczos coefficients. The results, shown in Fig. 3.B.1, are qualitatively consistent with those from the Ising model (Fig. 3.4): the linear growth rate depends only weakly on the interaction strength J as it goes to zero. Quantitative, a logarithmic dependence

$$\alpha \sim 1/\ln(1/J) \tag{3.86}$$

describes the numerical data well for vanishing J.

Large-q limit

In the large-q limit, (3.83) can be solved analytically. It is convenient to define the coupling constant [25, 41]

$$\mathcal{J}^2 := 2^{1-q}\, q\, J^2\,. \tag{3.87}$$

It is then known [25, 41] that $C(t)$ admits a $1/q$ expansion

$$C(t) = 1 + \frac{1}{q}\mathcal{C}(t) + \mathrm{O}(1/q^2)\,, \tag{3.88}$$

where the leading non-trivial term satisfies the following differential equation:

$$\mathcal{C}''(t) = -2\mathcal{J}^2 e^{\mathcal{C}(t)}\,,\; \mathcal{C}(0) = \mathcal{C}'(0) = 0\,, \tag{3.89}$$

whose solution is

$$C(t) = 1 + \frac{2}{q} \ln \operatorname{sech}(\mathcal{J} t) + \mathrm{O}(1/q^2) \,. \tag{3.90}$$

The corresponding moments

$$\mu_{2n} = \frac{2}{q} \mathcal{J}^{2n} T_{n-1} + \mathrm{O}(1/q^2) \,,\, n > 0 \,, \tag{3.91}$$

where $(T_n)_{n=0}^{\infty} = (1, 2, 16, 272, 7936, \dots)$ are the tangent numbers [105]. The generating function of T_n admits a continued fraction expansion [105]:

$$\sum_{n=0}^{\infty} T_n x^n = \cfrac{1}{1 - \cfrac{1 \times 2x}{1 - \cfrac{2 \times 3x}{1 - \cfrac{3 \times 4x}{1 - \ddots}}}} \tag{3.92}$$

Using this, one can obtain the following Lanczos coefficients for the large-q SYK model

$$b_n^{\mathrm{SYK}} = \begin{cases} \mathcal{J} \sqrt{2/q} + \mathrm{O}(1/q) & n = 1 \\ \mathcal{J} \sqrt{n(n-1)} + \mathrm{O}(1/q) & n > 1 \,. \end{cases} \tag{3.93}$$

It is not hard to check using (3.8) that the wavefunction on the semi-infinite chain is

$$\varphi_n(t) = \begin{cases} 1 + \dfrac{2}{q} \ln \operatorname{sech}(\mathcal{J} t) + \mathrm{O}(1/q^2) & n = 0 \\ \tanh(\mathcal{J} t) \sqrt{\dfrac{2}{nq}} + \mathrm{O}(1/q^2) & n > 0 \,. \end{cases} \tag{3.94}$$

The corresponding probability distribution is identical to the operator size distribution (see Eq. (5.11) of Ref. [41]):

$$P_s(t) = |\varphi_n(t)|^2 \,,\, s = 1 + n(q-2) \,. \tag{3.95}$$

The large-q SYK model is also solvable at any finite temperature [25]. The temperature T is parametrized by $v \in (0, 1)$ via

$$\frac{T}{\mathcal{J}} = \frac{\cos \frac{\pi v}{2}}{\pi v} \,. \tag{3.96}$$

The limits $T \to \infty$ and $T \to 0$ correspond to $v \to 0$ and $v \to 1$, respectively. The Lyapunov exponent is then

$$\lambda_{L,T} = 2 v \pi T \,, \tag{3.97}$$

and the autocorrelation under the Wightman inner product (3.60) is

$$C^W_\beta(t) = 1 + \frac{2}{q}\ln\operatorname{sech}(vt\pi T) + \mathrm{O}(1/q^2)\,. \tag{3.98}$$

Comparing to (3.90), we see immediately that

$$b^{(W)}_{n,T} = \begin{cases} v\pi T\sqrt{2/q} + \mathrm{O}(1/q) & n = 1 \\ v\pi T\sqrt{n(n-1)} + \mathrm{O}(1/q) & n > 1\,. \end{cases} \tag{3.99}$$

Therefore the finite-T growth rate with the Wightman inner product is

$$\alpha^{(W)}_T = v\pi T \tag{3.100}$$

at any temperature. Thus, the bound $\lambda_{L,T} \le 2\alpha^{(W)}_T$ is saturated at all temperature in the SYK model, whereas the bound $\lambda_{L,T} \le 2\pi T$ is only so in the zero-temperature limit (see Fig. 3.8). Using the relation between growth rate and spectral function decay rate (3.15) and the relation (3.61) between spectral functions of different inner products, it is not hard to obtain the growth rate with the standard inner product from (3.100):

$$\alpha^{(S)}_T = \frac{v\pi T}{1-v}\,. \tag{3.101}$$

Using (3.96), we obtain the limits $\alpha(T)_S \to \mathcal{J}\pi/2$ as $T \to 0$ and $\alpha^{(S)}_T \to \mathcal{J}$ as $T \to \infty$. We notice that $\alpha(T)_S$ *increases* at low temperatures while, in contrast, $\alpha^{(W)}_T$ decreases.

3.C Numerical Details for 1d Spin Chains

This section discusses the numerical details involved in computing the Lanczos coefficients and Krylov basis vectors in 1D spin chains. We work directly in the thermodynamic limit of a chain with $N \to \infty$ sites. However, bookkeeping will reduce this to finite-dimensional matrix multiplication.

Suppose we have a translation-invariant k-local Hamiltonian $H = \sum_n h_n$ and an ℓ-local operator $\widehat{\mathcal{O}} = \sum_n \widehat{\mathcal{O}}_n$. Here h_n and $\widehat{\mathcal{O}}_m$ are operators starting on sites n or m respectively. (For instance, we might have $\widehat{\mathcal{O}}_2 = \cdots \otimes I_1 \otimes X_2 \otimes Z_3 \otimes I_4 \otimes \cdots$.) We normalize the operators so that $(h_n|h_n) = 1 = (\widehat{\mathcal{O}}_m|\widehat{\mathcal{O}}_m)$. At minor additional computational cost, we can work with an operator at a finite wavevector q:

$$\widehat{\mathcal{O}}_q = \sum_n \widehat{\mathcal{O}}_n e^{iqn}. \tag{3.102}$$

The crucial point is that applying the Liouvillian to $\widehat{\mathcal{O}}_q$ is another operator at wavevector q by using translation-invariance to re-index the sum at the cost of phase factors. Explicitly,

$$[H, \widehat{\mathcal{O}}_q] = \sum_{m,n} [h_n, \widehat{\mathcal{O}}_m] e^{iqm} = \sum_m \widehat{\mathcal{O}}'_m e^{iqm} \tag{3.103}$$

where

$$\widehat{\mathcal{O}}'_m = \sum_{n=m-k+1}^{m-\ell+1} e^{iqs_{nm}}[h_{n+s_{nm}}, \widehat{\mathcal{O}}_{m+s_{nm}}] \tag{3.104}$$

where the shift is s_{nm} is the index of the first non-identity site of $[h_n, \widehat{\mathcal{O}}_m]$ minus m, which is needed to keep track of how much the support of the operator shifted due to the commutator. One can check that $\widehat{\mathcal{O}}'_m$ starts on site m.

Therefore we only need to keep track of operators starting on a single site, say site 0. We adopt the basis of Pauli strings and, following, e.g. [106], we adopt a representation which minimizes the computational cost of taking commutators. Since $iY = ZX$, we may adopt a representation

$$i^\delta(-1)^\epsilon Z_1^{v_1} X_1^{w_1} \otimes \cdots \otimes Z_n^{v_n} X_n^{w_n} \tag{3.105}$$

where $\delta, \epsilon, v_k, w_k \in \{0, 1\}$, i.e. a Pauli string of length n may be represented by two binary vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ of length n and two binary digits. So if $\tau_1 = i^{\delta_1}(-1)^{\epsilon_1} Z^{\boldsymbol{v}_1} X^{\boldsymbol{w}_2}$ and $\tau_2 = i^{\delta_2}(-1)^{\epsilon_2} Z^{\boldsymbol{v}_2} X^{\boldsymbol{w}_2}$, then their commutator is a string $\tau' = [\tau_1, \tau_2]$ with

$$\begin{aligned}
\delta' &= \delta_1 + \delta_2, \\
\epsilon' &= \epsilon_1 + \epsilon_2 + \delta_1\delta_2 + \boldsymbol{w}_1 \cdot \boldsymbol{v}_2, \\
\boldsymbol{v}' &= \boldsymbol{v}_1 + \boldsymbol{v}_2, \\
\boldsymbol{w}' &= \boldsymbol{w}_1 + \boldsymbol{w}_2.
\end{aligned} \tag{3.106}$$

All additions are performed over $\mathbb{Z}_2$.

With this setup, the Lanczos coefficients can be computed in a similar way to matrix-free exact diagonalization codes. A translation-invariant operator can be stored as a hash map of Pauli strings starting on site zero with complex coefficients. The Liouvillian is applied by combining (3.103), (3.104), and (3.106). Of course, it is not necessary to take $\widehat{\mathcal{O}}$ to be translation invariant. One could equally well take a small single-site operator and apply the same technique without the sum over all sites. We note that the Lanczos algorithm (3.4) only requires the storage of three operators at any time. In practice the method described here allows a few dozen Lanczos coefficients to be computed in a few minutes on a modern laptop and is generally memory-limited by the exponential increase in the number of Pauli strings required.

Once the Lanczos coefficients and Krylov vectors have been computed, it is possible to understand how the operators $\widehat{\mathcal{O}}_n$ grow in physical space. One way to characterize this is in terms of the distribution of string lengths in each $\widehat{\mathcal{O}}_n$. If $\widehat{\mathcal{O}}_n = \sum_a c_a \sigma^a$, where the sum runs over all Pauli strings a, then the distribution is defined by $P_n(s) = \sum_{a:|a|=s} |c_a|^2$. This distribution is shown for the Hamiltonian H_1 with the parameters given in Fig. 3.4. The mean and variance of the distribution grow with n. We have observed that the distribution $P_n(s)$ appears to be highly model-dependent. This makes it difficult to translate information about the exponential spreading of the wavefunction in the semi-infinite chain back to physical space.

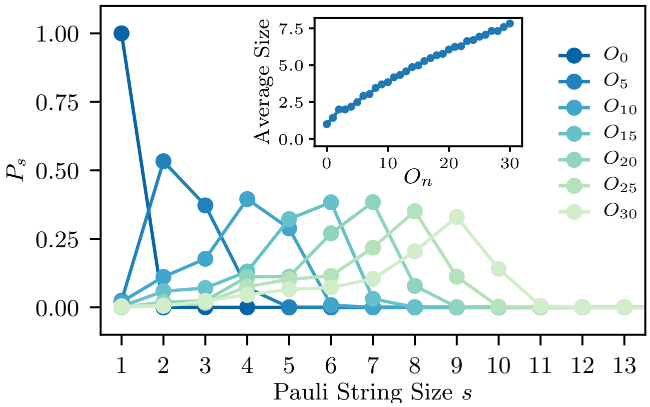


Figure 3.C.1: The size distribution of the Pauli strings in the Krylov vectors $\widehat{\mathcal{O}}_n$ for the Hamiltonian H_1 with parameters and initial operator as in Fig. 3.4. Though the distribution drops quickly after its peak, $P_n(s)$ is supported on $[0, \lfloor n/2 \rfloor + 2]$.

3.D A Family of Exact Solution with Linear Growth

This section will provide a derivation for the exact solution (3.25) of the 1d quantum mechanics problem with Lanczos coefficients

$$b_n = \alpha\sqrt{n(n-1+\eta)}\,. \tag{3.107}$$

To solve this problem, notice that our infinite, tri-diagonal matrix is actually quite a familiar setup. If instead we had $b_n = \sqrt{n}$, then L would be the matrix representing the Hamiltonian for the quantum harmonic oscillator in the basis of raising and lowering operators. So really this is just a 1d quantum mechanics problem, albeit not a standard one. In particular, it is known that the system described by L has very high symmetry, due to an infinite-dimensional representation of the Lie algebra $\mathfrak{su}(1,1)$, enabling us to find an exact solution [107, 108]. Indeed, there is a rich mathematical literature on the close connections between representations of $\mathfrak{su}(1,1)$, the combinatorics of Motzkin paths, and Meixner orthogonal polynomials [109, 110]. Our solution will be a simple application of these mathematical results.

We start with some generalities on orthogonal polynomials. Define $L_{(n)} = L_{0\leq i<n, 0\leq j<n}$ to be the $n \times n$ matrix in the upper-left block of L. For example,

$$L_{(3)} = \begin{pmatrix} 0 & b_1 & 0 \\ b_1 & 0 & b_2 \\ 0 & b_2 & 0 \end{pmatrix}. \tag{3.108}$$

We then define polynomials for each n via

$$Q_n(z;\alpha,\eta) = \det\left(z - L_{(n)}\right). \tag{3.109}$$

By performing a cofactor expansion for the determinant on the nth row, the Q's admit a three-term recursion relation

$$Q_{n+1}(z) = zQ_n(z) - b_n^2 Q_{n-1}(z), \tag{3.110}$$

together with initial conditions $Q_0(z) = 1$ and $Q_{-1}(z) = 0$. Eq. (3.110) should be compared with

$$Le_n = b_{n+1}e_{n+1} + b_n e_{n-1}\,, \tag{3.111}$$

where $\{e_n\}$ is the natural orthonormal basis of L. In fact, (3.110) and (3.111) are equivalent, under the identification:

$$Q_n(z) = \left[\prod_{k=1}^{n} b_k\right] e_n\,,\; z^n = L^n e_0\,. \tag{3.112}$$

Therefore, the polynomials $Q_n(z)$ are orthogonal, but not normalized. Instead they are monic, i.e., the highest order coefficient is unity: $Q_n(z) = z^n + \mathrm{O}(z^{n-1})$.

By construction, both $\{Q_k(z)\}$ and $\{z^n\}$ are a basis of $\mathbb{C}[z]$ and can be related by a triangular linear transform with matrix elements $\mu_{n,k}$:

$$z^n = \sum_{k=0}^{n} \mu_{n,k} Q_k(z)\,. \tag{3.113}$$

Combined with (3.112), and by orthonormality of $\{e_n\}$, we have

$$(e_d|L^n|e_0) = \mu_{n,d} \prod_{k=1}^{d} b_k\,, \tag{3.114}$$

and therefore

$$(e_d|e^{iLt}|e_0) = \prod_{k=1}^{d} b_k \sum_{n=0}^{\infty} \frac{(it)^n}{n!} \mu_{n,d}\,. \tag{3.115}$$

The statements so far are general and apply to any set of Lanczos coefficients.

In the specific case $b_n = \sqrt{n(n-1+\eta)}$ (the extra overall factor α in (3.107) can be recovered by a simple time rescaling), one may recognize from the recursion relation (3.110) that Q_n's are a special case of the Meixner polynomials of the second kind [111]. They are a non-classical family of orthogonal polynomials defined by the following three-term recursion: [112, 113]

$$\begin{aligned} M_{n+1}(z;\delta,\eta) &= (z-\lambda_n)M_n(z;\delta,\eta) - b_n^2 M_{n-1}(z), \\ \lambda_n &= (2n+\eta)\delta, \\ b_n^2 &= \left(\delta^2+1\right) n(n-1+\eta). \end{aligned} \tag{3.116}$$

In particular, $Q_n(z) = M_n(z; \delta = 0, \eta)$. For these polynomials, the matrix elements $\mu_{n,d}$ have been exactly calculated, in terms of the following generating function [110]:

$$\begin{aligned}&\sum_{n=0}^{\infty}\sum_{d=0}^{n}\mu_{n,d}w^d\frac{\tau^n}{n!}\\ &= \frac{\sec(\tau)^\eta}{(1-\delta\tan(\tau))^\eta}\exp\left(w\frac{\tan(\tau)}{1-\delta\tan(\tau)}\right). \end{aligned} \tag{3.117}$$

As a side note, we mention that the above generating function, referred to as that of the "inverse polynomials" in the theory of orthogonal polynomial, is closely related to the generating function of Meixner polynomials themselves. The latter has also a closed form expression, known to be of Sheffer type [109, 112]:

$$\begin{aligned}&\sum_{n\geq 0} M_n(z;\delta,\eta)\frac{\tau^n}{n!} \\ &= \left[(1+\tau\delta)^2+\tau^2\right]^{-\eta/2}\exp\left(z\arctan\left(\frac{\tau}{1+\tau\delta}\right)\right). \end{aligned} \tag{3.118}$$

Now, taking $\delta = 0$ and the series coefficient of w^d in (3.117), we have

$$\sum_{n=0}^{\infty}\mu_{n,d}\frac{\tau^n}{n!} = \frac{1}{d!}\sec(\tau)^\eta\tan(\tau)^d\,.$$

Applying this to (3.115), and recalling $b_n = \sqrt{n(n-1+\eta)}$, we obtain the wavefunction solution

$$(e_n|e^{iLt}|e_0) = i^n\sqrt{\frac{(\eta)_n}{n!}}\tanh(t)^n\,\text{sech}(t)^\eta, \tag{3.119}$$

where $(\eta)_n = \eta(\eta+1)\cdots(\eta+n-1)$ is the Pochhammer symbol. The general solution for $b_n = \alpha\sqrt{n(n-1+\eta)}$ can be obtained by a simple rescaling $t \mapsto \alpha t$, and is precisely Eq. (3.25) of the main text where, of course, $(\widehat{\mathcal{O}}_n|e^{i\mathcal{L}t}|\widehat{\mathcal{O}}_0) = (e_n|e^{iLt}|e_0)$. The special case $\eta = 1$ of this family of solutions is well-known [10, 38]. To the best of our knowledge, the general solution (3.119) has not been applied to the recursion method.

3.E Derivation of the q-Complexity Bound

This Appendix will derive Eq. (3.33), $(\mathcal{Q})_t \leq C(n)_t$ for $C = 2M$. The main idea of is that the definition of $\mathcal{Q}$ guarantees that the eigenbasis of $\mathcal{Q}$ is dilated by a factor of at most C compared to the Krylov basis.

We first show that the Krylov basis vectors have a bounded number of components in the $\mathcal{Q}$ basis due to the dilation property. For any operator Φ where there is an $R > 0$ such that $(q_a|\Phi) = 0$ for $q_a > R$, the hypothesis (3.28b) implies that $(q_a|\mathcal{L}|\Phi) = 0$ for $q_a > R+M$.

Using (3.28c), as a base case for induction, we have $(q_a|\mathcal{L}^n|\widehat{\mathcal{O}}) = 0$ for $q_a > M(n+1)$ and, in particular, for $q_a > Cn$. By the construction of the Krylov basis,

$$(q_a|\widehat{\mathcal{O}}_n) = 0 \quad \text{if } q_a > Cn. \tag{3.120}$$

We claim that (3.120) implies

$$(\Phi|\mathcal{Q}|\Phi) \leq C(\Phi|n|\Phi) \tag{3.121}$$

for any operator wavefunction Φ; taking $\Phi = \widehat{\mathcal{O}}(t)$, we obtain (3.33).

To show (3.121), we introduce projectors to large spectral values in the Krylov and $\mathcal{Q}$ bases, respectively:

$$\mathcal{P}_n^K = \sum_{m\geq n} |\widehat{\mathcal{O}}_m)\,(\widehat{\mathcal{O}}_m|\,, \quad \mathcal{P}_q^Q = \sum_{a\,:\,q_a\geq q} |q_a)\,(q_a|\,. \tag{3.122}$$

Then, we have for $n = q/C$,

$$\mathcal{P}_q^Q(1-\mathcal{P}_{n=q/c}^K) = \sum_{a\,:\,q_a\geq q}\sum_{m<n} |q_a)\,(q_a|O_m)\,(\widehat{\mathcal{O}}_m| = 0,$$

because $m < n = q/C \leq q_a/C$, $(q_a|O_m) = 0$ by (3.120). Equivalently,

$$\mathcal{P}_q^Q\mathcal{P}_{q/c}^K = \mathcal{P}_q^Q\,. \tag{3.123}$$

Applying this equation and its Hermitian conjugate, we have

$$\begin{aligned}(\Phi|\mathcal{P}_q^Q|\Phi) &= (\Phi|\mathcal{P}_q^Q\mathcal{P}_{q/C}^K|\Phi)\\ &= (\Phi|\mathcal{P}_{q/C}^K\mathcal{P}_q^Q\mathcal{P}_{q/C}^K|\Phi)\\ &\leq (\Phi|\mathcal{P}_{q/C}^K\mathcal{P}_{q/C}^K|\Phi)\\ &= (\Phi|\mathcal{P}_{q/C}^K|\Phi).\end{aligned} \tag{3.124}$$

where the inequality follows from the fact that $\mathcal{P}_q^Q$ is a projector. Finally we need a standard integration-by-parts identity that converts the expectation value to an integral over the projectors:

$$\begin{aligned}(\Phi|\mathcal{Q}^k|\Phi) &= \int_0^\infty dq\; kq^{k-1}(\Phi|\mathcal{P}_q^Q|\Phi)\,,\\ (\Phi|n^k|\Phi) &= \int_0^\infty dn\; kn^{k-1}(\Phi|\mathcal{P}_n^K|\Phi)\end{aligned} \tag{3.125}$$

for any $k = 1, 2, 3, \ldots$. Combining the case $k = 1$ and (3.124), we obtain

$$\begin{aligned}(\Phi|\mathcal{Q}|\Phi) &= \int_0^\infty dq\;(\Phi|\mathcal{P}_q^Q|\Phi)\\ &\leq \int_0^\infty dq\;(\Phi|\mathcal{P}_{q/C}^K|\Phi)\\ &= C(\Phi|n|\Phi)\,,\end{aligned} \tag{3.126}$$

which finishes the proof. More generally, for any k, we have

$$(\mathcal{Q}^k)_t \leq C^k (n^k)_t \,. \tag{3.127}$$

This is useful as a bound on the growth rate of higher moments of the q-complexity superoperator. See Section 3.6 for an application.

3.F Geometric Origin of the Upper Bounds

In this appendix we derive the geometric upper bound for the Lanczos coefficients in one-dimensional quantum systems. The main object of our analysis will be the growth of the moments $\mu_{2n} = (\widehat{\mathcal{O}}|\mathcal{L}^{2n}|\widehat{\mathcal{O}}) = ||\mathcal{L}^n \widehat{\mathcal{O}}||^2$. Moments and Lanczos coefficients are equivalent, and Appendix 3.A details how to translate between them.

To warm up, we first show a bound corresponding to linear growth (using essentially the same argument as in [46, 49]). This is asymptotically tight in $d > 1$. Suppose we have a 2-local Hamiltonian $H = \sum_x h_x$ and a 1-local operator $\widehat{\mathcal{O}}$ (the general case of r-local h_x and r-local $\widehat{\mathcal{O}}$ can be reduced to the previous case by a block renormalization step that groups consecutive sites into renormalized sites). The Liouvillian becomes a sum of terms $\mathcal{L} = \sum_x \ell_x$ with $\ell_x = [h_x, \cdot]$. We suppose that the local terms are uniformly bounded, i.e., for all x, $||h_x|| \leq \mathcal{E}$. Now, the moment μ_{2n} is the norm-squared of the sum

$$\mathcal{L}^n \widehat{\mathcal{O}} = \sum_{x_1, x_2, \ldots, x_n} \ell_{x_n} \cdots \ell_{x_2} \ell_{x_1} \widehat{\mathcal{O}}. \tag{3.128}$$

This sum is highly constrained by the spatial structure of the spin chain. The operator $\widehat{\mathcal{O}}$ is supported only on one site, and the applications of the Liouvillian grow that support at the edges. Each term in (3.128) can be visualized as a discrete quantum circuit, where each gate $\ell_{x_{k+1}}$ must act on at least one site that is already in the support of $\ell_{x_k} \cdots \ell_{x_1} \widehat{\mathcal{O}}$ — otherwise the term vanishes due to the commutator. This condition is satisfied by at most $(k+1) \leq 2k$ positions x_k, so the total number of non-zero terms in (3.128) is at most $2^n n!$ for large n. The value of each non-zero term is itself bounded due to the finite local bandwidth $\mathcal{E}$, so $||\ell_{x_n} \cdots \ell_{x_1} \widehat{\mathcal{O}}||^2 \leq (2\mathcal{E})^{2n}$. By the triangle inequality, we have

$$\mu_{2n} = ||\mathcal{L}^n \widehat{\mathcal{O}}||^2 \leq (n!)^2 (4\mathcal{E})^{2n}. \tag{3.129}$$

By Stirling's formula, the right hand side has the same asymptotics as (3.21), which corresponds to linear growth of the b_n's. Hence (3.129) implies that the Lanczos coefficients can grow at most linearly in any dimension.

Notice that, the bound comes essentially from counting the number of sequences $x_1, \ldots, x_n$ that give rise to a nonzero contribution to (3.128). In what follows we show that, in one dimension, there is a sharper upper bound on this number, leading to the sub-linear growth announced in Section 3.4. For this, we suppose without loss of generality that $\widehat{\mathcal{O}}$ is supported on site 0 and h_x on sites x and $x+1$. Then it is not hard to see that a $\ell_{x_n} \cdots \ell_{x_2} \ell_{x_1} \widehat{\mathcal{O}} \neq 0$ *only if* for all $k = 1, \dots, n$,

$$\begin{aligned} & L_k \leq x_k \leq R_k \,, \text{ where} \\ & L_k := \min\{x_1, \dots, x_{k-1}, 0\} - 1 \,, \\ & R_k := \max\{x_1, \dots, x_{k-1}, -1\} + 1 \,. \end{aligned} \tag{3.130}$$

We define $\mathcal{P}_n$ to be the set of $(x_1, \dots, x_n)$'s that satisfy (3.130) and denote its size by $P_n := |\mathcal{P}_n|$. Then, similarly to (3.129), we have

$$\mu_{2n} \leq P_n^2 (2\mathcal{E})^{2n} \,. \tag{3.131}$$

Hence bounding μ_{2n} reduces to bounding P_n, which is a completely combinatorial problem.

To produce this combinatorial bound, we partition the set $\mathcal{P}_n$ as follows

$$\begin{aligned} & \mathcal{P}_n = \bigcup_{\ell=1}^{n} \mathcal{P}_{n,\ell} \,, \text{ where} \\ & \mathcal{P}_{n,\ell} := \{(x_1, \dots, x_n) \in \mathcal{P}_n : \ell = L_n - R_n\} \,. \end{aligned} \tag{3.132}$$

Intuitively, if the support of the operator grows to size $\ell + 1$ after n applications of Liouvillian, then $(x_1, \dots, x_n) \in \mathcal{P}_{n,\ell}$. By "size", we mean the distance between the endpoints, disregarding the "holes" between them. In the 1d case, the operator size can only grow in two places: the left and right sides. Therefore, for any $(x_1, \dots, x_n) \in \mathcal{P}_{n,\ell}$, $x_k = L_k$ or $x_k = R_k$ must hold for ℓ values of k among $1, \dots, n$: for each of such k's, one has only two choices for x_k. For the remaining $n - \ell$, there are (at most) ℓ choices (by (3.130), minus 2 boundary choices). Summarizing, we have

$$|\mathcal{P}_{n,\ell}| \leq \binom{n}{\ell} 2^\ell \ell^{n-\ell} \leq 4^n \ell^{n-\ell} \,, \tag{3.133}$$

where the binomial coefficient counts the choices of the ℓ values. Combining this with (3.132), we have

$$P_n \leq n 4^n \max_{\ell \in [0,n]} \ell^{n-\ell} \,. \tag{3.134}$$

In the limit $n \gg 1$, the maximum is attained at $\ell = n/W(n)$ where W is the product-log function defined by $z = W(ze^z)$. For large n, $W(n) = \ln n - \ln\ln n + o(1)$, so

$$P_n \leq n 4^n \left(\frac{n}{W(n)}\right)^{n - \frac{n}{W(n)}} = \frac{n! 4^n}{(\ln n)^n} e^{o(n)}. \tag{3.135}$$

where we used $n/W(n) = e^{W(n)}$ and Stirling's formula. Therefore

$$\mu_{2n} \leq (4\mathcal{E})^{2n} \frac{(n!)^2}{(\ln n)^{2n}} e^{o(n)} \,, \tag{3.136}$$

which grows more slowly than the moment asymptotics corresponding to a linear growth with rate α (3.75), $b_n \ll \left(\frac{4n\alpha}{e\pi}\right)^{2n}$, for *any* $\alpha > 0$. So the Lanczos coefficients corresponding to (3.136) must be sub-linear.

What, then, is the fastest possible growth of the b_n's in 1D? Although we cannot bound the individual Lanczos coefficients in a useful way from the bound on the moments, we can use the bound on their cumulative product $\ln \prod_{k=1}^{n} b_k^2 \leq \ln \mu_{2n}$ (3.73) and differentiate with respect to n. As a result, we find

$$b_n = A\frac{n}{W(n)} = Ae^{W(n)} \sim \frac{An}{\ln n}\,. \tag{3.137}$$

The bound (3.73) (together with (3.136)) is satisfied asymptotically by the above choice of b_n if and only if $A \leq 4\mathcal{E}/e$. Therefore, $b_n = ae^{W(n)}$ captures the correct asymptotic behavior of the upper-bound in the moments, and qualifies as the maximal growth rate of Lanczos coefficients in 1d.

Chapter 4

Local Matrix Product Operators

4.1 iMPOs Introduction

This chapter will focus on the structure and practical computation of local operators. In the previous chapter we explored the connection between local operators and chaos. One theme was that the growth in quantity of computational resources needed to store an operator grows in time “as quickly as possible” in a chaotic system. This leads to a natural question: given a local operator, what is the “best” or “cheapest” way to represent it? This chapter is devoted to answering this question in the computationally convenient setting of Matrix Product Operators (MPO)s. Our main result is an algorithm for compressing operators, which provides the most accurate representation of an operator for a set quantity of resources. We will see that not only is this result useful for understanding the structure of time-evolution and quantum chaos, but it also enables one to compute the ground states of long-range 2D systems via matrix product state methods.

While it is now well understood how matrix product states (MPS) can approximate 1d ground states [114, 115, 116, 117, 118, 119, 120], matrix-product representations of operators (MPOs) remain less understood[1]. MPOs feature prominently in modern implementations of the density matrix renormalization group (DMRG) [115], yet we lack a complete understanding of the resources required for an MPO approximation of a complicated (but local) operator, an important ingredient for several problems of current interest. For instance, DMRG calculations of 1d systems with long-ranged interactions or 2d cylinder geometries are hampered by the large bond dimension of MPO representations of the Hamiltonian. Complex operators also arise during the Heisenberg evolution of simpler ones, so efficient numerical representations would have wide ranging applications in the study of quantum thermalization and the emergence of hydrodynamics.

While an MPO can formally be treated as an MPS in a doubled Hilbert space, this neglects the special structure of operators like Hamiltonians: they are a sum of local terms,

[1] The material in this chapter is mainly drawn from [2], which is joint work with Michael Zaletel and Xiangyu Cao.

$\widehat{H} = \sum_j \widehat{H}_j$, where $\widehat{H}_j$ is localized around site j. If the standard MPS compression algorithm via Schmidt decomposition (i.e., singular value decomposition) is directly applied to operators, this structure leads to an ill-conditioned thermodynamic limit, in which some of the Schmidt values become infinite. In 1d, locality gives rise to the following simple property that is the basis for our results. When a 1d system is partitioned into left and right halves, any local operator can be written as:

$$\widehat{H} = \widehat{H}_L \otimes \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \otimes \widehat{H}_R + \sum_a h_{ab} \widehat{h}^a_L \otimes \widehat{h}^b_R \,. \tag{4.1}$$

where $\widehat{h}^a_{L/R}$ run over traceless operators localized on the left/right halves respectively, with coefficients h_{ab}. The first two terms contain the part of the operator supported on strictly one or the other side of the partition, whose magnitude grows linearly with system size, while the third term contains the terms in the operator straddling the partition. This immediately suggests a compression scheme: approximate the intensive part h_{ab} using a singular value decomposition (SVD), whose rank will determine the bond dimension of the MPO, while leaving the extensive terms untouched. Doing so manifestly preserves locality, which will allow us to take the limit of infinite system size, addressing the long-standing problem of efficiently representing operators in the thermodynamic limit [121, 122, 123, 124, 125]. This idea was discussed in Ref. [122]. However, the coefficients h_{ab}, and the resulting singular value spectrum, depend on the choice of operators $\widehat{h}^a_{L/R}$, and *a priori* there is no reason SVD truncation should be optimal. In this work we provide the simple 'fix' which makes the procedure optimal: the compression is performed only after the MPO is brought to a *canonical form* in which $\mathrm{Tr}[\hat{h}^a_{L/R} \hat{h}^b_{L/R}] \propto \delta_{ab}$. The main result of this work is an compression algorithm for both finite and infinite MPOs (iMPOs) which works for physical Hamiltonians with virtually any type of interaction.

Canonical forms play a crucial rôle in MPS compression and many other algorithms, but the naive generalization of the MPS definition to MPOs fails to capture the locality structure of Eq. (4.1) (for this reason, naive SVD truncation of an MPO in the same manner as an MPS generically destroys locality.) We therefore adapt the MPS technology of "canonicalization" and compression algorithms to the class of "first degree" MPOs, which includes short and long ranged Hamiltonians. As a byproduct, we provide a rigorous analysis of the convergence of well-known iterative "canonicalization" algorithms for infinite MPSes. We also present a non-iterative compression algorithm specific to the type of iMPOs that occur in DMRG calculations, which exploits their upper-triangular structure to efficiently handle MPOs with bond dimensions on the order of 100,000. Finally, we detail an intriguing connection to notions from control theory: our compression scheme is a generalization of Kung's method for model-order reduction via balanced truncation[126]. Whenever possible, we provide rigorous proofs of our statements. Our results apply to both finite MPOs and infinite matrix product operators, although we put more emphasis on the infinite case.

This chapter is organized into two parts: the first three sections are a "practical handbook" for compressing finite MPOs, followed by a more sophisticated treatment of infinite

MPOs. The practical handbook starts with an overview of the key ideas of MPO compression in Section 4.2 and Section 4.3 reviews standard facts about MPOs to set notation. We then provide all the concepts and algorithms needed for finite MPO compression in Section 4.4, along with a quick numerical example. After this, we transition to the bulk of the chapter on infinite MPOs. Infinite MPOs require a

We then transition to infinite MPOs, which require a somewhat more detailed and mathematical treatment. Section 4.5 specifies the class of "first degree" MPOs our method applies to, and shows their Jordan block structure is completely fixed by locality. Sections 4.6 is devoted to canonical forms and algorithms to compute them. We give the algorithm for compressing infinite MPOs in Section 4.7. Section 4.8 reveals the peculiar structure of the operator entanglement of local MPOs, which we use to show the error from our compression scheme is ε-close to optimal. We also show that the change in the sup norm is small under compression. Section 4.9 goes on to reinterpret our compression algorithm within control theory. We provide a few examples of iMPO compression in Section 4.10: compressing operators with long-ranged interactions and computing Lanczos coefficients for operator dynamics. We conclude in Section 4.13. The Appendices prove statements from the main text and describe how all elementary algebra operations can be performed on MPOs.

