


Local matrix product operators: Canonical form, compression, and control theory

Daniel E. Parker¹,* Xiangyu Cao¹,† and Michael P. Zaletel¹‡
 Department of Physics, University of California, Berkeley, California 94720, USA

 (Received 5 December 2019; revised 29 June 2020; accepted 30 June 2020; published 23 July 2020)

We present a new method for compressing matrix product operators (MPOs) which represent sums of local terms, such as Hamiltonians. Just as with area law states, such local operators may be fully specified with a small amount of information per site. Standard matrix product state (MPS) tools are ill-suited to this case, due to extensive Schmidt values that coexist with intensive ones, and Jordan blocks in the transfer matrix. We ameliorate these issues by introducing an “almost Schmidt decomposition” that respects locality. Our method is “ ϵ -close” to the accuracy of MPS-based methods for finite MPOs, and extends seamlessly to the thermodynamic limit, where MPS techniques are inapplicable. In the framework of control theory, our method generalizes Kung’s algorithm for model order reduction. Several examples are provided, including an all-MPO version of the operator recursion method (Lanczos algorithm) directly in the thermodynamic limit. All results are accompanied by practical algorithms, well-suited for the large MPOs that arise in DMRG for long-range or quasi-2D models.

DOI: [10.1103/PhysRevB.102.035147](https://doi.org/10.1103/PhysRevB.102.035147)

I. INTRODUCTION

While it is now well understood how matrix product states (MPS) can approximate 1d ground states [1–7], matrix-product representations of operators (MPOs) remain less understood. MPOs feature prominently in modern implementations of the density matrix renormalization group (DMRG) [2], yet we lack a complete understanding of the resources required for an MPO approximation of a complex (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 a MPO can formally be treated as a MPS in a doubled Hilbert space, this neglects the special structure of operators like Hamiltonians: they are a sum of local terms, $\hat{H} = \sum_j \hat{H}_j$, where \hat{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

$$\hat{H} = \hat{H}_L \otimes \hat{1}_R + \hat{1}_L \otimes \hat{H}_R + \sum_a h_{ab} \hat{h}_L^a \otimes \hat{h}_R^b, \quad (1)$$

where $\hat{h}_{L/R}^a$ 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 [8–12]. This idea was discussed in Ref. [9]. However, the coefficients h_{ab} , and the resulting singular value spectrum, depend on the choice of operators $\hat{h}_{L/R}^a$, 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 $\text{Tr}[\hat{h}_{L/R}^a \hat{h}_{L/R}^b] \propto \delta_{ab}$. The main result of this work is a 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. (1) (for this reason, naive SVD truncation of an MPO in the same manner as 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 noniterative compression algorithm specific to the type of iMPOs that occur in DMRG calculations, which exploits their upper triangular structure to

*daniel_parker@berkeley.edu

†xiangyu.cao@berkeley.edu

‡mikezaletel@berkeley.edu

efficiently handle MPOs with bond dimensions on the order of several thousands. 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 [13]. 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 work 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 Secs. II and III reviews standard facts about MPOs to set notation. We then provide all the concepts and algorithms needed for finite MPO compression in Sec. IV, along with a quick numerical example. We then transition to infinite MPOs, which require a somewhat more detailed and mathematical treatment. Section V specifies the class of “first degree” MPOs our method applies to, and shows their Jordan block structure is completely fixed by locality. Section VI is devoted to canonical forms and algorithms to compute them. We give the algorithm for compressing infinite MPOs in Sec. VII. Section VIII 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 IX goes on to reinterpret our compression algorithm within control theory. We provide a few examples of iMPO compression in Sec. X: compressing operators with long-ranged interactions and computing Lanczos coefficients for operator dynamics. We conclude in Sec. XI. Appendices prove statements from the main text and describe how all elementary algebra operations can be performed on MPOs.

II. 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

$$\begin{aligned} \widehat{H} &= \widehat{H}_L \otimes \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \otimes \widehat{H}_R + \sum_{a,b=1}^{\chi} M_{ab} \widehat{h}_L^a \otimes \widehat{h}_R^a \\ &= (\widehat{\mathbb{1}}_L \widehat{\mathbf{h}}_L \widehat{H}_L) \begin{pmatrix} 1 & & \\ & \mathbf{M} & \\ & & 1 \end{pmatrix} (\widehat{H}_R \widehat{\mathbf{h}}_R \widehat{\mathbb{1}}_R)^T, \end{aligned} \quad (2)$$

where we have introduced vectors of operators $\widehat{\mathbf{h}}_{L/R}$ on the left and right, and the matrix \mathbf{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 (2) be a Schmidt decomposition by ensuring that \mathbf{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

$$\widehat{H}_{e.g.} = \sum_{n=1}^N J \widehat{X}_n \widehat{X}_{n+1} + K \widehat{X}_n \widehat{Z}_{n+1} \widehat{X}_{n+2} + h \widehat{Z}_n, \quad (3)$$

where \widehat{X}_n and \widehat{Z}_n are operators acting on lattice site n . $H_{e.g.}$ 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 (nonuniquely) as

$$\begin{aligned} \widehat{\mathbf{h}}_L &= (\widehat{X}_n, \widehat{X}_n, \widehat{X}_{n-1} \widehat{Z}_n), \\ \widehat{\mathbf{h}}_R &= (\widehat{X}_{n+1}, \widehat{Z}_{n+1} \widehat{X}_{n+2}, \widehat{X}_{n+1}), \\ \mathbf{M} &= \text{diag}(J, K, K), \\ \widehat{H}_L &= \sum_{k=1}^n J \widehat{X}_{k-1} \widehat{X}_k + K \widehat{X}_{k-2} \widehat{Z}_{k-1} \widehat{X}_k + h \widehat{Z}_k, \end{aligned} \quad (4)$$

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 1. A local operator in regular form Eq. (2), is in left canonical form if

$$\langle \widehat{h}_L^a, \widehat{h}_L^b \rangle = \delta^{ab}, \quad 0 \leq a, b \leq \chi, \quad (5)$$

where $\langle \widehat{A}, \widehat{B} \rangle := \text{Tr}[\widehat{A}^\dagger \widehat{B}] / \text{Tr}[\widehat{\mathbb{1}}]$ is the inner-product for operators and $\widehat{h}_L^0 := \widehat{\mathbb{1}}_L$. 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 $\mathbf{M} = \mathbf{U} \mathbf{S} \mathbf{V}^\dagger$.

Definition 2. 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

$$\widehat{H} = \widehat{H}_L \otimes \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \otimes \widehat{H}_R + \sum_{a=1}^{\chi} s_a \widehat{h}_L^a \otimes \widehat{h}_R^a, \quad (6)$$

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

This is not a true Schmidt decomposition because we have excluded $\widehat{H}_{L/R}$; $\langle \widehat{h}_{L/R}^a, \widehat{H}_{L/R} \rangle$ is generically nonzero. 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. VIII). 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. (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.

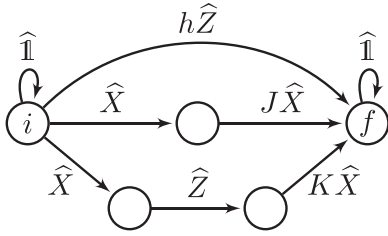


FIG. 1. A finite-state machine that generates Eq. (3).

III. 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 [2,14], which we illustrate with an example.

Consider $\widehat{H}_{\text{e.g.}}$ from Eq. (3) again. All of the Pauli strings needed to generate $\widehat{H}_{\text{e.g.}}$ can be described by a finite state machine, shown in Fig. 1. (We will see below this machine can be improved.) The MPO itself is the adjacency matrix of the finite state machine:

$$\widehat{W}_{\text{e.g.}} = \left(\begin{array}{ccc|c} \widehat{\mathbb{1}} & \widehat{X} & \widehat{X} & 0 \\ \hline & 0 & 0 & 0 \\ & 0 & 0 & \widehat{Z} \\ & 0 & 0 & 0 \\ \hline & & & \widehat{\mathbb{1}} \end{array} \right), \quad (7)$$

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

$$\widehat{H}_{\text{e.g.}} = \boldsymbol{\ell} \underbrace{\widehat{W}_{\text{e.g.}} \widehat{W}_{\text{e.g.}} \cdots \widehat{W}_{\text{e.g.}}}_{N \text{ matrices}} \boldsymbol{r}, \quad (8)$$

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 (8) is a matrix product in the *auxiliary space* and a tensor product in the *physical space*, such that physical indices of the n th matrix in (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

$$\widehat{H} = \boldsymbol{\ell} \widehat{W}^{(1)} \widehat{W}^{(2)} \cdots \widehat{W}^{(N)} \boldsymbol{r}. \quad (9)$$

where $\widehat{W}^{(1)}, \dots, \widehat{W}^{(N)}$ 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. (2) above. This property is encoded by the distinguished nodes i and f in the finite state machine Fig. 1, and is reflected by the block structure of the MPO (7). We therefore restrict ourselves to a special class of (i)MPOs which manifestly maintain this local structure.

Definition 3. An (i)MPO is in regular form if each matrix has the block upper triangular structure

$$\widehat{W} = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\boldsymbol{c}} & \widehat{\boldsymbol{d}} \\ 0 & \widehat{A} & \widehat{\boldsymbol{b}} \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}, \quad (10)$$

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

$$\boldsymbol{\ell} = (1 \ * \ *) , \ \boldsymbol{r}^\dagger = (* \ * \ 1) \quad (11)$$

where $*$ denotes an arbitrary block.

The shape of \widehat{W} in (10) 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 (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: $a, b, c \cdots = 0; 1, 2, \dots \chi; \chi + 1$.

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 E 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 $L^{(0)}, \dots, L^{(N)}$ that satisfy the interlacing conditions

$$\widehat{W}^{(n)'} L^{(n)} = L^{(n-1)} \widehat{W}^{(n)}, \ \boldsymbol{\ell}' L^{(0)} = \boldsymbol{\ell}, \ \boldsymbol{r}' = L^{(N)} \boldsymbol{r}. \quad (12)$$

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

$$\widehat{W}' L = L \widehat{W}. \quad (13)$$

To preserve the regular form (10), all gauge matrices must be block triangular,

$$L = \begin{pmatrix} 1 & \boldsymbol{t} & \boldsymbol{r} \\ 0 & \boldsymbol{L} & \boldsymbol{s} \\ 0 & 0 & 1 \end{pmatrix}. \quad (14)$$

Note that L need not be square, but only shaped to be compatible with (12) or (13).² In particular, \widehat{W}' and \widehat{W} may have different bond dimensions.

¹Structurally, $\widehat{\boldsymbol{d}}$ is a single operator, and $\widehat{\boldsymbol{c}}$ and $\widehat{\boldsymbol{b}}$ are operator-valued vectors.

²Some 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.

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

$$\widehat{W}'_{\text{e.g.}} = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{X} & 0 & h\widehat{Z} \\ & 0 & Z & J\widehat{X} \\ & 0 & 0 & K\widehat{X} \\ & & & \widehat{\mathbb{1}} \end{pmatrix}, \quad (15)$$

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

IV. 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.

A. 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 4. An MPO $\widehat{H} = \ell\widehat{W}^{(1)} \dots \widehat{W}^{(N)}\mathbf{r}$ is in left canonical form if, for each $n > 1$, the upper left block of $\widehat{W}^{(n)}$,

$$\widehat{V}^{(n)} := \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\mathbf{c}}^{(n)} \\ 0 & \widehat{\mathbf{A}}^{(n)} \end{pmatrix}, \quad (16)$$

has orthonormal columns:

$$\forall b, c \leq \chi^{(n)}, \sum_{a=0}^{\chi} \langle \widehat{W}_{ab}^{(n)}, \widehat{W}_{ac}^{(n)} \rangle = \delta_{bc}. \quad (17)$$

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

An MPO is in right canonical form if, and only if, its *mirror*³ is in left canonical form. Right canonical forms are always directly analogous, so we focus on the left-handed case.

Let us now see why left canonical MPOs describe left canonical operators, in the sense of definition 1.⁴ If we split an

³A MPO is mirrored by (I) transposing each matrix $\widehat{W}^{(n)}$, (II) exchanging $\ell^\dagger \leftrightarrow \mathbf{r}$, (III) reversing all auxiliary indices ($0 \leftrightarrow \chi + 1, 1, \dots, \chi \leftrightarrow \chi, \dots, 1$), and (IV) reversing the physical positions.

⁴Actually the two definitions are entirely equivalent, but we show only one implication for concision.

MPO in left canonical form at a bond n , then we can multiply the matrices together to put the operator into regular form:

$$\begin{aligned} \widehat{H}_W &= (\ell\widehat{W}^{(1)} \dots \widehat{W}^{(n)})(\widehat{W}^{(n+1)} \dots \widehat{W}^{(N)}\mathbf{r}) \\ &= (\widehat{\mathbb{1}}_L^{(n)} \widehat{\mathbf{h}}_L^{(n)} \widehat{H}_L^{(n)})(\widehat{H}_R \widehat{\mathbf{h}}_R \widehat{\mathbb{1}}_R)^T. \end{aligned} \quad (18)$$

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

$$(\widehat{\mathbb{1}}_L^{(n-1)} \widehat{\mathbf{h}}_L^{(n-1)})\widehat{V}^{(n)} = (\widehat{\mathbb{1}}_L^{(n)} \widehat{\mathbf{h}}_L^{(n)}). \quad (19)$$

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

$$\langle \widehat{h}_{L,a}, \widehat{h}_{L,b} \rangle = \delta_{ab}, \quad 0 \leq a, b \leq \chi^{(n)}, \quad (20)$$

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 $(\widehat{\mathbb{1}}_L, \widehat{\mathbf{h}}_L, \widehat{H}_L)$ are orthonormal—and imposes no constraint whatsoever on \widehat{H}_L . So MPO canonical form implies operator canonical form, definition 1.

Now that we have defined canonical forms for MPOs, our next task is compute them. One can always find a gauge transform, Eq. (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 (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

$$\widehat{V} = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\mathbf{c}} \\ 0 & \widehat{\mathbf{A}} \end{pmatrix} \stackrel{QR}{=} \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\mathbf{c}} \\ 0 & \widehat{\mathbf{A}} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{t} \\ 0 & \mathbf{R} \end{pmatrix}, \quad (21)$$

where \mathbf{R} is upper triangular.

Definition 5. Define the block-respecting \widehat{QR} decomposition of \widehat{W} as

$$\widehat{QR}[\widehat{W}] = \widehat{Q}\mathbf{R} \quad (22)$$

with

$$\widehat{Q} := \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{\mathbf{c}} & \widehat{\mathbf{d}} \\ 0 & \widehat{\mathbf{A}} & \widehat{\mathbf{b}} \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}, \quad \mathbf{R} := \begin{pmatrix} 1 & \mathbf{t} & 0 \\ 0 & \mathbf{R} & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (23)$$

where the upper left block comes from (21). Therefore \widehat{Q} is in left canonical form, and \mathbf{R} is upper triangular.

