MACHINE LEARNING
Science and Technology

PAPER

A quantum inspired approach to learning dynamical laws from data—block-sparsity and gauge-mediated weight sharing

J Fuksa[2,5,*] , M Götte[1,5], I Roth[3] and J Eisert[2,4]

[1] Institute of Mathematics, Technische Universität Berlin, Berlin, Germany
[2] Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, Berlin, Germany
[3] Quantum Research Center, Technology Innovation Institute, Abu Dhabi, United Arab Emirates
[4] Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[5] These authors contributed equally to this work.
* Author to whom any correspondence should be addressed.

E-mail: jonas.fuksa@fu-berlin.de

Abstract

Recent years have witnessed an increased interest in recovering dynamical laws of complex systems in a largely data-driven fashion under meaningful hypotheses. In this work, we propose a scalable and numerically robust method for this task, utilizing efficient block-sparse tensor train representations of dynamical laws, inspired by similar approaches in quantum many-body systems. Low-rank tensor train representations have been previously derived for dynamical laws of one-dimensional systems. We extend this result to efficient representations of systems with K-mode interactions and controlled approximations of systems with decaying interactions. We further argue that natural structure assumptions on dynamical laws, such as bounded polynomial degrees, can be exploited in the form of block-sparse support patterns of tensor-train cores. Additional structural similarities between interactions of certain modes can be accounted for by weight sharing within the ansatz. To make use of these structure assumptions, we propose a novel optimization algorithm, block-sparsity restricted alternating least squares with gauge-mediated weight sharing. The algorithm is inspired by similar notions in machine learning and achieves a significant improvement in performance over previous approaches. We demonstrate the performance of the method numerically on three one-dimensional systems—the Fermi–Pasta–Ulam–Tsingou system, rotating magnetic dipoles and point particles interacting via modified Lennard–Jones potentials, observing a highly accurate and noise-robust recovery.

1. Introduction

Discovering dynamical laws that govern the time evolution of dynamical systems has been a central task in physics and engineering for centuries, from practical as well as fundamental perspective. Historically, this task has been approached from two directions—firstly by using expert physical knowledge and intuition, and secondly by using data obtained by measuring the time evolution of the dynamical system in question. With the ever growing availability of large amounts of computational power and data, the second approach is becoming increasingly accessible [SL09, BPK16, GKES19, GGR+20, IMW+20, CPSW21, KBK22, CDA+21, MBPK16]. However, since a crucial aspect of using data to learn dynamical laws is choosing the right hypothesis class, physical intuition about the system is essential in developing efficient data-driven algorithms.

One prominent recent approach is the *sparse identification of non-linear dynamics* (SINDy) algorithm [BPK16, SBK21, dSCQ+20]. Here the learning task is phrased as a linear inversion problem for a chosen function dictionary. To arrive at a physically motivated hypothesis class, the authors impose sparsity of the recovered dynamical laws with respect to this dictionary—implementing the principle of Occam's razor.

The remarkable success of the SINDy algorithm demonstrates that imposing structure in learning dynamical laws is immensely powerful. This is in spite of the fact that Occam's razor is a general heuristic principle that is not linked to any specific physical properties of the system at hand.

Building on the ideas of SINDy and its variant MANDy [GKES19], in [GGR+20] the authors proposed to use *locality* in one dimension as the structure imposing physical principle. They have shown that the resultant hypothesis class consists of low rank *tensor trains* (TT) [BSU16], a specific type of an efficient tensor network representation of multivariate functions [CSS15, SS16, LYCS17]. A similar result is widely known in the quantum many-body literature [CPGSV21, Oru14, VC04, HMOV14], where it has been shown that low rank TTs (known in this context as *matrix product states* (MPS)) parametrize ground states of one-dimensional local Hamiltonians [VC06], as well as other states of physical importance [ECP10, SWVC08, VMC08].