4.2 The Idea of Compression

To introduce the key ideas, we first present them on the level of operators, then later translate them into the language of MPOs. Consider a local operator $\widehat{H}$ on N sites. As mentioned in the introduction, we can split the system into left and right halves at some bond, which gives the **regular form** of an operator

$$\begin{aligned}\widehat{H} &= \widehat{H}_L \otimes \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \otimes \widehat{H}_R + \sum_{a,b=1}^{\chi} \mathsf{M}_{ab} \widehat{h}_L^a \otimes \widehat{h}_R^a \qquad (4.2)\\ &= \begin{pmatrix} \widehat{\mathbb{1}}_L & \widehat{\boldsymbol{h}}_L & \widehat{H}_L \end{pmatrix} \begin{pmatrix} 1 & & \\ & \mathsf{M} & \\ & & 1 \end{pmatrix} \begin{pmatrix} \widehat{H}_R & \widehat{\boldsymbol{h}}_R & \widehat{\mathbb{1}}_R \end{pmatrix}^T ,\end{aligned}$$

where we have introduced vectors of operators $\widehat{\boldsymbol{h}}_{L/R}$ on the left and right, and the matrix M keeps track of the coefficients which straddle the cut. This decomposition is not unique — we can insert basis transformations to the left / right. So, roughly speaking, we will require (4.2) be a Schmidt decomposition by ensuring that M is diagonal and that the components of the vectors are mutually orthogonal. One can then compress $\widehat{H}$ by truncating the Schmidt spectrum — but there is a slight wrinkle due to locality.

To understand the extra structure present in a local operator, let's consider an example. Let

$$\widehat{H}_{\text{e.g}} = \sum_{n=1}^{N} J\widehat{X}_n\widehat{X}_{n+1} + K\widehat{X}_n\widehat{Z}_{n+1}\widehat{X}_{n+2} + h\widehat{Z}_n, \qquad (4.3)$$

where $\widehat{X}_n$ and $\widehat{Z}_n$ are operators acting on lattice site n. $H_{\text{e.g.}}$ is a linear combination of strings, such as $\cdots \otimes \widehat{\mathbb{1}}_1 \otimes \widehat{\mathbb{1}}_2 \otimes X_3 \otimes X_4 \otimes \widehat{\mathbb{1}}_5 \otimes \widehat{\mathbb{1}}_6 \otimes \cdots$. If we split $\widehat{H}_{e.g.}$ across a bond n in the middle, we can write it in regular form (non-uniquely) as

$$\begin{aligned}
\widehat{\boldsymbol{h}}_L &= (\widehat{X}_n, \widehat{X}_n, \widehat{X}_{n-1}\widehat{Z}_n) \\
\widehat{\boldsymbol{h}}_R &= (\widehat{X}_{n+1}, \widehat{Z}_{n+1}\widehat{X}_{n+2}, \widehat{X}_{n+1}) \\
\mathsf{M} &= \operatorname{diag}(J, K, K) \\
\widehat{H}_L &= \sum_{k=1}^{n} J\widehat{X}_{k-1}\widehat{X}_k + K\widehat{X}_{k-2}\widehat{Z}_{k-1}\widehat{X}_k + h\widehat{Z}_k,
\end{aligned} \tag{4.4}$$

and with $\widehat{H}_R$ similar to $\widehat{H}_L$. We see $H_{L/R}$ differs from the $\widehat{h}_{L/R}$ in two respects: first, it's norm diverges linearly with system size (it is *extensive*) and second, it contains terms arbitrarily far from the partition. So in order for the Schmidt compression to be well defined in the thermodynamic limit and preserve locality, it is eminently reasonable to single out $\widehat{H}_{L/R}$ and treat them separately in a Schmidt decomposition.

This motivates the generalization and modification of canonical forms and Schmidt decompositions for the case of local operators.

Definition 8. A local operator in regular form Eq. (4.2), is in **left canonical form** if

$$\langle \widehat{h}^a_L, \widehat{h}^b_L \rangle = \delta^{ab}, \quad 0 \le a, b \le \chi, \tag{4.5}$$

where $\langle \widehat{A}, \widehat{B} \rangle := \operatorname{Tr}[\widehat{A}^\dagger \widehat{B}] / \operatorname{Tr}[\widehat{\mathbb{1}}]$ is the inner-product for operators and $\widehat{h}^0_L := \widehat{\mathbb{1}}_L$. **Right canonical form** is the same with $L \leftrightarrow R$.

Notice that we have excluded $\widehat{H}_{L/R}$ from the definition. If an operator is both left canonical and right canonical on a bond, then we can form the "almost" Schmidt decomposition by an SVD decomposition $\mathsf{M} = \mathsf{USV}^\dagger$.

Definition 9. Suppose $\widehat{H}$ is a local operator and suppose it is both left and right canonical at a bond. Then the **almost-Schmidt decomposition** of $\widehat{H}$ is

$$\widehat{H} = \widehat{H}_L \otimes \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \otimes \widehat{H}_R + \sum_{a=1}^{\chi} s_a \widehat{h}^a_L \otimes \widehat{h}^a_R, \tag{4.6}$$

for some real numbers $s_1 \ge s_2 \ge \cdots \ge s_\chi$.

This is not a true Schmidt decomposition because we have excluded $\widehat{H}_{L/R}$; $\langle h^a_{L/R}, H_{L/R} \rangle$ is generically non-zero. This seeming imperfection will actually prove to be a feature, leading to concise algorithms and an truncation error ε-close to optimal with respect to *both* the Frobenius and operator (induced) norms (see Sec. 4.8.) Once we know the almost-Schmidt decomposition of an operator, compressing it to a bond dimension $\chi' < \chi$ is easy: simply restrict the sum in Eq. (4.6) to run from 1 to χ' instead of χ. Our task is now to translate this idea from the level of operators to concrete computations and algorithms in the language of MPOs.

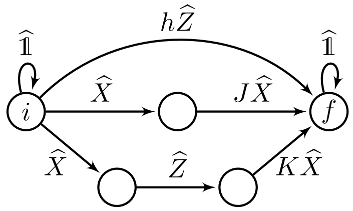


Figure 4.1: A finite-state machine that generates Eq. 4.3.

4.3 Review of MPOs

Matrix product operators (MPOs) arise in DMRG as a pithy representation of 1d Hamiltonians. This section will review a few essential facts about finite and infinite MPOs for the reader's convenience and to set notation. The well-known construction of MPOs comes from viewing a Hamiltonian as a finite-state machine [127, 115], which we illustrate with an example.

Consider $\widehat{H}_{e.g.}$ from Eq. (4.3) again. All of the Pauli strings needed to generate $\widehat{H}_{e.g.}$ can be described by a finite state machine, shown in Fig. 4.1. (We will see below this machine can be improved.) The MPO itself is the adjacency matrix of the finite state machine:

$$\widehat{W}_{\text{e.g}} = \left(\begin{array}{c|ccc|c} \widehat{\mathbb{1}} & \widehat{X} & \widehat{X} & 0 & h\widehat{Z} \\ \hline & 0 & 0 & 0 & J\widehat{X} \\ & 0 & 0 & \widehat{Z} & 0 \\ & 0 & 0 & 0 & K\widehat{X} \\ \hline & & & & \widehat{\mathbb{1}} \end{array}\right), \tag{4.7}$$

where the hat on the matrix $\widehat{W}_{\text{e.g}}$ indicates that its components are operator-valued. The Hamiltonian on the open chain $[1, N]$ then has the compact representation

$$\widehat{H}_{\text{e.g.}} = \boldsymbol{\ell} \underbrace{\widehat{W}_{\text{e.g.}} \widehat{W}_{\text{e.g.}} \cdots \widehat{W}_{\text{e.g.}}}_{N \text{ matrices}} \boldsymbol{r}, \tag{4.8}$$

where $\boldsymbol{\ell} := (1 \;\; \mathbf{0}_3 \;\; 0)$ and $\boldsymbol{r}^\dagger := (0 \;\; \mathbf{0}_3 \;\; 1)$ are c-number vectors, also called "boundary conditions". They encode the instructions "start at node i" and "end at node f". The multiplication of MPOs in (4.8) is a matrix product in the *auxiliary space* and a tensor product in the *physical space*, such that physical indices of the nth matrix in (4.8) acts on lattice site n.

The example above is a so-called **infinite MPO (iMPO)**: the whole operator only depends on one matrix $\widehat{W}$, regardless of the system size. A regular **MPO** is made of inhomogenous matrices

$$\widehat{H} = \boldsymbol{\ell} \widehat{W}^{(1)} \widehat{W}^{(2)} \cdots \widehat{W}^{(N)} \boldsymbol{r}\,. \tag{4.9}$$

where $\widehat{W}^{(1)}, \ldots, \widehat{W}^{(N)}$ are distinct matrices and need not be square, with $\widehat{W}^{(n)}$ of size $\chi^{(n-1)} \times \chi^{(n)}$ so that matrix multiplication makes sense.

In a local Hamiltonian, each term begins and ends with strings of identities, which gives rise to the first two terms in the regular form of an operator, Eq. (4.2) above. This property is encoded by the distingished nodes i and f in the finite state machine Fig. 4.1, and is reflected by the block structure of the MPO (4.7). We therefore restrict ourselves to a special class of (i)MPOs which manifestly maintain this local structure.

Definition 10. An (i)MPO is in **regular form** if each matrix has the block upper triangular structure

$$\widehat{W} = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}} & \widehat{d} \\ 0 & \widehat{A} & \widehat{\boldsymbol{b}} \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}, \tag{4.10}$$

where the first and last blocks have dimension 1 for both rows and columns.[2] Furthermore, we require that the boundary conditions are of the form

$$\boldsymbol{\ell} = \begin{pmatrix} 1 & * & * \end{pmatrix}, \; \boldsymbol{r}^\dagger = \begin{pmatrix} * & * & 1 \end{pmatrix}, \tag{4.11}$$

where $*$ denotes an arbitrary block.

The shape of $\widehat{W}$ in (4.1) is thus entirely determined by the shape of $\widehat{A}$. For iMPOs, $\widehat{A}$ is a square matrix of size $\chi \times \chi$ where χ is called the *bond dimension.* (Some authors instead define the bond dimension as the size of $\widehat{W}$, $\chi+2$.) Operators in regular form are represented by (i)MPOs in regular form, and all (i)MPOs in this work will be in regular form.

The usual diagram notation for tensor networks cannot capture the block structure of (4.10), so we simply work with equations, making them index-free whenever possible. In the rare exceptions, the auxiliary space is indexed by Latin letters starting from zero to highlight the block structure: $a, b, c \dots = 0; 1, 2, \dots \chi; \chi + 1$.

The class of (i)MPOs in regular form is closed under addition, scalar multiplication, and operator multiplication. These constructions are computationally straightforward and more-or-less well-known. They are collected in Appendix 4.D for the reader's convenience.

Physical operators admit many distinct MPO representations; MPOs have a large *gauge freedom.* An operator $\widehat{H} = \boldsymbol{\ell}\widehat{W}^{(1)} \cdots \widehat{W}^{(N)}\boldsymbol{r}$ can also be represented by $\widehat{H} = \boldsymbol{\ell}'\widehat{W}^{(1)'} \cdots \widehat{W}^{(N)'}\boldsymbol{r}'$ whenever there are matrices $L^{(0)}, \ldots, L^{(N)}$ that satisfy the interlacing conditions

$$\widehat{W}^{(n)'}L^{(n)} = L^{(n-1)}\widehat{W}^{(n)}, \boldsymbol{\ell}'L^{(0)} = \boldsymbol{\ell}, \boldsymbol{r}' = L^{(N)}\boldsymbol{r}. \tag{4.12}$$

In the infinite case, all the $L^{(n)}$'s are equal to some L, so the gauge transformation resembles a similarity transform:

$$\widehat{W}'L = L\widehat{W}. \tag{4.13}$$

[2] Structurally, $\widehat{d}$ is a single operator, and $\widehat{\boldsymbol{c}}$ and $\widehat{\boldsymbol{b}}$ are operator-valued vectors.

To preserve the regular form (4.10), all gauge matrices must be block triangular,

$$L = \left(\begin{array}{c|c|c} 1 & \boldsymbol{t} & r \\ \hline 0 & \mathsf{L} & \boldsymbol{s} \\ \hline 0 & 0 & 1 \end{array}\right) . \tag{4.14}$$

Note that L need not be square, but only shaped to be compatible with (4.12) or (4.13)[3]. In particular, $\widehat{W}'$ and $\widehat{W}$ may have different bond dimensions.

For instance, we can gauge transform $\widehat{W}_{\text{e.g.}}$ to

$$\widehat{W}'_{\text{e.g}} = \left(\begin{array}{c|cc|c} \widehat{\mathbb{1}} & \widehat{X} & 0 & h\widehat{Z} \\ \hline & 0 & Z & J\widehat{X} \\ & 0 & 0 & K\widehat{X} \\ \hline & & & \widehat{\mathbb{1}} \end{array}\right) \tag{4.15}$$

which encodes $\widehat{H}_{\text{e.g.}}$ more simply than $\widehat{W}_{\text{e.g}}$. This previews our end goal: given an MPO (and an error tolerance), how do we compute the smallest MPO that encodes the same operator?

4.4 Finite MPO Compression

Now that we have reviewed MPOs, we give a "practical handbook" for compressing finite matrix product operators. We proceed expeditiously: first upgrading canonical forms and "sweeps" to MPOs, then giving the compression algorithm, and lastly a brief numerical example. Readers familiar with matrix product states will find that our compression method amount to a small — yet conceptually significant — modification of standard MPS algorithms. As the subsequent treatment of iMPOs will revisit all the concepts here in greater detail, many technical details are postponed for later sections.

MPO Canonical Forms

Just as with matrix product states, the main tool for manipulating matrix product operators is the idea of *canonical forms*. They are choices of gauge that make the rows or columns of the matrix $\widehat{W}$ orthogonal, an essential step for controlling the errors from compression or carrying out the DMRG algorithm.

We define canonical forms in terms of a condition on the matrix itself, then show that canonical MPOs represent canonical operators.

Definition 11. An MPO $\widehat{H} = \boldsymbol{\ell}\widehat{W}^{(1)} \cdots \widehat{W}^{(N)}\boldsymbol{r}$ is in **left canonical form** if, for each $n > 1$, the upper left block of $\widehat{W}^{(n)}$,

$$\widehat{V}^{(n)} := \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}}^{(n)} \\ 0 & \widehat{A}^{(n)} \end{pmatrix} , \tag{4.16}$$

[3] Some authors define a less general class of invertible gauge transformation $\widehat{W}' = L\widehat{W}L^{-1}$, which precludes L from changing the bond dimension.

has orthonormal columns:

$$\forall b, c \leq \chi^{(n)}, \sum_{a=0}^{\chi} \langle \widehat{W}^{(n)}_{ab}, \widehat{W}^{(n)}_{ac} \rangle = \delta_{bc}. \tag{4.17}$$

For $n = 1$ we instead require $\langle [\boldsymbol{\ell}\widehat{W}^{(1)}]_b, [\boldsymbol{\ell}\widehat{W}^{(1)}]_c \rangle = \delta_{bc}$ for all $b, c \leq \chi^{(1)}$.

An MPO is in **right canonical form** if, and only if, its *mirror*[4] is in left canonical form. Right canonical forms are always directly analagous, so we focus on the left-handed case.

Let us now see why left canonical MPOs describe left canonical operators, in the sense of Defn. 8.[5] If we split an MPO in left canonical form at a bond n, then we can multiply the matrices together to put the operator into regular form:

$$\begin{aligned} \widehat{H}_W &= \left(\boldsymbol{\ell}\widehat{W}^{(1)} \cdots \widehat{W}^{(n)}\right) \left(\widehat{W}^{(n+1)} \cdots \widehat{W}^{(N)} \boldsymbol{r}\right) \\ &= \begin{pmatrix} \widehat{\mathbb{1}}^{(n)}_L & \widehat{\boldsymbol{h}}^{(n)}_L & \widehat{H}^{(n)}_L \end{pmatrix} \begin{pmatrix} \widehat{H}_R & \widehat{\boldsymbol{h}}_R & \widehat{\mathbb{1}}_R \end{pmatrix}^T . \end{aligned} \tag{4.18}$$

Standard form for MPOs implies that the vectors of operators are related by the recursion relation

$$\begin{pmatrix} \widehat{\mathbb{1}}^{(n-1)}_L & \widehat{\boldsymbol{h}}^{(n-1)}_L \end{pmatrix} \widehat{V}^{(n)} = \begin{pmatrix} \widehat{\mathbb{1}}^{(n)}_L & \widehat{\boldsymbol{h}}^{(n)}_L . \end{pmatrix} \tag{4.19}$$

If the MPO's are in regular form, then $\widehat{V}^{(1)}, \ldots \widehat{V}^{(n)}$ have orthonormal columns, so by induction,

$$\langle \widehat{h}_{L,a}, \widehat{h}_{L,b} \rangle = \delta_{ab},\, 0 \leq a, b \leq \chi^{(n)}\,, \tag{4.20}$$

where $\widehat{h}_{L,0} := \widehat{\mathbb{1}}_L$. Left canonical form for MPOs therefore ensures that all components but the last of the vector $(\widehat{\mathbb{1}}_L, \widehat{\boldsymbol{h}}_L, \widehat{H}_L)$ are orthonormal — and imposes no constraint whatsoever on $\widehat{H}_L$. So MPO canonical form implies operator canonical form, Defn. 8.

Now that we have defined canonical forms for MPOs, our next task is compute them. One can always find a gauge transform, Eq. (4.12), to bring a finite MPO to left canonical form and, just as in the MPS situation, we can compute the change of gauge via a QR decomposition. Suppose $\widehat{W}$ is an MPO in regular form of dimensions $(1+\chi+1)$ by $(1+\chi'+1)$ with $\widehat{V}$ given by (4.10). If we group indices as $V_{(\alpha a)b}$, where $0 \leq \alpha < d^2$ indexes the standard orthonormal basis of $\mathcal{A}$, then $\widehat{V}$ can be interpreted as a matrix with shape $d^2(1+\chi) \times (1+\chi')$. Performing a (thin) QR decomposition gives

$$\widehat{V} = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}} \\ 0 & \widehat{A} \end{pmatrix} \overset{QR}{=} \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}}' \\ 0 & \widehat{A}' \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{t} \\ 0 & \mathsf{R} \end{pmatrix}, \tag{4.21}$$

where R is upper-triangular.

[4] A MPO is mirrored by (I) transposing each matrix $\widehat{W}^{(n)}$, (II) exchanging $\boldsymbol{\ell}^\dagger \leftrightarrow \boldsymbol{r}$, (III) reversing all auxiliary indices $(0 \leftrightarrow \chi + 1,\, 1, \ldots, \chi \leftrightarrow \chi, \ldots, 1)$, and (IV) reversing the physical positions.

[5] Actually the two definitions are entirely equivalent, but we show only one implication for concision.

Algorithm 1 Left Canonical Form for finite MPOs

```
procedure MPOLeftCan($\{\boldsymbol{\ell}, \{\widehat{W}^{(n)}\}_{n=1}^{N}, \boldsymbol{r}\}$)
    $R^{(0)} \leftarrow \boldsymbol{\ell}$
    for $n \in [1, N]$ do
        $(\widehat{Q}^{(n)}, R^{(n)}) \leftarrow \widehat{QR}[R^{(n-1)}\widehat{W}^{(n)}]$        ▷ Eq. (4.21)
    return $\{\boldsymbol{\ell}, \{\widehat{Q}^{(n)}\}_{n=1}^{N}, R^{(N)}\boldsymbol{r}\}, \{R^{(n)}\}$
```

Definition 12. Define the **block-respecting $\widehat{QR}$ decomposition** of $\widehat{W}$ as

$$\widehat{QR}[\widehat{W}] = \widehat{Q}R \tag{4.22}$$

with

$$\widehat{Q} := \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}}' & \widehat{d} \\ 0 & \widehat{A}' & \widehat{\boldsymbol{b}} \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}, \; R := \begin{pmatrix} 1 & \boldsymbol{t} & 0 \\ 0 & \mathsf{R} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.23}$$

where the upper-left block comes from (4.21). Therefore, $\widehat{Q}$ is in left canonical form, and R is upper-triangular.

With this, we can define a sweeping procedure to put a finite MPO into left canonical form.

$$
\begin{aligned}
&\boldsymbol{\ell}\widehat{W}^{(1)}\widehat{W}^{(2)}\widehat{W}^{(3)}\cdots && (4.24)\\
\overset{QR}{=}\;& \boldsymbol{\ell}\left[\widehat{Q}^{(1)}R^{(1)}\right]\widehat{W}^{(2)}\widehat{W}^{(3)}\cdots && (4.25)\\
=\;& \boldsymbol{\ell}\widehat{Q}^{(1)}\left[R^{(1)}\widehat{W}^{(2)}\right]\widehat{W}^{(3)}\cdots && (4.26)\\
\overset{QR}{=}\;& \boldsymbol{\ell}\widehat{Q}^{(1)}\left[\widehat{Q}^{(2)}R^{(2)}\right]\widehat{W}^{(3)}\cdots && (4.27)\\
=\;& \boldsymbol{\ell}\widehat{Q}^{(1)}\widehat{Q}^{(2)}\left[R^{(2)}\widehat{W}^{(3)}\right]\cdots && (4.28)\\
& && (4.29)
\end{aligned}
$$

By the definitition of the block QR decomposition, the first $1 + \chi^{(n)}$ columns of each $\widehat{Q}^{(n)}$ are indeed orthonormal. Moreover, $\{R^{(1)}, \ldots, R^{(N)}\}$ specifies a gauge transform from $\{\boldsymbol{\ell}, W^{(n)}, \boldsymbol{r}\}$ to $\{\boldsymbol{\ell}, Q^{(n)}, R^{(N)}\boldsymbol{r}\}$. We summarize the procedure as Algorithm 1.

Note that Algorithm 1 is almost identical to a standard "right-sweep" that brings an MPS to its left-canonical form, except that the block-respecting $\widehat{QR}$ decomposition is used *in lieu* of normal QR.

Finite MPO Compression

We can now give the compression procedure for finite MPOs. Suppose we have a finite MPO on sites $[1, N]$. We first bring the whole chain to right canonical form

$$\widehat{H}_W = \boldsymbol{\ell}\,\widehat{W}_R^{(1)}\,\widehat{W}_R^{(2)}\ldots\widehat{W}_R^{(N)}\,\boldsymbol{r}\,,$$

by the mirror of Algorithm 1. To truncate at bond $(n, n+1)$, we first bring sites $[1, n]$ to left canonical form

$$\begin{aligned}&\boldsymbol{\ell}\,\widehat{W}_R\,\widehat{W}_R\,\cdots\,\widehat{W}_R\,\widehat{W}_R\,\cdots\widehat{W}_R\,\boldsymbol{r}\\ =&\boldsymbol{\ell}\,\widehat{W}_L\,R\,\widehat{W}_R\,\cdots\,\widehat{W}_R\,\widehat{W}_R\cdots\widehat{W}_R\,\boldsymbol{r}\\ &\vdots\\ =&\;\boldsymbol{\ell}\,\underbrace{\widehat{W}_L\,\widehat{W}_L\,\cdots\,\widehat{W}_L}_{\text{sites }[1,n]}\,R^{(n)}\,\underbrace{\widehat{W}_R\,\cdots\,\widehat{W}_R}_{\text{sites }[n+1,N]}\,\boldsymbol{r}.\end{aligned}$$

(Superscripts have been suppressed for clarity.) The block structure of $R^{(n)}$ is fixed by block QR decomposition, Eq. (4.23), and we can always decompose it as[6]

$$R^{(n)} = MR'\,,\; M = \begin{pmatrix}1 & 0 & 0\\ 0 & \mathsf{M} & 0\\ 0 & 0 & 1\end{pmatrix}\; R' = \begin{pmatrix}1 & \boldsymbol{t} & 0\\ 0 & \mathrm{Id}_\chi & 0\\ 0 & 0 & 1\end{pmatrix}. \tag{4.30}$$

We then perform an singular value decomposition of M and write

$$M = USV^\dagger\,,\; \mathsf{S} = \mathrm{diag}(s_1 \geq s_2 \geq \cdots \geq s_\chi)\,, \tag{4.31}$$

where the middle blocks are unitary: $\mathsf{U}^\dagger\mathsf{U} = \mathsf{V}^\dagger\mathsf{V} = \mathrm{Id}_\chi$. Combining (4.30) through (4.31), we obtain

$$\widehat{H}_W = \cdots\widehat{W}_L^{(n-1)}\widehat{Q}^{(n)}S\widehat{P}^{(n+1)}\widehat{W}_R^{(n+1)}\cdots \tag{4.32}$$

where

$$\widehat{Q}^{(n)} := \widehat{W}_L^{(n)}U, \quad \widehat{P}^{(n+1)} := V^\dagger R'\widehat{W}_R^{(n+1)} \tag{4.33}$$

are still left and right canonical, respectively.[7] Therefore Eq. (4.32) is left canonical on the left, right canonical on the right, and the central matrix S is diagonal — so it is an almost-Schmidt decomposition, Eq. (4.6), as desired.

We can now reduce the bond dimension by dropping the smallest singular values, as well as the corresponding columns of $\widehat{Q}$ and rows of $\widehat{P}$. The compression scheme is summarized in Algorithm 2. The truncation is combined with a left-sweep, so the returned MPO is left canonical.

[6] Here and below, we use the short hand $\mathrm{diag}(1, \mathsf{A}, 1) = A$ for block diagonal matrices, with sans-serif letters for the middle block.

[7] Right-canonical form is preserved because R' only affects the top row while leaving the bottom $\chi + 1$ rows orthonormal, as required for right-canonical form.

Algorithm 2 MPO Compression

```
procedure COMPRESS($\{\boldsymbol{\ell}, \widehat{W}^{(n)}, \boldsymbol{r}\}, \eta$)                    ▷ Cutoff $\eta$
    $\boldsymbol{\ell}, \{\widehat{W}_R^{(n)}\}, \boldsymbol{r} \leftarrow$ RIGHTCAN$[\boldsymbol{\ell}, \{\widehat{W}^{(n)}\}, \boldsymbol{r}]$
    $R \leftarrow \boldsymbol{\ell}$
    for $n = 1, \dots, N-1$ do
        $(\widehat{W}_L^{(n)}, R) \leftarrow \widehat{QR}[R\widehat{W}_R^{(n)}]$                    ▷ Eq. (4.23)
        $(M, R') \leftarrow R$                                        ▷ Eq. (4.30)
        $(U, S, V^\dagger) \leftarrow \mathrm{SVD}[M]$
        $\chi' \leftarrow \max\{a : s_a > \eta\}$; $I \leftarrow \{0, 1, \dots, \chi', \chi+1\}$.
        $\widehat{Q}^{(n)} \leftarrow [\widehat{W}_L^{(n)} U]_{0:\chi+1,I}$
        $R \leftarrow [V^\dagger R']_{I,0:\chi+1}$
    $(\widehat{Q}^{(N)}, R) \leftarrow \widehat{QR}[R\widehat{W}_R^{(N)}]$
    return $\boldsymbol{\ell}, \{\widehat{Q}^{(n)}\}, R\boldsymbol{r}$
```

Due the presence of "sweeps" in the algorithms, it is not immediately clear how to generalize them to the infinite case, nor is the precise relation to truncations by "true" Schmidt decompositions clear. We will address these points in Sections 4.7 and 4.8 below. We note that our compression scheme is ε-close to optimal, in a sense we make clear below.

An Example

To demonstrate the utility of our compression scheme, we give a brief numerical example. Specifically, we compress a Hamiltonian with long-ranged interactions and show our method is quite comperable to the standard "MPS" compression technique, i.e. treating the operator like an MPS in a doubled Hilbert space. We note, however, that our "MPO" compression technique outscales the naive "MPS" technique because it contains only intensive values in the entanglement spectrum.

It is well known that a two body interaction $V\left(i-j\right)\widehat{\mathcal{O}}_i\widehat{\mathcal{O}}_j$, where $V(r) = \sum_{j=1}^{\chi} a_j \lambda_j^r$ is a sum of χ exponentials has an exact MPO representation with bond dimension χ.[8] Our algorithm will *automatically* discover this structure even if the MPO is initially presented in a non-optimal form.

We therefore select a more challenging example with power-law interactions:

$$H_1 = \sum_{k,n,m=1}^{N} J_{kn} J_{nm} \widehat{Z}_k \widehat{Z}_n \widehat{Z}_m + J'_{nm} \widehat{Z}_n \widehat{Z}_m \tag{4.34}$$

[8] See Eq. 4.42 for an example.

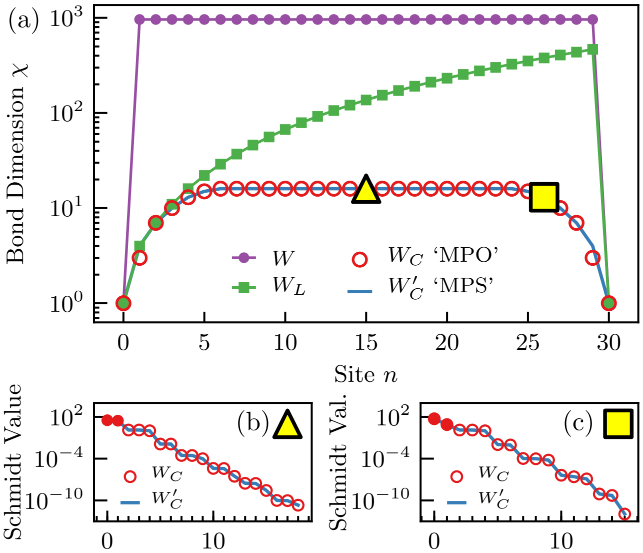


Figure 4.2: Compression of a finite MPO representing the Hamiltonian (4.34). (a) The bond dimensions for: $\widehat{W}$, the naive MPO representation of H_1; $\widehat{W_L}$, the left-canonical representation by Alg. 1; $\widehat{W_C}$, the compressed MPO by Alg. 2, and $\widehat{W'_C}$, the result of the standard MPS compression. (b,c) The Schmidt spectra of $\widehat{W'_C}$ and almost-Schmidt spectra of $\widehat{W_C}$ at the sites denoted by the triangle and square, respectively. The numerical precision was taken to be $\varepsilon_{\text{can}} = 10^{-12}$ for canonicalization and $\varepsilon_C = 10^{-4}$ for compression.

where $J_{nm} = |n-m|^{-2}$ and $J'_{nm} = |n-m|^{-4}$. In (4.34) and below, we include a three-body term to test our algorithms beyond the domain of two-body Hamiltonians, which was addressed in previous work [125]. The results are shown in Fig. 4.2.

The compression in Fig. 4.2 follows Algorithm 2, and takes place in two stages. First, a right-sweep with block-QR decomposition (Algorithm 1) performs a preliminary bond reduction: it only reduces bond dimensions if columns are linearly dependent. Then a left-sweep of almost Schmidt value truncation results in a more significant compression. We compare the resulting bond dimensions with those obtained from a standard MPS compression (which does not preserve the block structure) and find them essentially identical. In fact, the whole entanglement spectrum from the almost-Schmidt decomposition closely matches the one from the true Schmidt decomposition. The only difference is the first two Schmidt values are extensive and not present in the almost-Schmidt spectrum.[9]. We return to this point in Section 4.8 below, when we discuss operator entanglement.

This concludes our discussion of compressing finite MPOs. We now move on to infinite matrix product operators.

[9] Such an precise match of the spectra holds only for simple Hamiltonians; in general, however, we have the interlacing relations (4.87).

4.5 Local Infinite Matrix Product Operators

We now transition to infinite matrix product operators. The discussion proceeds analogously to the finite case above. However, working with infinite operators requires additional care, and our discussion will become corresponding more precise and detailed. Indeed, before we can define and compute canonical forms, we must examine exactly what it means for an infinite MPO to be local. We will precisely define and characterize a good class of operators — operators of "first degree" — which (1) includes local physical Hamiltonians and (2) are described by "local" iMPOs.

Locality is a non-trivial requirement for a physical operator. It is accompanied by a host of properties, such as an extensive norm, and that spatially-separated terms should commute. For Hamiltonians, perhaps the most important consequence of locality, however, is the existence of thermodynamic limits: the ground state energy and other thermodynamic observables grow as first order polynomials in the size of the system, i.e. extensively. We would like to be able to work with and compress all such local Hamiltonians. As characterizing the class of iMPOs with extensive ground states is quite difficult, we will instead work with a class of operators characterized by an extensive norm, which includes virtually all local physical Hamiltonians. As an analogy, just as local Hamiltonians of interest contribute a constant amount of energy per site, we work with operators that are described be a constant amount of "information per site". We will often call such operators "local as iMPOs" or simply "local".

Norm and Transfer Matrices

The norm of an operator is a starkly different object than that of a state. States, of course, are normalized, so the norm of a generic iMPS should be 1 in the limit $N \to \infty$. This is rooted in the iMPS transfer matrix, where a standard result [116] shows that the largest eigenvalue is non-degenerate with eigenvalue $\lambda = 1$, after normalization. In contrast, the space of operators admits many different norms, and this choice must often be resolved by physical considerations. When one is interested in ground state energies and static expectation values, the sup norm is usually the correct choice. However, for questions of quantum dynamics in the common setting of infinite temperature, the Frobenius (aka Hilbert-Schmidt) norm is the natural one, which is relatively easy to compute.

In this work, our "default" norm will be a Frobenius norm per unit length. For a translation-invariant operator $\widehat{H}$, call its restriction to N sites $\widehat{H}_N$ and define

$$||\widehat{H}||_F^2 := \lim_{N\to\infty} \langle \widehat{H}_N, \widehat{H}_N \rangle = \lim_{N\to\infty} \frac{\mathrm{Tr}[\widehat{H}_N^\dagger \widehat{H}_N]}{\mathrm{Tr}[\widehat{\mathbb{1}}^N]}. \tag{4.35}$$

where the subscript "F" is a reminder that this is essentially the Frobenius norm.[10] The norm is normalized so that $||I||_F = 1$, unlike the usual Frobenius norm where the norm

[10] We note that this norm is not submultiplicative: see Appendix D.

of the identity is the dimension of the space. We will be interested in iMPOs where this norm is extensive. Despite this choice of norm, we prove in Section 4.8 that our compression algorithm behaves well with respect to the sup norm as well — so our choice of norm is suitable for both dynamics and statics applications. We will therefore refer to (4.35) as *the* norm of an operator in this work.