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

$$\ell\widehat{W}^{(1)}\widehat{W}^{(2)}\widehat{W}^{(3)} \dots \quad (24)$$

$$\stackrel{QR}{=} \ell[\widehat{Q}^{(1)}\mathbf{R}^{(1)}]\widehat{W}^{(2)}\widehat{W}^{(3)} \dots \quad (25)$$

$$= \ell\widehat{Q}^{(1)}[\mathbf{R}^{(1)}\widehat{W}^{(2)}]\widehat{W}^{(3)} \dots \quad (26)$$

$$\stackrel{QR}{=} \ell\widehat{Q}^{(1)}[\widehat{Q}^{(2)}\mathbf{R}^{(2)}]\widehat{W}^{(3)} \dots \quad (27)$$

$$= \ell\widehat{Q}^{(1)}\widehat{Q}^{(2)}[\mathbf{R}^{(2)}\widehat{W}^{(3)}] \dots \quad (28)$$

By the definition of the block QR decomposition, the first $1 + \chi^{(n)}$ columns of each $\widehat{Q}^{(n)}$ are indeed orthonormal.

Moreover, $\{R^{(1)}, \dots, R^{(N)}\}$ specifies a gauge transform from $\{\ell, W^{(n)}, \mathbf{r}\}$ to $\{\ell, \widehat{Q}^{(n)}, R^{(N)}\mathbf{r}\}$. We summarize the procedure as algorithm 1.

Algorithm 1. Left Canonical Form for finite MPOs.

```

1: Procedure MPOLEFTCAN  $\{\ell, \{\widehat{W}^{(n)}\}_{n=1}^N, \mathbf{r}\}$ 
2:    $R^{(0)} \leftarrow \ell$ 
3:   for  $n \in [1, N]$  do
4:      $(\widehat{Q}^{(n)}, R^{(n)}) \leftarrow \widehat{Q}R[R^{(n-1)}\widehat{W}^{(n)}]$  ▷ Eq. (21)
5:   return  $\{\ell, \{\widehat{Q}^{(n)}\}_{n=1}^N, R^{(N)}\mathbf{r}\}, \{R^{(n)}\}$ 

```

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 $\widehat{Q}R$ decomposition is used *in lieu* of normal QR.

B. 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

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

by the mirror of algorithm 1. To truncate at bond $(n, n+1)$, we first bring sites $[1, n]$ to left canonical form

$$\begin{aligned} & \ell \widehat{W}_R \widehat{W}_R \dots \widehat{W}_R \widehat{W}_R \dots \widehat{W}_R \mathbf{r} \\ = & \ell \widehat{W}_L R \widehat{W}_R \dots \widehat{W}_R \widehat{W}_R \dots \widehat{W}_R \mathbf{r} \\ & \vdots \\ = & \ell \underbrace{\widehat{W}_L \widehat{W}_L \dots \widehat{W}_L}_{\text{sites } [1, n]} R^{(n)} \underbrace{\widehat{W}_R \dots \widehat{W}_R}_{\text{sites } [n+1, N]} \mathbf{r}. \end{aligned}$$

(Superscripts have been suppressed for clarity.) The block structure of $R^{(n)}$ is fixed by block QR decomposition, Eq. (23), and we can always decompose it as⁵

$$R^{(n)} = MR', \quad M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & 1 \end{pmatrix} R' = \begin{pmatrix} 1 & \mathbf{t} & 0 \\ 0 & \text{Id}_\chi & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (30)$$

We then perform a singular value decomposition of M and write

$$M = USV^\dagger, \quad \mathbf{S} = \text{diag}(s_1 \geq s_2 \geq \dots \geq s_\chi), \quad (31)$$

where the middle blocks are unitary: $U^\dagger U = V^\dagger V = \text{Id}_\chi$. Combining (30) through (31), we obtain

$$\widehat{H}_W = \dots \widehat{W}_L^{(n-1)} \widehat{Q}^{(n)} S \widehat{P}^{(n+1)} \widehat{W}_R^{(n+1)} \dots \quad (32)$$

where

$$\widehat{Q}^{(n)} := \widehat{W}_L^{(n)} U, \quad \widehat{P}^{(n+1)} := V^\dagger R' \widehat{W}_R^{(n+1)} \quad (33)$$

are still left and right canonical, respectively.⁶ Therefore Eq. (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. (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.

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 Secs. VII and VIII below. We note that our compression scheme is ε -close to optimal, in a sense we make clear below.

C. 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 outscals the naive “MPS” technique because it contains only intensive values in the entanglement spectrum.

Algorithm 2. MPO Compression

```

1: procedure Compress  $\{\ell, \widehat{W}^{(n)}, \mathbf{r}\}, \eta$  ▷ Cutoff  $\eta$ 
2:    $\ell, \{\widehat{W}_R^{(n)}\}, \mathbf{r} \leftarrow \text{RIGHTCAN}[\ell, \{\widehat{W}^{(n)}\}, \mathbf{r}]$ 
3:    $R \leftarrow \ell$ 
4:   for  $n = 1, \dots, N-1$  do
5:      $(\widehat{W}_L^{(n)}, R) \leftarrow \widehat{Q}R[R\widehat{W}_R^{(n)}]$  ▷ Eq. (23)
6:      $(M, R') \leftarrow R$  ▷ Eq. (30)
7:      $(U, S, V^\dagger) \leftarrow \text{SVD}[M]$ 
8:      $\chi' \leftarrow \max\{a : s_a > \eta\}; I \leftarrow \{0, 1, \dots, \chi', \chi + 1\}$ .
9:      $\widehat{Q}^{(n)} \leftarrow [\widehat{W}_L^{(n)} U]_{0:\chi+1, I}$ 
10:     $R \leftarrow [V^\dagger R']_{I, 0:\chi+1}$ 
11:     $(\widehat{Q}^{(n)}, R) \leftarrow \widehat{Q}R[R\widehat{W}_R^{(n)}]$ 
12:  return  $\ell, \{\widehat{Q}^{(n)}\}, R\mathbf{r}$ 

```

It is well known that a two body interaction $V(i-j)\widehat{O}_i\widehat{O}_j$, 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 .⁷ Our algorithm will *automatically* discover this structure even if the MPO is initially presented in a nonoptimal form.

We therefore select a more challenging example with power-law interactions:

$$H_1 = \sum_{k,n,m=1}^N J_{kn} J_{nm} \widehat{Z}_k \widehat{Z}_n \widehat{Z}_m + J'_{nm} \widehat{Z}_n \widehat{Z}_m, \quad (34)$$

⁵Here and below, we use the short hand $\text{diag}(1, \mathbf{A}, 1) = A$ for block diagonal matrices, with sans-serif letters for the middle block.

⁶Right 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.

⁷See Eq. (42) for an example.

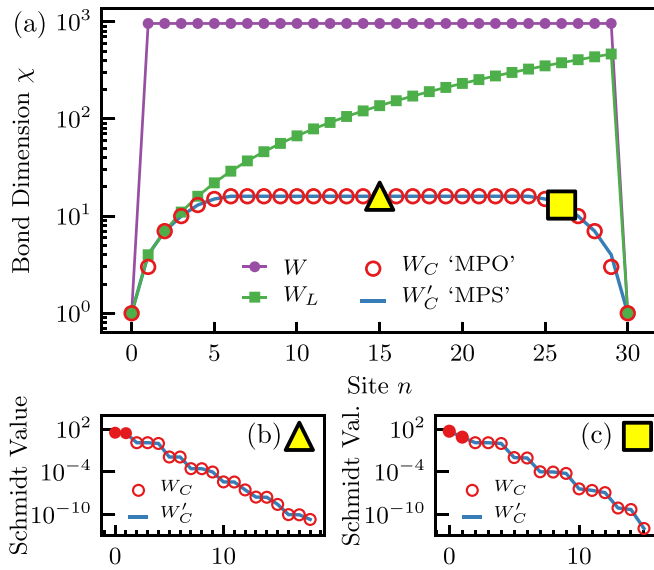


FIG. 2. Compression of a finite MPO representing the Hamiltonian (34). (a) The bond dimensions for: \widehat{W} , the naive MPO representation of H_1 ; \widehat{W}_L , the left canonical representation by algorithm 1; \widehat{W}_C , the compressed MPO by algorithm 2, and \widehat{W}'_C , the result of the standard MPS compression. [(b) and (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 $J_{mm} = |n - m|^{-2}$ and $J'_{mm} = |n - m|^{-4}$. In (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 [12]. The results are shown in Fig. 2.

The compression in Fig. 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.⁸ We return to this point in Sec. VIII below, when we discuss operator entanglement.

This concludes our discussion of compressing finite MPOs. We now move on to infinite matrix product operators.

V. LOCAL INFINITE MATRIX PRODUCT OPERATORS

We now transition to infinite matrix product operators. The discussion proceeds analogously to the finite case above.

⁸Such an precise match of the spectra holds only for simple Hamiltonians; in general, however, we have the interlacing relations (90).

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 nontrivial 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 by a constant amount of “information per site.” We will often call such operators “local as iMPOs” or simply “local.”

A. 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 $N \rightarrow \infty$. This is rooted in the iMPS transfer matrix, where a standard result [3] shows that the largest eigenvalue is nondegenerate 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

$$\|\widehat{H}\|_F^2 := \lim_{N \rightarrow \infty} \langle \widehat{H}_N, \widehat{H}_N \rangle = \lim_{N \rightarrow \infty} \frac{\text{Tr}[\widehat{H}_N^\dagger \widehat{H}_N]}{\text{Tr}[\mathbb{1}^N]}, \quad (35)$$

where the subscript “ F ” is a reminder that this is essentially the Frobenius norm.⁹ The norm is normalized so that $\|\mathbb{1}\|_F = 1$, unlike the usual Frobenius norm where the norm 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 Sec. VIII 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 static applications. We will therefore refer to (35) as *the* norm of an operator in this work.

⁹We note that this norm is not submultiplicative: see Appendix D.

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 $\mathcal{A} = \text{span}\{\widehat{O}_\alpha : 0 \leq \alpha < d\}$ (indexed by Greek letters α, β, \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, $\text{Tr}[\widehat{c}^\dagger \widehat{c}] = \text{Tr}[\widehat{n}^\dagger \widehat{n}] = 1$, 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

$$\widehat{W} = \sum_{\alpha} \widehat{O}_\alpha W_\alpha, \quad (W_\alpha)_{ab} := \langle \widehat{O}_\alpha, \widehat{W}_{ab} \rangle. \quad (36)$$

Definition 6. 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

$$T_W := \sum_{\alpha} \overline{W}_\alpha \otimes W_\alpha, \quad (37)$$

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

$$X T_W = \sum_{\alpha} W_\alpha^\dagger X W_\alpha, \quad (38)$$

where W_α^\dagger is the Hermitian conjugate as usual. By Choi's theorem [15], transfer matrix are always positive operators: whenever X is positive semidefinite, 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

$$\|\widehat{H}_N\|_F^2 = (\ell\ell)(T_W)^N(\mathbf{r}\mathbf{r}). \quad (39)$$

where $\ell\ell := \bar{\ell} \otimes \ell$ and $\mathbf{r}\mathbf{r} := \bar{\mathbf{r}} \otimes \mathbf{r}$.

The only way that (39) can give rise to an extensive norm, (35), is if the iMPO transfer matrix T_W (37) is dominated by some nontrivial Jordan block with eigenvalue 1.

To build intuition, we first consider the simple example

$$\widehat{H} = \sum_i \widehat{d}_i \text{ with } \widehat{W} = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{d} \\ 0 & \widehat{\mathbb{1}} \end{pmatrix}, \quad (40)$$

such that $\langle \widehat{\mathbb{1}}, \widehat{d} \rangle = 0$ and $\langle \widehat{d}, \widehat{d} \rangle = \rho$. Of course, $\|\widehat{H}_N\|_F^2 = N\rho$. Then the transfer matrix T_W is a 4×4 matrix

$$T_W = \begin{pmatrix} 1 & 0 & 0 & \rho \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \sim \left(\begin{array}{cc|cc} 1 & \rho & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array} \right), \quad (41)$$

where \sim denotes a similarity transform (but not a gauge transform). Taking powers T_W^N , (39) shows that the Jordan

block is clearly responsible for the extensive norm.¹⁰ This behavior should be generic; all local operators should have an extensive norm. However, not all iMPOs in regular form satisfy (35) because, even though such a Jordan block always exists, it may not dominate the norm (39) as $N \rightarrow \infty$. The remedy is to precisely define the what it means for iMPOs to be “local as an iMPO.”

B. First degree operators

This section will carefully define the class of first degree operators. Before giving the mathematical definition, let us provide some motivation.

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. 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 7. 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 (35). Specifically, we make a condition on the spectral properties of the \widehat{A} block of their iMPO representation.

Definition 8. Suppose \widehat{W} is an iMPO in regular form (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 .¹¹

The name “first degree” anticipates proposition 2, 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 (1) where the operators $h_{L/R}^a$ fall off with exponentially-localized tails.

By definition 8, 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

$$\widehat{H}_{\text{FD}} = \sum_i \sum_{k=0}^{\infty} \widehat{X}_i \left[\prod_{j=i+1}^{i+k} \alpha \widehat{Z}_j \right] \widehat{Y}_{i+k+1}, \quad (42)$$

¹⁰The other two blocks do not contribute to the extensive norm, but can become relevant when the operator has an extensive trace; see Appendix A for details.

¹¹We note our definition is closely akin to the idea of an “interaction” in the mathematical physics literature. See, e.g., Chap. 6 of Ref. [16].

whose iMPO representation is

$$\widehat{W}_{\text{FD}} = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{X} & 0 \\ 0 & \alpha \widehat{Z} & \widehat{Y} \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}. \quad (43)$$

The only eigenvalue of T_A is $|\alpha|^2$, so

$$\|\widehat{H}_{\text{FD},N}\|_F^2 \sim \begin{cases} N & |\alpha| < 1 \\ N^2 & |\alpha| = 1 \\ |\alpha|^{2N} & |\alpha| > 1 \end{cases} \quad (44)$$

so \widehat{W}_{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}_{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 1 in Appendix 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. In general, we expect all physical Hamiltonians are first degree operators, but not all first degree operators are physical Hamiltonians. For example, projectors are first degree operators which do *not* make sense as Hamiltonians. See Appendix E for an example of another non-Hamiltonian first degree operators.

A slight drawback of our broad definition 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 E) 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 Sec. X below).

C. 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 (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 Sec. VI below.

Proposition 1. Suppose that \widehat{W} is a first degree iMPO and consider its upper left block

$$\widehat{V} := \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{c} \\ 0 & \widehat{A} \end{pmatrix}. \quad (45)$$

Then the transfer matrix T_V has a unique dominant left eigenvalue of unity with an eigenvector X of the form

$$XT_V = X, \quad X = \begin{pmatrix} 1 & \mathbf{x} \\ \mathbf{x}^\dagger & \mathbf{X} \end{pmatrix}. \quad (46)$$

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.¹² Moreover, it is block upper triangular in that basis:

$$T_V = \begin{pmatrix} 1 & * & * & * \\ 0 & \bar{A}_0 & 0 & * \\ 0 & 0 & A_0 & * \\ 0 & 0 & 0 & T_A \end{pmatrix}, \quad A_0 = (\widehat{\mathbb{1}}, \widehat{A}), \quad (47)$$

so the eigenvalues of T_V are those of the diagonal blocks.