Another structural observation made by the quantum many-body community is that the TT representations of quantum states that are symmetric under local action of some symmetry group have tensor cores with a characteristic sparse support, a property dubbed *block-sparsity* [SPV10]. Recently, it has been proven that functions with bounded polynomial degree (or an equivalent notion suitable for the chosen function dictionary) also admit block-sparse TT representations [GST21, BGP21]. Being able to control the polynomial degree of non-linear functions is in the context of learning dynamical laws a promising primitive. On a technical level, limiting the total polynomial degree offers a natural truncation of a multivariate function space via Taylor's theorem. On a conceptual level, high-degree polynomials in dynamical laws correspond to terms that vary quickly with respect to many modes. The appearance of such terms in dynamical laws is generally a sign that a better choice of coordinates parametrizing the state space could be made. Similar arguments also apply to other dictionaries, e.g. for trigonometric functions limiting the total 'degree' corresponds to neglecting fast oscillating terms in the dynamical law.

Bounding a notion of total degree suitable for a chosen multi-variate function dictionary, block-sparse TTs arise as a natural, physically well-motivated efficient restriction of the ansatz class for dynamical laws. In another context, block-sparsity has been observed to improve the performance of tensor network optimization methods in terms of computational, memory and sample complexity [GST21, BGP21].

The contribution of this work is three-fold: (1) we broaden the range of physical principles that lead to efficient TT representations of dynamical laws. In particular, we show that systems with K-mode interactions admit an efficient TT representation and that systems with algebraically decaying interactions can be approximated with bounded error by a system with efficient TT representation. (2) We utilize block-sparsity in the context of learning dynamical laws—demonstrating a significantly improved performance of the resulting method compared to previous work. (3) We use self-similarity between certain modes in the system to further restrict the search space and develop a new optimization algorithm for the resultant hypothesis class, referred to as *ALS optimization with gauge mediated weight sharing (ALS-GMWS)*, which is inspired by similar notions in machine learning [RCR+20]. We show numerically that (2) and (3) improve scalability of the method, allowing us to learn dynamical laws of systems more than three times larger than the systems presented in [GGR+20].

The remainder of this work is structured as follows: Our setting of learning dynamical laws is formally introduced in section 2. In section 3 we give a brief primer into tensor networks. Section 4 shows how a natural truncation of the function space leads to an efficient parametrization via block sparse TTs. In section 5 we prove that low rank TTs parametrize dynamical laws of systems obeying generic physical principles. The concept of self-similarity and the ALS-GMWS algorithm is presented in section 6. Finally, we numerically demonstrate the performance of our method in section 7, before concluding in section 8.

2. Setup

We will use the notation $[N] := \{1, \dots, N\}$ for any integer N. Consider a dynamical system with state space $\mathcal{S} \subset \mathbb{R}^d$, such that its state is described by a real d-dimensional vector $x = (x^{(1)}, \dots, x^{(d)})$. We call each of the degrees of freedom $x^{(i)}$ a *mode*. The time evolution is a smooth curve $x(t) : \mathbb{R} \to \mathcal{S}$, which is, given initial conditions x_0, generated by

$$\dot{x}(t) = f(x(t)) \quad \text{or} \quad \ddot{x}(t) = f(x(t)), \tag{1}$$

where we have restricted our attention to time-independent systems. We assume that it is known which of these forms governs the dynamical system of interest, and we will refer to both equation (1) and the function $f : \mathcal{S} \to \mathbb{R}^d$ as the *dynamical law*. The first form of equation (1) can arise, e.g. from Hamilton's equations, where $f(x) = \{x, H\}$ with H the Hamiltonian and $\{\cdot, \cdot\}$ the Poisson bracket, while the second form appears, e.g. in Newton's equations, where $f_k(x)$ is the total force acting on the k-th mode in the state x. By *learning a dynamical law* we mean identifying $f(x)$ given data pairs $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^d$ for $i \in [M]$ with the relation

$y_i(x_i) \approx f(x_i)$. The data may be coming from time-series measurements of a trajectory with the gradients y_i approximated, e.g. by the method of finite differences.