To compute the norm of an operator expressed as an iMPO, we must recall the definition of the **transfer matrix**. The space of single site operators forms an algebra $\mathcal{A}$ with an inner product $\langle \cdot, \cdot \rangle$ such that $\langle \widehat{\mathbb{1}}, \widehat{\mathbb{1}} \rangle = 1$. We fix an orthonormal basis $\mathcal{A} = \operatorname{span}\{\widehat{O}_\alpha \ : \ 0 \leq \alpha < d\}$ (indexed by Greek letters $\alpha, \beta, \dots$) starting with $\widehat{O}_0 = \widehat{\mathbb{1}}$. For example, one might take the algebra of spin-$\frac{1}{2}$ operators with the basis of Pauli operators $\{\widehat{\mathbb{1}}, \widehat{X}, \widehat{Y}, \widehat{Z}\}$. Then the real algebra over this basis gives Hermitian operators and the complex algebra gives all operators. For Fermions, $\mathrm{Tr}[\widehat{c}^\dagger \widehat{c}] = \mathrm{Tr}[\widehat{n}^\dagger \widehat{n}] = 1$, so one orthonormal basis is $\{\widehat{\mathbb{1}}, \sqrt{2}\widehat{c}^\dagger, \sqrt{2}\widehat{c}, \widehat{Z} = \widehat{\mathbb{1}} - 2\widehat{n}\}$ with complex coefficients. In such a single site basis, any operator-valued matrix $\widehat{W}$ becomes equivalent to an vector of c-number matrices $\{W_\alpha\}$ defined via

$$\widehat{W} = \sum_\alpha \widehat{O}_\alpha W_\alpha \,, \quad (W_\alpha)_{ab} := \langle \widehat{O}_\alpha, \widehat{W}_{ab} \rangle. \tag{4.36}$$

Definition 13. Suppose $\widehat{W}$ is an operator-valued square matrix that acts on the auxiliary vector space $\mathcal{V}$ of dimension χ. Then the $\widehat{W}$**-transfer matrix** is a linear operator on $\mathcal{V} \otimes \mathcal{V}$, defined as

$$T_W := \sum_\alpha \overline{W}_\alpha \otimes W_\alpha, \tag{4.37}$$

where the bar denotes complex conjugation.

It is sometimes convenient to identify $\mathcal{V} \otimes \mathcal{V}$ with the space of square matrices. Then T_W acts on matrices $X \in \mathcal{V} \otimes \mathcal{V}$ on the left by

$$X T_W = \sum_\alpha W_\alpha^\dagger X W_\alpha, \tag{4.38}$$

where $W_\alpha^\dagger$ is the Hermitian conjugate as usual. By Choi's Theorem [128], transfer matrix are always postive operators: whenever X is positive semi-definite, so is $X T_W$.

The transfer matrix gives a simple formula for the norm of an operator in terms of its MPO representation. On a lattice of N sites, the norm squared is

$$||\widehat{H}_N||_F^2 = (\boldsymbol{\ell\ell}) \left(T_W\right)^N (\boldsymbol{rr}) \,. \tag{4.39}$$

where $\boldsymbol{\ell\ell} := \overline{\boldsymbol{\ell}} \otimes \boldsymbol{\ell}$ and $\boldsymbol{rr} := \overline{\boldsymbol{r}} \otimes \boldsymbol{r}$.

The only way that (4.39) can give rise to an extensive norm, (4.35), is if the iMPO transfer matrix T_W (4.37) is dominated by some nontrivial Jordan block with eigenvalue 1.

To build intuition, we first consider the simple example

$$\widehat{H} = \sum_i \widehat{d}_i \text{ with } \widehat{W} = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{d} \\ 0 & \widehat{\mathbb{1}} \end{pmatrix} \,, \tag{4.40}$$

such that $\langle \widehat{\mathbb{1}}, \widehat{d} \rangle = 0$ and $\langle \widehat{d}, \widehat{d} \rangle = \rho$. Of course, $||H_N||_F^2 = N\rho$. Then the transfer matrix T_W is a 4×4 matrix

$$T_W = \begin{pmatrix} 1 & 0 & 0 & \rho \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \sim \left(\begin{array}{cc|c|c} 1 & \rho & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right), \tag{4.41}$$

where $\sim$ denotes a similarity transform (but not a gauge transform). Taking powers T_W^N, (4.39) shows that the Jordan block is clearly responsible for the extensive norm [11]. This behavior should be generic; all local operators should have an extensive norm. However, not all iMPOs in regular form satisfy (4.35) because, even though such a Jordan block always exists, it may not dominate the norm (4.39) as $N \to \infty$. The remedy is to precisely define the what it means for iMPOs to be "local as an iMPO".

First Degree Operators

This section will carefully define the class of first degree operators. Before giving the mathematical definition, let us provide some motiviation.

A natural class of iMPOS which are local by any reasonable criterion are those whose finite state machines do not involve any loops, such as Fig. 4.1. Such iMPOs represent operators where each term has identities on all sites except on a contiguous block of at most χ sites. This structure implies that the ground state must be extensive. These operators can be readily characterized as follows:

Definition 14. An iMPO $\widehat{W}$ is **strictly local** if its $\widehat{A}$ block is strictly upper-triangular.

However, this definition has important drawbacks: the property of $\widehat{A}$ being strictly upper-triangular is neither gauge invariant, nor robust under small perturbations — which inevitably arise as numerical errors from compression. This definition is therefore an inadequate starting point to define a good class of local operators.

As mentioned earlier, the cure is actually to consider a *larger* class of operators. We will define this class first in terms of the transfer matrix and we will show by the end of the section that these are the operators with extensive norms (4.35). Specifically, we make a condition on the spectral properties of the $\widehat{A}$ block of their iMPO representation.

Definition 15. Suppose $\widehat{W}$ is an iMPO in regular form (4.10), and T_A is the transfer matrix corresponding to its $\widehat{A}$ block. $\widehat{W}$ is called **first degree** if $|\lambda| < 1$ for all eigenvalues λ of T_A.[12]

[11] The other two blocks do not contribute to the extensive norm, but can become relevant when the operator has an extensive trace; see Appendix 4.A for details.

[12] We note our definition is closely akin to the idea of an "interaction" in the mathematical physics literature. See e.g. Chapter 6 of [129].

Property	**SL**	**FD**	**Gen**
Leading eig.val. of T_A	$\lambda = 0$	$\lambda < 1$	$\lambda < \infty$
Norm $\|\|H_N\|\|_F^2$	$\sim N$	$\sim N$	$\sim \lambda^N$
Open Set	✗	✓	✓
Closed under commutation.	✓	✗	✓
Canonical form (see Sec. 4.6)	✓	✓	✗

Table 4.1: Properties of different set of iMPOs: strictly local (SL), first degree (FD), and the set of general (Gen) iMPOs without restriction.

The name "first degree" anticipates Prop. 17, which states that first degree operators have extensive norm $||\widehat{O}_N||_F^2 = O(N)$. Physically, this definition amounts to the requirement that there is a decomposition (4.1) where the operators $h^a_{L/R}$ fall off with exponentially-localized tails.

By Definition 15, the set of first degree iMPOs is a topologically open set, and is therefore numerically robust, but also a superset of strictly local iMPOs. Indeed, strict locality implies that the T_A matrix is also strictly upper-triangular and thus nilpotent (all $\lambda = 0$). To give an example of an first degree iMPO which is not strictly local, consider

$$\widehat{H}_{\text{FD}} = \sum_i \sum_{k=0}^{\infty} \widehat{X}_i \left[\prod_{j=i+1}^{i+k} \alpha \widehat{Z}_j \right] \widehat{Y}_{i+k+1} \,. \tag{4.42}$$

whose iMPO representation is

$$\widehat{W}_{\text{FD}} = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{X} & 0 \\ 0 & \alpha\widehat{Z} & \widehat{Y} \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix} . \tag{4.43}$$

The only eigenvalue of T_A is $|\alpha|^2$, so

$$||\widehat{H}_{FD,N}||_F^2 \sim \begin{cases} N & |\alpha| < 1 \\ N^2 & |\alpha| = 1 \\ |\alpha|^{2N} & |\alpha| > 1 \end{cases} \tag{4.44}$$

so $\widehat{W}_{\text{FD}}$ is first degree if and only if $|\alpha| < 1$. In this sense, the definition of first degree operators is tight. (Note that $\widehat{H}_{\text{FB}}$ only has extensive ground state energy for $\alpha < 1$.) The spectral nature of the definition also makes the class of first degree iMPOs invariant under iMPO gauge transforms (see Lemma 25 in App. 4.A).

We caution that the class of first degree Hamiltonians is quite vast. It includes all operators that are usually classified as "local Hamiltonians". For instance, it include all

k-local Hamiltonians, but also Hamiltonians with long ranged interactions with exponential falloff. However, there are many operators which are first degree — such as projectors — which do *not* make sense as Hamiltonians. See Appendix 4.D for an example.

A slight drawback of first degree operators is that — unlike strictly local operators — they are not closed under commutation (the commutator of two first degree operators can be "second degree"). Nevertheless, one can show (see Appendix 4.D) that if $\widehat{W}$ is first degree and $\widehat{W}'$ strictly local, the commutator $[\widehat{W}, \widehat{W}']$ is still first degree. This is sufficient for our applications, including operator dynamics (see Section 4.10 below).

The dominant Jordan block of T_W

We now show that the transfer matrix of first degree iMPOs have the dominant Jordan block structure required for an extensive norm (4.35). From the finite state machine picture, we know that the iMPO always maps the initial state to the initial state, and the final state to the final state. Intuitively, the dominant Jordan block encodes the fact that these are the "most important processes" in the state machine, rather than running around loops in intermediate states.

We begin with an intermediate result which will be crucial to establish canonical forms in Section 4.6 below.

Proposition 16. *Suppose that $\widehat{W}$ is a first degree iMPO and consider its upper-left block*

$$\widehat{V} := \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}} \\ 0 & \widehat{A} \end{pmatrix} \tag{4.45}$$

Then the transfer matrix T_V has a unique dominant left eigenvalue of unity with an eigenvector X of the form

$$XT_V = X, \quad X = \begin{pmatrix} 1 & \boldsymbol{x} \\ \boldsymbol{x}^\dagger & \mathsf{X} \end{pmatrix}. \tag{4.46}$$

All other eigenvalues λ satisfy $|\lambda| < 1$.

Proof. Since $\widehat{V}$ has block sizes $(1, \chi)$, the transfer matrix T_V has block sizes $(1, \chi, \chi, \chi^2)$ in the natural basis.[13] Moreover, it is block upper-triangular in that basis:

$$T_V = \begin{pmatrix} 1 & * & * & * \\ 0 & \overline{A}_0 & 0 & * \\ 0 & 0 & A_0 & * \\ 0 & 0 & 0 & T_A \end{pmatrix}, \ A_0 = \langle \widehat{\mathbb{1}}, \widehat{A} \rangle, \tag{4.47}$$

so the eigenvalues of T_V are those of the diagonal blocks.

[13] Schematically, $(\overline{1} \oplus \overline{\chi}) \otimes (1 \oplus \chi) \cong (1 \oplus \chi \oplus \chi \oplus \chi^2)$.

By first degreeness, all eigenvalues λ of the T_A block have $|\lambda| < 1$. A technical linear algebra fact, Lemma 26 from App. 4.A, shows the same is true for the A_0 and $\overline{A}_0$ blocks. The dominant eigenvalue of T_V is therefore $\lambda = 1$ from the trivial upper-left block of T_V.

To find the eigenvector, we compute XT_V, which yields

$$\begin{pmatrix} 1 & \boldsymbol{c}_0 + \boldsymbol{x} A_0 \\ \boldsymbol{c}_0^\dagger + A_0^\dagger \boldsymbol{x}^\dagger & \sum_\alpha \boldsymbol{c}_\alpha^\dagger \boldsymbol{c} + \boldsymbol{c}_\alpha^\dagger \boldsymbol{x} A_\alpha + A_\alpha^\dagger \boldsymbol{x}^\dagger \boldsymbol{c}_\alpha + A_\alpha^\dagger \mathsf{X} A_\alpha \end{pmatrix}. \tag{4.48}$$

So $\boldsymbol{x}$ and X are determined by

$$\boldsymbol{x}[I - A_0] = \boldsymbol{c}_0 \tag{4.49a}$$

$$\mathsf{X}[\mathrm{Id} - T_A] = Q \tag{4.49b}$$

$$Q := \sum_\alpha \boldsymbol{c}_\alpha^\dagger \boldsymbol{c}_\alpha + \boldsymbol{c}_\alpha^\dagger \boldsymbol{x} A_\alpha + A_\alpha^\dagger \boldsymbol{x}^\dagger \boldsymbol{c}_\alpha. \tag{4.49c}$$

As the eigenvalues λ of A_0 and T_A satisfy $|\lambda| < 1$, the operators on the left-hand sides of (4.49) are invertible and solutions $\boldsymbol{x}$ and X exist. The dominant eigenvalue therefore has the form (4.46). □

Intuitively, in terms of the state machine, the leading eigenvector of T_V is dominated by the "initial to initial" process. It is worth noting that (4.49) can be written as $Y - \sum_\alpha A_\alpha^\dagger Y A_\alpha = Q$, which is reminiscent of the discrete Lyapunov equation $Y - A^\dagger Y A = Q$ which occurs in control theory. This is a first indication of a nice connection, which we shall detail in Section 4.9 below.

We now "enlarge" the leading eigenvector of T_V to form the dominant Jordan block of T_W, which is responsible for the extensive norm, Eq. (4.35).

Proposition 17. *Suppose $\widehat{W}$ is an first degree iMPO for $\widehat{H}$ with order-unity trace:* $\mathrm{tr}[\widehat{H}] = O(1)$*. Then there is a vector $\boldsymbol{z}$ such that the matrices*

$$Z = \begin{pmatrix} X & \boldsymbol{z} \\ \boldsymbol{z}^\dagger & 0 \end{pmatrix}, \text{ and } Z' = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \tag{4.50}$$

[with the same X from Eq. (4.46)*] span the dominant Jordan block of T_W:*

$$\begin{pmatrix} Z\,T_W & Z' T_W \end{pmatrix} = \begin{pmatrix} Z & Z' \end{pmatrix} \begin{pmatrix} 1 & \rho \\ 0 & 1 \end{pmatrix}, \tag{4.51}$$

for some real number $\rho \geq 0$. The norm, Eq. (4.35)*, is then extensive with $||\widehat{H}_N||_F^2 \to \rho N$ as $N \to \infty$.*

This proposition is easily generalized to traceful operators at the cost of a more complex Jordan block structure. The proof, given in Appendix 4.A, is similar to the one for Prop 16, but somewhat more technical.

We note that X, $\boldsymbol{z}$ and ρ can be calculated from $\widehat{W}$, but computational tractable formulas use canonical forms, and await us in Sec. 4.6. Intuitively, the reason for the extensive norm is that the overlaps of $\boldsymbol{\ell\ell}$ with Z and $\boldsymbol{rr}$ with Z' are both 1, so if $\widehat{W}$ is first degree then

$$\boldsymbol{\ell\ell} T_W^N \boldsymbol{rr} \sim \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ 0 & 1 \end{pmatrix}^N \begin{pmatrix} 0 \\ 1 \end{pmatrix} = N\rho. \tag{4.52}$$

Therefore first degree operators, as anticipated by their name, have (Frobenius) norm which is a first degree polynomial in N.

In summary, we have identified a well-behaved class of local iMPOs — first degree operators — that are general enough to contain most operators of interest, and satisfy the physical requirements of an extensive norm. Crucially, first degree iMPOs are qualitatively distinct from generic infinite MPSes: their transfer matrix do not have a unique dominant eigenvalue, but rather a dominant Jordan block (whose eigenvalue is fixed to unity without normalization). Table 4.1 recapitulates these results. The distinction between a unique dominant eigenvalue versus a Jordan block is of paramount importance as we upgrade canonical forms from states to operators.

4.6 Canonical forms for Infinite MPOs

This section discusses canonical forms for infinite matrix product operators. We first show that canonical forms *exist*: any first degree iMPO admits a choice of gauge that brings it to left canonical form. Actually *computing* such a gauge transform is rather subtle. We first give a general-purpose algorithm, based on QR iteration, with fast convergence for generic iMPOs. Most iMPOs constructed to represent an analytical formula have a special property: they are upper triangular. In this case, canonicalization can be done by an more efficient, iteration free method. We also show that once an operator is in canonical form, it is easy to read off its norm. To our knowledge, canonical forms for operators have not been defined before, perhaps because of the non-trivial first degree requirement.

Existence of iMPO Canonical Forms

The definition of canonical form is much the same as in the finite case.

Definition 18. An iMPO $\widehat{W}$ is in **left-canonical form** if its upper-left block $\widehat{V}$ has orthonormal columns: $\forall b, c \leq \chi'$,

$$\sum_{a=0}^{\chi} \langle \widehat{W}_{ab}, \widehat{W}_{ac} \rangle = \delta_{bc}. \tag{4.53}$$

An iMPO is in **right canonical form** if its mirror is left canonical.

Defn. 18, the definition of iMPO canonical form, is closely related to the MPS case. Precisely, $\widehat{W}$ is left canonical as an iMPO if, and only if, $\widehat{V}$ is left canonical *as an MPS*. We can thus import many properties from the case of states. For example, (4.53) can be written in terms of the transfer matrix (defined in (4.38)) as

$$\mathrm{Id}_{[0,\chi]}\, T_V = \sum_\alpha V_\alpha^\dagger V_\alpha = \mathrm{Id}_{[0,\chi]}\,. \tag{4.54}$$

So $\widehat{W}$ is left-canonical whenever $\mathrm{Id}_{[0,\chi]}$ is a left eigenvector of T_V with eigenvalue 1. This fact is exactly what allows us to prove that canonical forms exist.

Proposition 19. *Let $\widehat{W}$ be a first degree iMPO. Then there exists a matrix L that which specifies a gauge transform*

$$\widehat{W}_L L = L\widehat{W} \tag{4.55}$$

so that $\widehat{W}_L$ is left canonical.

The proof itself is given in Appendix 4.B, but we briefly outline the idea. Prop. 16 tells us that, for any first degree $\widehat{W}$, the dominant eigenvector of T_V is $XT_V = X$. Suppose that we could take the "square root decomposition" $X = K^\dagger K$ with some invertible matrix K. Then we could enlarge K to $L = \mathrm{diag}(K\ \ 1)$ and use it as a gauge transform $\widehat{W}_L = L\widehat{W}L^{-1}$. Such a $\widehat{W}_L$ is left-canonical:

$$\begin{aligned}\mathrm{Id}\, T_{V_L} &= \sum_\alpha (K^{-1})^\dagger \widehat{V}_\alpha^\dagger K^\dagger K \widehat{V}_\alpha K^{-1} \\ &= (K^{-1})^\dagger X K^{-1} = \mathrm{Id}\end{aligned}$$

where $\widehat{V}_L = K\widehat{V}K^{-1}$ is the upper-left part of $\widehat{W}_L$. To turn this into a genuine proof, one must deal carefully with the case when L is *not* invertible — and this is precisely what we do in Appendix 4.B.

To demonstrate the utility of canonical forms, we now give a simple formula for the norm of an (i)MPO. For any traceless operator, we can easily "improve" the canonical form via the gauge transform

$$L_{lc} := \begin{pmatrix} 1 & & \\ & I & \boldsymbol{s} \\ & & 1 \end{pmatrix}, \boldsymbol{s} := [A_0 - I]^{-1}\boldsymbol{b}_0. \tag{4.56}$$

(Also see Lemma 28.) This will gauge away the identity components of the last so that:

$$\langle \widehat{\mathbb{1}}, \widehat{d}\rangle = \langle \widehat{\mathbb{1}}, \widehat{\boldsymbol{b}}_a\rangle = 0, \quad \forall 1 \le a \le \chi\,, \tag{4.57}$$

Doing this makes the dominant Jordan block particularly simple.

Proposition 20. *Suppose $\widehat{W}$ is an iMPO for $\widehat{H}$ in left-canonical form where* (4.57) *holds. Then the dominant Jordan block of $\widehat{W}$ is given by* (4.50) *and* (4.51) *with $X = \mathrm{Id}_{[0,\chi]}$, $\boldsymbol{z}_a = \langle \widehat{A}_{ab}, \widehat{\boldsymbol{b}}_b\rangle + \langle \widehat{\boldsymbol{c}}_a, \widehat{d}\rangle$, and*

$$\lim_{N\to\infty} ||H_N||_F^2/N = \rho = \langle \widehat{d}, \widehat{d}\rangle + \sum_{a=1}^{\chi} \langle \widehat{b}_a, \widehat{b}_a\rangle. \tag{4.58}$$

The proof is immediate from matrix multiplication. In practice, then, one should compute the intensive norm of an iMPO by bringing it to left canonical form, gauging away identities in $\widehat{\boldsymbol{b}}$ by (4.155), and applying (4.58). The intuitive reason this works is that, in left canonical form, orthonormality pushes all the weight in each term to the last site (e.g. $0.3X_1Y_2Z_3 \to X_1Y_2[0.3Z_3]$.) The norm is then simply the sums of the squares of the weights of the ending sites. The condition (4.57) ensures that all the edges incident to "f" in the automata are identity-free, i.e. no terms can "end prematurely".

The finite case is directly analogous. A finite operator H whose MPO is left-canonical with each $\widehat{W}^{(n)}$ also identity-free in the last column has norm

$$||H_N||_F^2 = \sum_{n=1}^{N} \left[\langle \widehat{d}^{(n)}, \widehat{d}^{(n)}\rangle + \sum_{a=1}^{\chi^{(n)}} \langle \widehat{b}_a^{(n)}, \widehat{b}_a^{(n)}\rangle \right]. \tag{4.59}$$

QR Iteration

We now present a general-purpose algorithm to gauge an iMPO $\widehat{W}$ into left canonical form. Recall that if we can decompose the dominant eigenvector $XT_V = X$ as $X = R^\dagger R$, then R is exactly the gauge transform we need. Any algorithm along these lines must follow the strategy: (I) find X, (II) decompose it to find R, and (III) deal with the case where R is not invertible. We will see that (I) and (II) are straightforward, but (III) requires considerable care.

Because X is the dominant eigenvector, it is simple to compute using the power method. If $X_{n+1} := X_n T_V$, then $X_n \to X$ as $n \to \infty$. The speed of convergence is controlled by the gap to the second-largest eigenvalue. Unlike in the MPS case, the second-largest eigenvalue is typically far less than 1, so X_n converges quite fast. We have therefore achieved (I).

To decompose X, we need to take the square-root. Simply taking the matrix square-root of X via eigendecomposition or Cholesky decomposition will severely reduce the precision (from 10^{-16} to 10^{-8} with the standard floating point), which is undesirable. To sidestep this, we use the technique of QR iteration, wherein each application of T_V is performed by taking a QR decomposition. Precisely, let $\widehat{W}_0 := \widehat{W}$ and for $n \geq 1$ inductively define

$$\widehat{Q}_n R_n := \widehat{QR}[\widehat{W}_{n-1}], \quad \widehat{W}_n := R_n \widehat{W}. \tag{4.60}$$

Algorithm 3 iMPO Left Can. Form: Iterated QR

```
procedure LeftCanQRIter($\widehat{W}, \eta$)                ▷ $\eta$: desired precision
    $L \leftarrow \mathrm{Id}_{[0,\chi+1]}$
    $\varepsilon \leftarrow \infty$                                   ▷ Current error
    while $\varepsilon > \eta$ do                          ▷ Repeat until convergence
        $(\widehat{Q}, R) \leftarrow \widehat{QR}(\widehat{W})$                       ▷ Eq. (4.21)
        $\widehat{W} \leftarrow R\widehat{Q}$
        $L \leftarrow R\,L$
        $\varepsilon \leftarrow ||R - \mathrm{Id}||$ if $R$ is square else $\infty$
    return $\widehat{Q}, L$
```

Let $\widetilde{R}_n$ denote the restriction of $R_n = \mathrm{diag}(\widetilde{R}_n\ 1)$ to the upper left blocks (and similarly for $\widetilde{Q}_n$). We have

$$\sum_\alpha \widehat{V}_\alpha^\dagger \widetilde{R}_{n-1}^\dagger \widetilde{R}_{n-1} \widehat{V}_\alpha = \sum_\alpha \widetilde{R}_n^\dagger \left(\widetilde{Q}_n\right)_\alpha^\dagger \left(\widetilde{Q}_n\right)_\alpha \widetilde{R}_n, \tag{4.61}$$

so $\left(\widetilde{R}_{n-1}^\dagger \widetilde{R}_{n-1}\right) T_V = \widetilde{R}_n^\dagger \widetilde{R}_n$. This computes the application of the transfer matrix while maintaining the factorized form, giving the limit:

$$\widetilde{R}_n^\dagger \widetilde{R}_n = X_n \xrightarrow{n\to\infty} X = \widetilde{R}^\dagger \widetilde{R}. \tag{4.62}$$

One could then gauge-transform by $R = \mathrm{diag}(\widetilde{R}\ 1)$ as $\widehat{W}_L R = RW$ to find a left canonical $\widehat{W}_L$. We have now achieved (II).

The above procedure is no more than a simple adaption of a well-known standard method in the iMPS context [118], and suffices to compute canonical forms for generic iMPOs. However there are many reasonable iMPOs for which it fails badly (we will encounter them in the application discussed in Section 4.10, Fig. 4.4 below). The essential problem is that convergence $X_n \to X$ does *not* guarentee $\widetilde{R}_n \to \widetilde{R}$, especially when X is a singular matrix. This is the main obstruction to achiving (III).

Algorithm 3 presents the "practical solution" to this conundrum. The idea is to apply a gauge transformation after every QR step, i.e.:

$$\widehat{W}_0 = \widehat{Q}_1 R_1\,,\ \widehat{W}_1 = R_1 \widehat{Q}_1\,,\ \widehat{W}_1 = \widehat{Q}_2 R_2\,,\ \widehat{W}_2 = R_2 \widehat{Q}_2 \ldots$$

Then $\widehat{W}_1$ is related to $\widehat{W}_0$ by a gauge transform $R_1 \widehat{W}_0 = \widehat{W}_1 R_1$, and $\widehat{W}_2$ to $\widehat{W}_0$ by $R_2 R_1 \widehat{W}_0 = \widehat{W}_1 R_2 R_1$, etc. The desired gauge transform to a canonical form will be approached by the product $L_n = R_n R_{n-1} \ldots R_1$. An important advantage of this method comes from bond dimension reduction: to see this, suppose that $\widehat{W}_0$ has bond dimension χ_0 but linearly dependent columns, so that $\widehat{Q}_1, R_1$ can have shape $(\chi_0 + 2) \times (\chi_1 + 2), (\chi_1 + 2) \times (\chi_0 + 2)$

respectively, with $\chi_1 < \chi_0$.[14] As a result, $\widehat{W}_1$ will have a smaller bond dimension χ_1. Thus, the first few iterations will reduce the bond dimension of $\widehat{W}$. Eventually, the bond dimension will stabilize, and R_n will become a square matrix, and invertible in most situations, thereby ameliorating the problem (III).

Unfortunately, there are still pathological cases where this algorithm will fail as well, but it gives a good balance between speed, applicability, and ease-of-implementation. Appendix 4.B proves the conditions under which Alg. 3 converges, supplies non-converging counterexamples, and a more complex algorithm which we prove *always* converges (Algorithm 8). We reiterate that Algorithm 3 will work almost always in practice, and the fool-proof algorithm is only used to handle rare exceptions.

We remark that the above discussion on iMPO canonical forms (including Appendix 4.B) can also be regarded as a careful treatment of iMPS canonical forms. To our knowledge, the subtlety involved in the convergence of QR iteration has not been thoroughly discussed previously, since it appears that the matrices encountered in iMPS calculations are always in a generic class for which any QR iteration scheme converges.

Upper Triangular Algorithm

When an iMPO is an upper-triangular operator-valued matrix — as is often the case when MPOs are constructed to represent an analytical Hamiltonian — it is possible to put it into canonical form with a non-iterative algorithm. In some sense, algorithms for canonical forms are a generalization of the Gram-Schmidt algorithm, where elementary row- and column-operations are replaced by gauge transforms. In the upper-triangular case, however, gauge transformations are so close to elementary row/column operations that we can adapt Gram-Schmidt directly. The result is a non-iterative algorithm that uses an upper-triangular solver to compute the gauge transform one column at a time.

Suppose we have an upper-triangular MPO

$$\widehat{W}_{M-1} = \begin{pmatrix} \widehat{\mathbb{1}} & \vert & \vert & \vert & \cdots \\ & \widehat{\boldsymbol{w}}_1 & \vert & \vert & \cdots \\ & & \widehat{\boldsymbol{w}}_2 & \vert & \cdots \\ & & & \widehat{\boldsymbol{w}}_3 & \cdots \\ & & & & \ddots \end{pmatrix}. \tag{4.63}$$

and assume, for induction, that the first M column vectors $\widehat{\boldsymbol{w}}_0, \cdots \widehat{\boldsymbol{w}}_{M-1}$ are already orthonormal. We want to modify $\widehat{\boldsymbol{w}}_M \to \widehat{\boldsymbol{w}}'_M$ to be orthogonal to all previous columns. To do

[14] This is known as "rank-revealing" QR, and can be done by removing vanishing rows of R and the corresponding columns of $\widehat{Q}$ after running some standard QR routine, for example.

this, we apply a gauge transformation which is the identity except for the Mth column:

$$R_M = \begin{pmatrix} 1 & 0 & & r_0 & & \\ & \ddots & & \vdots & & \\ & & 1 & r_{M-1} & & \\ & & & s_M & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}. \tag{4.64}$$

The transformation $\widehat{W}_M = R_M \widehat{W}_{M-1} R_M^{-1}$ is then easily computed[15] and maintains the upper-triangular form, while only affecting columns M and beyond. In particular, setting $s_M = 1$ temporarily,

$$\widehat{\boldsymbol{w}}'_M = \widehat{\boldsymbol{w}}_M - \sum_{a=0}^{M-1} r_a \widehat{\boldsymbol{w}}_a + \sum_{a=0}^{M-1} r_a \widehat{d}_M \boldsymbol{e}_a \tag{4.65}$$

where $\boldsymbol{e}_a$ is the standard basis vector $(\boldsymbol{e}_a)_b = \delta_{ba}$ and $\widehat{d}_M := (\widehat{\boldsymbol{w}}_M)_M = \widehat{W}_{MM}$ is the diagonal component of the Mth column. In Gram-Schmidt, the last term is absent, and one would simply set $r_b = \langle \widehat{\boldsymbol{w}}_b, \widehat{\boldsymbol{w}}_M \rangle$ to orthogonalize the columns. We need only make a slight modification to account for the last term.

Orthogonality against column $b < M$ is the condition

$$0 \equiv \langle \widehat{\boldsymbol{w}}_b, \widehat{\boldsymbol{w}}_M \rangle + \sum_{a=0}^{M-1} \Big(- \langle \widehat{\boldsymbol{w}}_b, \widehat{\boldsymbol{w}}_a \rangle + \langle \widehat{\boldsymbol{w}}_b, \widehat{d}_M \boldsymbol{e}_a \rangle \Big) r_a. \tag{4.66}$$

This is just a linear equation $K\boldsymbol{r} = \boldsymbol{c}$ where

$$\begin{aligned} K_{ba} &= \delta_{ba} - \langle \widehat{W}_{ba}, \widehat{W}_{MM} \rangle & (4.67\text{a}) \\ c_b &= \langle \widehat{\boldsymbol{w}}_b, \widehat{\boldsymbol{w}}_M \rangle , & (4.67\text{b}) \end{aligned}$$

the Kronecker-δ comes from the induction hypothesis $\langle \widehat{\boldsymbol{w}}_b, \widehat{\boldsymbol{w}}_a \rangle = \delta_{ba}$, and K is lower-triangular. Therefore we can easily solve for $\boldsymbol{r} = K^{-1}\boldsymbol{c}$ by back-substitution to find the r_b's, giving an $\widehat{\boldsymbol{w}}'_M$ orthogonal to previous columns. We can use the final free parameter, s_M, to normalize. The effect of s_M on column M is

$$\widehat{\boldsymbol{w}}'_M \to \widehat{\boldsymbol{w}}''_M = \frac{1}{s_M}(\widehat{\boldsymbol{w}}'_M - \widehat{d}_M \boldsymbol{e}_M) + \widehat{d}_M \boldsymbol{e}_M, \tag{4.68}$$

The normalization condition $1 \equiv \langle \widehat{\boldsymbol{w}}''_N, \widehat{\boldsymbol{w}}''_N \rangle$ implies

$$s_M = \sqrt{\frac{\langle \widehat{\boldsymbol{w}}_M, \widehat{\boldsymbol{w}}_M \rangle}{1 - \langle \widehat{d}_M, \widehat{d}_M \rangle}}. \tag{4.69}$$

[15] The inverse R_M^{-1} has the same form as R_M but with $r_a \to -r_a$ and $s_M \to 1/s_M$.

Exponential locality ensures the denominator is non-zero.

We have thus solved for the gauge transformation R_M to orthonormalize column M against the previous columns. Of course, this gauge will modify the columns beyond M, but those are treated in subsequent steps. The procedure is summarized in Algorithm 4 and has a total cost of $O(\chi^3)$ operations. In each loop, we perform a triangular solve and a matrix multiplication. The triangular solve costs $O(\chi^2)$ and, since R is almost the identity matrix, we can apply it in time $O(\chi^2)$ as well. With the outer loop of size χ, we have a total cost of $O(\chi^3)$.

Algorithm 4 iMPO Left Can. Form: Triangular

```
Require: $\widehat{W}$ upper-triangular
procedure LeftCanTriangular($\widehat{W}$)
    $R_T \leftarrow I_{1+\chi+1}$
    for $M \in [1, \chi]$ do
        $K_{ba} = \delta_{ba} - \langle \widehat{W}^\dagger_{ab}, \widehat{W}_{MM} \rangle\,, \quad m, k \in [0, M-1]$
        $c_b = \sum_{a=0}^{M-1} \langle \widehat{W}^\dagger_{bM}, \widehat{W}_{aM} \rangle\,, \quad m \in [0, M-1]$
        $\boldsymbol{r} \leftarrow K^{-1}\boldsymbol{c}$                         ▷ $O(\chi^2)$ triangular solve
        $R \leftarrow \mathrm{Id}_{1+\chi+1},\ R_{bM} \leftarrow r_b, \quad m \in [0, M-1]$
        $\widehat{W} \leftarrow R\widehat{W}R^{-1},\ R_T \leftarrow RR_T$                    ▷ only $O(\chi^2)$
        $s \leftarrow$ Eq. (4.69)
        $R \leftarrow \mathrm{Id}_{1+\chi+1},\ R_{MM} \leftarrow s$
        $\widehat{W} \leftarrow R\widehat{W}R^{-1},\ R_T \leftarrow LR_T$                    ▷ only $O(\chi)$
    return $\widehat{W}, R_T$
```

Several remarks are in order. First, this algorithm has an easily-curable instability, which arises when s_M in (4.69) is vanishingly small. This means $\widehat{w}'_M - \widehat{d}_M \boldsymbol{e}_M$ is also vanishing. Consequently, in terms of the state machine, the Mth state cannot be reached from the initial state, so one should simply discard the Mth row and column of $\widehat{W}$ (as well as the Mth row of the gauge matrix), and carry on.

In this section we have shown that first degree iMPOs can always be brought to canonical forms. We then gave two algorithms for computing them, one which converges well for almost all local iMPOs, and one which is specialized to upper-triangular iMPOs. Appendix 4.B gives a yet-more-general algorithm, which is guarenteed to converge for *all* first degree iMPOs. We now proceed to compression of infinite MPOs which, unlike canonicalization, hews closely to the finite case.

4.7 Compression of iMPOs

We now explain how to compress infinite MPOs. The algorithm is directly analagous to the finite case: use canonical forms to make an almost-Schmidt decomposition of the operator, then truncate the almost-Schmidt values. Subsequently, Section 4.8 will show it is virtually optimal by bounding its error and Section 4.9 will link operator compression to problems in control theory.