By first degreeeness, all eigenvalues λ of the T_A block have $|\lambda| < 1$. A technical linear algebra fact, lemma 2 from Appendix A, shows the same is true for the A_0 and \bar{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

$$\begin{pmatrix} 1 & \mathbf{c}_0 + \mathbf{x}A_0 \\ \mathbf{c}_0^\dagger + A_0^\dagger \mathbf{x}^\dagger & \sum_\alpha \mathbf{c}_\alpha^\dagger \mathbf{c}_\alpha + \mathbf{c}_\alpha^\dagger \mathbf{x}A_\alpha + A_\alpha^\dagger \mathbf{x}^\dagger \mathbf{c}_\alpha + A_\alpha^\dagger \mathbf{X}A_\alpha \end{pmatrix}. \quad (48)$$

So \mathbf{x} and \mathbf{X} are determined by

$$\mathbf{x}[I - A_0] = \mathbf{c}_0, \quad (49a)$$

$$\mathbf{X}[\text{Id} - T_A] = Q, \quad (49b)$$

$$Q := \sum_\alpha \mathbf{c}_\alpha^\dagger \mathbf{c}_\alpha + \mathbf{c}_\alpha^\dagger \mathbf{x}A_\alpha + A_\alpha^\dagger \mathbf{x}^\dagger \mathbf{c}_\alpha. \quad (49c)$$

As the eigenvalues λ of A_0 and T_A satisfy $|\lambda| < 1$, the operators on the left-hand sides of (49) are invertible and solutions \mathbf{x} and \mathbf{X} exist. The dominant eigenvalue therefore has the form (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 (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 Sec. IX 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. (35).

Proposition 2. Suppose \widehat{W} is a first degree iMPO for \widehat{H} with order-unity trace: $\text{tr}[\widehat{H}] = O(1)$. Then there is a vector \mathbf{z} such that the matrices

$$Z = \begin{pmatrix} X & \mathbf{z} \\ \mathbf{z}^\dagger & 0 \end{pmatrix}, \quad \text{and } Z' = \begin{pmatrix} 0 & \rho \\ 0 & 1 \end{pmatrix}, \quad (50)$$

[with the same X from Eq. (46)] span the dominant Jordan block of T_W :

$$(Z T_W \quad Z' T_W) = (Z \quad Z') \begin{pmatrix} 1 & \rho \\ 0 & 1 \end{pmatrix}, \quad (51)$$

for some real number $\rho \geq 0$. The norm, Eq. (35), is then extensive with $\|\widehat{H}_N\|_F^2 \rightarrow \rho N$ as $N \rightarrow \infty$.

This proposition is easily generalized to traceful operators at the cost of a more complex Jordan block structure. The proof, given in Appendix A, is similar to the one for proposition 1, but somewhat more technical.

¹²Schematically, $(\bar{\mathbb{1}} \oplus \bar{\chi}) \otimes (1 \oplus \chi) \cong (1 \oplus \chi \oplus \chi \oplus \chi^2)$.

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

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	\times	\checkmark	\checkmark
Closed under commutation.	\checkmark	\times	\checkmark
Canonical form (see Sec. VI)	\checkmark	\checkmark	\times

We note that X , z , and ρ can be calculated from \widehat{W} , but computational tractable formulas use canonical forms, and await us in Sec. VI. Intuitively, the reason for the extensive norm is that the overlaps of $\ell\ell$ with Z and rr with Z' are both 1, so if \widehat{W} is first degree then

$$\ell\ell T_W^N rr \sim (1 \ 0) \begin{pmatrix} 1 & \rho \\ 0 & 1 \end{pmatrix}^N \begin{pmatrix} 0 \\ 1 \end{pmatrix} = N\rho. \quad (52)$$

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 I 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.

VI. 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 nontrivial first degree requirement.

A. Existence of iMPO canonical forms

The definition of canonical form is much the same as in the finite case.

Definition 9. An iMPO \widehat{W} is in left canonical form if its upper left block \widehat{V} has orthonormal columns: $\forall b, c \leq \chi'$,

$$\sum_{a=0}^{\chi} \langle \widehat{W}_{ab}, \widehat{W}_{ac} \rangle = \delta_{bc}. \quad (53)$$

An iMPO is in right canonical form if its mirror is left canonical.

Definition 9, 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, (53) can be written in terms of the transfer matrix [defined in (38)] as

$$\text{Id}_{[0, \chi]} T_V = \sum_{\alpha} V_{\alpha}^{\dagger} V_{\alpha} = \text{Id}_{[0, \chi]}. \quad (54)$$

So \widehat{W} is left canonical whenever $\text{Id}_{[0, \chi]}$ is a left eigenvector of T_V with eigenvalue 1. This fact is exactly what allows us to prove that canonical forms exist.

Proposition 3. Let \widehat{W} be a first degree iMPO. Then there exists a matrix L that which specifies a gauge transform

$$\widehat{W}_L L = L \widehat{W} \quad (55)$$

so that \widehat{W}_L is left canonical.

The proof itself is given in Appendix B, but we briefly outline the idea. Proposition 1 tells us that, for any first degree \widehat{W} , the dominant eigenvector of T_V is $X T_V = X$. Suppose that we could take the “square root decomposition” $X = K^{\dagger} K$ with some invertible matrix K . Then we could enlarge K to $L = \text{diag}(K \ 1)$ and use it as a gauge transform $\widehat{W}_L = L \widehat{W} L^{-1}$. Such a \widehat{W}_L is left canonical:

$$\begin{aligned} \text{Id } T_{V_L} &= \sum_{\alpha} (K^{-1})^{\dagger} \widehat{V}_{\alpha}^{\dagger} K^{\dagger} K \widehat{V}_{\alpha} K^{-1} \\ &= (K^{-1})^{\dagger} X K^{-1} = \text{Id}, \end{aligned}$$

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 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

$$L_{lc} := \begin{pmatrix} 1 & & \\ & I & s \\ & & 1 \end{pmatrix}, \quad s := [A_0 - I]^{-1} \mathbf{b}_0. \quad (56)$$

(Also see lemma 4.) This will gauge away the identity components of the last so that

$$\langle \widehat{\mathbb{1}}, \widehat{d} \rangle = \langle \widehat{\mathbb{1}}, \widehat{\mathbf{b}}_a \rangle = 0, \quad \forall 1 \leq a \leq \chi, \quad (57)$$

Doing this makes the dominant Jordan block particularly simple.

Proposition 4. Suppose \widehat{W} is an iMPO for \widehat{H} in left canonical form where (57) holds. Then the dominant Jordan block of \widehat{W} is given by (50) and (51) with $X = \text{Id}_{[0, \chi]}$, $\mathbf{z}_a = \langle \widehat{A}_{ab}, \widehat{\mathbf{b}}_b \rangle +$

$\langle \widehat{c}_a, \widehat{d} \rangle$, and

$$\lim_{N \rightarrow \infty} \|H_N\|_F^2/N = \rho = \langle \widehat{d}, \widehat{d} \rangle + \sum_{a=1}^X \langle \widehat{b}_a, \widehat{b}_a \rangle. \quad (58)$$

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{b} by (A8), and applying (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., $0.3X_1Y_2Z_3 \rightarrow X_1Y_2[0.3Z_3]$.) The norm is then simply the sums of the squares of the weights of the ending sites. The condition (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

$$\|H_N\|_F^2 = \sum_{n=1}^N \left[\langle \widehat{d}^{(n)}, \widehat{d}^{(n)} \rangle + \sum_{a=1}^{\chi^{(n)}} \langle \widehat{b}_a^{(n)}, \widehat{b}_a^{(n)} \rangle \right]. \quad (59)$$

B. 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 $X_n \rightarrow X$ as $n \rightarrow \infty$. 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

$$\widehat{Q}_n R_n := \widehat{Q} R [\widehat{W}_{n-1}], \quad \widehat{W}_n := R_n \widehat{W}. \quad (60)$$

Let \widetilde{R}_n denote the restriction of $R_n = \text{diag}(\widetilde{R}_n, 1)$ to the upper left blocks (and similarly for \widetilde{Q}_n). We have

$$\sum_{\alpha} \widehat{V}_{\alpha}^{\dagger} \widetilde{R}_{n-1}^{\dagger} \widetilde{R}_{n-1} \widehat{V}_{\alpha} = \sum_{\alpha} \widetilde{R}_n^{\dagger} (\widetilde{Q}_n)_{\alpha}^{\dagger} (\widetilde{Q}_n)_{\alpha} \widetilde{R}_n, \quad (61)$$

so $(\widetilde{R}_{n-1}^{\dagger} \widetilde{R}_{n-1}) T_V = \widetilde{R}_n^{\dagger} \widetilde{R}_n$. This computes the application of the transfer matrix while maintaining the factorized form, giving the limit:

$$\widetilde{R}_n^{\dagger} \widetilde{R}_n = X_n \xrightarrow{n \rightarrow \infty} X = \widetilde{R}^{\dagger} \widetilde{R}. \quad (62)$$

One could then gauge-transform by $R = \text{diag}(\widetilde{R}, 1)$ as $\widehat{W}_L R = R W$ to find a left canonical \widehat{W}_L . We have now achieved (II).

Algorithm 3. iMPO Left Can. Form: Iterated QR

```

1: procedure LeftCanQRIter  $\widehat{W}, \eta$  ▷  $\eta$ : desired precision
2:    $L \leftarrow \text{Id}_{[0, \chi+1]}$ 
3:    $\varepsilon \leftarrow \infty$  ▷ Current error
4:   while  $\varepsilon > \eta$  do ▷ Repeat until convergence
5:      $(\widehat{Q}, R) \leftarrow \widehat{Q} R (\widehat{W})$  Eq. (21)
6:      $\widehat{W} \leftarrow R \widehat{Q}$ 
7:      $L \leftarrow R L$ 
8:      $\varepsilon \leftarrow \|R - \text{Id}\|$  if  $R$  is square else  $\infty$ 
9:   return  $\widehat{Q}, L$ 

```

The above procedure is no more than a simple adaption of a well-known standard method in the iMPS context [5], 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 Sec. X, Fig. 4 below). The essential problem is that convergence $X_n \rightarrow X$ does *not* guarantee $\widetilde{R}_n \rightarrow \widetilde{R}$, especially when X is a singular matrix. This is the main obstruction to achieving (III). Algorithm 3 presents the “practical solution” to this conundrum. The idea is to apply a gauge transformation after every QR step, i.e.,

$$\widehat{W}_0 = \widehat{Q}_1 R_1, \quad \widehat{W}_1 = R_1 \widehat{Q}_1, \quad \widehat{W}_1 = \widehat{Q}_2 R_2, \quad \widehat{W}_2 = R_2 \widehat{Q}_2 \dots$$

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}_1 by $R_2 R_1 \widehat{W}_1 = \widehat{W}_2 R_2 R_1$, etc. The desired gauge transform to a canonical form will be approached by the product $L_n = R_n R_{n-1} \dots R_1$. 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 $\chi_1 < \chi_0$.¹³ 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 B proves the conditions under which algorithm 3 converges, supplies nonconverging counterexamples, and a more complex algorithm which we prove *always* converges (algorithm 6). 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 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.

¹³This 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.

C. 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 noniterative 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 noniterative algorithm that uses an upper triangular solver to compute the gauge transform one column at a time.

Suppose we have an upper triangular MPO

$$\widehat{W}_{M-1} = \begin{pmatrix} \mathbb{1} & & & & \cdots \\ & \widehat{w}_1 & & & \cdots \\ & & \widehat{w}_2 & & \cdots \\ & & & \widehat{w}_3 & \cdots \\ & & & & \ddots \end{pmatrix} \quad (63)$$

and assume, for induction, that the first M column vectors $\widehat{w}_0, \dots, \widehat{w}_{M-1}$ are already orthonormal. We want to modify $\widehat{w}_M \rightarrow \widehat{w}'_M$ to be orthogonal to all previous columns. To do this, we apply a gauge transformation which is the identity except for the M th column:

$$R_M = \begin{pmatrix} 1 & 0 & & r_0 & & \\ & \ddots & & \vdots & & \\ & & 1 & r_{M-1} & & \\ & & & s_M & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}. \quad (64)$$

The transformation $\widehat{W}_M = R_M \widehat{W}_{M-1} R_M^{-1}$ is then easily computed¹⁴ and maintains the upper triangular form, while only affecting columns M and beyond. In particular, setting $s_M = 1$ temporarily,

$$\widehat{w}'_M = \widehat{w}_M - \sum_{a=0}^{M-1} r_a \widehat{w}_a + \sum_{a=0}^{M-1} r_a \widehat{d}_M e_a \quad (65)$$

where e_a is the standard basis vector $(e_a)_b = \delta_{ba}$ and $\widehat{d}_M := (\widehat{w}_M)_M = \widehat{W}_{MM}$ is the diagonal component of the M th column. In Gram-Schmidt, the last term is absent, and one would simply set $r_b = \langle \widehat{w}_b, \widehat{w}_M \rangle$ 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

$$0 \equiv \langle \widehat{w}_b, \widehat{w}'_M \rangle + \sum_{a=0}^{M-1} (-\langle \widehat{w}_b, \widehat{w}_a \rangle + \langle \widehat{w}_b, \widehat{d}_M e_a \rangle) r_a. \quad (66)$$

This is just a linear equation $Kr = c$ where

$$K_{ba} = \delta_{ba} - \langle \widehat{W}_{ba}^\dagger, \widehat{W}_{MM} \rangle, \quad (67a)$$

$$c_b = \langle \widehat{w}_b, \widehat{w}'_M \rangle, \quad (67b)$$

the Kronecker δ comes from the induction hypothesis $\langle \widehat{w}_b, \widehat{w}_a \rangle = \delta_{ba}$, and K is lower-triangular. Therefore we can easily solve for $r = K^{-1}c$ by back-substitution to find the r_b 's, giving an \widehat{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

$$\widehat{w}'_M \rightarrow \widehat{w}''_M = \frac{1}{s_M} (\widehat{w}'_M - \widehat{d}_M e_M) + \widehat{d}_M e_M, \quad (68)$$

The normalization condition $1 \equiv \langle \widehat{w}''_M, \widehat{w}''_M \rangle$ implies

$$s_M = \sqrt{\frac{\langle \widehat{w}_M, \widehat{w}_M \rangle}{1 - \langle \widehat{d}_M, \widehat{d}_M \rangle}}. \quad (69)$$

The first-order condition ensures the denominator is nonzero.