To learn f, we choose a *function dictionary* $\{\Psi_i : \mathbb{R} \to \mathbb{R}\}_{i \in [p]}$ of linearly independent functions. Forming a product basis

$$\Phi_{i_1,\ldots,i_d}(x) = \Psi_{i_1}\left(x^{(1)}\right)\ldots\Psi_{i_d}\left(x^{(d)}\right) \tag{2}$$

we obtain the space

$$\mathcal{A} = \mathrm{span}\left\{\Phi_{i_1,\ldots,i_d} : i_k \in [p] \ \forall k \in [d]\right\}, \tag{3}$$

with elements mapping $\mathbb{R}^d$ to $\mathbb{R}$, so that $\mathcal{A}^d$ can be used as a search space for $f : \mathbb{R}^d \to \mathbb{R}^d$.

Elements $g(x) \in \mathcal{A}$ are labeled by tensors $\Theta \in \mathbb{R}^{p^d}$ via

$$g(x) = \sum_{i_1,\ldots,i_d \in [p]^d} \Theta_{i_1,\ldots,i_d} \Phi_{i_1,\ldots,i_d}(x). \tag{4}$$

Hence, the dimension of $\mathcal{A}$ is exponential in the system size d, limiting the scalability of learning a function within $\mathcal{A}$. This is an example of the notorious *curse of dimensionality*. In order to overcome the curse, we use structural constraints on Θ due to known physical properties of the system to identify a physically relevant subspace $\mathcal{H} \subset \mathcal{A}$ with $\dim(\mathcal{H}) \in O(\mathrm{poly}(d))$, defining the *physical corner* of solutions. Specifically, in this work, we show that for many dynamical systems of interest a suitable $\mathcal{H}$ is the set of low rank block-sparse TTs, as defined in section 3.

3. A primer to tensor networks

To set the stage and notation, we start with a brief introduction to tensor networks. For a more thorough treatment of this topic, see [BSU16, BC17]. The curse of dimensionality, as discussed in the previous section, renders high-order tensors hard to work with. Not only are they hard to optimize over, even storing them and performing basic operations on them quickly becomes impractical as the order grows. *Tensor networks* are a way of decomposing high-order tensors in terms of contractions of smaller tensors. Limiting the ranks of these contractions we define subspaces of tensor spaces with dimensions scaling polynomially in the order, averting the curse of dimensionality. Crucially for us, such subspaces turn out to contain physically relevant tensors in many cases of interest.

Tensor networks can be visually represented by *tensor network diagrams*. Suppose an order d tensor T, which is a contraction of tensors $\{C_i\}_i$, called *tensor cores*. To draw a tensor network diagram for T, we draw a node for each tensor core C_i. For each of the indices of the tensor cores we draw an edge and connect those edges that represent indices that are contracted over in T. The indices represented by connected edges are referred to as *virtual indices*. The remaining d unconnected edges represent the d *physical indices* of T.

A widely studied example of such a network is the *tensor train*, which is represented by the tensor network

$$T_{i_1,\ldots,i_d} = \underset{\substack{\bullet \quad \bullet \\ i_1 \quad i_2}}{\overset{C_1 \quad C_2}{\rule{0pt}{0pt}}} \quad \ldots \quad \underset{\substack{\bullet \quad \bullet \\ i_{d-1} \quad i_d}}{\overset{C_{d-1} \quad C_d}{\rule{0pt}{0pt}}} \ . \tag{5}$$

This contraction of tensors can be written in matrix notation as

$$T_{i_1,\ldots,i_d} = \sum_{\substack{j_1,\ldots,j_{d-1} \\ \in [r_1]\times\ldots\times[r_{d-1}]}} (C_1)^{i_1}_{j_1} (C_2)^{i_2}_{j_1,j_2} \ldots (C_d)^{i_d}_{j_{d-1}}, \tag{6}$$

where the superscripts on the tensor cores are the physical indices, while the subscripts are the virtual indices. Any tensor $T \in \mathbb{R}^{d^p}$ can be represented as a TT, if we allow r_k to scale exponentially with d. To obtain an efficient ansatz class that does not suffer from the curse of dimensionality, we have to bound r_k by a polynomial in d. We call $r = \min_{k \in [d]} r_k$ the TT rank of such a decomposition of T.