Suppose $\widehat{W}_R$ is an iMPO in right canonical form. Using the gauge from Lemma 28 we may impose $c_0 = \langle \widehat{\mathbb{1}}, \widehat{\boldsymbol{c}} \rangle = 0$ without loss of generality.[16] There is then a gauge transform between right and left canonical form,

$$C\widehat{W}_R = \widehat{W}_L C, \tag{4.70}$$

and $c_0 = 0$ implies $C = \mathrm{diag}(1\ \mathsf{C}\ 1)$ is block-diagonal. (To ease bookkeeping, we treat $\widehat{W}_R$ and $\widehat{W}_L$ as square matrices of the same dimension, though the algorithm works equally well for non-square iMPOs.) The SVD of $C = USV^\dagger$, now implies

$$USV^\dagger \widehat{W}_R = \widehat{W}_L USV^\dagger\,, \tag{4.71}$$

where U and V are unitary. Therefore, we can use them to gauge transform $\widehat{W}_{L,R}$ into

$$\widehat{Q} := U^\dagger \widehat{W}_L U \text{ and } \widehat{P} := V^\dagger \widehat{W}_R V\,, \tag{4.72}$$

which are left and right canonical, respectively. Furthermore, (4.71) implies that they are related by the gauge transform

$$\widehat{Q}S = S\widehat{P}\,. \tag{4.73}$$

Consequently, we obtain a **mixed canonical form** for the iMPO:

$$\begin{aligned}
\widehat{H}_W &= \cdots \widehat{W}_R \widehat{W}_R \widehat{W}_R \widehat{W}_R \cdots \\
&= \cdots \widehat{W}_L \widehat{W}_L C \widehat{W}_R \widehat{W}_R \cdots \\
&= \cdots \widehat{W}_L \widehat{W}_L USV^\dagger \widehat{W}_R \widehat{W}_R \cdots \\
&= \cdots \widehat{Q}\widehat{Q}S\widehat{P}\widehat{P} \cdots\,.
\end{aligned} \tag{4.74}$$

In the second line above, we inserted an L matrix at $-\infty$ and moved to the center using (4.70); in the fourth line, U and V are moved to $-\infty$ and $+\infty$ respectively[17].

Compression of iMPOs must be done on all bonds simultaneously and self-consistently, otherwise errors are incurred even when the compression is exact. To ensure this self-consistency, suppose for now that only $\chi' < \chi$ singular values are non-vanishing.[18] Then

$$S = \mathbb{P}\mathbb{P}^\dagger S = S\mathbb{P}\mathbb{P}^\dagger = \mathbb{P}S'\widehat{P}^\dagger \tag{4.75}$$

[16] Actually we only need the $\boldsymbol{t}$ part and set $\boldsymbol{s} = 0.$ in Eq. (4.155).

[17] These operations incur O(1) errors near the boundary, which are negligible for an iMPO.

[18] In this case, the optimal compression error is zero, but the procedure itself is identical to the case where the singular values are numerically small.

Algorithm 5 iMPO Compression

```
procedure iCompress($\widehat{W}, \eta$)                                   ▷ Cutoff $\eta$
    $\widehat{W}_R \leftarrow$ RightCan[$\widehat{W}$]
    $\widehat{W}_R \leftarrow R\widehat{W}_R R^{-1}$ so that $\widehat{\boldsymbol{c}}_0 = 0$         ▷ Use $\boldsymbol{t}$ from Lem. (28)
    $\widehat{W}_L, C \leftarrow$ LeftCan[$\widehat{W}_R$]
    $(U, S, V^\dagger) \leftarrow$ SVD[$C$]
    $\widehat{Q}, \widehat{P} \leftarrow U^\dagger \widehat{W}_L U\ ,\ V^\dagger \widehat{W}_R V$
    $\chi' \leftarrow \max\{a \in [1, \chi] : s_a > \eta\}$                     ▷ Defines $\mathbb{P}$ (4.76)
    $\widehat{Q}, S, \widehat{P} \leftarrow \mathbb{P}^\dagger \widehat{Q}\mathbb{P}, , \mathbb{P}^\dagger S\mathbb{P}, \mathbb{P}^\dagger \widehat{P}\mathbb{P}$
    return $\widehat{P}$                                             ▷ One could also return $\widehat{Q}$.
```

where $\mathbb{P}$ is the projection matrix to the first χ' indices in the middle block

$$\mathbb{P}_{ab} = \begin{cases} \delta_{ab} & a \in \{0, 1, \ldots, \chi', \chi + 1\} \\ 0 & \text{otherwise,} \end{cases} \tag{4.76}$$

and $\mathsf{S}' = \operatorname{diag}(s_1, \ldots, s_{\chi'})$. We can then us the fact that, in mixed canonical form, the position of S can be freely translated to any site using (4.73). We can then use (4.75) to "conjure" up projectors at every bond of (4.74):

$$\begin{aligned} H_W &= \cdots \widehat{Q}\widehat{Q}S\widehat{P}\widehat{P} \cdots && (4.77) \\ &= \cdots \mathbb{P}\mathbb{P}^\dagger \widehat{Q}\mathbb{P}\mathbb{P}^\dagger \widehat{Q}\mathbb{P}S'\mathbb{P}^\dagger \widehat{P}\mathbb{P}\mathbb{P}^\dagger \widehat{P}\mathbb{P}\mathbb{P}^\dagger \cdots \\ &= \cdots \widehat{Q}'\widehat{Q}'S'\widehat{P}'\widehat{P}' \cdots && (4.78) \end{aligned}$$

where $\widehat{Q}' = \mathbb{P}^\dagger \widehat{Q}\mathbb{P}$ and $\widehat{P}' = \mathbb{P}^\dagger \widehat{P}\mathbb{P}$ now have bond dimension χ'. Either $\widehat{Q}'$ or $\widehat{P}'$ can be returned as a compression of the original iMPO; one may make a choice keeping in mind that $\widehat{Q}'$ and $\widehat{P}'$ are approximately left and right canonical, respectively. Since we have assumed that the singular values beyond χ' vanish exactly, this is an exact compression. When this is not true, there will be some finite error (see Sec. 4.8) but the procedure is unchanged. Algorithm 5 gives an implementation, which we reiterate works also for non-square matrices.

4.8 Operator Entanglement and Error Bounds

In this section we discuss the error resulting from compressing an operator. The first stage in our analysis will be to show that, just as the singular values of an MPS are closely related to the entanglement, the almost-Schmidt values of an MPO are closely related to the *operator* entanglement entropy. We will immediately apply this relation to answer a practical question: how accurate is our compression algorithm? We will derive a quantitative bound on the error and show the algorithm is ϵ-close to optimal. Finally, we will show that the change in the sup norm is small under compression and hence our compression algorithm is suitable to use when finding ground states.

Relation to Operator Entanglement

To assess the accuracy of our MPO compression scheme, we require a point of comparison. For this, we recall that all MPO's can be thought of as (non-injective) MPSes, and can be compressed via the true Schmidt decomposition. We will refer to this as the "MPS" compression method. For iMPOs, the iMPS method will simply fail, due to the Jordan block structure and the reasons detailed in Section 4.6, as well as below, so our compression scheme has no obvious competitor in the infinite case. On a finite chain, however, both methods are valid, and it is meaningful to compare the MPO and "MPS" methods.

It is well-known that the matrix product compression of a state is intimately related to its bipartite entanglement spectrum. The same notion can be defined for an operator $\widehat{H}$ viewed as a state. If we consider a finite chain $[1, N]$ and make an entanglement cut on bond $(n, n+1)$, then the (true) operator **Schmidt decomposition** is

$$\widehat{H} = \sum_{a=-1}^{\chi} \lambda_a \widehat{\mathcal{O}}_L^a \otimes \widehat{\mathcal{O}}_R^a \, , \; \mathrm{Tr}[\widehat{\mathcal{O}}_L^{a\dagger} \widehat{\mathcal{O}}_L^b] = \delta^{ab} \tag{4.79}$$

(and the same for R), where the $\widehat{\mathcal{O}}_L$'s and $\widehat{\mathcal{O}}_R$'s act only on the left or right of the cut respectively. The Schmidt values $\lambda_{-1} \geq \lambda_0 \geq \cdots \lambda_\chi > 0$ are unique and positive.[19] Note that we do *not* normalize $\sum_a \lambda_a^2$ to unity.

The reason the MPO compression scheme works is the close, quantitative, resemblance between the almost-Schmidt decomposition, Eq. (4.6), and the true Schmidt decomposition, Eq. (4.79). To see this, we start with the almost-Schmidt decomposition and convert it to the true one. Suppose we have an almost-Schmidt decomposition (Definition 9):

$$\begin{aligned} \widehat{H} &= \widehat{H}_L \otimes \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \otimes \widehat{H}_R + \sum_a s_a \widehat{h}_L^a \otimes \widehat{h}_R^a \\ &= \begin{pmatrix} \widehat{\mathbb{1}}_L & \widehat{\boldsymbol{h}}_L & \widehat{H}_L \end{pmatrix} \begin{pmatrix} 1 & & \\ & \mathsf{S} & \\ & & 1 \end{pmatrix} \begin{pmatrix} \widehat{H}_R & \widehat{\boldsymbol{h}}_R & \widehat{\mathbb{1}}_R \end{pmatrix}^T . \end{aligned} \tag{4.80}$$

where $\mathsf{S} = \mathrm{diag}(s_1 \geq \cdots \geq s_\chi)$ is a diagonal matrix built from the almost-Schmidt values and $\{\widehat{\mathbb{1}}_{L/R}, \widehat{h}^1_{L/R}, \ldots, \widehat{h}^\chi_{L/R}\}$ are already orthonormal. All we need to do to get to the true Schmidt decomposition is to add $\widehat{H}_{L/R}$ to the list and orthonormalize. Explicitly, we apply

[19] The irregular index convention for the λ_a's will prove convenient below.

a Gram-Schmidt update:

$$\begin{pmatrix}\widehat{\mathbb{1}}_L & \widehat{\boldsymbol{h}}_L & \widehat{H}_L\end{pmatrix} = \begin{pmatrix}\widehat{\mathbb{1}}_L & \widehat{\boldsymbol{h}}_L & \widehat{H}'_L\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & \mathrm{Id} & \boldsymbol{p}_L\\ 0 & 0 & \mathcal{N}_L\end{pmatrix}, \tag{4.81}$$

where $p^a_L := \langle \widehat{h}^a_L, \widehat{H}_L\rangle$ ensures orthogonality and $\mathcal{N}_L := ||\widehat{H}_L||^2_F - ||\boldsymbol{p}_L||^2_F$ enforces normalization, so that $\{\widehat{\mathbb{1}}, \widehat{h}^1_L, \dots, \widehat{h}^\chi_L, \widehat{H}'_L\}$ are now orthonormal. Doing the same on the right side, the operator now becomes

$$\widehat{H} = \begin{pmatrix}\widehat{\mathbb{1}}_L & \widehat{\boldsymbol{h}}_L & \widehat{H}'_L\end{pmatrix}\underbrace{\begin{pmatrix}\mathcal{N}_R & \boldsymbol{p}_R & 0\\ 0 & \mathsf{S} & \boldsymbol{p}_L\\ 0 & 0 & \mathcal{N}_L\end{pmatrix}}_{M:=}\begin{pmatrix}\widehat{H}'_R & \widehat{\boldsymbol{h}}_R & \widehat{\mathbb{1}}_R\end{pmatrix}^T. \tag{4.82}$$

It follows that the true Schmidt values, i.e. the entanglement spectrum, is given by the singular values of the matrix M. The essential point is that M and S are almost the same matrix — and so their spectra are as well. We compute the precise relation between the singular values of M and its matrix elements in Appendix 4.C with rank-one updates, and import those results to here for show the optimality of our method.

The dominant feature of the entanglement spectrum is the separation of scales between extensive and intensive values. Suppose $\widehat{H}$ comes from a translation-invariant MPO on $N \gg 1$ sites, and our entanglement cut is at some bond $(n, n+1)$ near the middle. Then the matrix elements of M have a separation of scales:

$$\mathcal{N}_L, \mathcal{N}_R \in \Theta(N), \quad s_a, \boldsymbol{p}_L, \boldsymbol{p}_R \in O(1). \tag{4.83}$$

Without the $\boldsymbol{p}$'s, M would be diagonal. There would then be two extensive singular values, namely $\mathcal{N}^2_L$ and $\mathcal{N}^2_R$, and χ intensive ones, $s^2_1, \dots, s^2_\chi$. Appendix 4.C shows that the extensive/intensive separation remains after the $\boldsymbol{p}$'s have been taken into account:

$$\lambda^2_{-1}, \lambda^2_0 \in \Theta(N)\,,\ \lambda_a \in \mathrm{O}(1)\,,\ a = 1, \dots \chi\,. \tag{4.84}$$

This result illustrates again why the MPS compression scheme must fail with iMPOs: the extensive Schmidt values diverge in the thermodynamic limt. Normalizing the Schmidt values, that is, considering $\sigma_a := \lambda_a/\sqrt{\sum_b \lambda^2_b}$, would not be helpful: for any $a > 0$, $\sigma_a \in \mathrm{O}(1/N)$ vanishes in the thermodynamic limit, so that the normalized spectrum retains no nontrivial information about the operator.

Intuitively, the separation of scales is a consequence of locality. Indeed, the two extensive Schmidt vectors are very close to $\widehat{H}_L \otimes \widehat{\mathbb{1}}_R$ and $\widehat{\mathbb{1}}_L \otimes \widehat{H}_R$ — exactly the operators that the block structure of our MPOs keeps track of "for free". In other words, the local MPO construction automatically keeps track of the extensive part of the spectrum (to a good approximation), and we need only deal with the *intensive* part. This is precisely the role of the almost-Schmidt decomposition.

Comparison of MPO and "MPS-style" Compression

We now make to a quantitative comparison between MPO and "MPS-style" methods on a finite chain. If we compress an operator from bond dimension χ down to χ' with either scheme, the new operators are

$$\widehat{H}_{\mathrm{MPS}} = \sum_{a=-1}^{\chi'} \lambda_a \widehat{\mathcal{O}}_L^a \otimes \widehat{\mathcal{O}}_R^a \,,$$

$$\widehat{H}_{\mathrm{MPO}} = \widehat{H}_L \otimes \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \otimes \widehat{H}_R + \sum_{a=1}^{\chi'} s_a \widehat{h}_L^a \otimes \widehat{h}_R^a \,,$$

respectively. The orthogonality properties of the decompositions tell us

$$||\widehat{H} - \widehat{H}_{\mathrm{MPS}}||_F^2 = \sum_{a=\chi'+1}^{\chi} \lambda_a^2 := \varepsilon_{\mathrm{MPS}}(\chi'), \tag{4.85}$$

$$||\widehat{H} - \widehat{H}_{\mathrm{MPO}}||_F^2 = \sum_{a=\chi'+1}^{\chi} s_a^2 := \varepsilon_{\mathrm{MPO}}(\chi'). \tag{4.86}$$

To compare these, we use the eigenvalue interlacing relation (derived in Appendix 4.C)

$$s_a \geq \lambda_a \geq s_{a+2} \,, \forall\, a \in [1, \chi - 2] \,. \tag{4.87}$$

We can therefore conclude

$$\varepsilon_{\mathrm{MPS}}(\chi') \leq \varepsilon_{\mathrm{MPO}}(\chi') \leq \varepsilon_{\mathrm{MPS}}(\chi' - 2) \,. \tag{4.88}$$

This means the difference between our scheme and the MPS scheme is within *two* Schmidt values, which is negligible, since in practice one always truncates sufficiently deep into the spectrum that $s_{\chi'}$ is small.

Since the MPS truncation scheme is known to be optimal[120], we can make the error from our MPO scheme ε-close to optimal, by truncating at χ' large enough that $|s_{\chi'} - s_{\chi'-2}| < \varepsilon$. There is no strict guarantee that this is possible, but for physical operators the entanglement spectrum usually becomes a continuum with increasingly small separation. It is in this sense that our truncation scheme is ε-close to optimal. We remark that the error analysis above applies to the truncation of a finite MPO on an individual bond. It would be interesting to analyze the global error of an iMPO compression, but we expect it to be almost exactly the same as the iMPS case.

In summary, the MPO compression scheme only captures the intensive Schmidt values, avoiding the pathological, extensive parts of the entanglement spectrum. As a result, we obtain an excellent approximation to the optimal "MPS" compression while preserving the locality structure.

4.9 Relation to Control Theory

Remarkably, our MPO canonicalization procedure is a generalization of an extremely well-studied problem in the field of control theory known as "model order reduction." With this connection in mind, one can use highly optimized libraries from that community to compute MPOs compressions for general two-body Hamiltonians. The relation to control theory was noted previously in Refs [124, 125]. Morally, one can think of writing the interaction potential as a sum of decaying exponentials. The MPO, in turn, can then be written as the sum of the small bond dimension MPOs for each operator. Our compression procedure is a strict generalization of this technique: if the input to our algorithm is a two-body interaction, then it *automatically* reproduces the sum of exponentials technique. On the other hand, higher-body Hamiltonians do not obviously map to the problem solved in control theory, so it would be interesting to pursue whether our procedure has useful implications for control theory.

The control systems setting is a "state-space" system: a dynamical system whose state is parameterized by a χ-dimensional vector $\boldsymbol{x}(t)$ with linear dynamics in discrete time. The dynamics are defined by the update rule

$$\begin{aligned}\boldsymbol{x}(t) &= A\boldsymbol{x}(t-1) + B\boldsymbol{u}(t)\\ \boldsymbol{y}(t) &= C\boldsymbol{x}(t) + D\boldsymbol{u}(t)\end{aligned} \tag{4.89}$$

where $\boldsymbol{u}(t)$ is an n_i-dimensional vector of possible "input" perturbations, $\boldsymbol{y}(t)$ a n_o dimensional vector of "outputs," and A is a matrix of size $\chi \times \chi$, B is $\chi \times n_i$, C is $n_o \times \chi$, and D is $n_o \times n_i$. The data can thus be bundled into a $(n_o+\chi)\times(n_i+\chi)$ matrix $\begin{pmatrix} C & D \\ A & B \end{pmatrix}$, which was the motivation for our MPO block conventions. One also defines **transfer function** of the system, $G(t) := CA^tB$, an $n_o \times n_i$ matrix which describes the linear input-output response at time t.

Two fundamental questions arise in the control theory setting. (I) Given a set of observations $G(t)$, what state-space system (A, B, C, D) can reproduce the observations? (II) Given a state-space system of dimension χ, can we produce a state-space system of lower order $\chi' < \chi$ which approximates $G(t)$? This problem could arise, for example, when modelling a complex electrical circuit, where $\boldsymbol{x}(t)$ parameterizes the voltages on wire segments, which we wish to approximate by a simpler "lumped element" circuit with fewer components.

It is easy to see that a state-space system is equivalent to an MPO in the particular case of a *two-body* Hamiltonian. A two-body interaction takes the general form

$$\widehat{H} = \sum_{x>y} \sum_{\alpha,\beta=1}^{n_o,n_i} \widehat{O}_x^{\alpha} V^{\alpha\beta}(x-y) \widehat{P}_y^{\beta} \tag{4.90}$$

where $\{\widehat{\mathbb{1}}\} \cup \{\widehat{O}_x^{\alpha}\}_{\alpha=1}^{n_0}$ and $\{\widehat{\mathbb{1}}\} \cup \{\widehat{P}_y^{\beta}\}_{\beta=1}^{n_i}$ are orthonormal sets of operators on sites x and y respectively. On the other hand, each set of matrices A, B, C as in (4.89) define an MPO in regular form via

$$\widehat{\boldsymbol{c}} = \widehat{\boldsymbol{O}}C\,,\ \widehat{\boldsymbol{b}} = B\widehat{\boldsymbol{P}}\,,\ \widehat{A} = A\widehat{\mathbb{1}}\,,\ \widehat{d} = 0\,,$$

where $\widehat{\boldsymbol{O}} = (\widehat{O}^{\alpha})_{\alpha=1}^{n_0}$ and $\widehat{\boldsymbol{P}} = (\widehat{P}^{\beta})_{\beta=1}^{n_0}$. It is not hard to check this MPO represents the Hamiltonian (4.90) if and only if

$$[CA^rB]^{ab} = V^{ab}(r)\,.$$

This data is in precise agreement with that of state-space system, with the transfer matrix $G(t)$ of the state-space encoding the two-body interaction $V(r)$. One could easily include on-site terms as well, in which case $\widehat{d}$ would be non-zero.

With this mapping, we see that problems (I) and (II) are equivalent to finding an MPO which reproduces a desired two-body interaction, and approximating an MPO by one of lower bond dimension. In the control theory literature, (I) has been solved by an algorithm of Kung [126], and (II) by "balanced truncation"[130], which we focus on here.

The starting point of the balanced truncation algorithm is the "controllability" Gramian X and the "observability" Gramian Y,

$$X \equiv \sum_{k=0}^{\infty} A^k BB^\dagger \left(A^\dagger\right)^k \tag{4.91a}$$

$$Y \equiv \sum_{k=0}^{\infty} \left(A^\dagger\right)^k C^\dagger C A^k \tag{4.91b}$$

They are determined by the discrete Lyapunov equations

$$AXA^\dagger = X - BB^\dagger \tag{4.92}$$

$$A^\dagger YA = Y - C^\dagger C \tag{4.93}$$

We can identify these as the fixed point condition for the left/right eigenvectors of the right/left transfer matrix $T_{R/L}$ of $\widehat{W}$ (c.f. T_V above) in the particular case that $\widehat{A} = A\mathbb{1}$. The controllability Gramian X is nothing other than the relevant block of the dominant eigenvector of the transfer matrix, and similarly for T_L and Y.

The idea of balanced truncation is to use the gauge freedom $A \to gAg^{-1}, C \to Cg^{-1}, B \to gB$, under which the Gramians transform as $X \to gXg^\dagger, Y \to {g^\dagger}^{-1}Yg^{-1}$, to demand that the Gramians be equal and diagonal: $X = Y = \mathrm{diag}(\Sigma)$. This is called the *balanced* condition. The Σ are called the "Hankel singular values" for reasons we will explain shortly. In operator language, this is nothing other than the almost-Schmidt decomposition Eq. (4.6) with values $s_a = \Sigma_a$. In balanced truncation, the model is then reduced by keeping the largest Σ_a, which is known to be optimal with respect to a particular norm, the "Hankel norm" [131].

Indeed, with this mapping in mind, the balanced truncation algorithms found in the literature are equivalent to the canonicalization procedure discussed here: solve the Lyapunov equations for the Gramians X, Y (equivalent to finding the dominant eigenvector of transfer matrix), compute the Cholesky decompositions $X = RR^\dagger$ and $Y = LL^\dagger$, and then SVD $U\Sigma V = L^\dagger R$, and let $g = \Sigma^{-1/2} V R^{-1}$.

Why are they called Hankel singular values? This brings us to Kung's algorithm, which obtains an approximate state-space representation given the desired output $G(t) \sim V(r)$. For simplicity, let's consider the simplest $n_i = n_o = 1$ case, arising for instance from a density-density interaction $\widehat{H} = \sum_{i,r>0} \widehat{n}_{i+r} V(r) \widehat{n}_i$. It is easy to see that in the mixed-canonical form at bond $(0, 1)$, the left / right operators can be chosen to be $\widehat{h}^i_L = \widehat{n}_{-i}, \widehat{h}^i_R = \widehat{n}_{i+1}$ for $i \geq 0$ so that $H = \sum_{i,j} \widehat{h}^i_L V(i+j+1) \widehat{h}^j_R$. The middle tensor then takes the form

$$M = \begin{pmatrix} & V(3) & V(2) & V(1) \\ \cdots & V(4) & V(3) & V(2) \\ & V(5) & V(4) & V(3) \\ \iddots & & \vdots & \end{pmatrix}, \tag{4.94}$$

which is by definition a “Hankel matrix,” with singular values $M = U\Sigma V$ consequently referred to as the Hankel singular values.

The connection results in highly optimized routines to compute the optimal A, B, C from the desired V using the Hankel structure. These are provided, for example, in the MATLAB Control Systems Toolbox as `balred, imp2ss` and in the SLICOT library [20] as `AB09AD`. The latter has a convenient Python API provided in the “`control`” library [21], which we have used with great success for quantum Hall DMRG [132].

While the equivalence is clear in the two-body case, what is the control theory interpretation of canonicalizing and truncating a more general MPO? This seems like an interesting question.

4.10 iMPO Examples

This section provides two numerical examples of iMPO compression. This is where our almost-Schmidt compression scheme truly shines, as the standard “MPS”-type truncation schemes do not work at all in this regime. Indeed, to our knowledge, our algorithm is the only one known to work for general iMPOs. We first give a “proof-of-concept” example for long-ranged Hamiltonians and then give an iMPO implementation of the Lanczos algorithm.

We consider the three-body Hamiltonian

$$\widehat{H}_2 = \sum_{n\in\mathbb{Z}} \sum_{x,y>0} \widehat{Z}_{n-x} \widehat{X}_n \widehat{Z}_{n+y} J_x J_y, \quad J_r = r^{-2}, \tag{4.95}$$

with power-law interaction. To encode the Hamiltonian (which has a formally infinite bond-dimension), we give the power-law interaction a large spatial cutoff R: $J_r := 0$ for $r > R$,

[20] See `http://slicot.org`.

[21] See `http://python-control.org`.

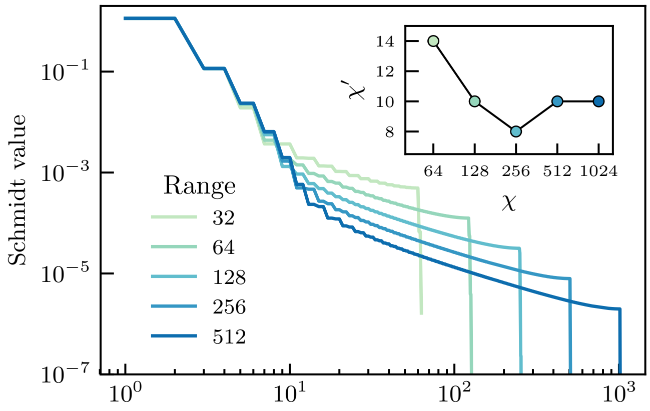


Figure 4.3: Compression of the iMPO representing (4.95). Main: The almost-Schmidt spectra of iMPOs representing $\widehat{H}_2$ with spatial cutoff R ranging from 32 to 512. As $R \to \infty$, the largest s_a converge to a point-wise limit, while the long tails rapidly decays (so the latter are finite-R artifacts). Inset: the bond dimensions of the iMPO before and afterwards with a cutoff of $\varepsilon = 10^{-4}$. Other numerical thresholds are the same as Fig. 4.2.

which we vary, so that the pre-compression bond dimension is $\chi = 2R$. The pre-compressed iMPOs have a block structure specific to three-body interaction; for example, when $R = 3$, we have

$$\left(\begin{array}{c|ccc|ccc|c}
\widehat{\mathbb{1}} & \widehat{Z} & 0 & 0 & & & & \\
\hline
 & 0 & \widehat{\mathbb{1}} & 0 & J_1\widehat{X} & 0 & 0 & \\
 & 0 & 0 & \widehat{\mathbb{1}} & J_2\widehat{X} & 0 & 0 & \\
 & 0 & 0 & 0 & J_3\widehat{X} & 0 & 0 & \\
\hline
 & & & & 0 & \widehat{\mathbb{1}} & 0 & J_1\widehat{Z} \\
 & & & & 0 & 0 & \widehat{\mathbb{1}} & J_2\widehat{Z} \\
 & & & & 0 & 0 & 0 & J_3\widehat{Z} \\
\hline
 & & & & & & & \widehat{\mathbb{1}}
\end{array}\right) . \tag{4.96}$$

We then compress them the iMPO compression routine (Algorithm 5) which calls the upper-triangular canonical form subroutine (Algorithm 4). The results are given in Figure 4.3. For any reasonable tolerance, as $R \to \infty$, the compressed bond dimension stabilizes to a tiny value, thanks to the rapid decay of the almost Schmidt values. It is also interesting to examine a compressed MPO (from $\chi = 2R = 256$ to $\chi' = 4$):

$$\left(\begin{array}{c|cc|cc|c}
\widehat{\mathbb{1}} & \widehat{Z} & \widehat{Z} & & & \\
\hline
 & 0.178\widehat{\mathbb{1}} & & .749\widehat{X} & .114\widehat{X} & \\
 & & .742\widehat{\mathbb{1}} & .11\widehat{X} & .0117\widehat{X} & \\
\hline
 & & & 0.178\widehat{\mathbb{1}} & & \widehat{Z} \\
 & & & & .742\widehat{\mathbb{1}} & \widehat{Z} \\
\hline
 & & & & & \widehat{\mathbb{1}}
\end{array}\right) . \tag{4.97}$$

Remarkably, while the strict locality of the uncompressed MPO is compromised, the block triangular structure of (4.96) is intact. We can clearly see that each power-law is approximated by a sum of exponential decays governed by the 2 × 2 matrix on the diagonal block of (4.97). Here we have applied a gauge transform after compression to make the MPO upper-triangular, but this is not possible in general as the SVD step will destroy "triangularizability".

Our final example is somewhat more involved: an iMPO implementation of the Lanczos algorithm. The Lanczos algorithm is originally from numerical linear algebra, where it is used to tri-diagonalize a matrix. However, it was recognized in the 1980s [31] that it provides an exact mapping from many-body dynamics problems to 1d quantum mechanics problems on a semi-infinite tight-binding model. (This is known as the "recursion method", see [10] for a review.) Recent work by some of us [1] has found there are deep connections between the Lanczos algorithm, thermalization, operator complexity, and quantum chaos.

The Lanczos algorithm is a simple iteration. Suppose $\widehat{H}$ is a Hamiltonian and $\widehat{O}$ is a Hermitian operator. Conceptually, the Lanczos algorithm constructs the Krylov subspace $\mathrm{span}\{\widehat{O}, [\widehat{H}, \widehat{O}], [\widehat{H}, [\widehat{H}, \widehat{O}]], \dots\}$ and iteratively orthonormalizes it. More precisely, we start from $\widehat{O}_{-1} = 0$, $\widehat{O}_0 := \widehat{O}$, $b_0 := 0$, and for $n > 0$, we define recursively

$$\begin{aligned} \widehat{A}_n &:= [\widehat{H}, \widehat{O}_n] - b_{n-1}\widehat{O}_{n-2} \\ \widehat{O}_n &:= b_n^{-1}\widehat{A}_n \text{ where } b_n := ||\widehat{A}_n||^{1/2} \,. \end{aligned} \tag{4.98a}$$

The b_n's are known as the *Lanczos coefficients*, and it is well-known that $\{\widehat{O}_0, \dots, \widehat{O}_n\}$ is an orthonormal basis of the n-dimensional Krylov subspace. These objects are highly relevant for the operator dynamics $\widehat{O}(t) = e^{i\widehat{H}t}\widehat{O}e^{-i\widehat{H}t}$, and it is desirable to compute as many of them as possible.

For generic many-body problems, exactly computing n Lanczos coefficients requires $\mathrm{O}(e^{Cn})$ resources. Now, whenever $\widehat{H}$ and $\widehat{O}$ are representable as iMPOs, the whole Lanczos algorithm can be implemented using iMPOs using elementary operations from Appendix 4.D and the intensive norm formula (4.58). If $\widehat{O}_0$ is first degree and $\widehat{H}$ is strictly local, all iMPOs generated in the process will be first degree, so our compression scheme can potentially reduce the computation cost of the Lanczos algorithm.

We benchmarked our iMPO implementation of the Lanczos algorithm, with the paradigmatic chaotic Ising chain, see Fig. 4.4. Remarkably, we observe that the resulting bond dimension of the operators $\widehat{O}_n$ grows only polynomially:

$$\chi[\widehat{O}_n] = \mathrm{O}(n^a) \,,\, a \approx 2 \,, \tag{4.99}$$

shown in Fig. 4.4 (c), while one would naively expect exponential growth. This means that, in principle, one could reach $n = 60 - 80$ with moderate hardware, far beyond $30 - 40$ by the exact method [1].

Practically, however, numerical precision becomes a limiting issue. Due to the iterative nature of the algorithm, any small compression error in $\widehat{O}_n$ is magnified on subsequent

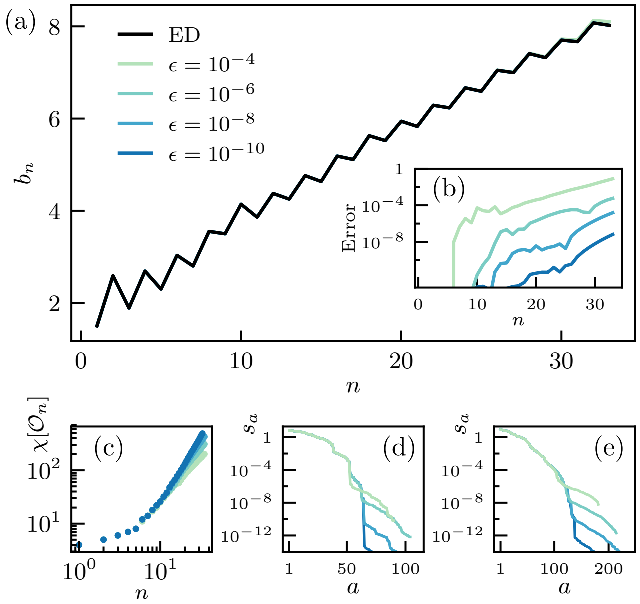


Figure 4.4: Results of an iMPO implementation of the Lanczos algorithm, applied to $\widehat{H} = \frac{1}{2}\sum_n Z_n Z_{n+1} - 1.05 Z_n + 0.5 X_n$ and $\widehat{O} = \sum_n Z_n$. (a) The Lanczos coefficients b_n computed by the iMPO implementation with SVD truncation threshold ε, compared to the exact method ("ED") of Ref. [1]. (b) Error in the b_n's at precision ε (compared to "ED" values). (c) Bond dimension of the operators $\widehat{O}_n$. The growth rate is roughly $\mathrm{O}(n^2)$. (d) The almost Schmidt spectra of $\widehat{O}_{10}$. A large gap is visible at $a \sim 60$ where s_a drops by $\sim 10^{-6}$. (e) The almost Schmidt spectra of $\widehat{O}_{20}$. The gap is barely visible even with the smallest ε; the error starts to grow rapidly around the same n.

steps. One can see from Fig. 4.4 (d) and (e) that the $\widehat{O}_n$'s singular value spectrum has a gap where the almost Schmidt values fall off by several orders of magnitude. A truncation targeted at the gap will be essentially lossless. However, the smallest singular value above the gap decreases rapidly with n, eventually reaching machine precision. Beyond that point, the singular value spectrum will look continuous with no apparent gap, and any further truncation will induce errors that grow quickly — as shown in Fig. 4.4 (b). One could account for this by dynamically increasing the working precision along with n. Although this is harder to implement and slower, the resource cost would still grow only polynomially with n, a qualitative improvement over the exact method, so long as (4.99) continues to hold. It will be very interesting to elucidate the reason of such an advantageous bond dimension scaling.