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

1: **procedure** LeftCanTriangular \widehat{W}

2: $R_T \leftarrow I_{1+\chi+1}$

3: **for** $M \in [1, \chi]$ **do**

4: $K_{ba} = \delta_{ba} - \langle \widehat{W}_{ab}^\dagger, \widehat{W}_{MM} \rangle$, $m, k \in [0, M-1]$

5: $c_b = \sum_{a=0}^{M-1} \langle \widehat{W}_{bM}^\dagger, \widehat{W}_{aM} \rangle$, $m \in [0, M-1]$

6: $r \leftarrow K^{-1}c$ $\triangleright O(\chi^2)$ triangular solve

7: $R \leftarrow \text{Id}_{1+\chi+1}$, $R_{bM} \leftarrow r_b$, $m \in [0, M-1]$

8: $\widehat{W} \leftarrow R\widehat{W}R^{-1}$, $R_T \leftarrow RR_T$ \triangleright only $O(\chi^2)$

9: $s \leftarrow \text{Eq. (69)}$

10: $R \leftarrow \text{Id}_{1+\chi+1}$, $R_{MM} \leftarrow s$

11: $\widehat{W} \leftarrow R\widehat{W}R^{-1}$, $R_T \leftarrow LR_T$ \triangleright only $O(\chi)$

12: **return** \widehat{W} , R_T

Several remarks are in order. First, this algorithm has an easily curable instability, which arises when s_M in (69) is vanishingly small. This means $\widehat{w}'_M - \widehat{d}_M e_M$ is also vanishing. Consequently, in terms of the state machine, the M th state cannot be reached from the initial state, so one should simply discard the M th row and column of \widehat{W} (as well as the M th row of the gauge matrix), and carry on.

Second, most upper triangular MPOs encountered in practice in DMRG have no diagonal components. In this case, \widehat{W} is often *strictly* upper triangular, whereupon $K = I$,

¹⁴The inverse R_M^{-1} has the same form as R_M but with $r_a \rightarrow -r_a$ and $s_M \rightarrow 1/s_M$.

the linear system becomes trivial, and the algorithm essentially reduces to normal Gram-Schmidt.¹⁵

Third, this algorithm is easily generalized to the case of extended unit cells, which occur frequently in applications to 2d DMRG. Let us sketch how this extension works. An iMPO with an extended unit cell with N sites is composed of repeating blocks $[\widehat{W}^{(1)}\widehat{W}^{(2)}\dots\widehat{W}^{(N)}]$. Gauge transforms are now collections of matrices R_1, \dots, R_N which satisfy intertwining relations $R_{n-1}\widehat{W}^{(n')} = \widehat{W}^{(n)}R_n$ for $n \in \mathbb{Z}/N\mathbb{Z}$. So we must now carry out the algorithm where the matrices R_n and R_{n-1} on each side are not the same. After a gauge transformation for column M using R_n 's with the same form as (64),

$$\widehat{w}_M^{(n')} = \widehat{w}_M^{(n)} - \sum_{a=0}^{M-1} r_a^{(n)} \widehat{w}_a + r_a^{(n-1)} \widehat{d}_M^{(n)} e_c. \quad (70)$$

The new orthogonality condition is the linear equation $\mathcal{M}\mathcal{R} = \mathcal{C}$ where \mathcal{M} is a $\chi \times \chi$ lower-triangular block matrix where each block is $N \times N$:

$$\mathcal{M} := \begin{pmatrix} M_{11} & & & \\ M_{21} & M_{22} & & \\ \vdots & \ddots & \ddots & \\ M_{\chi 1} & \dots & \dots & M_{\chi\chi} \end{pmatrix}, \quad (71a)$$

$$M_{ab} := \begin{pmatrix} 1 & & & M_{ab}^{(1)} \\ M_{ab}^{(2)} & \ddots & & \\ & \ddots & & 1 \\ & & M_{ab}^{(N-1)} & 1 \end{pmatrix}, \quad (71b)$$

$$\mathcal{C} := (\mathbf{c}_1 \mathbf{c}_2 \dots \mathbf{c}_\chi)^T, \quad (71c)$$

$$\mathbf{c}_a := (\mathbf{c}_a^{(1)} \quad \mathbf{c}_a^{(2)} \quad \dots \quad \mathbf{c}_a^{(N)})^T, \quad (71d)$$

$$M_{ba}^{(n)} := -\langle \widehat{W}_{ab}^{(n)}, \widehat{W}_{MM}^{(n)} \rangle, \quad (71e)$$

$$\mathbf{c}_a^{(n)} := \langle \widehat{w}_a^{(n)}, \widehat{w}_M^{(n)} \rangle. \quad (71f)$$

Again, when $\widehat{W}^{(n)}$ are all strictly upper triangular, the system is trivial and $\mathcal{R} = \mathcal{C}$. In general, however, this linear system is solvable in $O(N\chi^2)$ operations by exploiting the special structure of \mathcal{M} . Specifically, as each M_{ab} is almost tridiagonal, one may solve $M_{ab}\mathbf{x} = \mathbf{c}$ in $O(N)$ with a combination of forwards and backwards substitution.¹⁶ This allows \mathcal{M} to be solved by forwards substitution as

$$\mathcal{R} := (\mathbf{r}_1 \quad \mathbf{r}_2 \quad \dots \quad \mathbf{r}_\chi)^T, \quad (72a)$$

$$\mathbf{r}_a = M_{aa}^{-1} \left[\mathbf{c}_a - \sum_{b=1}^{a-1} M_{ab} \mathbf{r}_b \right]. \quad (72b)$$

¹⁵We caution that it is still necessary to row-transform the $j > M$ columns at each step, so the algorithm is distinct from a simple QR factorization.

¹⁶In particular, let α_k and β_k be such that $x_k = \alpha_k x_1 + \beta_k$. Put $(\alpha_1, \beta_1) := (1, 0)$ and recursively compute $\alpha_{k+1} = -(M_{ab}^{(k)}/1)\alpha_k$, $\beta_{k+1} = \beta_k + (c_k/1)$. Then $x_1 = \beta_{N+1}/(1 - \alpha_{N+1})$ and the other x_k 's follow from $x_k = \alpha_k x_1 + \beta_k$.

One can thus solve for \mathcal{R} in $O(N\chi^2)$ operations.

The other parts of the algorithm are simple to adapt, and the total cost to find the left canonical form is $O(N\chi^3)$, linear in the unit cell size N . This is a highly practical algorithm for compressing the iMPOs that appear in 2d DMRG.

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 B gives a yet-more-general algorithm, which is guaranteed to converge for *all* first degree iMPOs. We now proceed to compression of infinite MPOs which, unlike canonicalization, hews closely to the finite case.

VII. COMPRESSION OF iMPOs

We now explain how to compress infinite MPOs. The algorithm is directly analogous to the finite case: use canonical forms to make an almost-Schmidt decomposition of the operator, then truncate the almost-Schmidt values. Subsequently, Sec. VIII will show it is virtually optimal by bounding its error and Sec. IX 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 4, we may impose $c_0 = \langle \widehat{1}, \widehat{\mathbf{c}} \rangle = 0$ without loss of generality.¹⁷ There is then a gauge transform between right and left canonical form,

$$C\widehat{W}_R = \widehat{W}_L C, \quad (73)$$

and $c_0 = 0$ implies $C = \text{diag}(1 \quad \mathbf{C} \quad 1)$ 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 nonsquare iMPOs.) The SVD of $C = USV^\dagger$, now implies

$$USV^\dagger \widehat{W}_R = \widehat{W}_L USV^\dagger, \quad (74)$$

where U and V are unitary. Therefore we can use them to gauge transform $\widehat{W}_{L,R}$ into

$$\widehat{Q} := U^\dagger \widehat{W}_L U \quad \text{and} \quad \widehat{P} := V^\dagger \widehat{W}_R V, \quad (75)$$

which are left and right canonical, respectively. Furthermore, (74) implies that they are related by the gauge transform

$$\widehat{Q}S = S\widehat{P}. \quad (76)$$

Consequently, we obtain a mixed canonical form for the iMPO:

$$\begin{aligned} \widehat{H}_W &= \dots \widehat{W}_R \widehat{W}_R \widehat{W}_R \widehat{W}_R \dots \\ &= \dots \widehat{W}_L \widehat{W}_L C \widehat{W}_R \widehat{W}_R \dots \\ &= \dots \widehat{W}_L \widehat{W}_L USV^\dagger \widehat{W}_R \widehat{W}_R \dots \\ &= \dots \widehat{Q} \widehat{Q} S \widehat{P} \widehat{P} \dots \end{aligned} \quad (77)$$

¹⁷Actually we only need the t part and set $s = 0$. in Eq. (A8).

In the second line above, we inserted an L matrix at $-\infty$ and moved to the center using (73); in the fourth line, U and V are moved to $-\infty$ and $+\infty$, respectively.¹⁸

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 nonvanishing.¹⁹ Then

$$S = \mathbb{P}\mathbb{P}^\dagger S = S\mathbb{P}\mathbb{P}^\dagger = \mathbb{P}S'\widehat{P}^\dagger, \quad (78)$$

where \mathbb{P} is the projection matrix to the first χ' indices in the middle block

$$\mathbb{P}_{ab} = \begin{cases} \delta_{ab} & a \in \{0, 1, \dots, \chi', \chi + 1\} \\ 0 & \text{otherwise} \end{cases} \quad (79)$$

and $S' = \text{diag}(s_1, \dots, s_{\chi'})$. We can then use the fact that, in mixed canonical form, the position of S can be freely translated to any site using (76). We can then use (78) to “conjure” up projectors at every bond of (77):

$$H_W = \dots \widehat{Q}\widehat{Q}S\widehat{P}\widehat{P} \dots \quad (80)$$

$$= \dots \mathbb{P}\mathbb{P}^\dagger \widehat{Q}\mathbb{P}\mathbb{P}^\dagger \widehat{Q}\mathbb{P}S'\mathbb{P}^\dagger \widehat{P}\mathbb{P}\mathbb{P}^\dagger \widehat{P}\mathbb{P}\mathbb{P}^\dagger \dots$$

$$= \dots \widehat{Q}\widehat{Q}'S'\widehat{P}'\widehat{P} \dots, \quad (81)$$

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. VIII) but the procedure is unchanged. algorithm 5 gives an implementation, which we reiterate works also for nonsquare matrices.

Algorithm 5. iMPO Compression

1: procedure iCompress \widehat{W} , η	▷ Cutoff η
2: $\widehat{W}_R \leftarrow \text{RIGHTCAN}[\widehat{W}]$	
3: $\widehat{W}_R \leftarrow R\widehat{W}_R R^{-1}$ so that $\widehat{c}_0 = 0$	▷ Use t from Lem. (18)
4: $\widehat{W}_L, C \leftarrow \text{LEFTCAN}[\widehat{W}_R]$	
5: $(U, S, V^\dagger) \leftarrow \text{SVD}[C]$	
6: $\widehat{Q}, \widehat{P} \leftarrow U^\dagger \widehat{W}_L U, V^\dagger \widehat{W}_R V$	
7: $\chi' \leftarrow \max\{a \in [1, \chi] : s_a > \eta\}$	▷ Defines \mathbb{P} (79)
8: $\widehat{Q}, S, \widehat{P} \leftarrow \mathbb{P}^\dagger \widehat{Q}\mathbb{P}, \mathbb{P}^\dagger S\mathbb{P}, \mathbb{P}^\dagger \widehat{P}\mathbb{P}$	
9: return \widehat{P}	▷ One could also return \widehat{Q} .

VIII. 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

¹⁸These operations incur $O(1)$ errors near the boundary, which are negligible for an iMPO.

¹⁹In this case, the optimal compression error is zero, but the procedure itself is identical to the case where the singular values are numerically small.

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.

A. 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 (noninjective) 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 VI, 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

$$\widehat{H} = \sum_{a=-1}^{\chi} \lambda_a \widehat{O}_L^a \otimes \widehat{O}_R^a, \quad \text{Tr}[\widehat{O}_L^{a\dagger} \widehat{O}_L^b] = \delta^{ab} \quad (82)$$

(and the same for R), where the \widehat{O}_L 's and \widehat{O}_R 's act only on the left or right of the cut respectively. The Schmidt values $\lambda_{-1} \geq \lambda_0 \geq \dots \lambda_\chi > 0$ are unique and positive.²⁰ 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. (6), and the true Schmidt decomposition, Eq. (82). 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 2):

$$\begin{aligned} \widehat{H} &= \widehat{H}_L \otimes \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \otimes \widehat{H}_R + \sum_a s_a \widehat{h}_L^a \otimes \widehat{h}_R^a \\ &= (\widehat{\mathbb{1}}_L \widehat{h}_L \widehat{H}_L \begin{pmatrix} 1 & & \\ & \mathbf{S} & \\ & & 1 \end{pmatrix} (\widehat{H}_R \widehat{h}_R \widehat{\mathbb{1}}_R)^T, \end{aligned} \quad (83)$$

where $\mathbf{S} = \text{diag}(s_1 \geq \dots \geq s_\chi)$ is a diagonal matrix built from the almost-Schmidt values and $\{\widehat{\mathbb{1}}_{L/R}, \widehat{h}_{L/R}^1, \dots, \widehat{h}_{L/R}^\chi\}$ 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

²⁰The irregular index convention for the λ_a 's will prove convenient below.

orthonormalize. Explicitly, we apply a Gram-Schmidt update:

$$(\widehat{\mathbb{1}}_L \widehat{\mathbf{h}}_L \widehat{H}'_L) = (\widehat{\mathbb{1}}_L \widehat{\mathbf{h}}_L \widehat{H}'_L) \begin{pmatrix} 1 & 0 & 0 \\ 0 & \text{Id} & \mathbf{p}_L \\ 0 & 0 & \mathcal{N}_L \end{pmatrix}, \quad (84)$$

where $\mathbf{p}_L^a := \langle \widehat{h}_L^a, \widehat{H}'_L \rangle$ ensures orthogonality and $\mathcal{N}_L := \|\widehat{H}'_L\|_F^2 - \|\mathbf{p}_L\|_F^2$ enforces normalization, so that $\{\widehat{\mathbb{1}}_L, \widehat{h}_L^1, \dots, \widehat{h}_L^\chi, \widehat{H}'_L\}$ are now orthonormal. Doing the same on the right side, the operator now becomes

$$\widehat{H} = (\widehat{\mathbb{1}}_L \widehat{\mathbf{h}}_L \widehat{H}'_L) \underbrace{\begin{pmatrix} \mathcal{N}_R & \mathbf{p}_R & 0 \\ 0 & \mathbf{S} & \mathbf{p}_L \\ 0 & 0 & \mathcal{N}_L \end{pmatrix}}_{M:=} (\widehat{H}'_R \widehat{\mathbf{h}}_R \widehat{\mathbb{1}}_R)^T. \quad (85)$$

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 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:

$$\mathcal{N}_L, \mathcal{N}_R \in \Theta(N), \quad s_a, \mathbf{p}_L, \mathbf{p}_R \in O(1). \quad (86)$$

Without the \mathbf{p} 's M would be diagonal. There would then be two extensive singular values, namely \mathcal{N}_L^2 and \mathcal{N}_R^2 , and χ intensive ones, s_1^2, \dots, s_χ^2 . Appendix C shows that the extensive/intensive separation remains after the \mathbf{p} 's have been taken into account:

$$\lambda_{-1}^2, \lambda_0^2 \in \Theta(N), \quad \lambda_a \in O(1), \quad a = 1, \dots, \chi. \quad (87)$$

This result illustrates again why the MPS compression scheme must fail with iMPOs: the extensive Schmidt values diverge in the thermodynamic limit. Normalizing the Schmidt values, that is, considering $\sigma_a := \lambda_a / \sqrt{\sum_b \lambda_b^2}$, would not be helpful: for any $a > 0$, $\sigma_a \in O(1/N)$ 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.