An important observation is that the TT decomposition is not unique. The *gauge transformation*

$$\begin{aligned} C_1 \mapsto C_1 A_1^{-1}, \qquad & C_d \mapsto A_{d-1} C_d, \\ C_\ell \mapsto A_{\ell-1} C_\ell A_\ell^{-1} \quad & \forall \ell \in \{2,\ldots,d-1\}, \end{aligned} \tag{7}$$

applying matrix multiplication by invertible matrices A_ℓ on the virtual indices of the tensor cores, leaves the TT invariant. Part of this *gauge freedom* is removed by imposing the so-called *left canonical condition*

$$
\begin{array}{c} C_i \\ \square \\ C_i^* \end{array} = \begin{array}{c} C_1 \\ \square \\ C_1^* \end{array} = \square \tag{8}
$$

$\forall i \in \{2, \ldots, d-1\}$. The remaining gauge freedom can be shown to be a unitary transformation.

4. Block-sparsity

In choosing the finite local dictionary, we are truncating the univariate function space, e.g. bounding the polynomial degree or a similar notion relevant to the dictionary at hand. However, if we now take the *d*-fold tensor product $\mathcal{A}$ of the univariate dictionaries as the search space for a multivariate f_k, we introduce terms that are of higher total degree than the degree imposing the local truncation. Thus, from the perspective of multivariate Taylor series, the multivariate dictionary appears inconsistently truncated. Here, it is more natural to work with complete function spaces of bounded *total degree*. In this section we show that such a natural restriction can be conveniently captured by enforcing a certain support pattern of the TT cores, yielding so-called *block-sparse* TTs. This structure allows us to allocate resources much more efficiently—we can enlarge the dictionary while keeping the problem-relevant expressivity of the ansatz class and the computational resources required constant.

More formally, we define a *degree map* $w : [p] \to \mathbb{N}_0$, which assigns the *degree* to the elements of a given function dictionary. Without loss of generality we will assume that w is a non-decreasing function. The degree map provides us with a natural definition of the Laplace-like multivariate degree operator $L : \mathbb{R}^{p^d} \to \mathbb{R}^{p^d}$, which acts on tensor representations of multivariate functions,

$$
L = \sum_{j=1}^{d} \mathrm{Id}_p^{\otimes j-1} \otimes \Omega \otimes \mathrm{Id}_p^{\otimes d-j}, \tag{9}
$$

where $\Omega = \mathrm{diag}(w(1), \ldots, w(p))$ and Id_p is the $p \times p$ identity matrix. This operator is analogous to the bosonic particle number operator in quantum physics.

Suppose a tensor ϕ with a TT representation with cores $\{C_\ell\}_{\ell=1}^{d}$. We define the *left and right interface operators*

$$
L^{<\ell} := \sum_{j=1}^{\ell-1} \mathrm{Id}^{\otimes j-1} \otimes \Omega \otimes \mathrm{Id}^{\otimes \ell-j-1}, \tag{10}
$$

$$
L^{>\ell} := \sum_{j=1}^{d-\ell} \mathrm{Id}^{\otimes j-1} \Omega \otimes \mathrm{Id}^{\otimes d-\ell-j}, \tag{11}
$$

and the *left and right interface tensors*

$$
\begin{array}{c}
\phi^{<\ell} := \quad \overset{C_1 \ \ C_2 \qquad\qquad C_{\ell-1}}{\bullet\!-\!\!\bullet \ \cdots \ -\!\!\bullet} \\[2mm]
\phi^{>\ell} := \quad \overset{C_{\ell+1} \ C_{\ell+2} \qquad\quad C_d}{\bullet\!-\!\!\bullet \ \cdots \ -\!\!\bullet}
\end{array} \tag{12}
$$

for each interface at the ℓth tensor core. When using matrix notation for the interface tensors, we will think of them as linear operators $\phi^{<\ell} : \mathbb{R}^{r_{\ell-1}} \to \mathbb{R}^{p^{\ell-1}}$ and $\phi^{>\ell} : \mathbb{R}^{p^{d-\ell}} \to \mathbb{R}^{r_\ell}$.

In [GST21], the following theorem has been shown.