4.11 Unit Cell MPOs

Many physical systems have translationally-invariant unit cells with more than one site or degree of freedom. As a simple example, putting a staggered magnetic field $(-1)^n\widehat{Z}_n$ on a spin chain requires a unit cell of two sites. To describe such Hamiltonians with an MPO, one can either enlarge the on-site Hilbert space so it contains all N sites in the unit cell, or using N tensors to describe the unit cell at the cost of the system only being invariant under translation by N sites instead of 1 sites. Computationally, the second solution is far more efficient, and such MPOS are called unit cell MPOs. This section will generalize our previous notions of compression and canonicalization to the unit cell case. Just as in previous sections, the compression algorithm is essentially straightforward, while the canonicalization algorithm requires the new idea of topological ordering.

By far the most important use-case for unit cell MPOs is when doing DMRG on 2D Hamiltonians. A standard technique is to use a 'thin cylinder' geometry of circumference L_y sites and length $L_x \to \infty$. One then chooses a linear (1D) ordering for the sites on the cylinder by wrapping around in a helical pattern. This effectively reduces the problem to a 1D chain, but at a cost: interactions at distance r in 2D can be as far as $L_y \times r$ in 1D. Furthermore, the resource cost grows hugely, as the matrix product state (MPS) bond dimension needed to accurately capture a 2D area law state grows as $\chi \sim e^{L_y}$. In practice, therefore, 2D DMRG is often limited to around $L_y = 6$, even with bond dimensions of $\chi \sim 10,000$ or more. For sufficiently long-range interactions in 2D, however, the bottleneck is not the MPS bond dimension, but rather the MPO bond dimension needed to encode the Hamiltonian. Naively, long-range interactions of range R result in iMPOs of bond dimension $D \sim R^3$, and thence DMRG scales as $D^3 \sim R^9$, which becomes quickly impractical. The algorithms given below allow one to proceed by finding the best approximation for the iMPO of bond dimension $D' \ll D$. For many physical Hamiltonians, this compression incurs only a minor penalty (say, 10^{-8}) in the precision of the eventual ground state. For problems where the naive bond dimension of the Hamiltonian is in the tens or hundreds of thousands, compression is not only convenient but absolutely necessary to run DMRG on today's supercomputers.

The rest of this Section is organized as follows. We first carefully define unit cell MPOs, the immediate present the compression algorithm, which is a straightforward generalization. We then turn to canonicalization, which requires a new concept of "topologically ordered" finite state machines, and present an efficient canonicalization algorithm suitable for 2D Hamiltonians.

Unit Cell MPOs

We now define MPOs with non-trivial unit cells (UCMPOs). The definitions are exact analogues of the ones in Section 4.5, with one more index. As before, suppose suppose we have an on-site operator space with an orthonormal basis $\{\widehat{\mathbb{1}}, \widehat{O}_2, \dots, \widehat{O}_d\}$ with inner product $\langle \widehat{O}_\alpha, \widehat{O}_\beta \rangle = \delta_{\alpha\beta}$. Any translation-invariant operator, with unit cell of size N, can be written as

$$\widehat{H} = \cdots \left[\widehat{W}^{(1)}\widehat{W}^{(2)}\cdots\widehat{W}^{(N)}\right]\left[\widehat{W}^{(1)}\widehat{W}^{(2)}\cdots\widehat{W}^{(N)}\right]\cdots \tag{4.100}$$

where each $\widehat{W}^{(n)} = \sum_{\alpha=1}^{d}[W^{(n)}]^{\alpha}\widehat{O}_{\alpha}$ is an operator-valued matrix of size $\chi^{(n)} \times \chi^{(n+1)}$. We denote the UCMPO as $\{\widehat{W}^{(n)}\}_{n=1}^{N}$ or simply $\{\widehat{W}^{(n)}\}$. We require each $\widehat{W}^{(n)}$ to have blocks of size $(1, \chi^{(n)} - 2, 1) \times (1, \chi^{(n+1)} - 2, 1)$[22] of the form

$$\widehat{W} = \left(\begin{array}{c|c|c} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}} & \widehat{d} \\ \hline & \widehat{A} & \widehat{\boldsymbol{b}} \\ \hline & & \widehat{\mathbb{1}} \end{array}\right) \tag{4.101}$$

This ensures that each operator is a sum of terms that are the identity far enough to the left or right — a physical and mathematical necessity for a local operator.

A UCMPO is said to be in **left-canonical form** if all but the last columns of each $\widehat{W}^{(n)}$ are mutually orthnormal:[23]

$$\sum_{a=1}^{\chi-1}\langle \widehat{W}_{ab}^{(n)}, \widehat{W}_{ac}^{(n)}\rangle = \delta_{bc}, \quad \forall n \in \mathbb{Z}/N\mathbb{Z}. \tag{4.102}$$

Similarly, an MPO is **right-canonical** if the rows of each tensor are orthogonal.

The representation (4.100) is not unique, which is a manifestation of gauge freedom. Two UCMPOs $\{\widehat{W}^{(n)}\}$ and $\{\widehat{W}^{(n)\prime}\}$ are **gauge equivalent** if there gauge exist matrices $\{G^{(n)}\}$ such that

$$G_{n-1}\widehat{W}^{(n)\prime} = \widehat{W}^{(n)}G_n, \quad n \in \mathbb{Z}/N\mathbb{Z}. \tag{4.103}$$

So long as $\widehat{H}$ is sufficiently local, one can show there exist a gauge where the $\widehat{W}^{(n)}$'s are left-canonical (and another gauge for right canonical).

The Unit Cell Compression Algorithm

The idea of the compression algorithm for unit cell MPOs $\{\widehat{W}^{(n)}\}$ has the same steps as the non-unit cell case:

1. Compute the right-canonical form $\{\widehat{W}_R^{(n)}\}$.

2. Find the gauge $\{G_n\}$ needed to transform to left-canonical form $\{\widehat{W}_L^{(n)}\}$.

3. Take the SVD decomposition of the gauge: $G_n = U_n S_n V_n^{\dagger}$ and absorb the unitaries into the $\widehat{W}$'s. This realizes the almost-Schmidt decomposition of Eq. (4.6)

[22] Note this is a slightly different convention from the above sections.

[23] More explicitly, if $\sum_{\alpha=1}^{d}\sum_{a,b=1}^{\chi-1}[W^{(n)}]_{ab}^{\alpha*}[W^{(n)}]_{ac}^{\alpha} = \delta_{bc}$ for each $\widehat{W}^{(n)}$.

4. Truncate the number of singular values in S_n from $\chi^{(n)}$ to $\chi^{(n)'}$ and reduce the bond dimension of the $\widehat{W}$'s, producing the compressed Hamiltonian.

This section will show that this same procedure is correct and fill in some of the details. It turns out that the most subtle part by far is canonicalization — the algorithm for putting an MPO into left/right canonical form. We therefore delay the discussion of canonicalization to Section 4.11 and for now simply assume it can be done.

We specialize to the case of $N = 2$ for concision, as larger unit cells are a direct generalization. Suppose that

$$R_{n-1}\widehat{W}_R^n = \widehat{W}^{(n)}R_n \tag{4.104}$$

is a gauge transformation so that the $\widehat{W}_R^{(n)}$'s are left-canonical. Then

$$\begin{aligned}\widehat{H} &= \cdots \widehat{W}^{(1)}\widehat{W}^{(2)}\widehat{W}^{(1)}\widehat{W}^{(2)}\cdots && (4.105)\\ &= \cdots \widehat{W}_L^{(1)}\widehat{W}_L^{(2)}\widehat{W}_L^{(1)}\widehat{W}_L^{(2)}\cdots && (4.106)\\ & && (4.107)\end{aligned}$$

by introducing R_2 at ∞ and sweeping to the right. We can then impose a further gauge transformation to make the first row of each $\widehat{W}_R^{(n)}$ simultaneously identity-free. This is done by

$$R'_n := \begin{pmatrix} 1 & \boldsymbol{t}_n & 0 \\ 0 & I & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{4.108}$$

where the t_n's are chosen such that $0 = \boldsymbol{c}_0^{(n)} + \boldsymbol{t}_n - \boldsymbol{t}_{n+1}A_0^{(n)}$ or

$$\begin{aligned}&\begin{pmatrix} \boldsymbol{c}_0^{(1)} & \cdots & \boldsymbol{c}_0^{(N)} \end{pmatrix} = \\ &\begin{pmatrix} \boldsymbol{t}_1 & \cdots & \boldsymbol{t}_N \end{pmatrix} \begin{pmatrix} I & & \cdots & -A_0^{(n)} \\ -A_0^{(1)} & I & & \\ & \ddots & \ddots & \\ & & -A_0^{(n-1)} & I \end{pmatrix}.\end{aligned} \tag{4.109}$$

To compute and apply this gauge iteratively, it is most efficient to iterate over the columns. For each column, the linear equation can be solved in $O(N)$ cost (as it is essentially upper-triangular), and then the gauge may be applied in $O(\chi)$ cost as it only affects a single row and column. The total cost is then $O(N\chi^2)$.

Imposing this gauge we may assume $\widehat{W}_R^{(n)}$ has no identity components in its first row. We may therefore find gauge transformations

$$C_{n-1}\widehat{W}_R^{(n)} = \widehat{W}_L^{(n)}C_n, \tag{4.110}$$

Algorithm 6 Unit Cell iMPO Compression

```
Require: $\{\widehat{W}^{(n)}\}$ is a first-order[2] unit cell iMPO.
procedure UnitCellCompress($\widehat{W}^{(n)}, \eta$)    ▷ Cutoff $\eta$
    $\widehat{W}_R^{(n)} \leftarrow \textsc{RightCan}[\widehat{W}^{(n)}]$
    $\widehat{W}_R^{(n)} \leftarrow R'_{n-1}\widehat{W}_R^{(n)}R_n'^{-1}$ so that $\widehat{\boldsymbol{c}}_0^{(n)} = 0$    ▷ Use $\boldsymbol{t_n}$ from Eq. (4.109).
    $\widehat{W}_L^{(n)}, C_n \leftarrow \textsc{LeftCan}[\widehat{W}_R^{(n)}]$
    $(U_n, S_n, V_n^\dagger) \leftarrow \mathrm{SVD}[C_n]$
    $\widehat{Q}^{(n)}, \widehat{P}^{(n)} \leftarrow U_{n-1}^\dagger\widehat{W}_L^{(n)}U_n\ ,\ V_{n-1}^\dagger\widehat{W}_RV$
    $\chi^{(n)} \leftarrow \max\{a \in [1, \chi^{(n)}] : S_{aa}^{(n)} > \eta\}$    ▷ Defines $\mathbb{P}_n$.
    $\widehat{Q}^{(n)}, S_n, \widehat{P}^{(n)} \leftarrow \mathbb{P}_{n-1}^\dagger\widehat{Q}_n\mathbb{P}_n, \mathbb{P}_n^\dagger S^{(n)}\mathbb{P}_n, \mathbb{P}_{n-1}^\dagger\widehat{P}^{(n)}\mathbb{P}_n$
    return $\widehat{Q}^{(n)}$
```

where $C_n = \mathrm{diag}(1\ \mathsf{C}_n\ 1)$ are each block diagonal. Putting in a C_2 matrix at $-\infty$ and sweeping it to the center, we arrive at a mixed canonical form

$$\widehat{H} = \cdots \widehat{W}_L^{(1)}\widehat{W}_L^{(2)}C_2\widehat{W}_R^{(1)}\widehat{W}_R^{(2)}\cdots \tag{4.111}$$

$$= \cdots \widehat{W}_L^{(2)}\widehat{W}_L^{(1)}C_1\widehat{W}_R^{(2)}\widehat{W}_R^{(1)}\cdots \tag{4.112}$$

As C_n are block diagonal, we can compute their SVD's

$$C_n = U_nS_nV_n^\dagger, \tag{4.113}$$

which will also be block-diagonal. Define

$$\widehat{Q}^{(n)} := U_{n-1}^\dagger\widehat{W}_L^{(n)}U_n, \tag{4.114}$$

$$\widehat{P}^{(n)} := V_{n-1}^\dagger\widehat{W}_R^{(n)}V_n, \tag{4.115}$$

for $n \in \mathbb{Z}/N\mathbb{Z}$. Then, since $U_nU_n^\dagger = I = V_nV_n^\dagger$, we have

$$\widehat{H} = \cdots \widehat{Q}^{(1)}\widehat{Q}^{(2)}S_2\widehat{P}^{(1)}\widehat{P}^{(2)}\cdots \tag{4.116}$$

$$= \cdots \widehat{Q}^{(2)}\widehat{Q}^{(1)}S_1\widehat{P}^{(2)}\widehat{P}^{(1)}\cdots \tag{4.117}$$

which is analgous to center-canonical form for MPS's. The center bond can be swept back and forth via the gauge relation

$$S_{n-1}\widehat{P}^{(n)} = \widehat{Q}^{(n)}S_n, \quad n \in \mathbb{Z}/N\mathbb{Z}. \tag{4.118}$$

To see how compression works, we adopt the technique of assuming that the operator can be represented *exactly* by an MPO of lower bond-dimension, i.e. that a number of the singular values vanish exactly. Finding the lower bond dimension MPO uses the same algorithm as compression when the small singular values are truncated, so this shows the correctness of the algorithm.

Thus we assume, temporarily, that only $\chi^{(n)'}$ of the $\chi^{(n)}$ singular values of S_n are non-zero. Hence there are projection operators $\mathbb{P}_n$ from bond dimension $\chi^{(n)}$ to bond dimension $\chi^{(n)'}$ with $\mathbb{P}_n\mathbb{P}_n^\dagger$ a projector and $\mathbb{P}_n^\dagger\mathbb{P}_n = I_{1+\chi^{(n)'}+1}$ and

$$S_n = S_n\mathbb{P}_n\mathbb{P}_n^\dagger = \mathbb{P}_n S_n'\mathbb{P}_n^\dagger = \mathbb{P}_n\mathbb{P}_n^\dagger S_N, \tag{4.119}$$

where S_n' is the projected diagonal matrix of non-zero singular values.

We can then introduce pairs of projectors on each bond:

$$\begin{aligned}\widehat{H} &= \cdots \widehat{Q}^{(1)}\widehat{Q}^{(2)}S_2\mathbb{P}_2\mathbb{P}_2^\dagger\widehat{P}^{(1)}\widehat{P}^{(2)}\cdots \\ &= \cdots \widehat{Q}^{(1)}S_1\widehat{P}^{(2)}\mathbb{P}_2\mathbb{P}_2^\dagger\widehat{P}^{(1)}\widehat{P}^{(2)}\cdots \\ &= \cdots \widehat{Q}^{(1)}S_1\mathbb{P}_1\mathbb{P}_1^\dagger\widehat{P}^{(2)}\mathbb{P}_2\mathbb{P}_2^\dagger\widehat{P}^{(1)}\widehat{P}^{(2)}\cdots \\ &= \cdots \mathbb{P}_2^\dagger\widehat{Q}^{(1)}\mathbb{P}_1\mathbb{P}_1^\dagger\widehat{P}^{(2)}\mathbb{P}_2S_2'\mathbb{P}_2^\dagger\widehat{P}^{(1)}\mathbb{P}_1\mathbb{P}_1^\dagger\widehat{P}^{(2)}\mathbb{P}_2\cdots \\ &= \cdots \mathbb{P}_1^\dagger\widehat{Q}^{(2)}\mathbb{P}_2\mathbb{P}_2^\dagger\widehat{P}^{(1)}\mathbb{P}_1S_1'\mathbb{P}_1^\dagger\widehat{P}^{(2)}\mathbb{P}_2\mathbb{P}_2^\dagger\widehat{P}^{(1)}\mathbb{P}_1\cdots\end{aligned} \tag{4.120}$$

It is now clear how to define a new representation for $\widehat{H}$ with a reduced bond dimension:

$$\widehat{P}^{(n)'} := \mathbb{P}_{n-1}\widehat{P}^{(n)}\mathbb{P}_n \tag{4.121}$$

$$\widehat{Q}^{(n)'} := \mathbb{P}_{n-1}\widehat{Q}^{(n)}\mathbb{P}_n \tag{4.122}$$

whereupon

$$\widehat{H} = \cdots \widehat{Q}^{(1)'}\widehat{Q}^{(2)'}S_2\widehat{P}^{(1)'}\widehat{P}^{(2)'}\cdots \tag{4.123}$$

is a representation of $\widehat{H}$ with lower bond dimension. Again, if we now relax the requirement that the truncated singular values were exactly zero, the strict equality of the new representation becomes approximate, but the algorithm is the same.

Canonicalization & Topological Sorting for Unit Cell MPOs

In this section we provide the "missing link" needed to complete the compression procedure: a canonicalization algorithm. Any unit cell MPO (UCMPO) can be put into left or right canonical form using QR iteration[2] with cost $O(N\chi^3)$. As many as 40 iterations can be necessary to reach high-precision, making this quite slow in practice. However, the MPOs for Hamiltonians in DMRG have a special property, a "topological ordering", which enables canonicalization to be performed with cost $O(N\chi^3)$ but *without iteration.* For large MPOs such as the one for BLG with $\chi \sim 100{,}000$, this is a crucial speed-up. We first define "topological order", then provide the canonicalization algorithm and a proof of its correctness and runtime. We conclude the section with a few remarks on practical implementation details.

An MPO can be thought of as a finite state machine (FSM) for placing on-site operators in a certain order [cite]. For MPOs with N tensors in a unit cell, the FSM gains an additional structure: the FSM has N parts, with the nodes of part n corresponding to the bond between $\widehat{W}^{(n-1)}$ and $\widehat{W}^{(n)}$ and edges between parts $n-1$ and n corresponding to tensor elements $\widehat{W}^{(n)}_{ab}$. See Fig 4.5 for an example.

When one writes down an (non-unit cell) MPO $\widehat{W}$ for a Hamiltonian "by hand", then the MPO generally has a special structure: $\widehat{W}$ is upper-triangular as a matrix. In [2], the upper triangular structure permitted a fast canonicalization algorithm. However, this does not immediately generalize to a unit cell MPO, for a simple reason: if a unit cell MPO $\{\widehat{W}^{(n)}\}$ has bond dimensions $\chi^{(1)}, \chi^{(2)} \dots \chi^{(n)}$, not all equal, then the matrices are rectangular and cannot all be upper trianglar. To find a good generalization of triangularity, we must look to the finite state machine.

A UCMPO $\{\widehat{W}^{(n)}\}_{n=1}^{N}$ is said to be **loop free** if its finite state machine contains no loops, aside from the initial and final nodes. One can check that for $N = 1$, then an MPO is loop free if, and only if, $\widehat{W}$ is upper-triangular. (We note that both the upper triangular and loop free conditions are gauge-*dependent*.) Furthermore, for any N, if each $\widehat{W}^{(n)}$ is square and upper-triangular, then the UCMPO is loop free. The converse is almost true as well; any loop free UCMPO is an upper-triangular MPO "in disguise". To see this, we need a definition, which will be at the heart of this section.

Definition 21. A **topological ordering** for a UCMPO $\{\widehat{W}^{(n)}\}_{n=1}^{N}$ is an ordering of the nodes of the FSM (excluding the initial and final nodes)

$$O = \{(a_1, n_1) \prec (a_2, n_2) \prec \cdots \prec (a_\chi, n_\chi)\} \tag{4.124}$$

such that

$$\widehat{W}^{(n)}_{ab} = 0 \text{ whenever } (a, n-1) \succeq (b, n), \tag{4.125}$$

where $n \in \mathbb{Z}/N\mathbb{Z}$ indexes the bond, $a \in \mathbb{N}$ indexes the node within the bond, and $\chi = \sum_n \chi^{(n)}$ is the total number of nodes.

If an MPO is loop free, then its finite state machine (excluding the initial and final nodes) is a directed acyclic graph, and thus contains at least one topological ordering. This is easily computed by Kahn's algorithm (a standard result in graph theory) with cost linear in the number of nodes plus edges in the FSM. With this, we can show that loop free UCMPOs are upper triangular ones "in disguise" and then use this ordering as the basis for an efficient canonicalization algorithm.

Lemma 22. *Suppse $\{\widehat{W}^{(n)}\}_{n=1}^{N}$ is a loop free MPO. Then, by inserting rows and columns of zeros and permuting the rows and columns of the matrices (which is a gauge transform), $\{\widehat{W}^{(n)}\}$ can be made upper triangular.*

Proof. Suppose the bond dimension of $\widehat{W}^{(n)}$ is $\chi^{(n-1)}$ on the left and $\chi^{(n)}$ on the right, with $\chi = \sum_n \chi^{(n)}$. Let O be a topological ordering for $\{\widehat{W}^{(n)}\}$ of the form 4.124. Define a gauge

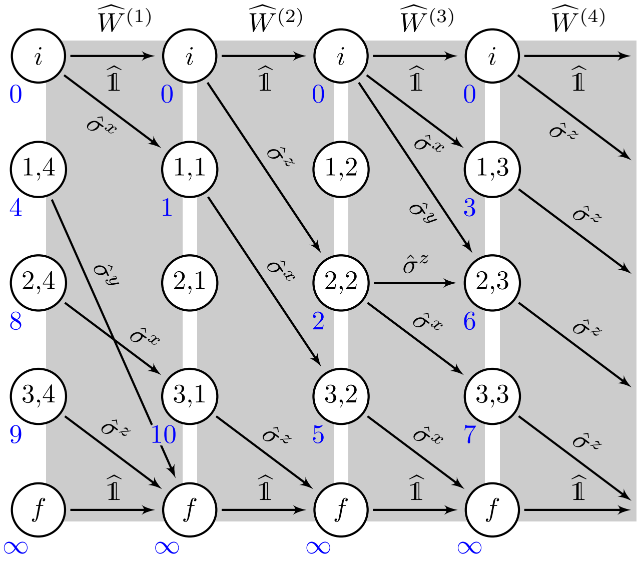


Figure 4.5: An example of the finite state machine for an MPO with unit cell size 4. One on-site operator is placed for each arrow, and the arrows wrap around from right to left. Each gray box represents the data stored in one tensor. The UCMPO has bond dimension $(\chi_4, \chi_1, \chi_2, \chi_3) = (5, 4, 4, 5)$ and is loop free. The blue numbers are a (non-unique) topological ordering for the nodes.

matrix $\mathbb{P}_n$ of dimension $\chi^{(n)} \times \chi$ with matrix elements

$$[\mathbb{P}_n]_{b,i} = \begin{cases} 1 & \text{if } (b, n) = O_i \\ 0 & \text{otherwise.} \end{cases} \tag{4.126}$$

This "blows up" $\widehat{W}^{(n)}$ on the right to bond dimension $\chi > \chi^{(n)}$ by inserting zeros, and puts the indices into topological order. One can check that $\mathbb{P}_n^\dagger \mathbb{P}_n = I_{\chi^{(n)}}$, so we may define $\widehat{W}^{(n)\prime} := \mathbb{P}_{n-1}^\dagger \widehat{W}^{(n)} \mathbb{P}_n$ of size $\chi \times \chi$ (which obeys $\mathbb{P}_{n-1} \widehat{W}^{(n)\prime} = \widehat{W}^{(n)} \mathbb{P}_n$, making it a gauge transform).

This is upper-triangular. Take $i \geq j$. Then either $\widehat{W}_{ij}^{(n)\prime} = W_{ab}^{(n)}$ for $O_i = (a, n-1)$ and $O_j = (b, n)$, or $\widehat{W}_{ij}^{(n)\prime} = 0$. But $O_i = (a, n-1) \succeq (b, n) = O_j$, so $W_{ab}^{(n)} = 0$ regardless. Therefore $\widehat{W}^{(n)\prime}$ is upper triangular. □

The algorithm for left canonicalization of loop free UCMPO is given in Alg. 7. We will now show that it is both correct and efficient.

Proposition 23. *Suppose $\{\widehat{W}^{(n)}\}_{n=1}^N$ is loop free. The output of Alg. 7 is a left canonical UCMPO.*

Algorithm 7 Unit Cell iMPO (Left) Canonicalization

```
Require: $\{\widehat{W}^{(n)}\}_{n=1}^{N}$ is a loop free UCMPO.
procedure UnitCellLeftCanonical($\widehat{W}^{(n)}, \eta$)
    $O = $ Kahn'sAlgorithm[FSM[$\{\widehat{W}^{(n)}\}$]]
    for $(b, n) \in O$ do
        $P \leftarrow \{a \,:\, O \ni (a, n) \prec (b, n)\}$
        $r_a \leftarrow \sum_c \langle \widehat{W}_{ca}, \widehat{W}_{cb}, , \rangle \quad \forall a \in P$
        $R \leftarrow I_{\chi_n}$, $R_{ab} \leftarrow r_a, \quad \forall a \in P$
        $\widehat{W}^{(n)} \leftarrow \widehat{W}^{(n)} R$, $\widehat{W}^{(n+1)} \leftarrow R^{-1} \widehat{W}^{(n+1)}$
        $R \leftarrow I_{\chi_n}$, $R_{bb} \leftarrow \left(\sum_c \langle \widehat{W}^{(n)}_{cb}, \widehat{W}^{(n)}_{cb} \rangle\right)^{-1/2}$
        $\widehat{W}^{(n)} \leftarrow \widehat{W}^{(n)} R$, $\widehat{W}^{(n+1)} \leftarrow R^{-1} \widehat{W}^{(n+1)}$
    return $\{\widehat{W}^{(n)}\}$
```

Proof. The main idea is to iterate over the columns of $\{\widehat{W}^{(n)}\}$ in topological order, orthogonalizing each column against all the previous ones as in Gram-Schmidt.

Let O be a topological order for the nodes as in (4.124). As each $\widehat{W}^{(n)}$ has the form (4.101), the first column of each is already orthonormal. We proceed by induction. Suppose that we have orthogonalized columns up to $(d, n) \in O$. Then for $(a, m), (b, m) \prec (d, n)$,

$$\sum_c \langle \widehat{W}^{(m)}_{ca}, \widehat{W}^{(m)}_{cb} \rangle = \delta_{ab}. \tag{4.127}$$

Let $P := \{(a, n) \,:\, (a, n) \prec (d, n)\}$ and for each $(a, n) \in P$, define the inner products with all previous columns as

$$r_a := \sum_c \langle \widehat{W}^{(m)}_{ca}, \widehat{W}^{(m)}_{cd} \rangle, \tag{4.128a}$$

$$R := I_{\chi^{(n)}} - \sum_{a \in L} r_a \boldsymbol{e}_{ad} \tag{4.128b}$$

where $\boldsymbol{e}_{ad}$ is the elementary matrix where entry ad is 1 and the rest are zero: $(\boldsymbol{e}_{ad})_{ij} = \delta_{ai}\delta_{dj}$. Here R is only non-identity in column d, and it performs elementary column operations when acting to the left and elementary row operations acting to the right. In particular, we have chosen it to perform one Gram-Schmidt step, orthogonalizing column d against previous columns of $\widehat{W}^{(n)}$. As R is invertible, $R^{-1} = I_{\chi^{(n)}} + \sum_{a \in L} r_a \boldsymbol{e}_{ad}$, so we can cast this Gram-Schmidt step as a gauge transform:

$$\widehat{W}^{(n)\prime} := \widehat{W}^{(n)} R, \tag{4.129a}$$

$$\widehat{W}^{(n+1)\prime} := R^{-1} \widehat{W}^{(n)}. \tag{4.129b}$$

We then have two things to show: (1) that this gauge transform really does orthogonalize column d of $\widehat{W}^{(n)}$ against previous columns and (2) that the gauge transform does not ruin the orthogonality condition of Eq. (4.127). Both are easy computations.

For (1), the effect of R acting on $\widehat{W}^{(n)}$ on the right is to add column c to column d with coefficient r_c:

$$\widehat{W}^{(n)} R = \widehat{W}^{(n)} - \sum_{c \in L} r_c \widehat{W}^{(n)} \boldsymbol{e}_{cd}.$$

However, $\widehat{W}^{(n)} \boldsymbol{e}_{cd} = \sum_b \left(\widehat{W}^{(n)}_{bc} \right) \boldsymbol{e}_{bd}$, so only column d is modified and

$$\sum_b \langle \widehat{W}^{(n)}_{ba}, \widehat{W}^{(n)} \boldsymbol{e}_{cd} \rangle = \delta_{ac}$$

by (4.127). Therefore, for any $(a,n) \prec (d,n)$,

$$\begin{aligned}\langle \widehat{W}^{(n)\prime}_{:,a}, \widehat{W}^{(n)\prime}_{:,d} \rangle &= \langle \widehat{W}^{(n)}_{:a}, \widehat{W}^{(n)}_{:d} \rangle - \sum_{c \in L} r_c \langle \widehat{W}^{(n)}_{:,a}, [\widehat{W}^{(n)} \boldsymbol{e}_{cd}]_{:,d} \rangle \\ &= r_a - \sum_{c \in L} r_c \delta_{ca} = 0.\end{aligned}$$

Therefore column d of $\widehat{W}^{(n)}$ is orthogonal to each previous column.

For (2), the effect of R^{-1} acting on the left of $\widehat{W}^{(n+1)}$ is to add row d to row c with coefficient r_c:

$$\widehat{W}^{(n+1)\prime} = \widehat{W}^{(n+1)} + \sum_{c \in L} r_c \sum_e \left(\widehat{W}^{(n+1)}_{de} \right) \boldsymbol{e}_{ce} =: \widehat{W} + \delta \widehat{W}.$$

Take $(a, n+1), (b, n+1) \prec (d, n)$. Then

$$[\delta \widehat{W}]_{:,a} = \sum_{c \in L} r_c \left(\widehat{W}^{(n+1)}_{da} \right) \boldsymbol{e}_{ca} = 0$$

since $\widehat{W}^{(n+1)}_{da} = 0$ as $(d,n) \succeq (a, n+1)$ by (4.125), and similarly $[\delta \widehat{W}]_{:,b} = 0$. Therefore,

$$\begin{aligned}\langle \widehat{W}^{(n+1)\prime}_{:,a}, \widehat{W}^{(n+1)\prime}_{:,b} \rangle &= \langle \widehat{W}^{(n)}_{:,a} + \delta \widehat{W}_{:,a}, \widehat{W}^{(n)}_{:,b} + \delta \widehat{W}_{:,b} \rangle \\ &= \langle \widehat{W}^{(n)}_{:,a}, \widehat{W}^{(n)}_{:,b} \rangle + 0 = \delta_{ab},\end{aligned}$$

so the induction hypothesis (4.127) holds for $\{\widehat{W}^{(n)\prime}\}$.

As the gauge transform R adds previous to column d of $\widehat{W}^{(n)}$ and adds row d of $\widehat{W}^{(n+1)}$ to previous rows, the transformed UCMPO is also loop free. Thus after this gauge transform, column d of $\widehat{W}^{(n)\prime}$ is orthogonal to all previous columns and all of the structure of the UCMPO is preserved.

A similar, simpler gauge transform

$$R = I_{\chi^{(n)}} + \Big(\sum_c \langle \widehat{W}^{(n)'}_{cd}, \widehat{W}^{(n)'}_{cd}\rangle\Big)^{-1/2} \boldsymbol{e}_{dd}$$

can then be used to normalize column d of $\widehat{W}^{(n)'}$. Repeating the previous arguments, one can show that this similarly does not disrupt the orthogonality of $\widehat{W}^{(n+1)'}$ or the loop free condition. Therefore we have made one more column orthonormal to the previous ones, completing the proof. □

Algorithm 7 is has cost $O(\sum_n \chi_n^3)$. This is somewhat surprising, as it seems we are doing $\chi = \sum_n \chi_n$ total gauge transformations, each of which is a matrix multiplication. However, the R matrices are particularly simple: they only differ from the identity in a single column. The transformations $\widehat{W}^{(n)'} = \widehat{W}^{(n)} R$ and $\widehat{W}^{(n+1)'} = R\widehat{W}^{(n+1)}$ to orthogonalize a column may be performed with rank-1 matrix updates whose cost is only $O(\chi_n^2)$. Similarly, the gauge transform to normalize a column, which simply scales a row or column, costs only $O(\chi_n)$. As we must iterate over every column of every tensor, the total cost in then $O(\sum_n \chi_n^3)$. However, each iteration requires only elementary matrix operations, for which highly optimized libraries are available, which gives a low constant factor on the algorithm. One can also employ these algorithms with charge-conserving MPOs, which vastly decreases the runtime in practice.

4.12 Properties of Compressed Hamiltonians

This section will show that compressed Hamiltonians are accurate approximations to the original Hamiltonian. This will give us guarantees that the (ground state) physics of interest in is unchanged by compression. In fact, just as with matrix product *states*, the error is controlled by the weight of the truncated singular values. We demonstrate three properties of the compressed Hamiltonian: (1) the change in the sup norm and ground state energy is small, (2) the fidelity of the compressed ground state versus the true ground state is high, and (3) ground state observables are accurate. We illustrate these properties with the example of the fraction quantum hall effect.

To frame the question, let us back up for a second. We envisage two common applications for our compression algorithm: compressing operators for use in infinite-temperature dynamics, and compressing Hamiltonians whose naive MPO bond dimensions are too large for DMRG. For the first, the figure of merit for the compression error is the change in the Frobenius norm of the operator — which we have already shown is small and proportional to the sum of the truncated singular values. For the second, however, the figure of merit is the change in the *sup* norm, an operator norm well-suited for ground state properties. If $\widehat{H}$ is an operator, then it's **sup norm** is given by

$$||\widehat{H}||^2 := \sup_{|\psi\rangle} \frac{\langle\psi|\widehat{H}\widehat{H}|\psi\rangle}{\langle\psi|\psi\rangle}. \tag{4.130}$$

For finite dimensional systems, the sup norm is the magnitude of the eigenvector furthest from zero.[24]

We will also use the non-scaled Frobenius norm, which we denote with a lowercase 'f':

$$||\widehat{H}||_f^2 := \mathrm{Tr}[\widehat{H}^\dagger \widehat{H}] = \mathrm{Tr}[I] \cdot ||\widehat{H}||_F^2. \tag{4.131}$$

Ground State Error Bound

Our task for this section is to show the change in the ground state energy is also small under compression[25]. As mentioned above, the class of first degree operators to which our algorithms apply is broader than the class of physically reasonable Hamiltonians. For instance, there are projectors which can be represented with small bond dimension MPOs which are first degree, but whose ground states energies are *not* extensive. If one feeds in an first degree operator which is a "non-Hamiltonian" with a superextensive ground state energy, then the error in the ground state energy may be very large. But, while this is mathematically true, such operators do not make sense as physical Hamiltonians. We therefore exclude them for consideration and restrict ourselves to operators which are sums of terms with support on at most k sites.[26] This allows us to give the following bound.

Proposition 24. *Suppose $\widehat{H}$ is an operator on N sites with on-site dimension d and at most k-body interactions. Suppose $\widehat{H}$ can be written in the form*

$$\widehat{H} = \widehat{H}_L \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \widehat{H}_R + \sum_{a,b=1}^{\chi} \widehat{h}_L^a M_{ab} \widehat{h}_R^b \tag{4.132}$$

where each $\widehat{h}_S^a$ is a unique tensor product of on-site operators (such as a Pauli string $XYXZ$ or $\hat{c}^\dagger \hat{c} \hat{c}^\dagger \hat{c}$ for fermions).