B. 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

$$\widehat{H}_{\text{MPS}} = \sum_{a=-1}^{\chi'} \lambda_a \widehat{O}_L^a \otimes \widehat{O}_R^a,$$

$$\widehat{H}_{\text{MPO}} = \widehat{H}_L \otimes \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \otimes \widehat{H}_R + \sum_{a=1}^{\chi'} s_a \widehat{h}_L^a \otimes \widehat{h}_R^a,$$

respectively. The orthogonality properties of the decompositions tell us

$$\|\widehat{H} - \widehat{H}_{\text{MPS}}\|_F^2 = \sum_{a=\chi'+1}^{\chi} \lambda_a^2 := \varepsilon_{\text{MPS}}(\chi'), \quad (88)$$

$$\|\widehat{H} - \widehat{H}_{\text{MPO}}\|_F^2 = \sum_{a=\chi'+1}^{\chi} s_a^2 := \varepsilon_{\text{MPO}}(\chi'). \quad (89)$$

To compare these, we use the eigenvalue interlacing relation (derived in Appendix C)

$$s_a \geq \lambda_a \geq s_{a+2}, \quad \forall a \in [1, \chi - 2]. \quad (90)$$

We can therefore conclude

$$\varepsilon_{\text{MPS}}(\chi') \leq \varepsilon_{\text{MPO}}(\chi') \leq \varepsilon_{\text{MPS}}(\chi' - 2). \quad (91)$$

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 [7], 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.

C. Ground state error bound

We conclude this section by discussing the special case of Hamiltonians. We will show that the change in the ground state of a Hamiltonian under compression is small and, just as one would expect, the error is proportional to the weight of the truncated singular values.

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

$$\|\widehat{O}\|_s^2 := \sup_{|\Psi\rangle} \frac{\langle \Psi | \widehat{O} \widehat{O} | \Psi \rangle}{\langle \Psi | \Psi \rangle}, \quad (92)$$

or, perhaps more physically, the ground state energy. Our task for this section is to show the change in the ground state energy is also small under compression²¹.

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 state energies are *not* extensive. If one feeds in a 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. However, 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.²² This allows us to give the following bound.

Proposition 5. 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

$$\widehat{H} = \widehat{H}_L \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \widehat{H}_R + \sum_{a,b=1}^{\chi} \widehat{h}_L^a M_{ab} \widehat{h}_R^b, \quad (93)$$

where each \widehat{h}_s^a is a unique tensor product of on-site operators (such as a Pauli string $XYXZ$ or $\widehat{c}^\dagger \widehat{c} \widehat{c}^\dagger \widehat{c}$ 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

$$\widehat{H}' := \widehat{H}_L \widehat{\mathbb{1}}_R + \widehat{\mathbb{1}}_L \widehat{H}_R + \sum_{a=1}^{\chi'} \widehat{O}_L^a s_a \widehat{O}_R^a, \quad (94)$$

where $\{s_1 \geq s_2 \geq \dots \geq s_{\chi'}\}$ are the singular values. Then the change in the ground state energy δE satisfies

$$\delta E \leq \|\widehat{H} - \widehat{H}'\|_s \leq \sqrt{d^k \sum_{a=\chi'}^{\chi} s_a^2} \leq d^{\frac{k}{2}} \|\widehat{H} - \widehat{H}'\|_F. \quad (95)$$

The proof is given in Appendix D.

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

²¹We note that small changes to the Hamiltonian can cause dramatic changes to the ground state wavefunction. 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. Far from phase boundaries, the ground state wavefunction and its expectation values should change continuously with the Hamiltonian.

²²Similar bounds apply to broader classes of Hamiltonians, but require greater technical complexity.

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.

IX. 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. [11, 12]. 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 $\mathbf{x}(t)$ with linear dynamics in discrete time. The dynamics are defined by the update rule

$$\begin{aligned} \mathbf{x}(t) &= A\mathbf{x}(t-1) + B\mathbf{u}(t), \\ \mathbf{y}(t) &= C\mathbf{x}(t) + D\mathbf{u}(t), \end{aligned} \quad (96)$$

where $\mathbf{u}(t)$ is an n_i -dimensional vector of possible “input” perturbations, $\mathbf{y}(t)$ 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, $G(t) := CA^t B$, 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 modeling a complex electrical circuit, where $\mathbf{x}(t)$ 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 a MPO in the particular case of a *two-body* Hamiltonian. A

two-body interaction takes the general form

$$\widehat{H} = \sum_{x>y} \sum_{\alpha, \beta=1}^{n_o, n_i} \widehat{O}_x^\alpha V^{\alpha\beta} (x-y) \widehat{P}_y^\beta, \quad (97)$$

where $\{\widehat{\mathbb{1}}\} \cup \{\widehat{O}_x^\alpha\}_{\alpha=1}^{n_o}$ 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 (96) define an MPO in regular form via

$$\widehat{c} = \widehat{O}C, \quad \widehat{b} = \widehat{B}P, \quad \widehat{A} = A\widehat{\mathbb{1}}, \quad \widehat{d} = 0,$$

where $\widehat{O} = (\widehat{O}_x^\alpha)_{\alpha=1}^{n_o}$ and $\widehat{P} = (\widehat{P}_y^\beta)_{\beta=1}^{n_i}$. It is not hard to check this MPO represents the Hamiltonian (97) if and only if

$$[CA^\dagger B]^{ab} = V^{ab}(r).$$

These data are 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 nonzero.

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 [13], and (II) by ‘‘balanced truncation’’ [17], which we focus on here.

The starting point of the balanced truncation algorithm is the ‘‘controllability’’ Gramian X and the ‘‘observability’’ Gramian Y ,

$$X \equiv \sum_{k=0}^{\infty} A^k B B^\dagger (A^\dagger)^k, \quad (98a)$$

$$Y \equiv \sum_{k=0}^{\infty} (A^\dagger)^k C^\dagger C A^k. \quad (98b)$$

They are determined by the discrete Lyapunov equations

$$AXA^\dagger = X - BB^\dagger, \quad (99)$$

$$A^\dagger YA = Y - C^\dagger C. \quad (100)$$

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 $\widehat{A} = A\widehat{\mathbb{1}}$. 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 $A \rightarrow gAg^{-1}$, $C \rightarrow Cg^{-1}$, $B \rightarrow gB$, under which the Gramians transform as $X \rightarrow gXg^\dagger$, $Y \rightarrow g^\dagger Y g^{-1}$, to demand that the Gramians be equal and diagonal: $X = Y = \text{diag}(\Sigma)$. 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. (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’’ [18].

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 $\widehat{h}_L^i = \widehat{n}_{-i}$, $\widehat{h}_R^i = \widehat{n}_{i+1}$ for $i \geq 0$ so that $H = \sum_{i,j} \widehat{h}_L^i V(i+j+1) \widehat{h}_R^j$. The middle tensor then takes the form

$$M = \begin{pmatrix} \dots & V(3) & V(2) & V(1) \\ \dots & V(4) & V(3) & V(2) \\ \dots & V(5) & V(4) & V(3) \\ \vdots & & & \vdots \end{pmatrix}, \quad (101)$$

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 SLITCOT library Ref. [19] as AB09AD. The latter has a convenient PYTHON API provided in the ‘‘CONTROL’’ library Ref. [20], which we have used with great success for quantum Hall DMRG [21].

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.

X. 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

$$\widehat{H}_2 = \sum_{n \in \mathbb{Z}} \sum_{x, y > 0} \widehat{Z}_{n-x} \widehat{X}_n \widehat{Z}_{n+y} J_x J_y, \quad J_r = r^{-2}, \quad (102)$$

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 $r > R$, which we vary, so that the pre-compression bond dimension is $\chi = 2R$. The precompressed iMPOs have a block structure specific to three-body interaction; for example, when $R = 3$,

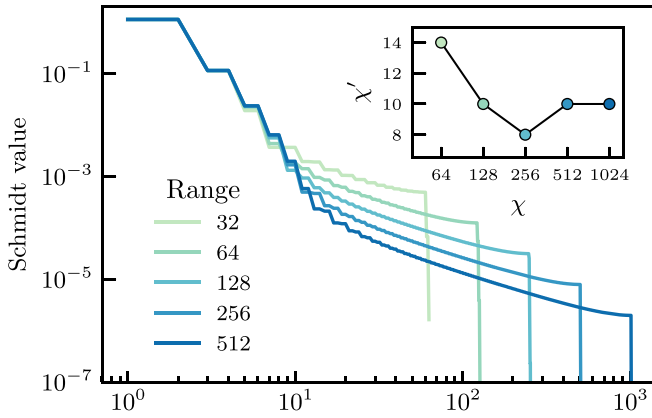


FIG. 3. Compression of the iMPO representing (102). The almost-Schmidt spectra of iMPOs representing \hat{H}_2 with spatial cutoff R ranging from 32 to 512. As $R \rightarrow \infty$, the largest s_n converge to a pointwise limit, while the long tails rapidly decays (so the latter are finite- R artifacts). (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. 2.

we have

$$\left(\begin{array}{ccc|ccc|} \widehat{\mathbb{1}} & \widehat{Z} & 0 & 0 & & & \\ \hline & 0 & \widehat{\mathbb{1}} & 0 & J_1 \widehat{X} & 0 & 0 \\ & 0 & 0 & \widehat{\mathbb{1}} & J_2 \widehat{X} & 0 & 0 \\ & 0 & 0 & 0 & J_3 \widehat{X} & 0 & 0 \\ \hline & & & & 0 & \widehat{\mathbb{1}} & 0 \\ & & & & 0 & 0 & \widehat{\mathbb{1}} \\ & & & & 0 & 0 & 0 \\ \hline & & & & & & J_1 \widehat{Z} \\ & & & & & & J_2 \widehat{Z} \\ & & & & & & J_3 \widehat{Z} \\ & & & & & & \widehat{\mathbb{1}} \end{array} \right). \quad (103)$$

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 Fig. 3. For any reasonable tolerance, as $R \rightarrow \infty$, 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 $\chi' = 4$):

$$\left(\begin{array}{ccc|ccc|} \widehat{\mathbb{1}} & \widehat{Z} & \widehat{Z} & & & \\ \hline & 0.178\widehat{\mathbb{1}} & & 0.749\widehat{X} & 0.114\widehat{X} & \\ & & 0.742\widehat{\mathbb{1}} & 0.11\widehat{X} & 0.0117\widehat{X} & \\ \hline & & & 0.178\widehat{\mathbb{1}} & & \widehat{Z} \\ & & & & 0.742\widehat{\mathbb{1}} & \widehat{Z} \\ \hline & & & & & \widehat{\mathbb{1}} \end{array} \right). \quad (104)$$

Remarkably, while the strict locality of the uncompressed MPO is compromised, the block triangular structure of (103) 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 (104). 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 tridiagonalize a matrix. However, it was recognized in the 1980s [22] 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 Ref. [23] for a review.) Recent work by some of us [24] has found there are deep connections between the Lanczos algorithm, thermalization, operator complexity, and quantum chaos.

The Lanczos algorithm is a simple iteration. Suppose \hat{H} is a Hamiltonian and \hat{O} is a Hermitian operator. Conceptually, the Lanczos algorithm constructs the Krylov subspace $\text{span}\{\hat{O}, [\hat{H}, \hat{O}], [\hat{H}, [\hat{H}, \hat{O}]], \dots\}$ and iteratively orthonormalizes it. More precisely, we start from $\hat{O}_{-1} = 0$, $\hat{O}_0 := \hat{O}$, $b_0 := 0$, and for $n > 0$, we define recursively

$$\begin{aligned} \widehat{A}_n &:= [\hat{H}, \widehat{O}_n] - b_{n-1}\widehat{O}_{n-2} \\ \widehat{O}_n &:= b_n^{-1}\widehat{A}_n \text{ where } b_n := \|\widehat{A}_n\|^{1/2}. \end{aligned} \quad (105a)$$

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 $\widehat{O}(t) = e^{i\hat{H}t}\widehat{O}e^{-i\hat{H}t}$, and it is desirable to compute as many of them as possible.

For generic many-body problems, exactly computing n Lanczos coefficients requires $O(e^{Cn})$ resources. Now, whenever \hat{H} and \hat{O} are representable as iMPOs, the whole Lanczos algorithm can be implemented using iMPOs using elementary operations from Appendix E and the intensive norm formula (58). If \widehat{O}_0 is first degree and \hat{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. Remarkably, we observe that the resulting bond dimension of the operators \widehat{O}_n grows only polynomially:

$$\chi[\widehat{O}_n] = O(n^a), \quad a \approx 2, \quad (106)$$

shown in Fig. 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 [24].

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 steps. One can see from Figs. 4(d) and 4(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(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 (106) continues to hold. It will be very interesting to elucidate the reason of such an advantageous bond dimension scaling.

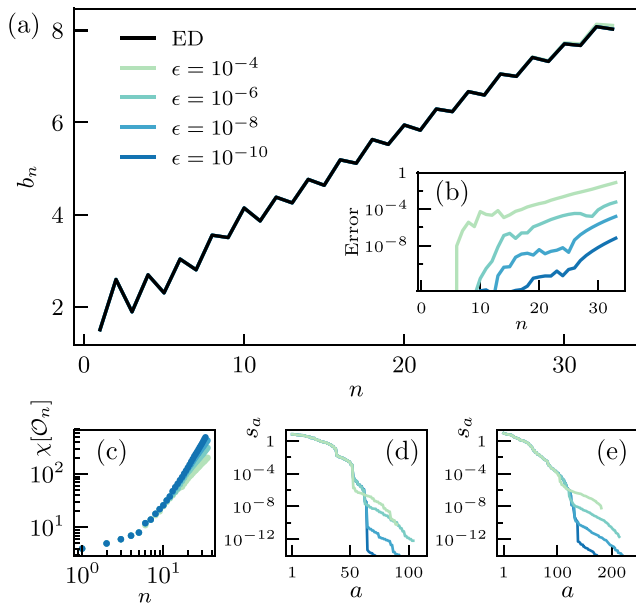


FIG. 4. Results of an iMPO implementation of the Lanczos algorithm, applied to $\hat{H} = \frac{1}{2} \sum_n Z_n Z_{n+1} - 1.05Z_n + 0.5X_n$ and $\hat{O} = \sum_n Z_n$. (a) The Lanczos coefficients b_n computed by the iMPO implementation with SVD truncation threshold ϵ , compared to the exact method (“ED”) of Ref. [24]. (b) Error in the b_n ’s at precision ϵ (compared to “ED” values). (c) Bond dimension of the operators \hat{O}_n . The growth rate is roughly $O(n^2)$. (d) The almost Schmidt spectra of \hat{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 \hat{O}_{20} . The gap is barely visible even with the smallest ϵ ; the error starts to grow rapidly around the same n .

XI. CONCLUSIONS

In this work, we have endeavored to promote matrix-product operators to “first-class citizens” among 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 [23,24]. 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

$$\hat{W}(k) = \begin{pmatrix} e^{ik} \hat{\mathbb{1}} & \hat{c} & \hat{d} \\ 0 & \hat{A} & \hat{b} \\ 0 & 0 & \hat{\mathbb{1}} \end{pmatrix}. \quad (107)$$

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 analog 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 nontrivial 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 “multilocal” operators that arise naturally as products such as $\hat{H}\hat{H}$ (used in computing energy fluctuations in DMRG [10])? 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.