If we take the singular value decomposition $M = USV^\dagger$ and define $O_L := \widehat{h}_L U_L$, $O_R := V_R^\dagger \widehat{h}_R$, then for $\chi' < \chi$, we can define the compressed Hamiltonian

$$\widehat{H}' := \widehat{H}_L \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \widehat{H}_R + \sum_{a=1}^{\chi'} \widehat{O}_L^a s_a \widehat{O}_R^a \tag{4.133}$$

[24] As one can freely shift the zero point of energy of a Hamiltonian by taking $\widehat{H} \to \widehat{H} + \lambda I$, we sometimes assume without loss of generality that the sup norm gives the ground state eigenvalue. For local Hamiltonians, the ground state energy is extensive in the number of sites N: $E_0 \propto \epsilon_0 N$. We therefore often work with the sup norm *per site*, with the same notation.

[25] We note that small changes to the Hamiltonian can cause dramatic changes to the ground state wave*function*. For example, if the Hamiltonian is ϵ-close to a first order phase transition, like $H = \sum(1 - \frac{\epsilon}{2})\widehat{Z} + \widehat{X}\widehat{X}$, then an ϵ change (such as $\epsilon\widehat{Z}$) in the Hamiltonian will completely alter the ground state, even though the change in the ground state energy will still be ϵ-small. Away from phase boundaries, the ground state wavefunction and its expectation values should change continuously with the Hamiltonian.

[26] Similar bounds apply to broader classes of Hamiltonians, but require greater technical complexity.

where $\{s_1 \geq s_2 \geq \cdots \geq s_\chi\}$ are the singular values. Then the change in the ground state energy δE satisfies

$$\delta E \leq ||\widehat{H} - \widehat{H}'||_s \leq \sqrt{d^k \sum_{a=\chi'}^{\chi} s_a^2} \leq d^{\frac{k}{2}} ||\widehat{H} - \widehat{H'}||_F. \tag{4.134}$$

The idea of the proof is that each term in the Hamiltonian, being local, can only change the ground state energy slightly. The total change in the energy is then bounded above by the number of terms times the size of each term, which, we know to be small since they have small singular values. One could prove analagous bounds broader classes of Hamiltonians, such as long-range interactions, but this might require a significant amount of "technology" to specify the class of operators under discussion. Nevertheless, we expect the essential point to remain unchanged: for Hamiltonian-class operators, the change in the ground state energy is $O(1)$ times the weight of the truncated singular values.

Proof of Prop. 24. With our default inner product, if $\hat{O}$ is an operator supported on $S \subset \mathbb{Z}$, a set of size $|S| = k$, then

$$\langle \hat{O} | \hat{O} \rangle_F = \frac{\mathrm{Tr}[\hat{O}^\dagger \hat{O}]}{\mathrm{Tr}[I]} = \frac{\mathrm{Tr}[\widehat{O}_S^\dagger \widehat{O}_S]}{\mathrm{Tr}[\widehat{\mathbb{1}}^{\otimes k}]} = \frac{||\widehat{O}||_f^2}{d^k}. \tag{4.135}$$

Quite generally, $||\widehat{O}||_s \leq ||\widehat{O}||_f$. So $\langle \widehat{O} \big| \widehat{O} \rangle = 1$ implies

$$||\widehat{O}||_s^2 \leq ||\widehat{O}||_f^2 = d^k \tag{4.136}$$

for an operator supported on k sites.

We will assume that $E_0 = ||\widehat{H}||_2$ is the ground state of the Hamiltonian, though in principle it could also be the ground state of $-\widehat{H}$. As each term is unique, the operators on the right and left sides are both orthonormal:

$$\langle \widehat{H}_S^a | \widehat{H}_S^b \rangle = \delta^{ab}, S \in \{L, R\}. \tag{4.137}$$

As each term is supported on at most k sites, it follows from (4.136) that

$$||\widehat{H}_L^a \widehat{H}_R^b||_f^2 = d^k. \tag{4.138}$$

It is a standard fact about eigenvalues that if $\widehat{H} = \widehat{H}' + \delta \widehat{H}$ and $E_0' := ||\widehat{H}'||_s$, then

$$\delta E := |E_0' - E_0| \leq ||\delta \widehat{H}||_s. \tag{4.139}$$

By hypothesis

$$||\delta \widehat{H}||_s^2 = || \sum_{a=\chi'}^{\chi} \widehat{O}_L^a s_a \widehat{O}_R^a ||_s^2 \leq \sum_{a=\chi'}^{\chi} s_a^2 ||\widehat{O}_L^a \widehat{O}_R^a||_s^2 \tag{4.140}$$

We can now separately bound each term in the sum using locality:

$$\begin{aligned}
||\widehat{O}_L^c\widehat{O}_R^c||_s^2 &\leq ||\widehat{O}_L^c\widehat{O}_R^c||_f^2 \\
&= ||\sum_{a,b=1}^{\chi} U^{ac}V^{cb}\widehat{H}_L^a\widehat{H}_R^b||_f^2 \\
&= \sum_{abef} U^{ac}V^{cb}(U^{ec})^*(V^{cf})^* \operatorname{Tr}[\widehat{H}_L^{a\dagger}\widehat{H}_R^{b\dagger}\widehat{H}_L^e\widehat{H}_R^f] \\
&= \sum_{abef} U^{ac}V^{cb}(U^{ec})^*(V^{cf})^* d^k\delta^{ae}\delta^{bf} \\
&= d^k\sum_{a=1}^{\chi}|U^{ac}|^2\sum_{b=1}^{\chi}|V^{cb}|^2 \\
&= d^k,
\end{aligned}$$

where we have used (4.136) several times and the last two equalities follow from orthogonality of the H's and orthogonality of the columns and rows of U and V, respectively.

Combining our inequalities, we have

$$||\delta\widehat{H}||_s^2 \leq \sum_{a=\chi'}^{\chi} s_a^2||\widehat{O}_L^a\widehat{O}_R^a||_s^2 \leq d^k\sum_{a=\chi'}^{\chi} s_a^2. \tag{4.141}$$

□

In other words, the change in the ground state energy from truncation is proportional to the truncated singular values. It is crucial that this error does not involve N, the number of sites, so one can easily take the thermodynamic limit to find that, in an infinite system, the change in the ground state energy from truncating on every bond is also small. We also note that, although we have expressed this bound in terms of operators for convenience, this bound also applies to our MPO compression algorithm. Thus one may take a Hamiltonian, write it in a suboptimal MPO representation with a large bond dimension, then compress it to a small bond dimension and run DMRG or other algorithms to find its ground state energy with only a small error. This is particularly useful in the case of long-ranged interactions or two-dimensional problems, where the MPO dimensions for the naive MPOs are can be impractically large.

Compressed Hamiltonians are Accurate

The accuracy of a compressed Hamiltonian is controlled by the weight of the truncated singular values in almost-Schmidt form, Eq. (4.6). Conceptually, one should think of truncation as introducing a small perturbation to the Hamiltonian. If the truncated weight is small, then the perturbation is small, and its effects to the ground state energy, the fidelity, and other observables are also small.

We know from the previous section that the change in the ground state energy is bounded by weight of the truncated singular values

$$\epsilon^2 := \sum_{a=\chi+1}^{\chi'} s_a^2, \tag{4.142}$$

In practice, the singular values for an Hamiltonian fall off quite quickly — often exponentially, or a power law at worse. So retaining only a small number of singular values can produce a highly accurate approximation for the ground state energy.

It is natural to assume that if the ground state energy is accurate, then the other ground state properties — such as expectation values of observables and even the entire ground state wavefunction — are accurate as well. Unfortunately, there is a rare but severe failure of this assumption. Near a first-order phase transition, a tiny perturbation to a Hamiltonian can push the system across the phase transition, changing the properties of the ground state in a discontinuous manner (except for the energy). On the other hand, first-order phase transitions are usually measure zero in phase space, so this is almost always irrelevent. We can therefore understand the generic case by simply assuming we are far from a phase transition and the ground state changes continuously.

To do this, we work in first order perturbation theory. Suppose $\widehat{H}$ is a k-body Hamiltonian with a unique ground state with gap ΔE. Suppose we write $\widehat{H} = \widehat{H}' + \delta\widehat{H}$ with truncated weight ϵ^2 as in (4.142), and consider an observable of interest $\widehat{O}$. Then we can write the new ground state as

$$\begin{aligned} |E_0(\delta)'\rangle &= |E_0\rangle + |\delta E_0\rangle + O(\epsilon^2), \\ |\delta E_0\rangle &= \sum_{\lambda\neq 0} \frac{\langle E_\lambda|\delta\widehat{H}|E_0\rangle}{E_\lambda - E_0} \, |E_\lambda\rangle \, . \end{aligned}$$

Then

$$\begin{aligned} \Delta O &:= |\langle E_0(\delta)'|\widehat{O}|E_0(\delta)'\rangle - \langle E_0|\widehat{O}|E_0\rangle| \\ &= 2|\mathrm{Re}\,\langle E_0|\widehat{O}|\delta E_0\rangle| + O(\epsilon^2), \end{aligned}$$

so

$$\begin{aligned} |\langle E_0|\widehat{O}|\delta E_0\rangle| &\leq \left| \sum_{\lambda\neq 0} \frac{\langle E_0|\delta\widehat{H}|E_\lambda\rangle \, \langle E_\lambda|\widehat{O}|E_0\rangle}{E_\lambda - E_0} \right| \\ &\leq \frac{1}{\Delta E} \left| \sum_{\lambda\neq 0} \langle E_0|\widehat{O}|E_\lambda\rangle \, \langle E_\lambda|\delta\widehat{H}|E_0\rangle \right| \\ &\leq \frac{1}{\Delta E} \left| \langle E_0|\widehat{O}\,\delta\widehat{H}|E_0\rangle - \langle E_0|\widehat{O}|E_0\rangle \, \langle E_0|\delta\widehat{H}|E_0\rangle \right| \\ &\leq \frac{2}{\Delta E} ||\widehat{O}|| \cdot ||\delta\widehat{H}|| \end{aligned}$$

where we have used $\sum_{\lambda\neq 0} |E_\lambda\rangle\langle E_\lambda| = I - |E_0\rangle\langle E_0|$ and submultiplicativity of the norm. Using (4.134), the change in the expectation value is bounded by

$$\Delta O \leq \frac{4d^{\frac{k}{2}}}{\Delta E}||\widehat{O}||\,\epsilon. \tag{4.143}$$

We may therefore conclude that the error in expectation values is small, provided that the uncompressed Hamiltonian is sufficiently far from a first-order phase transition. A physical example is given in Fig 4.6 below. The condition of a gapped ground state may be relaxed, in which case the error will be controlled by the matrix elements of $\widehat{O}$ between the ground state and low-lying excited states.

Compressed Hamiltonians have High Fidelity

We have now seen that the ground state energy and expectation values of observables are accurately captured by the approximate, compressed Hamiltonian. In fact, the entire ground state wavefunction $|\psi'\rangle$ of $\widehat{H}'$ is very close to the original ground state wavefunction $|\psi\rangle$ of $\widehat{H}$. This allows us to use structural properties of $|\psi'\rangle$, such as its correlation length as a function of MPS bond dimension, as an accurate stand-in for the true ones and use them to e.g. diagnose the scaling properties of phase transitions.

To see this, we again work in perturbation theory, this time to second order. Let $\widehat{H} = \widehat{H}' + \delta\widehat{H}$ and take the same assumptions as above. Then we write

$$|E_0(\delta')\rangle = |E_0\rangle + |\delta E_0\rangle + |\delta^2 E_0\rangle + O(\epsilon^3).$$

so

$$\langle E_0|E_0(\delta)'\rangle = 1 + 0 - \frac{1}{2}\sum_{\lambda\neq 0}\frac{\langle E_0|\delta\widehat{H}|E_\lambda\rangle\langle E_\lambda|\delta\widehat{H}|E_0\rangle}{(E_\lambda - E_0)^2}.$$

By the same argument as above, therefore, the error in the ground state fidelity is bounded by

$$|1 - \langle\psi'|\psi\rangle| \leq \frac{d^k}{\Delta E^2}\,\epsilon^2. \tag{4.144}$$

To conclude, we provide a physical example to illustrate the bounds we have given here: a model for the Fractional Quantum Hall effect (FHQE). This is a strongly-interacting 2D system, which we place on an infite cylinder. The MPO for the Hamiltonian has 4-body interactions and a naive bond dimension of $10,000$. We compress the MPO with the compression algorithm given above by truncating all singular values with weight less than a cutoff η. For each value of η, we run iDMRG [117] to find the ground state and compute ground state energies, expectations, and fidelities. The bond dimension required to reach the DMRG error floor is only a few hundred. These results are shown in Fig. 4.6.

More concretely, we consider a well-studied[133] psuedopotential model on a thin cylinder with circumference L_y and infinite length. In Landau gauge, y-momentum around the

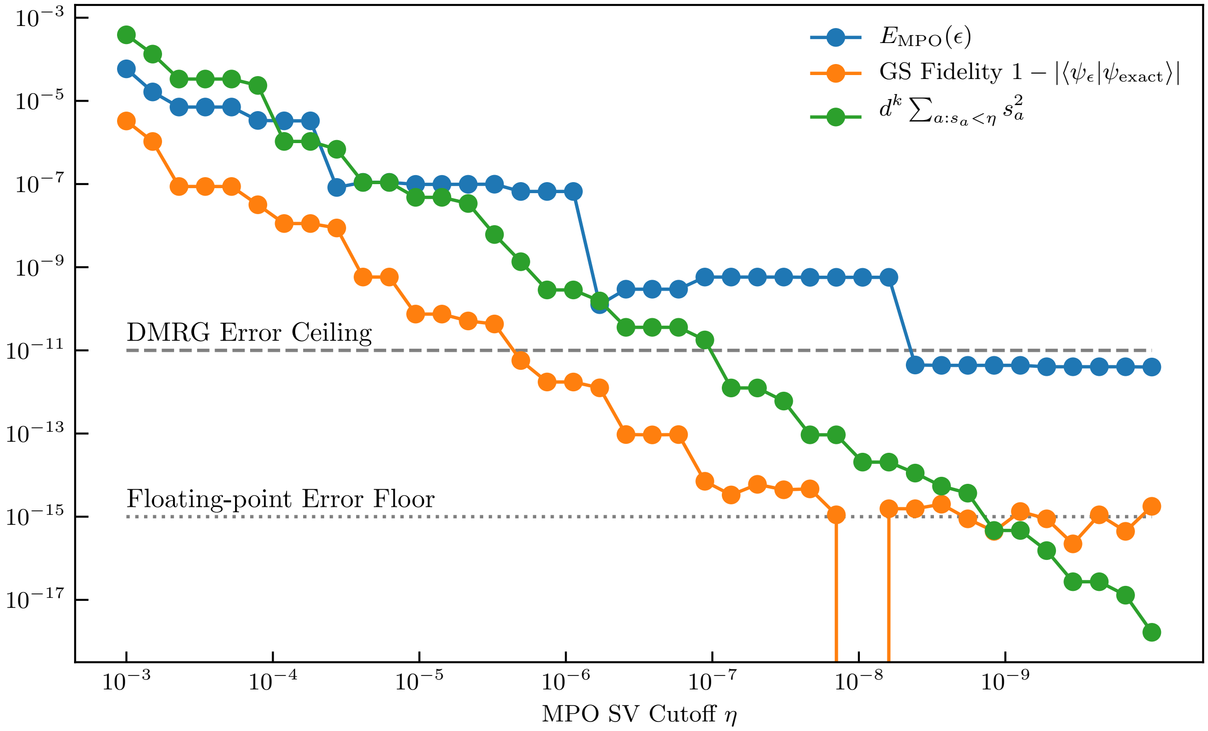


Figure 4.6: Ground state fidelity as a function of MPO singular value cutoff for the Fraction Quantum Hall effect Eq. (4.146).

cylinder is preserved. For each Fourier mode $k_n = 2\pi n/L$, the orbitals in the first Landau level are Gaussians centered at $k_n\ell_B^2$ where ℓ_B is the magnetic length. By mapping each orbital to a site, the entire cylinder model becomes a 1D fermion chain ($k_n \in \mathbb{Z}$ links the real-space position and y momentum, so moving along and around the cylinder requires only a single index). We then take the most general two-body interaction allowed by symmetry and consider a hard core Haldane psuedopotential

$$\widehat{H} = \sum_{n\in\mathbb{Z}} \sum_{k\geq|m|} V_{km} c^\dagger_{n+m} c^\dagger c_{n+m+k} c_n \tag{4.145}$$

$$V_{km} \propto (k^2 - m^2) \exp\left(-\frac{1}{2}(2\pi\ell_B/L)(k^2+m^2)\right). \tag{4.146}$$

In practice, one must cut off the interactions at some $|m| \leq R$. We take a unit cell of length two, and fill one orbital, appropriate for the $\nu = \frac{1}{2}$ state. It is known that ground state energy for this model is exactly zero, allowing us to compute the exact error in the ground state energy. Furthermore, the bond dimensions necessary to capture the ground state are relatively limited and thus easy to converge; we can be sure that any trends in the data are due to MPO compression rather than DMRG artifacts.

One can see that the truncated weight $\epsilon(\eta) := \sum_{a:s_a<\eta} s_a^2$ decreases quickly as a function of η. Meanwhile, the error in the ground state energy is proportional to $\epsilon(\eta)$ and decreases until reaching the DMRG error floor of 10^{-11}. Likewise, the error in ground state expectation values decreases roughly as $\epsilon(\eta)$, and the ground state fidelity decreases much fast, as $\epsilon(\eta)^2$. Therefore both the rigorous bound (4.134) and ones from perturbation theory (4.143) and (4.144) are borne out in practice.

4.13 Conclusions

In this chapter we have endevoured to promote matrix-product operators to "first-class citizens" amoung computational techniques. Our primary focus was the physically relevant case of local operators, operators that tend to the identity at spatial infinity. Locality of an operator imposes a constraint upon its matrix-product representations, namely a certain upper-triangular block structure. We then adapted the standard tools and techniques of matrix-product states to this framework. In particular, we generalized the notion of left and right canonical forms to the MPO case in a way that respects the local structure, and gave efficient algorithms for computing them. These lead naturally to a novel compression scheme for MPOs that also respects locality and is almost as optimal as SVD truncation is in the MPS case. We treated both the finite and infinite cases and proved the correctness of our techniques wherever possible. To showcase the utility of these new techniques, we included two brief applications: computing the Lanczos coefficients of operator dynamics, and compressing long-range (i)MPOs. In summary, this work enables all standard operations of matrix-product states to be performed on explicitly local matrix-product operators.

On a practical level, these results are applicable both to simulating quantum dynamics in 1d and solving strongly correlated systems in 2d. In 1d, this compression scheme should enable hydrodynamic coefficients, such as diffusion or conductivity, to be calculated using Krylov space techniques. The idea is that the Green's function $G(\omega, k)$ may be well-approximated by information contained in the Lanczos coefficients [1, 10]. Above we computed these for an example model at $k = 0$ (translation invariant sums), but one may work at arbitrary wavevector by slightly modifying the form of the MPO to

$$\widehat{W}(k) = \begin{pmatrix} e^{ik}\widehat{\mathbb{1}} & \widehat{\boldsymbol{c}} & \widehat{d} \\ 0 & \widehat{A} & \widehat{\boldsymbol{b}} \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}. \tag{4.147}$$

This application will be the focus of future work. In 2d, DMRG studies on infinite strips can be limited by the large bond dimension of the Hamiltonian operator. However, since these Hamiltonians are constructed "by-hand", it is reasonable to expect that, in many cases, they can be highly compressed. Moreover, as they have an upper-triangular form, this compression can be carried out quite efficiently. Alternatively, one could use an "over-compressed" Hamiltonian as a pre-conditioning step to find an approximate ground state before carrying out the full DMRG algorithm. In any event, the operator-centric tools developed in this work should bring immediate practical benefits to a variety of applications.

We wish to close with a few speculative remarks on our theoretical results. Operators are more than merely states in a doubled Hilbert space in at least two ways: (I) they have an algebraic structure and can thus be multiplied, and (II) they can be local. One perspective on this work is that local operators, as we have defined them, are the analogue of area law states, with a bounded amount of information per site. The standard notions of quantum information theory, especially the entanglement spectrum, struggle to capture the non-trivial local structure of operators — which is what led us to define the "almost-Schmidt decomposition". It is unclear how general this notion is. For example, how do we treat "second degree" and "multi-local" operators that arise naturally as products such as $\widehat{H}\widehat{H}$ (used in computing energy fluctuations in DMRG [123])? Can it be extended beyond 1d?

Curiously, the algebraic nature of operators is almost completely absent from this work. After all, locality is a by-product of the operator algebra, namely the condition that spatially-separated operators tend to commute. It is natural to speculate that a deeper "quantum information theory of operators" would be intimately connected to the operator algebra structure and yield greater benefits for computation.

Appendices

4.A Proofs for Local MPOs

This appendix proves statements about local MPOs from Section 4.5 of the main text. Our main goal is the proof of the forms of the dominant Jordan blocks, Prop 16, but we begin with a series of technical Lemmas.

Lemma 25. *Suppose $\widehat{W}$ and $\widehat{W}'$ are related by a gauge transform $L\widehat{W} = \widehat{W}'L$, and $\widehat{W}$ is first degree. Then $\widehat{W}'$ is also first degree.*

Proof. The block triangular form (4.14) of the gauge matrix L implies the sub-matrices $\widehat{A}$ and $\widehat{A}'$ are related by $\mathsf{L}\widehat{A} = \widehat{A}'\mathsf{L}$. Then, by the definition of the transfer matrix, we have

$$\begin{aligned}[\mathsf{L}^\dagger X \mathsf{L}]T_A &= \sum_\alpha A_\alpha^\dagger \mathsf{L}^\dagger X \mathsf{L} A_\alpha \\ &= \sum_\alpha \mathsf{L}^\dagger (A'_\alpha)^\dagger X A'_\alpha \mathsf{L} = \mathsf{L}^\dagger (X T_{A'}) \mathsf{L}\,. \end{aligned} \tag{4.148}$$

Now, suppose $\widehat{W}'$ is not first degree, then there is X such that $XT_{A'} = \lambda X$ with $|\lambda| \geq 1$. By (4.148), $Y := \mathsf{L}^\dagger X \mathsf{L}$ is an eigenvector of T_A with the same λ, which contradicts the first degree property of $\widehat{W}$. □

Lemma 26. *Suppose* spec(T_A) *is strictly inside the unit disk. Then so is* spec(A_0).

Proof. Suppose not. Then there is a (generalized) eigenvalue $\sigma \in \text{spec}(A_0)$ with $|\sigma| \geq 1$. This eigenvalue must be in some Jordan block

$$J = \begin{pmatrix} \sigma & 1 & & \\ & \ddots & \ddots & \\ & & \sigma & 1 \\ & & & \sigma \end{pmatrix} \tag{4.149}$$

with some (generalized) eigenvector $A_0\boldsymbol{v} = \sigma\boldsymbol{v}$. Then $\boldsymbol{w} := \boldsymbol{v}^\dagger \otimes \boldsymbol{v}$ is an eigenvector $T_{A_0}\boldsymbol{w} = |\sigma|^2\boldsymbol{w}$. So

$$\langle \boldsymbol{w}, T_{A^0}^N \boldsymbol{w}\rangle = |\sigma|^{2N} \geq 1\,,\ \forall N\,. \tag{4.150}$$

For each component A_α, $0 \leq \alpha < d^2$, of $\widehat{A}$, let $T_\alpha[X] := (A_\alpha)^\dagger\,[X]A_\alpha$. Each of these is a positive map and $T_A = \sum_\alpha T_\alpha$, so

$$T_A^N = T_0^N + \sum_{\substack{\alpha_1,\ldots,\alpha_N \\ \exists \alpha_i \neq 0}} T_{\alpha_1} \cdots T_{\alpha_N}, \tag{4.151}$$

Since the composition of positive maps is positive, $\langle \boldsymbol{w}, T_{\alpha_1} \cdots T_{\alpha_N} \boldsymbol{w} \rangle \geq 0$. So (4.151) implies $\langle \boldsymbol{w}, T_A^N \boldsymbol{w} \rangle \geq |\sigma|^{2N} \to \infty$. But all the eigenvalues of T_A are less than 1, so $\langle \boldsymbol{w}, T_A^N \boldsymbol{w} \rangle \to 0$, a contradiction. □

Lemma 27. *Let*

$$T = \begin{pmatrix} A & B \\ 0 & C \end{pmatrix} \tag{4.152}$$

be a block upper-triangular matrix such that A and C are square matrices. Let $\lambda \in \operatorname{spec}(A) \setminus \operatorname{spec}(C)$ and $(x\ y)T = \lambda(x\ y)$ be a left eigenvector. Then $x \neq 0$ and satisfies $xA = \lambda x$, so x is a left eigenvector of A.

Proof. $(x\ y)T = \lambda(x\ y)$ means $xA = \lambda x$ and $xB + y = \lambda y$. Now suppose $x = 0$. Then $y \neq 0$, and $yC = \lambda y$, so $\lambda \in \operatorname{spec}(C)$, a contradiction. □

Lemma 28. *Suppose $\widehat{W}$ is an first degree iMPO. Then there exists a gauge transform* [27]

$$\widehat{W}' = L\widehat{W}L^{-1} \textit{ where } L = \begin{pmatrix} 1 & \boldsymbol{t} & 0 \\ 0 & \mathrm{Id} & \boldsymbol{s} \\ 0 & 0 & 1 \end{pmatrix} \tag{4.153}$$

such that

$$W_0' = \langle \widehat{\mathbb{1}}, \widehat{W}' \rangle = \begin{pmatrix} 1 & 0 & d_0' \\ 0 & A_0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.154}$$

In fact, $\widehat{A}' = \widehat{A}$ is unchanged.

Proof. A direct computation shows

$$\boldsymbol{s} := (A_0 - \mathrm{Id})^{-1}\boldsymbol{b}_0\,,\ \boldsymbol{t} := \boldsymbol{c}_0(A_0 - \mathrm{Id})^{-1} \tag{4.155}$$

give the desired gauge transform. The inverse $(A_0 - \mathrm{Id})^{-1}$ exists by Lemma 26. □

We now have all the tools needed to unravel the Jordan block structure of MPOs. We prove Prop 17 in the special case where the operator has subextensive trace, and subsequently sketch the more general case.

[27] When using this to compute the norm via Eq. (4.58), one should only compute $\boldsymbol{s}$ and set $\boldsymbol{t} \equiv 0$ so that left-canonical form is preserved. When compressing iMPOs, one should instead set $\boldsymbol{s} = 0$ and use only $\boldsymbol{t}$.

Proof of Prop. 17. The idea of the proof is to explicitly find the dominant Jordan block (i.e. the Jordan block that gives the leading contribution to the norm) using the block structure of T_W. Unfortunately, just as in (4.41), there are two other “spurious” eigenvectors whose eigenvalue is also 1. Just as the dominant Jordan block is responsible for the extensive norm, they give rise to the extensive part of the trace. For a traceless operator, they form an invariant subspace that does not contribute to the extensive norm — hence the name “spurious”.

We first impose the condition of tracelessness. Without loss of generality, we work in the gauge of Lemma 28, and note that $\widehat{A}$ is unchanged so the first degree property is maintained. On a finite system of N sites, the trace is given by,

$$\begin{aligned}\mathrm{Tr}[\widehat{H}_N] &= \boldsymbol{\ell} W_0^N \boldsymbol{r} = (1\ \boldsymbol{\ell}'\ \ell_{\chi+1})\, W_0^N\, (r_0\ \boldsymbol{r}'\ 1)^T \\ &= r_0 + \ell_{\chi+1} + N d_0 + \boldsymbol{\ell}' A_0^N \boldsymbol{r}' \\ &= N d_0 + O(1)\,,\ N \to \infty\,, \end{aligned} \tag{4.156}$$

where we used the standard boundary conditions (4.11) and used Lemma 26 for the last asymptotic. Therefore

$$\lim_{N\to\infty} \frac{1}{N} \mathrm{Tr}[\widehat{H}_N] = 0 \Longleftrightarrow d_0 = 0 \text{ in gauge (4.153).} \tag{4.157}$$

We now exhibit all the generalized eigenvectors with eigenvalue 1. For concision, we rewrite $\widehat{W}$ as

$$\widehat{W} = \begin{pmatrix} \widehat{V} & \widehat{\boldsymbol{f}} \\ 0 & \widehat{\mathbb{1}} \end{pmatrix}\,,\ \widehat{\boldsymbol{f}} := \begin{pmatrix} \widehat{d} \\ \widehat{\boldsymbol{b}} \end{pmatrix}\,, \tag{4.158}$$

with block sizes $1+\chi$ and 1. Similarly to Eq. (4.47), we have

$$T_W = \left(\begin{array}{c|cc|cc|c} T_V & 0 & \overline{U} & 0 & U & F \\ \hline 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \overline{A}_0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \tag{4.159}$$

for some U, where the block sizes are $(1+\chi)^2, 1+\chi, 1+\chi, 1$, and

$$F := \sum_\alpha \overline{\boldsymbol{f}}_\alpha \otimes \boldsymbol{f}_\alpha\,. \tag{4.160}$$

We observe that T_W is the sum of “reduced” and “spurious” parts

$$T_W = \begin{pmatrix} T_V & \overline{U} & U & F \\ 0 & A_0 & 0 & 0 \\ 0 & 0 & \overline{A}_0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \oplus \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} =: T_{\mathrm{red}} \oplus T_{\mathrm{sp}} \tag{4.161}$$

The spurious block T_{sp} has eigenvectors E and E^T where $E_{ab} = \delta_{a0}\delta_{b,\chi+1}$ and, in particular, $E_{00} = 0$.

The dominant Jordan block comes from T_{red}. Consider the truncated operator

$$T_{\text{red}}^{\text{truncated}} = \begin{pmatrix} T_V & \overline{U} & U & 0 \\ 0 & A_0 & 0 & 0 \\ 0 & 0 & \overline{A}_0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} . \tag{4.162}$$

By Proposition 16 and Lemma 26, it has a unique eigenvalue 1 (the rest have $|\lambda| < 1$). By Lemma 27, the corresponding left eigenvector of T_{red} is (after rescaling)

$$Z' = \begin{pmatrix} X & \boldsymbol{z} \\ \overline{\boldsymbol{z}} & 0 \end{pmatrix} . \tag{4.163}$$

for some $\boldsymbol{z}$ and where X is the unique largest eigenvector of T_V from Eq. (4.46). Then we have

$$\begin{pmatrix} Z'T_W & ZT_W \end{pmatrix} = \begin{pmatrix} Z' & Z \end{pmatrix} \begin{pmatrix} 1 & \rho \\ 0 & 1 \end{pmatrix} , \, Z = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \tag{4.164}$$

where $\rho := XF = \sum_{a,b=1}^{\chi} X_{ab} \langle \widehat{\boldsymbol{f}}_a, \widehat{\boldsymbol{f}}_b \rangle$. (In practice, one should compute ρ using Eq. (4.58) which makes use of canonical form.) All the other eigenvalues of T_{red}, and indeed all other eigenvectors of T_W are those of A_0 and $\overline{A}_0$, and satisfy $|\lambda| < 1$ by first degreeness. We have thus found the dominant Jordan block of T_W, as well as the "spurious" eigenvectors.

We are now ready to compute the norm $||H||_N^2$ using the transfer matrix formula (4.39). We expand $\boldsymbol{\ell\ell}$ in the left generalized eigenbasis of T_W:

$$\boldsymbol{\ell\ell} = \underbrace{(aZ' + bZ)}_{\lambda=1 \text{ Jordan block}} + \underbrace{(cE + \overline{c}E^T)}_{\lambda=1 \text{ 'spurious'}} + \underbrace{S}_{|\lambda|<1} \tag{4.165}$$

where S is a linear combination of generalized left eigenvectors with eigenvalues $|\lambda| < 1$. It follows that

$$(\boldsymbol{\ell\ell}) T_W^N (\boldsymbol{rr}) = Na\rho(Z\boldsymbol{rr}) + \mathrm{O}(1) = Na\rho + \mathrm{O}(1) \tag{4.166}$$

as $N \to \infty$, since $\boldsymbol{r}_{\chi+1} = 1$ by the regular form. It remains to determine the coefficient a. For this we look at the 00-component of (4.165). First, $ST_W^N \longrightarrow 0$ by the definition of S. Meanwhile, (4.47) and (4.159) imply $(ST_W)_{00} = S_{00}$. Therefore, $S_{00} = 0$. For the other terms of the RHS, we have $Z'_{00} = 1$ by (4.163) and (4.46), $Z_{00} = 0$ by (4.164), and $E_{00} = 0$. On the LHS, the regular form (4.11) requires $(\boldsymbol{\ell\ell})_{00} = 1$. Therefore we have $a = 1$ and

$$||\widehat{H}_N||_F^2 = \boldsymbol{\ell\ell} T_W^N \boldsymbol{rr} = N\rho + \mathrm{O}(1) \, . \tag{4.167}$$

□

As noted above, the condition that the trace is sub-extensive can be lifted.

Suppose $\widehat{W}$ is an first degree iMPO for $\widehat{H}$. Then the transfer matrix T_W has maximum eigenvalue unity with a generalized eigenspace V_1 of dimension four. This may be Jordan decomposed as follows:

Case 1. $V_1 = J_3 \oplus J_1$ if $\mathrm{Tr}[\widehat{H}_N] = O(N)$, i.e. the trace is extensive

Case 2. $V_1 = J_2 \oplus J_1 \oplus J_1$ if $\mathrm{Tr}[\widehat{H}_N] = o(N)$, i.e. the trace is subextensive.

Without loss of generality, we adopt the gauge from Lemma 28. Define block matrices of size $\chi + 1 \times \chi + 1$

$$Z_i = \begin{pmatrix} X & \boldsymbol{z} \\ \boldsymbol{z}^\dagger & 0 \end{pmatrix}, \quad Z_t = \begin{pmatrix} 0 & \boldsymbol{t} \\ 0 & 0 \end{pmatrix}, \quad Z_f = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \tag{4.168}$$

and $Z_{t'} = Z_t^\dagger$ where X is the dominant eigenvalue of T_A, $\boldsymbol{z}$ is the same as above, and $\boldsymbol{t} = (1, 0, \dots, 0)$ is a vector of length χ. These span the dominant generalized eigenspace:

$$\begin{pmatrix} Z_i \\ Z_t \\ Z_{t'} \\ Z_n \end{pmatrix} T_W = \begin{pmatrix} Z_i \\ Z_t \\ Z_{t'} \\ Z_n \end{pmatrix} \underbrace{\begin{pmatrix} 1 & d & d & \rho \\ 0 & 1 & 0 & d \\ 0 & 0 & 1 & d \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{M^\dagger} \tag{4.169}$$

where d is the extensive part of the trace: $\mathrm{Tr}[\widehat{H}_N] = Nd$ and the dagger is because T_W acts on the right. The Jordan decomposition $M = SJS^{-1}$ is then

Case 1.

$$J = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 0 & 0 & 1 \\ -\frac{1}{2} & 0 & d & 0 \\ \frac{1}{2} & 0 & d & 0 \\ -\frac{\rho}{2d} & 2d^2 & \rho & 0 \end{pmatrix} \tag{4.170}$$

Case 2.

$$J = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & \frac{1}{\rho} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \tag{4.171}$$

Case 2 is, of course, the same as the above proof, where Z_t and $Z_{t'}$ span the 'spurious' dimensions and the Jordan block of size 2 is responsible for the extensive norm. In Case 1, however, those two dimensions are now mixed together. One can compute

$$(\boldsymbol{\ell\ell}) T_W^N (\boldsymbol{rr}) = N^2 d^2 + N(\rho - d) + O(1). \tag{4.172}$$

The Frobenius norm is then no longer extensive as is has been “polluted” with the trace. Nevertheless, the largest eigenvalue is still unity and the matrix Z_i overlaps with the dominant Jordan block.