ACKNOWLEDGMENTS

We thank Ehud Altman, Nick Bultinck, Joel E. Moore, Johannes Motruk, Frank Verstraete, and Laurens Vanderstraeten for helpful discussions. We thank the second anonymous referee whose careful comments greatly improved this work. We acknowledge support from the NSF Graduate Research Fellowship Program Grant No. NSF DGE 1752814 (D.E.P.), ERC synergy Grant UQUAM and DOE Grant No. DE-SC001938 (X.C.). M.P.Z. was supported by the DOE, office of Basic Energy Sciences under Contract No. DEAC02-05-CH11231.

APPENDIX A: PROOFS FOR LOCAL MPOs

This Appendix proves statements about local MPOs from Sec. V of the main text. Our main goal is the proof of the forms of the dominant Jordan blocks, proposition 1, but we begin with a series of technical lemmas.

Lemma 1. 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 (14) of the gauge matrix L implies the submatrices \widehat{A} and \widehat{A}' are related by $L\widehat{A} = \widehat{A}'L$. Then, by the definition of the transfer matrix, we have

$$\begin{aligned} [L^\dagger X L]T_A &= \sum_{\alpha} A_{\alpha}^{\dagger} L^{\dagger} X L A_{\alpha} \\ &= \sum_{\alpha} L^{\dagger} (A'_{\alpha})^{\dagger} X A'_{\alpha} L = L^{\dagger} (X T_{A'}) L. \end{aligned} \quad (\text{A1})$$

Now, suppose \widehat{W}' is not first degree, then there is X such that $X T_{A'} = \lambda X$ with $|\lambda| \geq 1$. By (A1), $Y := L^{\dagger} X L$ is an eigenvector of T_A with the same λ , which contradicts the first degree property of \widehat{W} . ■

Lemma 2. Suppose $\text{spec}(T_A)$ is strictly inside the unit disk. Then so is $\text{spec}(A_0)$.

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

$$J = \begin{pmatrix} \sigma & 1 & & \\ & \ddots & \ddots & \\ & & \sigma & 1 \\ & & & \sigma \end{pmatrix} \quad (\text{A2})$$

with some (generalized) eigenvector $A_0 \mathbf{v} = \sigma \mathbf{v}$. Then $\mathbf{w} := \mathbf{v}^{\dagger} \otimes \mathbf{v}$ is an eigenvector $T_{A_0} \mathbf{w} = |\sigma|^2 \mathbf{w}$. So

$$\langle \mathbf{w}, T_{A_0}^N \mathbf{w} \rangle = |\sigma|^{2N} \geq 1, \quad \forall N. \quad (\text{A3})$$

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

$$T_A^N = T_0^N + \sum_{\substack{\alpha_1, \dots, \alpha_N \\ \exists \alpha_i \neq 0}} T_{\alpha_1} \cdots T_{\alpha_N}, \quad (\text{A4})$$

Since the composition of positive maps is positive, $\langle \mathbf{w}, T_{\alpha_1} \cdots T_{\alpha_N} \mathbf{w} \rangle \geq 0$. So (A4) implies $\langle \mathbf{w}, T_A^N \mathbf{w} \rangle \geq |\sigma|^{2N} \rightarrow \infty$. But all the eigenvalues of T_A are less than 1, so $\langle \mathbf{w}, T_A^N \mathbf{w} \rangle \rightarrow 0$, a contradiction. ■

Lemma 3. Let

$$T = \begin{pmatrix} A & B \\ 0 & C \end{pmatrix} \quad (\text{A5})$$

be a block upper triangular matrix such that A and C are square matrices. Let $\lambda \in \text{spec}(A) \setminus \text{spec}(C)$ and $(x \ y)T = \lambda(x \ y)$ be a left eigenvector. Then $x \neq 0$ and satisfies $x A = \lambda x$, so x is a left eigenvector of A .

Proof. $(x \ y)T = \lambda(x \ y)$ means $x A = \lambda x$ and $x B + y C = \lambda y$. Now suppose $x = 0$. Then $y \neq 0$, and $y C = \lambda y$, so $\lambda \in \text{spec}(C)$, a contradiction. ■

Lemma 4. Suppose \widehat{W} is a first degree iMPO. Then there exists a gauge transform²³

$$\widehat{W}' = L \widehat{W} L^{-1} \text{ where } L = \begin{pmatrix} 1 & \mathbf{t} & 0 \\ 0 & \text{Id} & \mathbf{s} \\ 0 & 0 & 1 \end{pmatrix} \quad (\text{A6})$$

such that

$$W'_0 = \langle \widehat{\mathbb{1}}, \widehat{W}' \rangle = \begin{pmatrix} 1 & 0 & d'_0 \\ 0 & A_0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (\text{A7})$$

In fact, $\widehat{A}' = \widehat{A}$ is unchanged.

Proof. A direct computation shows

$$\mathbf{s} := (A_0 - \text{Id})^{-1} \mathbf{b}_0, \quad \mathbf{t} := \mathbf{c}_0 (A_0 - \text{Id})^{-1} \quad (\text{A8})$$

give the desired gauge transform. The inverse $(A_0 - \text{Id})^{-1}$ exists by lemma 2. ■

We now have all the tools needed to unravel the Jordan block structure of MPOs. We prove proposition 2 in the special case where the operator has subextensive trace, and subsequently sketch the more general case.

Proof of proposition 2. 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 (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 4, 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

$$\begin{aligned} \text{Tr}[\widehat{H}_N] &= \ell W_0^N \mathbf{r} = (\ell' \ell_{\chi+1}) W_0^N (r_0 \mathbf{r}' \ 1)^T \\ &= r_0 + \ell_{\chi+1} + N d_0 + \ell' A_0^N \mathbf{r}' \\ &= N d_0 + O(1), \quad N \rightarrow \infty, \end{aligned} \quad (\text{A9})$$

where we used the standard boundary conditions (11) and used lemma 2 for the last asymptotic. Therefore

$$\lim_{N \rightarrow \infty} \frac{1}{N} \text{Tr}[\widehat{H}_N] = 0 \iff d_0 = 0 \text{ in gauge (A6)}. \quad (\text{A10})$$

We now exhibit all the generalized eigenvectors with eigenvalue 1. For concision, we rewrite \widehat{W} as

$$\widehat{W} = \begin{pmatrix} \widehat{V} & \widehat{\mathbf{f}} \\ 0 & \widehat{\mathbb{1}} \end{pmatrix}, \quad \widehat{\mathbf{f}} := \begin{pmatrix} \widehat{d} \\ \widehat{\mathbf{b}} \end{pmatrix}, \quad (\text{A11})$$

²³When using this to compute the norm via Eq. (58), one should only compute s and set $\mathbf{t} \equiv 0$ so that left canonical form is preserved. When compressing iMPOs, one should instead set $s = 0$ and use only \mathbf{t} .

with block sizes $1 + \chi$ and 1 . Similarly to Eq. (47), we have

$$T_W = \begin{pmatrix} T_V & 0 & \bar{U} & 0 & U & F \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \bar{A}_0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (\text{A12})$$

for some U , where the block sizes are $(1 + \chi)^2$, $1 + \chi$, $1 + \chi$, 1 , and

$$F := \sum_{\alpha} \bar{f}_{\alpha} \otimes f_{\alpha}. \quad (\text{A13})$$

We observe that T_W is the sum of ‘‘reduced’’ and ‘‘spurious’’ parts

$$T_W = \begin{pmatrix} T_V & \bar{U} & U & F \\ 0 & A_0 & 0 & 0 \\ 0 & 0 & \bar{A}_0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \oplus \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} =: T_{\text{red}} \oplus T_{\text{sp}}. \quad (\text{A14})$$

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

$$T_{\text{red}}^{\text{truncated}} = \begin{pmatrix} T_V & \bar{U} & U & 0 \\ 0 & A_0 & 0 & 0 \\ 0 & 0 & \bar{A}_0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (\text{A15})$$

By proposition 1 and lemma 2, it has a unique eigenvalue 1 (the rest have $|\lambda| < 1$). By lemma 3, the corresponding left eigenvector of T_{red} is (after rescaling)

$$Z' = \begin{pmatrix} X & z \\ \bar{z} & 0 \end{pmatrix} \quad (\text{A16})$$

for some z and where X is the unique largest eigenvector of T_V from Eq. (46). Then we have

$$(Z'T_W ZT_W) = (Z' Z) \begin{pmatrix} 1 & \rho \\ 0 & 1 \end{pmatrix}, \quad Z = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad (\text{A17})$$

where $\rho := XF = \sum_{a,b=1}^{\chi} X_{ab}(\hat{f}_a, \hat{f}_b)$. (In practice, one should compute ρ using Eq. (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 \bar{A}_0 , and satisfy $|\lambda| < 1$ by first degreeeness. 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 $\|\widehat{H}_N\|_F^2$ using the transfer matrix formula (39). We expand $\ell\ell$ in the left generalized eigenbasis of T_W :

$$\ell\ell = \underbrace{(aZ' + bZ)}_{\lambda=1 \text{ Jordan block}} + \underbrace{(cE + \bar{c}E^T)}_{\lambda=1 \text{ 'spurious' }} + \underbrace{S}_{|\lambda|<1}, \quad (\text{A18})$$

where S is a linear combination of generalized left eigenvectors with eigenvalues $|\lambda| < 1$. It follows that

$$(\ell\ell)T_W^N(\mathbf{r}\mathbf{r}) = N\rho Z(\mathbf{r}\mathbf{r}) + O(1) = N\rho + O(1) \quad (\text{A19})$$

as $N \rightarrow \infty$, since $r_{\chi+1} = 1$ by the regular form. It remains to determine the coefficient a . For this we look at the 00-component of (A18). First, $ST_W^N \rightarrow 0$ by the definition of S . Meanwhile, (47) and (A12) imply $(ST_W)_{00} = S_{00}$. Therefore $S_{00} = 0$. For the other terms of the RHS, we have $Z'_{00} = 1$ by (A16) and (46), $Z_{00} = 0$ by (A17), and $E_{00} = 0$. On the LHS, the regular form (11) requires $(\ell\ell)_{00} = 1$. Therefore we have $a = 1$ and

$$\|\widehat{H}_N\|_F^2 = \ell\ell T_W^N \mathbf{r}\mathbf{r} = N\rho + O(1). \quad (\text{A20})$$

As noted above, the condition that the trace is subextensive can be lifted. ■

Suppose \widehat{W} is a 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 $\text{Tr}[\widehat{H}_N] = O(N)$, i.e., the trace is extensive.

Case 2. $V_1 = J_2 \oplus J_1 \oplus J_1$ if $\text{Tr}[\widehat{H}_N] = o(N)$, i.e., the trace is subextensive.

Without loss of generality, we adopt the gauge from lemma 4. Define block matrices of size $\chi + 1 \times \chi + 1$

$$Z_i = \begin{pmatrix} X & z \\ z^\dagger & 0 \end{pmatrix}, \quad Z_t = \begin{pmatrix} 0 & t \\ 0 & 0 \end{pmatrix}, \quad Z_f = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \quad (\text{A21})$$

and $Z_{t'} = Z_t^\dagger$ where X is the dominant eigenvalue of T_A , z is the same as above, and $t = (1, 0, \dots, 0)$ is a vector of length χ . These span the dominant generalized eigenspace:

$$\begin{pmatrix} Z_i \\ Z_t \\ Z_{t'} \\ Z_n \end{pmatrix} T_W = \begin{pmatrix} Z_i \\ Z_t \\ Z_{t'} \\ Z_n \end{pmatrix} \underbrace{\begin{pmatrix} 1 & d & d & \rho \\ 0 & 1 & 0 & d \\ 0 & 0 & 1 & d \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{M^\dagger}, \quad (\text{A22})$$

where d is the extensive part of the trace: $\text{Tr}[\widehat{H}_N] = Nd$ and the dagger is because T_W acts on the right. The Jordan decomposition $M = SJS^{-1}$ is then

(case 1)

$$J = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 0 & 0 & 1 \\ -\frac{1}{2} & 0 & d & 0 \\ \frac{1}{2} & 0 & d & 0 \\ -\frac{\rho}{2d} & 2d^2 & \rho & 0 \end{pmatrix}, \quad (\text{A23})$$

(case 2)

$$J = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & \frac{1}{\rho} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \quad (\text{A24})$$

Case 2 is, of course, the same as the above proof, where Z_i 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

$$(\ell\ell)T_W^N(\mathbf{r}\mathbf{r}) = N^2 d^2 + N(\rho - d) + O(1). \quad (\text{A25})$$

The Frobenius norm is then no longer extensive as it 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. (A12) to

$$T_W = \begin{pmatrix} T_V & dt & \bar{U} & dt & U & F \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \bar{A}_0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \quad (\text{A26})$$

APPENDIX 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 submatrix \widehat{V} of an iMPO \widehat{W} will be actively involved. Indeed, any gauge transform of the submatrix $L\widehat{V} = \widehat{V}'L$ can be easily promoted to the iMPO level:

$$\underbrace{\begin{pmatrix} L & \\ & 1 \end{pmatrix}}_{L_W} \underbrace{\begin{pmatrix} \widehat{V} & \widehat{f} \\ & \widehat{\mathbb{1}} \end{pmatrix}}_{\widehat{W}} = \underbrace{\begin{pmatrix} \widehat{V}' & L\widehat{f} \\ & \widehat{\mathbb{1}} \end{pmatrix}}_{\widehat{W}'} \underbrace{\begin{pmatrix} L & \\ & 1 \end{pmatrix}}_{L_W}. \quad (\text{B1})$$

Hence we focus on \widehat{V} and its gauge transforms.²⁴ From this point of view, the QR iteration algorithm 3 is defined by the following recursion:

$$R_0 := \text{Id}_{[0,\chi]}, \quad (\text{B2a})$$

$$\widehat{V}_{n-1} := \widehat{Q}_n R_n, \quad \forall n \geq 1, \quad (\text{B2b})$$

$$\widehat{V}_n := R_n \widehat{Q}_n, \quad (\text{B2c})$$

$$L_n := R_n \dots R_1, \quad (\text{B2d})$$

where (B2b) is a (normal) QR decomposition as defined in (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 5. QR iteration produces a sequence $\{\widehat{V}_n\}$ that are each related to \widehat{V} by a gauge transform:

$$L_n \widehat{V} = \widehat{V}_n L_n. \quad (\text{B3})$$

Proof. Eq. (B2b) implies the gauge transform $R_n \widehat{V}_{n-1} = \widehat{V}_n R_n$ for any $n > 0$. Then (B3) follows from Eq. (B2d) by composing the gauge transforms. ■

Algorithm 3 enjoys also a close relation to the “small” transfer matrix:

Lemma 6. For any $n \geq 0$,

$$\text{Id}_{[0,\chi]}(T_V)^n = L_n^\dagger L_n, \quad (\text{B4})$$

where \widehat{V} has bond dimension χ , that is, $(1 + \chi)$ rows and columns.

²⁴Accordingly, the notation in this Appendix will differ from the main text in that gauge matrices acting on \widehat{V} will not have an overline.

Proof. We again proceed by induction on n . The base case $n = 0$ is trivial. For $n > 0$, we have

$$\begin{aligned} \text{Id}_{[0,\chi]}(T_V)^n &= (L_{n-1}^\dagger L_{n-1}) T_V \\ &= \sum_{\alpha} V_{\alpha}^\dagger L_{n-1}^\dagger L_{n-1} V_{\alpha} \\ &= \sum_{\alpha} L_{n-1}^\dagger V_{n-1,\alpha}^\dagger V_{n-1,\alpha} L_{n-1} \\ &= \sum_{\alpha} L_{n-1}^\dagger R_n^\dagger Q_{n,\alpha}^\dagger Q_{n,\alpha} R_n L_{n-1} \\ &= \sum_{\alpha} L_n^\dagger Q_{n,\alpha}^\dagger Q_{n,\alpha} L_n \\ &= L_n^\dagger L_n, \end{aligned}$$

where we used the induction hypothesis, (38), (B3), (B2b), (B2d), 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 subroutine to be positive rank-revealing, in the following sense:

Definition 10. Suppose \widehat{V} have $1 + \chi$ columns and column rank $1 + \chi'$, where $0 \leq \chi' \leq \chi$. The QR decomposition routine $\widehat{Q}, R \leftarrow \text{QR}[\widehat{V}]$ 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:

$$R_{aa} > 0, \quad a = 0, \dots, \chi. \quad (\text{B5})$$

These requirements can be fulfilled, for example, by the Gram-Schmidt procedure applied to the columns of \widehat{V} .

Proposition 6. 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 after a few lemmas.

Lemma 7. 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

$$\mathbb{T}_m \ni L \mapsto L^\dagger L \in \mathcal{P}_m \quad (\text{B6})$$

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, $\chi_0 \geq \chi_1 \geq \dots$. However, when X is nonsingular, no strict bond dimension reduction can occur:

Lemma 8. Under the same hypotheses of proposition 6, all the \widehat{W}_n 's have the same bond dimension as \widehat{W} .

Proof. By the gauge transform (B3) and lemma 1, \widehat{W}_n is also first degree. So we can apply proposition 1 and let X_n be

the dominant eigenvector of $T_{V_n}: X_n T_{V_n} = X_n$. Then the gauge transform (B3) implies

$$[L_n^\dagger X_n L_n] T_V = L_n^\dagger X_n L_n, \quad (\text{B7})$$

similarly to (A1). This means that $L_n^\dagger X_n L_n = X$ by proposition 1 (the constant is fixed by the 00th element). For X to be nonsingular, L_n must be a square matrix, so the bond dimension does not change. ■

We remark on a useful consequence of lemma 8: 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 proposition 6. By the definition of positive rank-revealing QR, and lemma 8, for any $n \geq 1$, $R_n \in \mathbb{T}_{1+\chi}$, and thus $L_n \in \mathbb{T}_{1+\chi}$. Now, Lemma 6 and proposition 1 imply that

$$L_n^\dagger L_n = \text{Id}_{[0,\chi]}(T_V)^n \xrightarrow{n \rightarrow \infty} X = \begin{pmatrix} 1 & \mathbf{y} \\ \mathbf{y}^\dagger & Y \end{pmatrix}. \quad (\text{B8})$$

Note that $(L_n)_{00} = 1$ for all n . Eq. (B8) implies that X is positive semi-definite. Since we assume X is nonsingular, X is positive definite. Then, lemma 7 implies that $L_n \rightarrow L$ for some invertible L , and the QR iteration converges as follows:

$$\begin{aligned} \widehat{V}_n &= L_n \widehat{V} L_n^{-1} \rightarrow L \widehat{V} L^{-1} := \widehat{V}_L, \\ R_n &= L_{n+1}^{-1} L_n \rightarrow \text{Id}_{[0,\chi+1]}, \\ \widehat{Q}_n &= \widehat{V}_{n-1} R_n^{-1} \rightarrow \widehat{V}_L, \end{aligned}$$

so that \widehat{V}_L is a left canonical MPS. Promoting to the iMPO level using (B1) completes the proof. ■

Proposition 6 establishes the existence of left canonical for all “generic” first degree iMPOs, in the sense that X is nonsingular. We now treat the singular cases:

Proposition 7. Let \widehat{W} be a first degree iMPO and such that the leading eigenvector X of T_V is positive semi-definite of rank $1 + \chi' \leq 1 + \chi$. 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 eigendecomposition:

$$\begin{aligned} X &= \begin{pmatrix} 1 & \mathbf{x} \\ \mathbf{x}^\dagger & \mathbf{X} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \mathbf{x}^\dagger & \text{Id} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{X} - \mathbf{x}^\dagger \otimes \mathbf{x} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{x} \\ 0 & \text{Id} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ \mathbf{x}^\dagger & \mathbf{U} \end{pmatrix} X_1 \begin{pmatrix} 1 & \mathbf{x} \\ 0 & \mathbf{U} \end{pmatrix} =: L^\dagger X_1 L, \end{aligned} \quad (\text{B9})$$

where U is unitary and $X_1 = \text{diag}(1, \sigma_1, \dots, \sigma_\chi)$ where

$$\begin{aligned} \sigma_a &> 0 & \text{if } a \leq \chi', \\ \sigma_a &= 0 & \text{if } a > \chi'. \end{aligned} \quad (\text{B10})$$

Since L is invertible, we have the gauge transform

$$\widehat{V}_1 := L \widehat{V} L^{-1} \quad (\text{B11})$$

so that the leading eigenvector of T_{V_1} becomes the diagonal matrix X_1 . Thus the a ath component of the equation $X_1 T_{V_1} =$

X_1 becomes

$$\sigma_a = \sum_{b=0}^{\chi} \sigma_b \langle (\widehat{V}_1)_{ba}, (\widehat{V}_1)_{ba} \rangle. \quad (\text{B12})$$

When $a > \chi'$, $\sigma_a = 0$, so every term on the RHS must also vanish. Now for $b \leq \chi'$, $\sigma_b > 0$, so $(\widehat{V}_1)_{ba} = 0$. Namely, we showed that \widehat{V}_1 has the block-diagonal form:

$$\widehat{V}_1 = \begin{pmatrix} \widehat{V}'_{[0,\chi']} & 0 \\ 0 & * \end{pmatrix}, \quad (\text{B13})$$

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:

$$(\text{Id}_{[0,\chi']} \ 0) \widehat{V}_1 = \widehat{V}' (\text{Id}_{[0,\chi']} \ 0) \quad (\text{B14})$$

It is easy to check that $T_{V'}$ has leading eigenvector $X_2 = \text{diag}(1, \sigma_1, \dots, \sigma_{\chi'})$, which is nonsingular. Composing the two gauge transforms (B11) and (B14) 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 proposition 3. By proposition 7, 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 proposition 6. 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 $L = L_1 L_0$. ■

Algorithm 6. iMPO Left Can. Form: General

- 1: **procedure** Precondition \widehat{W}, η
 - 2: $X \leftarrow \text{EIGMAX}(T_V)$ ▷ Find max. eigenvector
 - 3: $\mathbf{x}, \mathbf{U}, \Sigma \leftarrow X$ ▷ Eq. (B9)
 - 4: $\chi' \leftarrow \max\{a : \sigma_a > \eta^2\}$
 - 5: $\mathbf{x}, \mathbf{U} \leftarrow [x_a]_{1 \leq a \leq \chi'}, [\mathbf{U}_{ab}]_{1 \leq a \leq \chi', 1 \leq b \leq \chi}$
 - 6:
 - 7: $L \leftarrow \begin{pmatrix} 1 & \mathbf{x} & 0 \\ & \mathbf{U} & 0 \\ 0 & & 1 \end{pmatrix}, L' \leftarrow \begin{pmatrix} 1 & -\mathbf{x} & 0 \\ & \mathbf{U}^\dagger & 0 \\ & & 1 \end{pmatrix}$
 - 8: **return** $L \widehat{W} L', L$
 - 9: **procedure** LeftCan \widehat{W}, η ▷ η : tolerance
 - 10: $\widehat{W}, L_0 \leftarrow \text{PRECONDITION}(\widehat{W}, \eta)$
 - 11: $\widehat{W}, L_1 \leftarrow \text{QRITER}(\widehat{W}, \eta)$ ▷ Alg. 3
 - 12: **return** $\widehat{W}, L_1 L_0$
-

Note that the above proof and that of proposition 7 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 6. 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 6 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 could not we prove the existence of left canonical form for all first degree iMPOs (proposition 3) 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 1. Consider the spin-half iMPO

$$\widehat{W} := \begin{pmatrix} \widehat{\mathbb{1}} & 0 & \widehat{Z} \\ \alpha \widehat{Z} & \widehat{X} & \widehat{\mathbb{1}} \end{pmatrix}, \quad (\text{B15})$$

where $|\alpha| < 1$ so that \widehat{W} is first degree. However, applying algorithm 3 to it will yield

$$\widehat{W}_n = \begin{pmatrix} \widehat{\mathbb{1}} & 0 & \widehat{Z} \\ \alpha \widehat{Z} & \alpha^n \widehat{X} & \widehat{\mathbb{1}} \end{pmatrix}, \quad L_n = \begin{pmatrix} 1 & 0 & 0 \\ \alpha^n & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (\text{B16})$$

Everything seems to converge, but $\lim_{n \rightarrow \infty} \widehat{W}_n$ is not left canonical! In fact, $\lim_{n \rightarrow \infty} L_n$ is singular, which makes the argument in the proof of proposition 6 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 6 does.) However, 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 proposition 6 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.

APPENDIX C: EXACT ESTIMATES OF SCHMIDT VALUES

We study the singular values of the matrix M defined in (85) (which form the entanglement spectrum of an MPO) by repeatedly applying a rank one perturbation.

First, we consider the submatrix

$$M_0 := \begin{pmatrix} \mathcal{N}_R & \mathbf{p}_R \\ 0 & \mathbf{S} \end{pmatrix}, \quad (\text{C1})$$

where $\mathbf{S} = \text{diag}(s_1 \geq \dots \geq s_\chi)$ so that

$$M_0^\dagger M_0 = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{S}^2 \end{pmatrix} + \begin{pmatrix} \mathcal{N}_R \\ \mathbf{p}_R^\dagger \end{pmatrix} (\mathcal{N}_R \mathbf{p}_R) \quad (\text{C2})$$

is a rank one perturbation of $\text{diag}(0, s_1^2, s_2^2, \dots)$. 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

$$\frac{\mathcal{N}_R^2}{\mu^2} + \sum_a \frac{|p_R^a|^2}{\mu^2 - s_a^2} = 1. \quad (\text{C3})$$

This implies the interlacing relation

$$\mu_0 \geq s_1 \geq \mu_1 \geq s_2 \geq \dots \geq s_\chi \geq \mu_\chi. \quad (\text{C4})$$

For the largest singular value, (C3) 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:

$$\mathcal{N}_R^2 + \|\mathbf{p}_R\|^2 + s_1^2 \geq \mu_0^2 \geq \mathcal{N}_R^2 + \|\mathbf{p}_R\|^2. \quad (\text{C5})$$

In particular, the separation of scales (86) implies $\mu_0^2 = \Theta(N)$ and $\mu_{a \geq 1}^2 = \text{O}(1)$.

In a very similar fashion, we now go back to the full matrix and consider

$$MM^\dagger = \begin{pmatrix} M_0 M_0^\dagger & \\ & 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{p}_L \\ \mathcal{N}_L \end{pmatrix} (0 \mathbf{p}_L^\dagger \mathcal{N}_L) \quad (\text{C6})$$

which is similar to

$$\begin{pmatrix} \mu_0^2 & & \\ & D_\mu & \\ & & 0 \end{pmatrix} + \begin{pmatrix} \mathbf{q}_L \\ \mathcal{N}_L \end{pmatrix} (\mathbf{q}_L^\dagger \mathcal{N}_L), \quad (\text{C7})$$

under conjugation where $D_\mu = \text{diag}(\mu_1^2, \dots, \mu_\chi^2)$, $\mathbf{q}_L = U(0 \mathbf{p}_L)^T$, U being a unitary matrix such that $UM_0 M_0^\dagger U^\dagger = \text{diag}(\mu_0^2, \mu_1^2, \dots, \mu_\chi^2)$. Applying rank one perturbation again to (C7), we obtain the following equation determining the singular values of M :

$$\frac{\mathcal{N}_L^2}{\lambda^2} + \frac{|q_L^0|^2}{\lambda^2 - \mu_0^2} + \sum_{a=1}^\chi \frac{|q_L^a|^2}{\lambda^2 - \mu_a^2} = 1. \quad (\text{C8})$$

This implies the interlacing relation

$$\lambda_{-1} \geq \mu_0 \geq \lambda_0 \geq \mu_1 \geq \dots \geq \mu_\chi \geq \lambda_\chi, \quad (\text{C9})$$

which, combined with (C4), gives (90) in the main text.

Similarly to (C5), we can bound λ_{-1} as follows:

$$\lambda_{-1}^2 \geq \mathcal{N}_L^2 + \|\mathbf{q}_L\|^2 = \mathcal{N}_L^2 + \|\mathbf{p}_L^2\|, \quad (\text{C10a})$$

$$\lambda_{-1}^2 \leq \mathcal{N}_L^2 + \|\mathbf{p}_L^2\| + \mu_0^2. \quad (\text{C10b})$$

Under the separation of scales (86), $\lambda_{-1} = \Theta(N)$ is extensive.

We also need a useful lower bound for largest singular value λ_0 . For this, we note that (C8) implies

$$\frac{\mathcal{N}_L^2}{\lambda_0^2} + \sum_{a=1}^\chi \frac{|q_L^a|^2}{\lambda_0^2} \leq 1 + \frac{|q_L^0|^2}{\mu_0^2 - \lambda_0^2} \quad (\text{C11})$$

which is a quadratic inequality (of λ_0^2). Its solution entails

$$2\lambda_0^2 \geq \mu_0^2 + \mathcal{N}_L^2 + \|\mathbf{q}_L\|^2 \quad (\text{C12})$$

$$\begin{aligned} & -\sqrt{(\mu_0^2 - \mathcal{N}_L^2 - \|\mathbf{q}_L\|^2)^2 + 4\mu_0^2 |q_L^0|^2} \\ & \geq 2 \min(\mu_0^2, \mathcal{N}_L^2 + \|\mathbf{q}_L\|^2) - 2\mu_0 |q_L^0|. \end{aligned} \quad (\text{C13})$$

Now, under (86), $\mu_0^2, \mathcal{N}_L^2 \in \Theta(N)$ and $\mathbf{q}_L \in \text{O}(1)$, so we conclude that $\lambda_0^2 \in \Theta(N)$ is also extensive.

APPENDIX D: HAMILTONIAN ERROR BOUND

This Appendix discusses the relation between the norm we have used throughout this work and the standard sup norm.