The proof for these statements is directly analogous to the above Proof with the single modification of Eq. (4.159) to

$$T_W = \left(\begin{array}{c|cc|cc|c} T_V & d\boldsymbol{t} & \overline{U} & d\boldsymbol{t} & U & F \\ \hline 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \overline{A}_0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right). \tag{4.173}$$

4.B Proofs for Canonical forms

This appendix provides a sufficient condition for the convergence of the QR iteration in Algorithm 3 for first degree MPOs, and proves the existence of left canonical forms.

It is clear from the definition of canonical forms that only the upper-left sub-matrix $\widehat{V}$ of an iMPO $\widehat{W}$ will be actively involved. Indeed, any gauge transform of the sub-matrix $L\widehat{V} = \widehat{V}'L$ can be easily promoted the iMPO level:

$$\underbrace{\begin{pmatrix} L & \\ & 1 \end{pmatrix}}_{L_W} \underbrace{\begin{pmatrix} \widehat{V} & \widehat{\boldsymbol{f}} \\ & \widehat{\mathbb{1}} \end{pmatrix}}_{\widehat{W}} = \underbrace{\begin{pmatrix} \widehat{V}' & L\widehat{\boldsymbol{f}} \\ & \widehat{\mathbb{1}} \end{pmatrix}}_{\widehat{W}'} \underbrace{\begin{pmatrix} L & \\ & 1 \end{pmatrix}}_{L_W}. \tag{4.174}$$

Hence we focus on $\widehat{V}$ and its gauge transforms. [28] From this point of view, the QR iteration Algorithm 3 is defined by the following recursion:

$$R_0 := \mathrm{Id}_{[0,\chi]}, \tag{4.175a}$$

$$\widehat{V}_{n-1} := \widehat{Q}_n R_n, \quad \forall n \geq 1 \tag{4.175b}$$

$$\widehat{V}_n := R_n \widehat{Q}_n, \tag{4.175c}$$

$$L_n := R_n \dots R_1, \tag{4.175d}$$

where (4.175b) is a (normal) QR decomposition as defined in (4.21).

We also point out a simple fact: two gauge transforms $L\widehat{W} = \widehat{W}'L$ and $L'\widehat{W}' = \widehat{W}''L'$ can be composed to obtain a new one: $L'L\widehat{W} = \widehat{W}''L'L$.

Lemma 29. *QR iteration produces a sequence $\{\widehat{V}_n\}$ that are each related to $\widehat{V}$ be a gauge transform:*

$$L_n\widehat{V} = \widehat{V}_n L_n\,. \tag{4.176}$$

[28] Accordingly, the notation in this appendix will differ form the main text in that gauge matrices acting on $\widehat{V}$ will not have an overline.

Proof. Eq. (4.175b) implies the gauge transform $R_m \widehat{V}_{m-1} = \widehat{V}_m R_m$ for any $m > 0$. Then (4.176) follows from Eq. (4.175d) by composing the gauge transforms. □

Algorithm 3 enjoys also a close relation to the 'small' transfer matrix:

Lemma 30. *For any $n \geq 0$,*

$$\mathrm{Id}_{[0,\chi]}(T_V)^n = L_n^\dagger L_n \,, \tag{4.177}$$

where $\widehat{V}$ has bond dimension χ, that is, $(1+\chi)$ rows and columns.

Proof. We again proceed by induction on n. The base case $n = 0$ is trivial. For $n > 0$, we have

$$\begin{aligned}
\mathrm{Id}_{[0,\chi]}(T_V)^n &= (L_{n-1}^\dagger L_{n-1}) T_V \\
&= \sum_\alpha V_\alpha^\dagger L_{n-1}^\dagger L_{n-1} V_\alpha \\
&= \sum_\alpha L_{n-1}^\dagger V_{n-1,\alpha}^\dagger V_{n-1,\alpha} L_{n-1} \\
&= \sum_\alpha L_{n-1}^\dagger R_n^\dagger Q_{n,\alpha}^\dagger Q_{n,\alpha} R_n L_{n-1} \\
&= \sum_\alpha L_n^\dagger Q_{n,\alpha}^\dagger Q_{n,\alpha} L_n \\
&= L_n^\dagger L_n
\end{aligned}$$

where we used the induction hypothesis, (4.38), (4.176), (4.175b), (4.175d), and the definition of QR, respectively. □

We now address the sufficient condition for the convergence of QR iteration. First we must remove some arbitrariness in QR decomposition. For instance, $\widehat{W} = \widehat{Q}R = (-\widehat{Q})(-R)$ are both valid, but such freedom can introduce unhelpful oscillations in n preventing convergence. To this end, we require our QR sub-routine to be positive rank-revealing, in the following sense:

Definition 31. Suppose $\widehat{V}$ have $1 + \chi$ columns and column rank $1 + \chi'$, where $0 \leq \chi' \leq \chi$. The QR decomposition routine $\widehat{Q}, R \leftarrow QR[\widehat{V}]$ is called positive rank-revealing when the following are guaranteed:

(I). **Rank-revealing**: $\widehat{Q}$ has $\chi' + 1$ columns and R has $\chi' + 1$ rows.

(II). **Positive**: if $\chi' = \chi$ (full column rank), R has positive diagonal elements:

$$R_{aa} > 0 \,,\; a = 0, \ldots, \chi \,. \tag{4.178}$$

These requirements can be fulfilled, for example, by the Gram-Schmidt procedure applied to the columns of $\widehat{V}$.

Proposition 32. *Let $\widehat{W}$ is a first degree iMPO of bond dimension χ, and let the sequence $(\widehat{W}_n, L_n, R_n)_{n\geq 1}$ be generated by positive, rank-revealing QR starting from $\widehat{W}$. Suppose further that the leading eigenvector X of T_V is an invertible $(1+\chi)\times(1+\chi)$ matrix. Then the iteration converges and brings $\widehat{W}$ to left canonical form.*

The proof will follow a after a few lemmas.

Lemma 33. *Let $m > 0$. Let $\mathbb{T}_m$ be the space of $m \times m$ upper-triangular matrices with positive diagonal elements and let $\mathcal{P}_m$ be the space of $m \times m$ positive definite matrices. Then*

$$\mathbb{T}_m \ni L \mapsto L^\dagger L \in \mathcal{P}_m \tag{4.179}$$

is a homeomorphism.

The continuous inverse is constructed explicitly in standard linear algebra textbooks.

In general, the QR iteration with rank revealing will produce a sequence of $\widehat{W}_n$'s with reducing bond dimensions, $\chi_0 \geq \chi_1 \geq \ldots$. However, when X is non-singular, no strict bond dimension reduction can occur:

Lemma 34. *Under the same hypotheses of Prop. 32, all the $\widehat{W}_n$'s have the same bond dimension as $\widehat{W}$.*

Proof. By the gauge transform (4.176) and Lemma 25, $\widehat{W}_n$ is also first degree. So we can apply Prop. 16 and let X_n be the dominant eigenvector of T_{V_n}: $X_n T_{V_n} = X_n$. Then the gauge transform (4.176) implies

$$[L_n^\dagger X_n L_n] T_V = L_n^\dagger X_n L_n \,, \tag{4.180}$$

similarly to (4.148). This means that $L_n^\dagger X_n L_n = X$ by Prop. 16 (the constant is fixed by the 00-th element). For X to be non-singular, L_n must be a square matrix, so the bond dimension does not change. □

We remark on a useful consequence of Lemma 34: since no rank reduction will happen, we only need the QR to be positive, not necessarily rank-revealing. This can be fulfilled by numerically stable implementations of QR based on Givens rotations or Householder reflections.

Proof of Prop. 32. By the definition of positive rank-revealing QR, and Lemma 34, for any $n \geq 1$, $R_n \in \mathbb{T}_{1+\chi}$, and thus $L_n \in \mathbb{T}_{1+\chi}$. Now, Lemma 4.B and Prop. 16 imply that

$$L_n^\dagger L_n = \mathrm{Id}_{[0,\chi]}(T_V)^n \xrightarrow{n\to\infty} X = \begin{pmatrix} 1 & \boldsymbol{y} \\ \boldsymbol{y}^\dagger & Y \end{pmatrix} . \tag{4.181}$$

Note that $(L_n)_{00} = 1$ for all n. Eq. (4.181) implies that X is positive semi-definite. Since we assume X is non-singular, X is positive definite. Then, Lemma 33 implies that $L_n \to L$ for some invertible L, and the QR iteration converges as follows:

$$\begin{aligned}\widehat{V}_n &= L_n \widehat{V} L_n^{-1} \to L\widehat{V}L^{-1} := \widehat{V}_L \\ R_n &= L_{n+1}^{-1} L_n \to \mathrm{Id}_{[0,\chi+1]} \\ \widehat{Q}_n &= \widehat{V}_{n-1} R_n^{-1} \to \widehat{V}_L \,,\end{aligned}$$

so that $\widehat{V}_L$ is a left canonical MPS. Promoting to the iMPO level using (4.174) completes the proof. □

Prop. 32 establishes the existence of left canonical for all "generic" first degree iMPOs, in the sense that X is non-singular. We now treat the singular cases:

Proposition 35. *Let $\widehat{W}$ be a first degree iMPO and such that the leading eigenvector X of T_V is positive semi-definite of rank $1+\chi' \le 1+\chi$. Then there is gauge transform $L\widehat{W} = \widehat{W}'L$ is such that $\widehat{W}'$ has bond dimension χ' and such that X' is positive definite.*

Proof. We will construct the gauge transform by composing two gauge transforms, and still work on the level of $\widehat{V}$.

First, we perform a Cholesky step followed by eigen-decomposition:

$$\begin{aligned}X = \begin{pmatrix}1 & \boldsymbol{x}\\ \boldsymbol{x}^\dagger & \mathsf{X}\end{pmatrix} &= \begin{pmatrix}1 & 0\\ \boldsymbol{x}^\dagger & \mathrm{Id}\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & \mathsf{X} - \boldsymbol{x}^\dagger \otimes \boldsymbol{x}\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{x}\\ 0 & \mathrm{Id}\end{pmatrix} \\ &= \begin{pmatrix}1 & 0\\ \boldsymbol{x}^\dagger & \mathsf{U}^\dagger\end{pmatrix} X_1 \begin{pmatrix}1 & \boldsymbol{x}\\ 0 & \mathsf{U}\end{pmatrix} =: L^\dagger X_1 L \end{aligned} \tag{4.182}$$

where U is unitary and $X_1 = \mathrm{diag}(1, \sigma_1, \dots, \sigma_\chi)$ where

$$\begin{cases}\sigma_a > 0 & \text{if } a \le \chi' \\ \sigma_a = 0 & \text{if } a > \chi'.\end{cases} \tag{4.183}$$

Since L is invertible, we have the gauge transform

$$\widehat{V}_1 := L\widehat{V}L^{-1} \tag{4.184}$$

so that the leading eigenvector of T_{V_1} becomes the diagonal matrix X_1. Thus, the aa-th component of the equation $X_1 T_{V_1} = X_1$ becomes

$$\sigma_a = \sum_{b=0}^{\chi} \sigma_b \left\langle (\widehat{V}_1)_{ba}, (\widehat{V}_1)_{ba} \right\rangle \,. \tag{4.185}$$

When $a > \chi'$, $\sigma_a = 0$, so every term on the RHS must also vanish. Now for $b \le \chi'$, $\sigma_b > 0$, so $(\widehat{V}_1)_{ba} = 0$. Namely, we showed that $\widehat{V}_1$ has the block-diagonal form:

$$\widehat{V}_1 = \begin{pmatrix} \widehat{V}'_{[0,\chi']} & 0 \\ 0 & * \end{pmatrix}, \tag{4.186}$$

where $\widehat{V}'$ has shape $(1+\chi') \times (1+\chi')$. This implies that $\widehat{V}_1$ can be gauge transformed to $\widehat{V}'$ by a projector:

$$\begin{pmatrix} \mathrm{Id}_{[0,\chi']} & 0 \end{pmatrix} \widehat{V}_1 = \widehat{V}' \begin{pmatrix} \mathrm{Id}_{[0,\chi']} & 0 \end{pmatrix} \tag{4.187}$$

It is easy to check that $T_{V'}$ has leading eigenvector $X_2 = \mathrm{diag}(1, \sigma_1, \ldots, \sigma_{\chi'})$, which is non-singular. Composing the two gauge transforms (4.184) and (4.187) and promoting them to the iMPO level completes the proof. □

Now we can finally prove the existence of left canonical form for all first degree iMPOs.

Proof of Prop. 19. By Prop 35, we find first a rank-reducing L_0 and $\widehat{W}'$ so that $L_0\widehat{W} = \widehat{W}'L_0$ and $\widehat{W}'$ satisfies the assumptions of Prop. 32. Then the QR iteration must converge and bring $\widehat{W}'$ to a left canonical $\widehat{W}_L$ by some gauge transform $L_1\widehat{W}' = \widehat{W}_L L_1$. Composing the gauge transforms gives $L\widehat{W} = \widehat{W}_L L$ with $L = L_1 L_0$. □

Note that the above proof and that of Lemma 35 provide a foolproof algorithm to compute the left canonical form: first precondition the MPO by reducing its rank, then use QR iteration. We provide an implementation in Algorithm 8. This algorithm is provably convergent for all first degree iMPOs, and has comparable numerical precision and stability to the QR iteration Algorithm 3. (Recall that any method of taking the square root of X directly reduces the precision from 10^{-16} to 10^{-8} with standard floating point; QR iteration is required for high precision.) The main drawback of Algorithm 8 is its efficiency: the preconditioning routine involves two eigenvalue problems: finding the leading eigenvector X, and (almost) diagonalizing it. It is often more expensive than the QR iteration itself. This brings us to a natural question: why couldn't we prove the existence of left canonical form for all first degree iMPOs (Prop. 19) directly using QR iteration? After all, the rank-revealing QR can also reduce bond dimension and potentially serve the rôle of the preconditioning step. The answer, unfortunately, is that there are first degree iMPOs for which the QR iteration fails.

Example 36. Consider the spin-half iMPO

$$\widehat{W} := \begin{pmatrix} \widehat{\mathbb{1}} & 0 & \widehat{Z} \\ & \alpha\widehat{Z} & \widehat{X} \\ & & \widehat{\mathbb{1}} \end{pmatrix}, \tag{4.188}$$

where $|\alpha| < 1$ so that $\widehat{W}$ is first degree. But applying Algorithm 3 to it will yield

$$\widehat{W}_n = \begin{pmatrix} \widehat{\mathbb{1}} & 0 & \widehat{Z} \\ & \alpha\widehat{Z} & \alpha^n\widehat{X} \\ & & \widehat{\mathbb{1}} \end{pmatrix}, \quad L_n = \begin{pmatrix} 1 & 0 & 0 \\ & \alpha^n & 0 \\ & 0 & 1 \end{pmatrix}. \tag{4.189}$$

Algorithm 8 iMPO Left Can. Form: General

```
procedure PRECONDITION($\widehat{W}, \eta$)
    $X \leftarrow$ EIGMAX($T_V$)                                        ▷ Find max. eigenvector
    $\boldsymbol{x}, \mathsf{U}, \Sigma \leftarrow X$                                        ▷ Eq. (4.182)
    $\chi' \leftarrow \max\{a : \sigma_a > \eta^2\}$
    $\boldsymbol{x}, \mathsf{U} \leftarrow [x_a]_{1\le a\le\chi'}, [\mathsf{U}_{ab}]_{1\le a\le\chi',1\le b\le\chi}$

    $L \leftarrow \begin{pmatrix} 1 & \boldsymbol{x} & 0 \\ & \mathsf{U} & 0 \\ & & 1 \end{pmatrix}, L' \leftarrow \begin{pmatrix} 1 & -\boldsymbol{x} & 0 \\ & \mathsf{U}^\dagger & 0 \\ & & 1 \end{pmatrix}$
    return $L\widehat{W}L', L$
procedure LEFTCAN($\widehat{W}, \eta$)                                        ▷ $\eta$: tolerance
    $\widehat{W}, L_0 \leftarrow$ PRECONDITION($\widehat{W}, \eta$)
    $\widehat{W}, L_1 \leftarrow$ QRITER($\widehat{W}, \eta$)                                        ▷ Alg. 3
    return $\widehat{W}, L_1 L_0$
```

Everything seems to converge, but $\lim_{n\to\infty} \widehat{W}_n$ is not left canonical! In fact, $\lim_{n\to\infty} L_n$ is singular, which makes the argument in the proof of Prop. 32 inapplicable. The origin of this failure is that, the middle state of the state machine is not reachable from the initial state, so the middle row and column can be removed altogether. (This is precisely what the PRECONDITION routine in Algorithm 8 does.) But the rank-revealing QR fails to detect this, because $\widehat{W}$ has full column rank.

We close this appendix by noting that the above theory for the convergence of QR iteration can be improved. Indeed the assumption of Prop. 32 can be certainly relaxed. It will be interesting to find a sufficient and necessary condition of convergence, and improve the efficiency of the preconditioning step.

4.C Exact estimates of Schmidt values

We study the singular values of the matrix M defined in (4.82) (which form the entanglement spectrum of an MPO) by repeatedly applying a rank one perturbation.

First, we consider the sub-matrix

$$M_0 := \begin{pmatrix} \mathcal{N}_R & \boldsymbol{p}_R \\ 0 & \mathsf{S} \end{pmatrix}, \tag{4.190}$$

where $\mathsf{S} = \mathrm{diag}(s_1 \geq \cdots \geq s_\chi)$ so that

$$M_0^\dagger M_0 = \begin{pmatrix} 0 & 0 \\ 0 & \mathsf{S}^2 \end{pmatrix} + \begin{pmatrix} \mathcal{N}_R \\ \boldsymbol{p}_R^\dagger \end{pmatrix} \begin{pmatrix} \mathcal{N}_R & \boldsymbol{p}_R \end{pmatrix} \tag{4.191}$$

is a rank one perturbation of $\mathrm{diag}(0, s_1^2, s_2^2, \dots)$. A standard result then shows that the singular values of M_0, denoted $\mu_0 \geq \mu_1 \geq \mu_2 \geq \dots \mu_\chi$, are given by the positive roots of the equation

$$\frac{\mathcal{N}_R^2}{\mu^2} + \sum_a \frac{|p_R^a|^2}{\mu^2 - s_a^2} = 1\,. \tag{4.192}$$

This implies the interlacing relation

$$\mu_0 \geq s_1 \geq \mu_1 \geq s_2 \geq \cdots \geq s_\chi \geq \mu_\chi. \tag{4.193}$$

For the largest singular value, (4.192) further implies

$$\frac{\mathcal{N}_R^2}{\mu_0^2} + \sum_a \frac{|p_R^a|^2}{\mu_0^2} \leq 1 \leq \frac{\mathcal{N}_R^2}{\mu_0^2 - s_1^2} + \sum_a \frac{|p_R^a|^2}{\mu_0^2 - s_1^2}\,,$$

leading to the following estimates:

$$\mathcal{N}_R^2 + ||\boldsymbol{p}_R||^2 + s_1^2 \geq \mu_0^2 \geq \mathcal{N}_R^2 + ||\boldsymbol{p}_R||^2\,. \tag{4.194}$$

In particular, the separation of scales (4.83) implies $\mu_0^2 = \Theta(N)$ and $\mu_{a\geq 1}^2 = \mathrm{O}(1)$.

In a very similar fashion, we now go back to the full matrix and consider

$$MM^\dagger = \begin{pmatrix} M_0 M_0^\dagger & \\ & 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \boldsymbol{p}_L \\ \mathcal{N}_L \end{pmatrix} \begin{pmatrix} 0 & \boldsymbol{p}_L^\dagger & \mathcal{N}_L \end{pmatrix} \tag{4.195}$$

which is similar to

$$\begin{pmatrix} \mu_0^2 & & \\ & D_\mu & \\ & & 0 \end{pmatrix} + \begin{pmatrix} \boldsymbol{q}_L \\ \mathcal{N}_L \end{pmatrix} \begin{pmatrix} \boldsymbol{q}_L^\dagger & \mathcal{N}_L \end{pmatrix}\,, \tag{4.196}$$

under conjugation where $D_\mu = \mathrm{diag}(\mu_1^2, \dots, \mu_\chi^2)$, $\boldsymbol{q}_L = U(0\ \boldsymbol{p}_L)^T$, U being a unitary matrix such that $UM_0M_0^\dagger U^\dagger = \mathrm{diag}(\mu_0^2, \mu_1^2, \dots, \mu_\chi^2)$. Applying rank one perturbation again to (4.196), we obtain the following equation determining the singular values of M:

$$\frac{\mathcal{N}_L^2}{\lambda^2} + \frac{|q_L^0|^2}{\lambda^2 - \mu_0^2} + \sum_{a=1}^{\chi} \frac{|q_L^a|^2}{\lambda^2 - \mu_a^2} = 1\,. \tag{4.197}$$

This implies the interlacing relation

$$\lambda_{-1} \geq \mu_0 \geq \lambda_0 \geq \mu_1 \geq \cdots \geq \mu_\chi \geq \lambda_\chi\,, \tag{4.198}$$

which, combined with (4.193), gives (4.87) in the main text.

Similarly to (4.194), we can bound λ_{-1} as follows:

$$\lambda_{-1}^2 \geq \mathcal{N}_L^2 + ||\boldsymbol{q}_L||^2 = \mathcal{N}_L^2 + ||\boldsymbol{p}_L^2|| \tag{4.199a}$$

$$\lambda_{-1}^2 \leq \mathcal{N}_L^2 + ||\boldsymbol{p}_L^2|| + \mu_0^2\,. \tag{4.199b}$$

Under the separation of scales (4.83), $\lambda_{-1} = \Theta(N)$ is extensive.

We also need a useful lower bound for largest singular value λ_0. For this, we note that (4.197) implies

$$\frac{\mathcal{N}_L^2}{\lambda_0^2} + \sum_{a=1}^{\chi} \frac{|q_L^a|^2}{\lambda_0^2} \leq 1 + \frac{|q_L^0|^2}{\mu_0^2 - \lambda_0^2} \tag{4.200}$$

which is a quadratic inequality (of λ_0^2). Its solution entails

$$2\lambda_0^2 \geq \mu_0^2 + \mathcal{N}_L^2 + ||\boldsymbol{q}_L||^2 - \sqrt{(\mu_0^2 - \mathcal{N}_L^2 - ||\boldsymbol{q}_L||^2)^2 + 4\mu_0^2|q_L^0|^2} \tag{4.201}$$

$$\geq 2\min(\mu_0^2, \mathcal{N}_L^2 + ||\boldsymbol{q}_L||^2) - 2\mu_0|q_L^0|\,. \tag{4.202}$$

Now, under (4.83), $\mu_0^2, \mathcal{N}_L^2 \in \Theta(N)$ and $\boldsymbol{q}_L \in \mathrm{O}(1)$, so we conclude that $\lambda_0^2 \in \Theta(N)$ is also extensive.

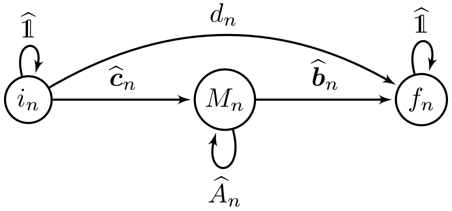
4.D Elementary operations

This Appendix discusses how to perform the standard algebraic operations — scalar multiplication, addition, multiplication, and commutation — for local MPOs. These are standard operations and are discussed in various places in the literature, but we review them here for completeness.

Suppose below that $\lambda \in \mathbb{R}$ is a scalar and operators $\widehat{\mathcal{O}}_1$ and $\widehat{\mathcal{O}}_2$ are represented by iMPOs

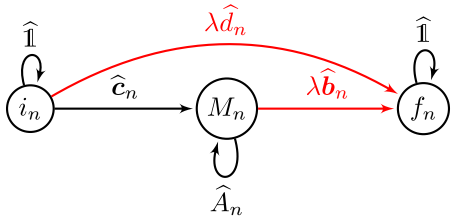
$$\widehat{W}[\widehat{\mathcal{O}}_1] = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}}_1 & \widehat{d}_1 \\ 0 & \widehat{A}_1 & \widehat{\boldsymbol{b}}_1 \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}, \widehat{W}[\widehat{\mathcal{O}}_2] = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}}_2 & \widehat{d}_2 \\ 0 & \widehat{A}_2 & \widehat{\boldsymbol{b}}_2 \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix} \tag{4.203}$$

respectively with finite-automata as follows.



Here $(i_n, M_n, f_n), n = 1, 2$ stand for the initial state, the χ middle states, and the final state.

The scalar product is straightforward: each term needs to be scaled exactly once as it moves through the automata. This can be done by scaling all the edges that are incident to the final (or initial) state.



At the matrix level:

$$\widehat{W}[\lambda\widehat{\mathcal{O}}_1] = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}}_1 & \lambda\widehat{d}_1 \\ 0 & \widehat{A}_1 & \lambda\widehat{\boldsymbol{b}}_1 \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbb{1}} & \lambda\widehat{\boldsymbol{c}}_1 & \lambda\widehat{d}_1 \\ 0 & \widehat{A}_1 & \widehat{\boldsymbol{b}}_1 \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}. \tag{4.204}$$

These two choices preserve left and right canonical forms respectively.

Addition of iMPOs is essentially the direct sum of the matrices:

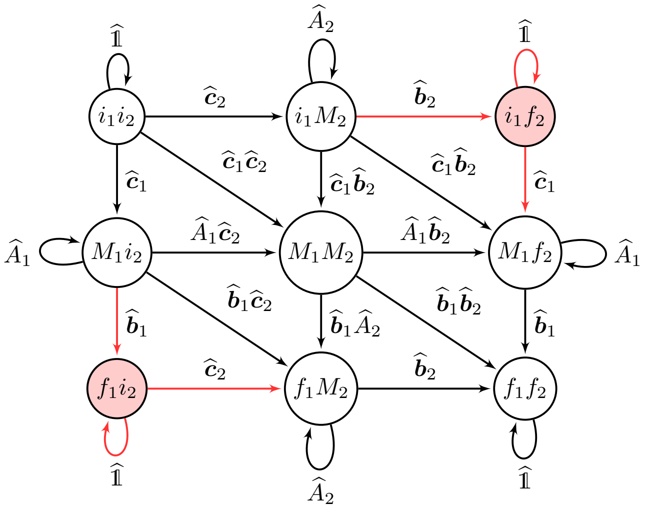
$$\widehat{W}[\widehat{\mathcal{O}}_1 + \widehat{\mathcal{O}}_2] = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}}_1 & \widehat{\boldsymbol{c}}_2 & \widehat{d}_1 + \widehat{d}_2 \\ 0 & \widehat{A}_1 & 0 & \widehat{\boldsymbol{b}}_1 \\ 0 & 0 & \widehat{A}_2 & \widehat{\boldsymbol{b}}_2 \\ 0 & 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}. \tag{4.205}$$

The operation of multiplication is more involved. The multiplication of two local operators, say $\widehat{\mathcal{O}}_1 = \sum_i \widehat{X}_i$ and $\widehat{\mathcal{O}}_2 = \sum_i \widehat{Y}_i$ is "bi-local" or "second degree", with arbitrarily long strings of identities between sites with information: $\widehat{\mathcal{O}}_1\widehat{\mathcal{O}}_2 = \sum_i \sum_{N=0}^{\infty} \widehat{X}_i \widehat{\mathbb{1}}^N \widehat{Y}_{i+N} + \cdots$. This is represented as an iMPO as

$$\widehat{W}[O_1 O_2] = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{X} & \widehat{Y} & i\widehat{Z} \\ 0 & \widehat{\mathbb{1}} & 0 & \widehat{Y} \\ 0 & 0 & \widehat{\mathbb{1}} & \widehat{X} \\ 0 & 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}. \tag{4.206}$$

The $\widehat{\mathbb{1}}$'s on the diagonal are an unavoidable consequence of being "second degree": $\widehat{W}[O_1 O_2]$ norm $\propto N^2$ in a system of size N.

It is insightful to look at the generic "product automata".


$\widehat{\mathbb{1}}$
$\widehat{A}_2$
$\widehat{\mathbb{1}}$
i_1i_2
$\widehat{\boldsymbol{c}}_2$
i_1M_2
$\widehat{\boldsymbol{b}}_2$
i_1f_2
$\widehat{\boldsymbol{c}}_1\widehat{\boldsymbol{c}}_2$
$\widehat{\boldsymbol{c}}_1\widehat{\boldsymbol{b}}_2$
$\widehat{\boldsymbol{c}}_1$
$\widehat{\boldsymbol{c}}_1\widehat{\boldsymbol{b}}_2$
$\widehat{\boldsymbol{c}}_1$
$\widehat{A}_1$
M_1i_2
$\widehat{A}_1\widehat{\boldsymbol{c}}_2$
M_1M_2
$\widehat{A}_1\widehat{\boldsymbol{b}}_2$
M_1f_2
$\widehat{A}_1$
$\widehat{\boldsymbol{b}}_1\widehat{\boldsymbol{c}}_2$
$\widehat{\boldsymbol{b}}_1\widehat{\boldsymbol{b}}_2$
$\widehat{\boldsymbol{b}}_1$
$\widehat{\boldsymbol{b}}_1\widehat{A}_2$
$\widehat{\boldsymbol{b}}_1$
$\widehat{\boldsymbol{c}}_2$
$\widehat{\boldsymbol{b}}_2$
f_1i_2
f_1M_2
f_1f_2
$\widehat{\mathbb{1}}$
$\widehat{A}_2$
$\widehat{\mathbb{1}}$

(We have dropped the $\widehat{d}$ terms and also the self-loop on M_1M_2 for clarity.) One should interpret the products on edges as the tensor products of the ancilla space but products in the physical space. For example, "$\widehat{\boldsymbol{b}}_1\widehat{A}_2$" has components

$$\left(\widehat{\boldsymbol{b}}_1\widehat{A}_2\right)^{\gamma}_{(a_1a_2),b_2} = \sum_{\alpha,\beta} f^{\gamma}_{\alpha\beta}(B_1)^{\alpha}_{a_1}\,(A_2)^{\beta}_{a_2,b_2} \tag{4.207}$$

where $f^{\gamma}_{\alpha\beta}$ are the structure constants of the on-site algebra $\mathcal{A}$.

The non-locality of the product comes only from the shaded parts of the automata. What if we were to simply remove the troublesome parts? This motivates a definition.

Definition 37. Suppose $\widehat{\mathcal{O}}_1$ and $\widehat{\mathcal{O}}_2$ are two strings of single site operators (Pauli strings in the spin-1/2 case) with support on sites $[a_1, b_1]$ and $[a_2, b_2]$ respectively. The **non-disjoint product** is

$$\widehat{\mathcal{O}}_1 \odot \widehat{\mathcal{O}}_2 = \begin{cases} \widehat{\mathcal{O}}_1\widehat{\mathcal{O}}_2 & \text{if } [a_1,b_1]\cap[a_2,b_2] \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \tag{4.208}$$

The definition extends to any local operators by linearity. At the MPO level, this is just the non-shaded part of the above diagram.

Terms with disjoint spatial support always commute, so the "non-disjoint commutator" is the same as the normal one:

$$[\widehat{\mathcal{O}}_1, \widehat{\mathcal{O}}_2] = \widehat{\mathcal{O}}_1 \odot \widehat{\mathcal{O}}_2 - \widehat{\mathcal{O}}_2 \odot \widehat{\mathcal{O}}_1. \tag{4.209}$$

This means that the commutator is local whenever $\widehat{\mathcal{O}}_1 \odot \widehat{\mathcal{O}}_2$ is. Therefore strictly local operators form a closed algebra under commutation.

First degree operators, however, are not closed even under commutation, as the following counter-example demonstrates. Suppose $\widehat{H}_l$ has an iMPO representation

$$\widehat{W}_l = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{X} & 0 \\ 0 & \widehat{\mathcal{O}} & \widehat{Y} \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix} \tag{4.210}$$

where $\widehat{\mathcal{O}} = \frac{c}{2}\left(\widehat{\mathbb{1}} + \widehat{Z}\right) = \begin{pmatrix} c & 0 \\ 0 & 0 \end{pmatrix}$ is an on-site projector matrix and take $c \in (2^{1/4}, 2^{1/2})$. The norm of H_l is $||H_l||^2 = \sum_{N=0}^{\infty} ||\widehat{\mathcal{O}}||^{2N} = \sum_{N=0}^{\infty}(c^2/2)^N < \infty$. However, the norm of the product diverges:

$$||H_l \odot H_l||^2 > \sum_{N=0}^{\infty} ||\widehat{\mathcal{O}}\widehat{\mathcal{O}}||^{2N} = \sum_{N=0}^{\infty} \left(c^4/2\right)^N = \infty, \tag{4.211}$$

since $c > 2^{1/4}$. The divergent terms here are not from the diagonal ones but from an eigenvalue $c^4/2 > 1$ of T_A. So not only can the product of two first degree iMPOs be strictly non-local, but the norm-per-unit-length is not even submultiplicative: there are cases where $||\widehat{\mathcal{O}}_1\widehat{\mathcal{O}}_2|| \nleq ||\widehat{\mathcal{O}}_1|| ||\widehat{\mathcal{O}}_2||$. It would be interesting to find the largest closed subalgebra of the first degree operators.

Thankfully, the commutator of a first degree operator with a strictly-local operator is well-controlled, which is what enables us to perform the Lanczos algorithm within first degree operator, so long as the Hamiltonian is strictly local — the most physically relevant case.

Proposition 38. *If $\widehat{\mathcal{O}}_1$ is strictly local and $\widehat{\mathcal{O}}_2$ is first degree, then $[\widehat{\mathcal{O}}_1, \widehat{\mathcal{O}}_2]$ is first degree.*

Proof. It is sufficient to show $\widehat{\mathcal{O}}_1 \odot \widehat{\mathcal{O}}_2$ is first degree.