Recall that the sup norm is

$$\|\widehat{H}\|_s^2 := \sup_{|\psi\rangle} \frac{\langle \psi | \widehat{H} \widehat{H} | \psi \rangle}{\langle \psi | \psi \rangle}. \quad (\text{D1})$$

For convenience, we work in this section with the nonscaled Frobenius norm, which we denote with a lowercase “ f ”:

$$\|\widehat{H}\|_f^2 := \text{Tr}[\widehat{H}^\dagger \widehat{H}] = \text{Tr}[I] \cdot \|\widehat{H}\|_F^2. \quad (\text{D2})$$

With our default inner product, if \widehat{O} is an operator supported on $S \subset \mathbb{Z}$, a set of size $|S| = k$, then

$$\langle \widehat{O} | \widehat{O} \rangle = \frac{\text{Tr}[\widehat{O}^\dagger \widehat{O}]}{\text{Tr}[I]} = \frac{\text{Tr}[\widehat{O}_S^\dagger \widehat{O}_S]}{\text{Tr}[\widehat{\mathbb{1}}^{\otimes k}]} = \frac{\|\widehat{O}\|_f^2}{d^k}. \quad (\text{D3})$$

Quite generally, $\|\widehat{O}\|_s \leq \|\widehat{O}\|_f$. So $\langle \widehat{O} | \widehat{O} \rangle = 1$ implies

$$\|\widehat{O}\|_s^2 \leq \|\widehat{O}\|_f^2 = d^k \quad (\text{D4})$$

for an operator supported on k sites.

We now prove proposition 5. The idea 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 analogous 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 proposition 5 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:

$$\langle \widehat{H}_S^a | \widehat{H}_S^b \rangle = \delta^{ab}, S \in \{L, R\}. \quad (\text{D5})$$

As each term is supported on at most k sites, it follows from (D4) that

$$\|\widehat{H}_L^a \widehat{H}_R^b\|_f^2 = d^k. \quad (\text{D6})$$

It is a standard fact about extremal eigenvalues that if $\widehat{H} = \widehat{H}' + \delta\widehat{H}$ and $E'_0 := \|\widehat{H}'\|_s$, then

$$\delta E := |E'_0 - E_0| \leq \|\delta\widehat{H}\|_s. \quad (\text{D7})$$

By hypothesis

$$\|\delta\widehat{H}\|_s^2 = \left\| \sum_{a=\chi'}^{\chi} \widehat{O}_L^a s_a \widehat{O}_R^a \right\|_s^2 \leq \sum_{a=\chi'}^{\chi} s_a^2 \|\widehat{O}_L^a \widehat{O}_R^a\|_s^2 \quad (\text{D8})$$

We can now separately bound each term in the sum using locality:

$$\begin{aligned} \|\widehat{O}_L^c \widehat{O}_R^c\|_s^2 &\leq \|\widehat{O}_L^c \widehat{O}_R^c\|_f^2 \\ &= \left\| \sum_{a,b=1}^{\chi} U^{ac} V^{cb} \widehat{H}_L^a \widehat{H}_R^b \right\|_f^2 \\ &= \sum_{abef} U^{ac} V^{cb} (U^{ec})^* (V^{cf})^* \text{Tr}[\widehat{H}_L^{a\dagger} \widehat{H}_R^{b\dagger} \widehat{H}_L^e \widehat{H}_R^f] \end{aligned}$$

$$\begin{aligned} &= \sum_{abef} U^{ac} V^{cb} (U^{ec})^* (V^{cf})^* d^k \delta^{ae} \delta^{bf} \\ &= d^k \sum_{a=1}^{\chi} |U^{ac}|^2 \sum_{b=1}^{\chi} |V^{cb}|^2 \\ &= d^k, \end{aligned}$$

where we have used (D4) 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

$$\|\delta\widehat{H}\|_s^2 \leq \sum_{a=\chi'}^{\chi} s_a^2 \|\widehat{O}_L^a \widehat{O}_R^a\|_s^2 \leq d^k \sum_{a=\chi'}^{\chi} s_a^2. \quad (\text{D9})$$

■

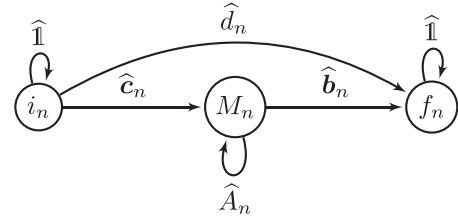
APPENDIX E: ELEMENTARY OPERATIONS

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

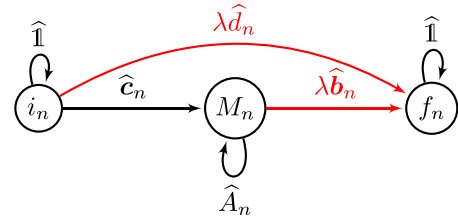
$$\widehat{W}[\widehat{\mathcal{O}}_1] = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{c}_1 & \widehat{d}_1 \\ 0 & \widehat{A}_1 & \widehat{b}_1 \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}, \widehat{W}[\widehat{\mathcal{O}}_2] = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{c}_2 & \widehat{d}_2 \\ 0 & \widehat{A}_2 & \widehat{b}_2 \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix} \quad (\text{E1})$$

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,

$$\widehat{W}[\lambda \widehat{\mathcal{O}}_1] = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{c}_1 & \lambda \widehat{d}_1 \\ 0 & \widehat{A}_1 & \lambda \widehat{b}_1 \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbb{1}} & \lambda \widehat{c}_1 & \lambda \widehat{d}_1 \\ 0 & \widehat{A}_1 & \widehat{b}_1 \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}. \quad (\text{E2})$$

These two choices preserve left and right canonical forms, respectively.

Addition of iMPOs is essentially the direct sum of the matrices:

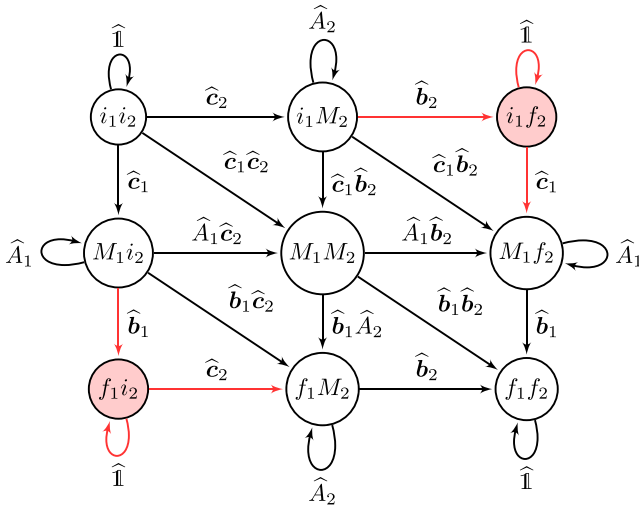
$$\widehat{W}[\widehat{O}_1 + \widehat{O}_2] = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{c}_1 & \widehat{c}_2 & \widehat{d}_1 + \widehat{d}_2 \\ 0 & \widehat{A}_1 & 0 & \widehat{b}_1 \\ 0 & 0 & \widehat{A}_2 & \widehat{b}_2 \\ 0 & 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}. \quad (\text{E3})$$

The operation of multiplication is more involved. The multiplication of two local operators, say $\widehat{O}_1 = \sum_i \widehat{X}_i$ and $\widehat{O}_2 = \sum_i \widehat{Y}_i$ is “bilocal” or “second degree”, with arbitrarily long strings of identities between sites with information: $\widehat{O}_1 \widehat{O}_2 = \sum_i \sum_{N=0}^{\infty} \widehat{X}_i \widehat{\mathbb{1}}^N \widehat{Y}_{i+N} + \dots$. This is represented as an iMPO as

$$\widehat{W}[O_1 O_2] = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{X} & \widehat{Y} & i\widehat{Z} \\ 0 & \widehat{\mathbb{1}} & 0 & \widehat{Y} \\ 0 & 0 & \widehat{\mathbb{1}} & \widehat{X} \\ 0 & 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix}. \quad (\text{E4})$$

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.”



(We have dropped the \widehat{d} terms and also the self-loop on $M_1 M_2$ 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, “ $\widehat{b}_1 \widehat{A}_2$ ” has components

$$(\widehat{b}_1 \widehat{A}_2)_{(a_1 a_2), b_2}^\gamma = \sum_{\alpha, \beta} f_{\alpha\beta}^\gamma (B_1)_{a_1}^\alpha (A_2)_{a_2, b_2}^\beta \quad (\text{E5})$$

where $f_{\alpha\beta}^\gamma$ are the structure constants of the on-site algebra \mathcal{A} .

The nonlocality 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 11. Suppose \widehat{O}_1 and \widehat{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 nondisjoint

product is

$$\widehat{O}_1 \odot \widehat{O}_2 = \begin{cases} \widehat{O}_1 \widehat{O}_2 & \text{if } [a_1, b_1] \cap [a_2, b_2] \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \quad (\text{E6})$$

The definition extends to any local operators by linearity. At the MPO level, this is just the nonshaded part of the above diagram.

Terms with disjoint spatial support always commute, so the “nondisjoint commutator” is the same as the normal one:

$$[\widehat{O}_1, \widehat{O}_2] = \widehat{O}_1 \odot \widehat{O}_2 - \widehat{O}_2 \odot \widehat{O}_1. \quad (\text{E7})$$

This means that the commutator is local whenever $\widehat{O}_1 \odot \widehat{O}_2$ is. Therefore strictly local operators form a closed algebra under commutation.

First degree operators are not closed under commutation, as we now demonstrate by counterexample. This is a consequence of the fact that the class of first degree operators includes operators which do not make sense as local or physical Hamiltonians. These can have bizarre properties from a ground state perspective, such as superextensive ground state energy, which lead in turn to other strange issues such as the nonclosure under commutation. The subset of first degree operators which *are* physical Hamiltonians should be free of these issues.

For the counterexample, suppose \widehat{H}_I has an iMPO representation

$$\widehat{W}_I = \begin{pmatrix} \widehat{\mathbb{1}} & \widehat{X} & 0 \\ 0 & \widehat{O} & \widehat{Y} \\ 0 & 0 & \widehat{\mathbb{1}} \end{pmatrix} \quad (\text{E8})$$

where $\widehat{O} = \frac{c}{2}(\widehat{\mathbb{1}} + \widehat{Z}) = \begin{pmatrix} c & 0 \\ 0 & 0 \end{pmatrix}$ is an on-site projector matrix and take $c \in (2^{1/4}, 2^{1/2})$. The norm of H_I is $\|H_I\|^2 = \sum_{N=0}^{\infty} \|\widehat{O}\|^{2N} = \sum_{N=0}^{\infty} (c^2/2)^N < \infty$. However, the norm of the product diverges:

$$\|H_I \odot H_I\|^2 > \sum_{N=0}^{\infty} \|\widehat{O}\widehat{O}\|^{2N} = \sum_{N=0}^{\infty} (c^4/2)^N = \infty, \quad (\text{E9})$$

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 nonlocal, but the norm-per-unit-length is not even submultiplicative: there are cases where $\|\widehat{O}_1 \widehat{O}_2\| \not\leq \|\widehat{O}_1\| \|\widehat{O}_2\|$. 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 8. If \widehat{O}_1 is strictly local and \widehat{O}_2 is first degree, then $[\widehat{O}_1, \widehat{O}_2]$ is first degree.

Proof. It is sufficient to show $\widehat{O}_1 \odot \widehat{O}_2$ is first degree.

Let the iMPOs for the operators be given by Eq. (E1). In particular, \widehat{A}_1 is strictly upper triangular. From the product automata above, we can see that the \widehat{A} block of $\widehat{O}_1 \odot \widehat{O}_2$ is

given by

$$\widehat{A} = \begin{pmatrix} \widehat{A}_2 & 0 & \widehat{c}_1\widehat{A}_2 & \widehat{c}_1\widehat{b}_2 & 0 \\ 0 & \widehat{A}_1 & \widehat{A}_1\widehat{c}_2 & 0 & \widehat{b}_1\widehat{c}_2 \\ 0 & 0 & \widehat{A}_1\widehat{A}_2 & \widehat{A}_1\widehat{b}_2 & \widehat{b}_1\widehat{A}_2 \\ 0 & 0 & 0 & \widehat{A}_1 & 0 \\ 0 & 0 & 0 & 0 & \widehat{A}_2 \end{pmatrix}, \quad (\text{E10})$$

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 its 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. (E7). It is advisable to compress the operator after each product and again after the difference. In circumstances where \widehat{O}_1 and \widehat{O}_2 are Hermitian or anti-Hermitian, the two nondisjoint products are related by a Hermitian conjugate and a sign, and need to be computed only once.

-
- [1] M. B. Hastings, *J. Stat. Mech.* (2007) P08024.
 [2] U. Schollwöck, *Ann. Phys.* **326**, 96 (2011).
 [3] I. P. McCulloch, *J. Stat. Mech.* (2007) P10014.
 [4] J. Hauschild and F. Pollmann, *SciPost Phys. Lect. Notes* **5**, 1 (2018).
 [5] L. Vanderstraeten, J. Haegeman, and F. Verstraete, *SciPost Phys. Lect. Notes* **7**, 1 (2019).
 [6] N. Schuch, M. M. Wolf, F. Verstraete, and J. I. Cirac, *Phys. Rev. Lett.* **100**, 030504 (2008).
 [7] F. Verstraete and J. I. Cirac, *Phys. Rev. B* **73**, 094423 (2006).
 [8] L. Michel and I. McCulloch, [arXiv:1008.4667](https://arxiv.org/abs/1008.4667) (2010).
 [9] G. K.-L. Chan, A. Keselman, N. Nakatani, Z. Li, and S. R. White, *J. Chem. Phys.* **145**, 014102 (2016).
 [10] C. Hubig, I. P. McCulloch, and U. Schollwöck, *Phys. Rev. B* **95**, 035129 (2017).
 [11] B. Pirvu, V. Murg, J. I. Cirac, and F. Verstraete, *New J. Phys.* **12**, 025012 (2010).
 [12] M. P. Zaletel, R. S. K. Mong, C. Karrasch, J. E. Moore, and F. Pollmann, *Phys. Rev. B* **91**, 165112 (2015).
 [13] S.-Y. Kung, in *Proceedings of 12th Asilomar Conference on Circuits, Systems and Computers* (Pacific Grove, CA, 1978).
 [14] G. M. Crosswhite and D. Bacon, *Phys. Rev. A* **78**, 012356 (2008).
 [15] M.-D. Choi, *Linear Algebra Appl.* **10**, 285 (1975).
 [16] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics* (Springer Science and Business Media, New York, 1996), Vol. 2.
 [17] L. M. Silverman and M. Bettayeb, in *Proceedings of the 1980 Joint Automatic Control Conference* (American Automatic Control Council, San Francisco, CA, 1980), p. 81.
 [18] U. Al-Saggaf and G. Franklin, *IEEE Trans. Automatic Control* **32**, 815 (1987).
 [19] See <http://slicot.org>.
 [20] See <http://python-control.org>.
 [21] M. P. Zaletel, R. S. K. Mong, F. Pollmann, and E. H. Rezayi, *Phys. Rev. B* **91**, 045115 (2015).
 [22] D. C. Mattis, in *Physics in One Dimension*, edited by J. Bernasconi and T. Schneider (Springer, Berlin, Heidelberg, 1981), Vol. 23.
 [23] V. Viswanath and G. Müller, *The Recursion Method: Applications to Many-body Dynamics* (Springer, Heidelberg, 2008).
 [24] D. E. Parker, X. Cao, A. Avdoshkin, T. Scaffidi, and E. Altman, *Phys. Rev. X* **9**, 041017 (2019).