Let the iMPOs for the operators be given by Eq. (4.203). In particular, $\widehat{A}_1$ is strictly upper triangular. From the product automata above, we can see that the $\widehat{A}$ block of $\widehat{\mathcal{O}}_1 \odot \widehat{\mathcal{O}}_2$ is given by

$$\widehat{A} = \begin{pmatrix} \widehat{A}_2 & 0 & \widehat{\boldsymbol{c}}_1\widehat{A}_2 & \widehat{\boldsymbol{c}}_1\widehat{\boldsymbol{b}}_2 & 0 \\ 0 & \widehat{A}_1 & \widehat{A}_1\widehat{\boldsymbol{c}}_2 & 0 & \widehat{\boldsymbol{b}}_1\widehat{\boldsymbol{c}}_2 \\ 0 & 0 & \widehat{A}_1\widehat{A}_2 & \widehat{A}_1\widehat{\boldsymbol{b}}_2 & \widehat{\boldsymbol{b}}_1\widehat{A}_2 \\ 0 & 0 & 0 & \widehat{A}_1 & 0 \\ 0 & 0 & 0 & 0 & \widehat{A}_2 \end{pmatrix}, \tag{4.212}$$

where "multiplications" such as $\widehat{A}_1\widehat{A}_2$ again stands for the tensor product in ancilla indices and multiplication in the physical indices. This is block-upper triangular, so the transfer matrix T_A is also block upper triangular, and it's spectrum is the union of the spectra of the transfer matrices of the diagonal blocks of $\widehat{A}$. Since $\widehat{A}_1$ and $\widehat{A}_1\widehat{A}_2$ are upper triangular with zeros on the diagonal, the maximal eigenvalue of their transfer matrices is also zero. Since $\widehat{A}_2$ is first degree, the maximal eigenvalue of its transfer matrix is some $\lambda < 1$, so the maximal eigenvalue of T_A is also λ. This completes the proof. □

As a practical matter, then, one should compute the commutator of two MPOs via Eq. (4.209). It is advisible to compress the operator after each product and again after the difference. In circumstances where $\widehat{\mathcal{O}}_1$ and $\widehat{\mathcal{O}}_2$ are Hermitian or anti-Hermitian, the two non-disjoint products are related by a Hermitian conjugate and a sign, and need to be computed only once.

Chapter 5

Conclusions

The theme of this work is the close connection between quantum chaos and local operators. We started in Chapter 2 by describing the Lanczos algorithm and its connection to physics. In Chapter 3 we proposed the “Universal Operator Growth Hypothesis”, which roughly says that operators in a chaotic quantum system grow “as fast as possible” and is formulated in terms of the Lanczos algorithm. This perspective on chaos has several practical consequences, chiefly (1) the notion of a Q-complexity which quantifies the rate at which an operator becomes more complex over time and (2) a efficient algorithm for computing the hydrodynamics of a chaotic quantum system via continued fraction methods. Finally, in Chapter 4 we described the structure of local MPOs and provided an efficient algorithm for compressing them, with applications for both dynamics and ground state physics. We may conclude that understanding the structure of local operators gives many insights into quantum dynamics and chaos.

However, as often occurs in science, any new ideas lead immediately to more questions. Let us conclude with a few of the most interesting, ranging from practical and straightforward to philosophical.

- Can one prove the operator growth hypothesis? We have several examples such as the SYK model and the model of Bouch where it is proven to hold analytically, but is it possible to prove in a more general setting? For instance, could one show it holds for a translation-invariant random matrix type model?

- Alternatively, should the Operator Growth Hypothesis be considered not as a theorem but instead as a deinfition for a class of “systems with fast operator growth” — an idea related to, but not synonymous with, chaos.

- How accurate is our algorithm for computing hydrodynamics from Section 3.7? Can one provide careful error bounds? Can this algorithm be used in more complex 1D systems or 2D systems?

- How can the ideas of operator growth be extended to finite temperature?

- How do these ideas of operator growth manifest in finite systems?

- What is the precise relationship between the notion of Q-complexity from Sec. 3.5 and quantum computational complexity? There are indications that there is a close relationship between our notion and definitions of operator complexity in the high-energy literature (see e.g. [134]).

- Similarly, our algorithm for operator compression works with local operators as *vectors* in the space of operators. However, operators form an *algebra*. So, in principle, one could imagine encoding a operator as a quantum circuit where the chief operator is multiplication rather than addition. Of course, this comes with the significant drawback that evaluating matrix elements of the operator becomes prohibitively difficult (at least on classical computers). Are there (efficient?) ways to find an optimal circuit which encodes a given many-body operator $O(t)$?

- How many resources does it require, in principle, to compute hydrodynamics in chaotic systems?

- Finally, what is quantum chaos?

Bibliography

[1] Daniel E Parker et al. “A Universal Operator Growth Hypothesis”. In: *arXiv:1812.08657* (2018).

[2] Daniel E Parker, Xiangyu Cao, and Michael P Zaletel. “Local Matrix Product Operators: Canonical Form, Compression, & Control Theory”. In: *arXiv preprint arXiv:1909.06341* (2019).

[3] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Matrix Computations. Johns Hopkins University Press, 2012. ISBN: 9781421408590. URL: `https://books.google.com/books?id=5U-l8U3P-VUC`.

[4] Barry A Cipra. “The best of the 20th century: Editors name top 10 algorithms”. In: *SIAM news* 33.4 (2000), pp. 1–2.

[5] Yousef Saad. *Iterative methods for sparse linear systems*. Vol. 82. siam, 2003.

[6] Gene H Golub and Gérard Meurant. *Matrices, moments and quadrature with applications*. Vol. 30. Princeton University Press, 2009.

[7] Richard Bruno Lehoucq. *Analysis and implementation of an implicitly restarted Arnoldi iteration*. Tech. rep. Rice University, Houston, Tx. Department of Computational and Applied Mathematics, 1995.

[8] Theodore S Chihara. *An introduction to orthogonal polynomials*. Gordon and Breach, Science Publishers, Inc., 1978.

[9] XG Viennot. “Une théorie combinatoire des polynômes orthogonaux”. In: *Lecture Notes, UQAM, Montréal* (1984).

[10] VS Viswanath and Gerhard Müller. *The Recursion Method: Applications to Many-body Dynamics*. Springer, 2008.

[11] Naum Iljic Aheizer and N Kemmer. *The classical moment problem and some related questions in analysis*. Oliver & Boyd Edinburgh, 1965.

[12] Gwo Dong Lin. “Recent developments on the moment problem”. In: *Journal of Statistical Distributions and Applications* 4.1 (2017), pp. 1–17.

[13] J. M. Deutsch. “Quantum statistical mechanics in a closed system”. In: *Phys. Rev. A* 43 (4 Feb. 1991), pp. 2046–2049. DOI: `10.1103/PhysRevA.43.2046`. URL: `https://link.aps.org/doi/10.1103/PhysRevA.43.2046`.

[14] Mark Srednicki. “Chaos and quantum thermalization”. In: *Phys. Rev. E* 50 (2 Aug. 1994), pp. 888–901. DOI: 10.1103/PhysRevE.50.888. URL: https://link.aps.org/doi/10.1103/PhysRevE.50.888.

[15] Marcos Rigol, Vanja Dunjko, and Maxim Olshanii. “Thermalization and its mechanism for generic isolated quantum systems”. In: *Nature* 452.7189 (2008), p. 854. URL: https://doi.org/10.1038/nature06838.

[16] Luca D’Alessio et al. “From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics”. In: *Advances in Physics* 65.3 (2016), pp. 239–362. DOI: 10.1080/00018732.2016.1198134. eprint: https://doi.org/10.1080/00018732.2016.1198134. URL: https://doi.org/10.1080/00018732.2016.1198134.

[17] Joshua M Deutsch. “Eigenstate thermalization hypothesis”. In: *Reports on Progress in Physics* 81.8 (2018), p. 082001. URL: http://stacks.iop.org/0034-4885/81/i=8/a=082001.

[18] C. W. von Keyserlingk et al. “Operator Hydrodynamics, OTOCs, and Entanglement Growth in Systems without Conservation Laws”. In: *Phys. Rev. X* 8 (2 Apr. 2018), p. 021013. DOI: 10.1103/PhysRevX.8.021013. URL: https://link.aps.org/doi/10.1103/PhysRevX.8.021013.

[19] Adam Nahum, Sagar Vijay, and Jeongwan Haah. “Operator spreading in random unitary circuits”. In: *Physical Review X* 8.2 (2018), p. 021014.

[20] Tibor Rakovszky, Frank Pollmann, and CW von Keyserlingk. “Diffusive hydrodynamics of out-of-time-ordered correlators with charge conservation”. In: *arXiv:1710.09827* (2017).

[21] Vedika Khemani, Ashvin Vishwanath, and David A. Huse. “Operator Spreading and the Emergence of Dissipative Hydrodynamics under Unitary Evolution with Conservation Laws”. In: *Phys. Rev. X* 8 (3 Sept. 2018), p. 031057. DOI: 10.1103/PhysRevX.8.031057. URL: https://link.aps.org/doi/10.1103/PhysRevX.8.031057.

[22] Sarang Gopalakrishnan et al. “Hydrodynamics of operator spreading and quasiparticle diffusion in interacting integrable systems”. In: *arXiv:1809.02126* (2018).

[23] Amos Chan, Andrea De Luca, and JT Chalker. “Solution of a minimal model for many-body quantum chaos”. In: *Physical Review X* 8.4 (2018), p. 041019.

[24] Juan Maldacena, Stephen H. Shenker, and Douglas Stanford. “A bound on chaos”. In: *Journal of High Energy Physics* 2016.8 (Aug. 2016), p. 106. ISSN: 1029-8479. DOI: 10.1007/JHEP08(2016)106. URL: https://doi.org/10.1007/JHEP08(2016)106.

[25] Juan Maldacena and Douglas Stanford. “Remarks on the Sachdev-Ye-Kitaev model”. In: *Phys. Rev. D* 94 (10 Nov. 2016), p. 106002. DOI: 10.1103/PhysRevD.94.106002. URL: https://link.aps.org/doi/10.1103/PhysRevD.94.106002.

[26] A. Kitaev. *A simple model of quantum holography.* 2015. URL: http://online.kitp.ucsb.edu/online/entangled15/kitaev/.

[27] Subir Sachdev and Jinwu Ye. "Gapless spin-fluid ground state in a random quantum Heisenberg magnet". In: *Phys. Rev. Lett.* 70 (21 May 1993), pp. 3339–3342. DOI: 10.1103/PhysRevLett.70.3339. URL: https://link.aps.org/doi/10.1103/PhysRevLett.70.3339.

[28] B. V. Fine et al. "Absence of exponential sensitivity to small perturbations in nonintegrable systems of spins 1/2". In: *Phys. Rev. E* 89 (1 Jan. 2014), p. 012923. DOI: 10.1103/PhysRevE.89.012923. URL: https://link.aps.org/doi/10.1103/PhysRevE.89.012923.

[29] Shenglong Xu and Brian Swingle. "Accessing scrambling using matrix product operators". In: *arXiv:1802.00801* (2018).

[30] Shenglong Xu and Brian Swingle. "Locality, quantum fluctuations, and scrambling". In: *arXiv:1805.05376* (2018).

[31] Daniel C Mattis. "How to reduce practically any problem to one dimension". In: *Physics in One Dimension*. Springer, 1981, pp. 3–10.

[32] Netanel H. Lindner and Assa Auerbach. "Conductivity of hard core bosons: A paradigm of a bad metal". In: *Phys. Rev. B* 81 (5 Feb. 2010), p. 054512. DOI: 10.1103/PhysRevB.81.054512. URL: https://link.aps.org/doi/10.1103/PhysRevB.81.054512.

[33] Ilia Khait et al. "Spin transport of weakly disordered Heisenberg chain at infinite temperature". In: *Phys. Rev. B* 93 (22 June 2016), p. 224205. DOI: 10.1103/PhysRevB.93.224205. URL: https://link.aps.org/doi/10.1103/PhysRevB.93.224205.

[34] Assa Auerbach. "Hall Number of Strongly Correlated Metals". In: *Phys. Rev. Lett.* 121 (6 Aug. 2018), p. 066601. DOI: 10.1103/PhysRevLett.121.066601. URL: https://link.aps.org/doi/10.1103/PhysRevLett.121.066601.

[35] D. A. McArthur, E. L. Hahn, and R. E. Walstedt. "Rotating-Frame Nuclear-Double-Resonance Dynamics: Dipolar Fluctuation Spectrum in CaF_2". In: *Phys. Rev.* 188 (2 Dec. 1969), pp. 609–638. DOI: 10.1103/PhysRev.188.609. URL: https://link.aps.org/doi/10.1103/PhysRev.188.609.

[36] M. Engelsberg and I. J. Lowe. "Free-inducation-decay measurements and determination of moments in CaF_2". In: *Phys. Rev. B* 10 (3 Aug. 1974), pp. 822–832. DOI: 10.1103/PhysRevB.10.822. URL: https://link.aps.org/doi/10.1103/PhysRevB.10.822.

[37] A A Lundin, A V Makarenko, and V E Zobov. "The dipolar fluctuation spectrum and the shape of the wings of nuclear magnetic resonance absorption spectra in solids". In: *Journal of Physics: Condensed Matter* 2.50 (1990), p. 10131. URL: http://stacks.iop.org/0953-8984/2/i=50/a=017.

[38] M. Howard Lee. “Ergodic Theory, Infinite Products, and Long Time Behavior in Hermitian Models”. In: *Phys. Rev. Lett.* 87 (25 Nov. 2001), p. 250601. DOI: 10.1103/PhysRevLett.87.250601. URL: https://link.aps.org/doi/10.1103/PhysRevLett.87.250601.

[39] Paul Fendley. “Free fermions in disguise”. In: *Journal of Physics A: Mathematical and Theoretical* 52.33 (2019), p. 335002.

[40] Tyler LeBlond et al. “Entanglement and matrix elements of observables in interacting integrable systems”. In: *Physical Review E* 100.6 (2019), p. 062134.

[41] Daniel A Roberts, Douglas Stanford, and Alexandre Streicher. “Operator growth in the SYK model”. In: *Journal of High Energy Physics* 2018.6 (2018), p. 122. DOI: 10.1007/JHEP06(2018)122. URL: https://doi.org/10.1007/JHEP06(2018)122.

[42] Edwin Huang. private communcation.

[43] Gabriel Bouch. “Complex-time singularity and locality estimates for quantum lattice systems”. In: *Journal of Mathematical Physics* 56.12 (2015), p. 123303.

[44] D. S. Lubinsky. “A survey of general orthogonal polynomials for weights on finite and infinite intervals”. In: *Acta Applicandae Mathematica* 10.3 (Nov. 1987), pp. 237–296. ISSN: 1572-9036. URL: https://doi.org/10.1007/BF00049120.

[45] A. Magnus. “The Recursion Method and Its Applications: Proceedings of the Conference, Imperial College, London, England September 13–14, 1984”. In: ed. by David G Pettifor and Denis L Weaire. Vol. 58. Springer Science & Business Media, 2012. Chap. 2, pp. 22–45.

[46] Dmitry A. Abanin, Wojciech De Roeck, and F. Huveneers. “Exponentially Slow Heating in Periodically Driven Many-Body Systems”. In: *Phys. Rev. Lett.* 115 (25 Dec. 2015), p. 256803. DOI: 10.1103/PhysRevLett.115.256803. URL: https://link.aps.org/doi/10.1103/PhysRevLett.115.256803.

[47] Niels Strohmaier et al. “Observation of Elastic Doublon Decay in the Fermi-Hubbard Model”. In: *Phys. Rev. Lett.* 104 (8 Feb. 2010), p. 080401. DOI: 10.1103/PhysRevLett.104.080401. URL: https://link.aps.org/doi/10.1103/PhysRevLett.104.080401.

[48] Itai Arad, Tomotaka Kuwahara, and Zeph Landau. “Connecting global and local energy distributions in quantum spin models on a lattice”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2016.3 (2016), p. 033301. URL: http://stacks.iop.org/1742-5468/2016/i=3/a=033301.

[49] Dmitry Abanin et al. “A Rigorous Theory of Many-Body Prethermalization for Periodically Driven and Closed Quantum Systems”. In: *Communications in Mathematical Physics* 354.3 (Sept. 2017), pp. 809–827. ISSN: 1432-0916. DOI: 10.1007/s00220-017-2930-x. URL: https://doi.org/10.1007/s00220-017-2930-x.

[50] Jian-Min Liu and Gerhard Müller. "Infinite-temperature dynamics of the equivalent-neighbor XYZ model". In: *Phys. Rev. A* 42 (10 Nov. 1990), pp. 5854–5864. DOI: 10.1103/PhysRevA.42.5854. URL: https://link.aps.org/doi/10.1103/PhysRevA.42.5854.

[51] O. Bohigas, M. J. Giannoni, and C. Schmit. "Characterization of Chaotic Quantum Spectra and Universality of Level Fluctuation Laws". In: *Phys. Rev. Lett.* 52 (1 Jan. 1984), pp. 1–4. DOI: 10.1103/PhysRevLett.52.1. URL: https://link.aps.org/doi/10.1103/PhysRevLett.52.1.

[52] D. Ullmo. "Bohigas-Giannoni-Schmit conjecture". In: *Scholarpedia* 11.9 (2016). revision #169195, p. 31721. DOI: 10.4249/scholarpedia.31721.

[53] Huzihiro Araki. "Gibbs states of a one dimensional quantum lattice". In: *Communications in Mathematical Physics* 14.2 (1969), pp. 120–157.

[54] Scott Aaronson. "The complexity of quantum states and transformations: from quantum money to black holes". In: *arXiv:1607.05256* (2016).

[55] Leonard Susskind. "Three Lectures on Complexity and Black Holes". In: *arXiv:1810.11563* (2018).

[56] Leonard Susskind. "Why do Things Fall?" In: *arXiv:1802.01198* (2018).

[57] T. Prosen. "Complexity and nonseparability of classical Liouvillian dynamics". In: *Phys. Rev. E* 83 (3 Mar. 2011), p. 031124. DOI: 10.1103/PhysRevE.83.031124. URL: https://link.aps.org/doi/10.1103/PhysRevE.83.031124.

[58] A. Politi. "Lyapunov exponent". In: *Scholarpedia* 8.3 (2013). revision #137286, p. 2722. DOI: 10.4249/scholarpedia.2722.

[59] Chaitanya Murthy and Mark Srednicki. "Bounds on chaos from the eigenstate thermalization hypothesis". In: *arXiv:1906.10808* (2019).

[60] Elliott H. Lieb and Derek W. Robinson. "The finite group velocity of quantum spin systems". In: *Comm. Math. Phys.* 28.3 (1972), pp. 251–257. URL: https://projecteuclid.org:443/euclid.cmp/1103858407.

[61] Dominic V Else et al. "An improved Lieb-Robinson bound for many-body Hamiltonians with power-law interactions". In: *arXiv:1809.06369* (2018).

[62] Uriel Frisch and Rudolf Morf. "Intermittency in nonlinear dynamics and singularities at complex times". In: *Phys. Rev. A* 23 (5 May 1981), pp. 2673–2705. DOI: 10.1103/PhysRevA.23.2673. URL: https://link.aps.org/doi/10.1103/PhysRevA.23.2673.

[63] HS Greenside et al. "A simple stochastic model for the onset of turbulence in Rayleigh-Bénard convection". In: *Physica D: Nonlinear Phenomena* 5.2-3 (1982), pp. 322–334.

[64] David E. Sigeti. “Exponential decay of power spectra at high frequency and positive Lyapunov exponents”. In: *Physica D: Nonlinear Phenomena* 82.1 (1995), pp. 136–153. ISSN: 0167-2789. DOI: https://doi.org/10.1016/0167-2789(94)00225-F. URL: http://www.sciencedirect.com/science/article/pii/016727899400225F.

[65] David E. Sigeti. “Survival of deterministic dynamics in the presence of noise and the exponential decay of power spectra at high frequency”. In: *Phys. Rev. E* 52 (3 Sept. 1995), pp. 2443–2457. DOI: 10.1103/PhysRevE.52.2443. URL: https://link.aps.org/doi/10.1103/PhysRevE.52.2443.

[66] Alexey Cheskidov, Michael Jolly, and E S. Van Vleck. “On a Relation between Lyapunov Exponents and the Radius of Analyticity”. In: *Indiana University Mathematics Journal* 57 (Jan. 2008), pp. 2663–2680. DOI: 10.1512/iumj.2008.57.3826.

[67] Tarek A. Elsayed, Benjamin Hess, and Boris V. Fine. “Signatures of chaos in time series generated by many-spin systems at high temperatures”. In: *Phys. Rev. E* 90 (2 Aug. 2014), p. 022910. DOI: 10.1103/PhysRevE.90.022910. URL: https://link.aps.org/doi/10.1103/PhysRevE.90.022910.

[68] J. E. Maggs and G. J. Morales. “Generality of Deterministic Chaos, Exponential Spectra, and Lorentzian Pulses in Magnetically Confined Plasmas”. In: *Phys. Rev. Lett.* 107 (18 Oct. 2011), p. 185003. DOI: 10.1103/PhysRevLett.107.185003. URL: https://link.aps.org/doi/10.1103/PhysRevLett.107.185003.

[69] David Ruelle. “Resonances of chaotic dynamical systems”. In: *Phys. Rev. Lett.* 56 (5 Feb. 1986), pp. 405–407. DOI: 10.1103/PhysRevLett.56.405. URL: https://link.aps.org/doi/10.1103/PhysRevLett.56.405.

[70] Gustavo J. Turiaci and Herman Verlinde. “On CFT and quantum chaos”. In: *Journal of High Energy Physics* 2016.12 (Dec. 2016), p. 110. ISSN: 1029-8479. DOI: 10.1007/JHEP12(2016)110. URL: https://doi.org/10.1007/JHEP12(2016)110.

[71] Mario Feingold and Asher Peres. “Regular and chaotic motion of coupled rotators”. In: *Physica D: Nonlinear Phenomena* 9.3 (1983), pp. 433–438. DOI: https://doi.org/10.1016/0167-2789(83)90282-8.

[72] Mario Feingold, Nimrod Moiseyev, and Asher Peres. “Ergodicity and mixing in quantum theory. II”. In: *Phys. Rev. A* 30 (1 July 1984), pp. 509–511. DOI: 10.1103/PhysRevA.30.509. URL: https://link.aps.org/doi/10.1103/PhysRevA.30.509.

[73] Yiyun Fan, Sven Gnutzmann, and Yuqi Liang. “Quantum chaos for nonstandard symmetry classes in the Feingold-Peres model of coupled tops”. In: *Phys. Rev. E* 96 (6 Dec. 2017), p. 062207. DOI: 10.1103/PhysRevE.96.062207. URL: https://link.aps.org/doi/10.1103/PhysRevE.96.062207.

[74] Kathleen T Alligood, Tim D Sauer, and James A Yorke. *Chaos.* Springer, 1996.

[75] Steven R. White. “Density matrix formulation for quantum renormalization groups”. In: *Phys. Rev. Lett.* 69 (19 Nov. 1992), pp. 2863–2866. DOI: 10.1103/PhysRevLett.69.2863. URL: https://link.aps.org/doi/10.1103/PhysRevLett.69.2863.

[76] Tobias J. Osborne and Michael A. Nielsen. "Entanglement, Quantum Phase Transitions, and Density Matrix Renormalization". In: *Quantum Information Processing* 1.1 (Apr. 2002), pp. 45–53. ISSN: 1573-1332. DOI: 10.1023/A:1019601218492. URL: https://doi.org/10.1023/A:1019601218492.

[77] Subroto Mukerjee, Vadim Oganesyan, and David Huse. "Statistical theory of transport by strongly interacting lattice fermions". In: *Phys. Rev. B* 73 (3 Jan. 2006), p. 035113. DOI: 10.1103/PhysRevB.73.035113. URL: https://link.aps.org/doi/10.1103/PhysRevB.73.035113.

[78] Marko Medenjak, Christoph Karrasch, and T. Prosen. "Lower bounding diffusion constant by the curvature of Drude weight". In: *Physical review letters* 119.8 (2017), p. 080602.

[79] Marko Ljubotina, Marko Znidaric, and Tomaz Prosen. "Spin diffusion from an inhomogeneous quench in an integrable system". In: *Nature communications* 8.1 (2017), pp. 1–6.

[80] Jacopo De Nardis, Denis Bernard, and Benjamin Doyon. "Diffusion in generalized hydrodynamics and quasiparticle scattering". In: *SciPost Phys.* 6 (4 2019), p. 49. DOI: 10.21468/SciPostPhys.6.4.049. URL: https://scipost.org/10.21468/SciPostPhys.6.4.049.

[81] Sarang Gopalakrishnan and Romain Vasseur. "Kinetic Theory of Spin Diffusion and Superdiffusion in XXZ Spin Chains". In: *Phys. Rev. Lett.* 122 (12 Mar. 2019), p. 127202. DOI: 10.1103/PhysRevLett.122.127202. URL: https://link.aps.org/doi/10.1103/PhysRevLett.122.127202.

[82] Jacopo De Nardis, Denis Bernard, and Benjamin Doyon. "Diffusion in generalized hydrodynamics and quasiparticle scattering". In: *arXiv:1812.00767* (2018).

[83] Jacopo De Nardis, Denis Bernard, and Benjamin Doyon. "Hydrodynamic Diffusion in Integrable Systems". In: *Phys. Rev. Lett.* 121 (16 Oct. 2018), p. 160603. DOI: 10.1103/PhysRevLett.121.160603. URL: https://link.aps.org/doi/10.1103/PhysRevLett.121.160603.

[84] Eyal Leviatan et al. "Quantum thermalization dynamics with matrix-product states". In: *arXiv:1702.08894* (2017).

[85] Christopher David White et al. "Quantum dynamics of thermalizing systems". In: *Phys. Rev. B* 97 (3 Jan. 2018), p. 035127. DOI: 10.1103/PhysRevB.97.035127. URL: https://link.aps.org/doi/10.1103/PhysRevB.97.035127.

[86] Johannes Hauschild et al. "Finding purifications with minimal entanglement". In: *Physical Review B* 98.23 (2018), p. 235163.

[87] Thomas Hartman, Sean A. Hartnoll, and Raghu Mahajan. "Upper Bound on Diffusivity". In: *Phys. Rev. Lett.* 119 (14 Oct. 2017), p. 141601. DOI: 10.1103/PhysRevLett.119.141601. URL: https://link.aps.org/doi/10.1103/PhysRevLett.119.141601.

[88] L. Faddeev. "Instructive History of the Quantum Inverse Scattering Method". In: *Quantum Field Theory: Perspective and Prospective.* Ed. by Cécile DeWitt-Morette and Jean-Bernard Zuber. Dordrecht: Springer Netherlands, 1999, pp. 161–177. ISBN: 978-94-011-4542-8. DOI: 10.1007/978-94-011-4542-8_8. URL: https://doi.org/10.1007/978-94-011-4542-8_8.

[89] N. Kitanine, J.M. Maillet, and V. Terras. "Form factors of the XXZ Heisenberg spin-12 finite chain". In: *Nuclear Physics B* 554.3 (1999), pp. 647–678. ISSN: 0550-3213. DOI: https://doi.org/10.1016/S0550-3213(99)00295-3. URL: http://www.sciencedirect.com/science/article/pii/S0550321399002953.

[90] J.M. Maillet and V. Terras. "On the quantum inverse scattering problem". In: *Nuclear Physics B* 575.3 (2000), pp. 627–644. ISSN: 0550-3213. DOI: https://doi.org/10.1016/S0550-3213(00)00097-3. URL: http://www.sciencedirect.com/science/article/pii/S0550321300000973.

[91] Hannes Bernien et al. "Probing many-body dynamics on a 51-atom quantum simulator". In: *Nature* 551.7682 (2017), pp. 579–584.

[92] C. J. Turner et al. "Weak ergodicity breaking from quantum many-body scars". In: *Nature Physics* 14.7 (2018), pp. 745–749. DOI: 10.1038/s41567-018-0137-5. URL: https://doi.org/10.1038/s41567-018-0137-5.

[93] Soonwon Choi et al. "Emergent SU(2) dynamics and perfect quantum many-body scars". In: *arXiv:1812.05561* (2018).

[94] Dmitry A Abanin et al. "Ergodicity, Entanglement and Many-Body Localization". In: *arXiv:1804.11065* (2018).

[95] Andrew Lucas. "Operator size at finite temperature and Planckian bounds on quantum dynamics". In: *arXiv:1809.07769* (2018).

[96] John Martyn and Brian Swingle. "Product Spectrum Ansatz and the Simplicity of Thermal States". In: *arXiv:1812.01015* (2018).

[97] Bruno Bertini, Pavel Kos, and T. Prosen. "Exact Spectral Form Factor in a Minimal Model of Many-Body Quantum Chaos". In: *Phys. Rev. Lett.* 121 (26 Dec. 2018), p. 264101. DOI: 10.1103/PhysRevLett.121.264101. URL: https://link.aps.org/doi/10.1103/PhysRevLett.121.264101.

[98] Terence Tao. *Topics in random matrix theory.* Vol. 132. American Mathematical Soc., 2012.

[99] Ramis Movassagh and Alan Edelman. "Density of States of Quantum Spin Systems from Isotropic Entanglement". In: *Phys. Rev. Lett.* 107 (9 Aug. 2011), p. 097205. DOI: 10.1103/PhysRevLett.107.097205. URL: https://link.aps.org/doi/10.1103/PhysRevLett.107.097205.

[100] A. Connes, H. Narnhofer, and W. Thirring. "Dynamical entropy of C^* algebras and von Neumann algebras". In: *Comm. Math. Phys.* 112.4 (1987), pp. 691–719. URL: `https://projecteuclid.org:443/euclid.cmp/1104160061`.

[101] R. Alicki and M. Fannes. "Defining quantum dynamical entropy". In: *Letters in Mathematical Physics* 32.1 (1994), pp. 75–82.

[102] Fabio Benatti. *Deterministic chaos in infinite quantum systems.* Springer Science & Business Media, 2012.

[103] T. Prosen. "Chaos and complexity of quantum motion". In: *Journal of Physics A: Mathematical and Theoretical* 40.28 (2007), p. 7881.

[104] Jonathan M Borwein, Peter B Borwein, and Karl Dilcher. "Pi, Euler numbers, and asymptotic expansions". In: *The American Mathematical Monthly* 96.8 (1989), pp. 681–687.

[105] N. J. A. Sloane. *The On-Line Encyclopedia of Integer Sequences.* Sequence A060338. 2018. URL: `https://oeis.org/A060338`.

[106] Jeroen Dehaene and Bart De Moor. "Clifford group, stabilizer states, and linear and quadratic operations over GF(2)". In: *Phys. Rev. A* 68 (4 Oct. 2003), p. 042318. DOI: `10.1103/PhysRevA.68.042318`. URL: `https://link.aps.org/doi/10.1103/PhysRevA.68.042318`.

[107] Andrew Hodges and CV Sukumar. "Bernoulli, Euler, permutations and quantum algebras". In: *Proc. Royal Soc. of London A.* Vol. 463. 2086. The Royal Society. 2007, pp. 2401–2414.

[108] CV Sukumar and Andrew Hodges. "Quantum algebras and parity-dependent spectra". In: *Proc. Royal Soc. of London A.* Vol. 463. 2086. The Royal Society. 2007, pp. 2415–2427.

[109] Gábor Hetyei. "Meixner polynomials of the second kind and quantum algebras representing su (1, 1)". In: *Proc. Royal Soc. of London A.* Vol. 466. 2117. The Royal Society. 2010, pp. 1409–1428.

[110] X Viennot. *Une théorie combinatoire des polynômes orthogonaux généraux, UQAM, Montréal, Québec.* 1983.

[111] N. J. A. Sloane. *The On-Line Encyclopedia of Integer Sequences.* Sequence A060338. 2018. URL: `https://oeis.org/A060338`.

[112] Mourad EH Ismail. "Classical and quantum orthogonal polynomials in one variable, with two chapters by Walter Van Assche, with a foreword by Richard A. Askey, reprint of the 2005 original". In: *Encyclopedia of Mathematics and its Applications* 98 (2009).

[113] Roelof Koekoek, Peter Lesky, and René Swarttouw. *Hypergeometric orthogonal polynomials and their q-analogues.* Springer, 2010.

[114] M B Hastings. "An area law for one-dimensional quantum systems". In: *Journal of Statistical Mechanics: Theory and Experiment* 2007.08 (Aug. 2007), P08024–P08024. DOI: `10.1088/1742-5468/2007/08/p08024`.

[115] Ulrich Schollwöck. "The density-matrix renormalization group in the age of matrix product states". In: *Annals of Physics* 326.1 (2011), pp. 96–192.

[116] Ian P McCulloch. "From density-matrix renormalization group to matrix product states". In: *Journal of Statistical Mechanics: Theory and Experiment* 2007.10 (2007), P10014.

[117] Johannes Hauschild and Frank Pollmann. "Efficient numerical simulations with Tensor Networks: Tensor Network Python (TeNPy)". In: *SciPost Phys. Lect. Notes* 5.10.21468 (2018).

[118] Laurens Vanderstraeten, Jutho Haegeman, and Frank Verstraete. "Tangent-space methods foGr uniform matrix product states". In: *SciPost Phys. Lect. Notes* (2019), p. 7. DOI: `10.21468/SciPostPhysLectNotes.7`. URL: `https://scipost.org/10.21468/SciPostPhysLectNotes.7`.

[119] Norbert Schuch et al. "Entropy scaling and simulability by matrix product states". In: *Physical review letters* 100.3 (2008), p. 030504.

[120] Frank Verstraete and J Ignacio Cirac. "Matrix product states represent ground states faithfully". In: *Physical Review B* 73.9 (2006), p. 094423.

[121] L Michel and IP McCulloch. "Schur forms of matrix product operators in the infinite limit". In: *arXiv:1008.4667* (2010).

[122] Garnet Kin-Lic Chan et al. "Matrix product operators, matrix product states, and ab initio density matrix renormalization group algorithms". In: *The Journal of chemical physics* 145.1 (2016), p. 014102.

[123] C Hubig, IP McCulloch, and U Schollwöck. "Generic construction of efficient matrix product operators". In: *Physical Review B* 95.3 (2017), p. 035129.

[124] Bogdan Pirvu et al. "Matrix product operator representations". In: *New Journal of Physics* 12.2 (2010), p. 025012.

[125] Michael P Zaletel et al. "Time-evolving a matrix product state with long-ranged interactions". In: *Physical Review B* 91.16 (2015), p. 165112.

[126] Sun-Yuan Kung. "A new identification and model reduction algorithm via singular value decomposition". In: *Proc. Twelfth Asilomar Conf. on Circuits, Systems and Computers* (1978).

[127] Gregory M. Crosswhite and Dave Bacon. "Finite automata for caching in matrix product algorithms". In: *Phys. Rev. A* 78 (1 July 2008), p. 012356. DOI: `10.1103/PhysRevA.78.012356`. URL: `https://link.aps.org/doi/10.1103/PhysRevA.78.012356`.

[128] Man-Duen Choi. “Completely positive linear maps on complex matrices”. In: *Linear algebra and its applications* 10.3 (1975), pp. 285–290.

[129] O. Bratteli and D. W. Robinson. *Operator algebras and quantum statistical mechanics.* Vol. Vol. 2. Springer, 1996.

[130] Leonard M Silverman and Maamar Bettayeb. “Optimal approximation of linear systems”. In: 17 (1980), p. 81.

[131] UBAIDM Al-Saggaf and GENEF Franklin. “An error bound for a discrete reduced order model of a linear multivariable system”. In: *IEEE transactions on Automatic Control* 32.9 (1987), pp. 815–819.

[132] Michael P. Zaletel et al. “Infinite density matrix renormalization group for multicomponent quantum Hall systems”. In: *Phys. Rev. B* 91 (4 Jan. 2015), p. 045115. DOI: 10.1103/PhysRevB.91.045115. URL: https://link.aps.org/doi/10.1103/PhysRevB.91.045115.

[133] Michael P Zaletel, Roger SK Mong, and Frank Pollmann. “Topological characterization of fractional quantum hall ground states from microscopic hamiltonians”. In: *Physical review letters* 110.23 (2013), p. 236801.

[134] JLF Barbón et al. “On the evolution of operator complexity beyond scrambling”. In: *Journal of High Energy Physics* 2019.10 (2019), p. 264.