Theorem 1 (Block-sparsity). *Suppose a TT ϕ with tensor cores $\{C_\ell\}_{\ell \in [d]}$ in left canonical form with minimal ranks $\{r_\ell\}_{\ell=1}^{d-1}$, such that*

$$L\phi = \lambda\phi, \tag{13}$$

where L is as in equation (9) and $\lambda \in \mathbb{N}_0$. Then there exist a unitary gauge transformation $\{A_\ell\}_{\ell \in [d-1]}$ acting via equation (7), such that for each interface $\ell \in [d-1]$ the transformed interface tensors satisfy

$$\phi^{>\ell} L^{>\ell} = \Lambda^{>\ell} \phi^{>\ell}, \tag{14}$$

$$L^{<\ell+1} \phi^{<\ell+1} = \phi^{<\ell+1} \left(\lambda \mathrm{Id} - \Lambda^{>\ell}\right) \tag{15}$$

for a set of diagonal matrices $\left\{\Lambda^{>\ell}\right\}_{\ell \in [d-1]}$ with non-increasing diagonal entries.

We provide a slightly simplified proof in appendix D.

Let us discuss the consequences of theorem 1 and see how it implies a block-sparse structure of the tensor cores. First, notice that equation (14) is an eigenvalue equation, i.e. it states that the rows of $\phi^{>\ell}$ are left eigenvectors of $L^{>\ell}$ with eigenvalues given by the corresponding diagonal element of $\Lambda^{>\ell}$. Similarly, equation (15) implies that the rows of $\phi^{<\ell}$ are right eigenvectors of $L^{<\ell}$ with eigenvalues given by the corresponding diagonal elements of $\Lambda^{<\ell} := \lambda \mathrm{Id} - \Lambda^{>\ell-1}$.

Choosing any interface $\ell \in \{2, \dots, d-1\}$, we can decompose the eigenvalue equation as

with $\Omega = \mathrm{diag}(w(1), \dots, w(p))$, where the double line collects multiple indices into a single edge. Since we assume that the TT ranks are minimal, $\phi^{<\ell}$ ($\phi^{>\ell}$) have full column (row) rank and the eigenvalue equation implies

If we fix the remaining physical index to $i \in [p]$, we obtain the matrix equation

$$(\lambda - w(i))(C_\ell)_i = \Lambda^{<\ell}(C_\ell)_i + (C_\ell)_i \Lambda^{>\ell}. \tag{18}$$

Therefore, for all $i \in [p]$, we can assign each block-row of $(C_\ell)_i$ to a (potentially degenerate) eigenvalue of $L^{<\ell}$ by which it is multiplied in the first term on the RHS of equation (18). Similarly, we can assign each block-column of $(C_\ell)_i$ to a (potentially degenerate) eigenvalue of $L^{>\ell}$ by which it is multiplied in the second term on the RHS of equation (18). The block sizes are given by the degeneracy of the corresponding eigenvalues of the interface operators. Now, equation (18) tells us that blocks of $(C_\ell)_i$, which correspond to eigenvalues $\lambda^{<\ell}, \lambda^{>\ell}$ of $L^{<\ell}$ and $L^{>\ell}$ respectively, can be non-zero only if

$$\lambda^{>\ell} + \lambda^{<\ell} + w(i) = \lambda. \tag{19}$$

Hence, each non-zero block of C_ℓ connects eigenvectors of $L^{<\ell}$ with eigenvectors of $L^{>\ell}$ to fulfil the eigenvalue equation $L\phi = \lambda\phi$. In this way theorem 1 implies block-sparsity of the tensor cores of an eigenvector of L.

In order to analyze the maximum block sizes, consider a block of C_ℓ that corresponds to the eigenvalues $\lambda^{<\ell}$ and $\lambda^{>\ell}$ of the interface operators. Now we need to bound the degeneracy of $\lambda^{<\ell}$ and $\lambda^{>\ell}$ in the spectra of the interface operators. The interface operators are diagonal, with elements given by $\sum_{\ell'=1}^{\ell-1} w(i_{\ell'})$ and $\sum_{\ell'=\ell+1}^{d} w(j_{\ell'})$, respectively, so the maximum size $s_\ell \times t_\ell$ of the considered block, is the number of solutions $(i_1,\ldots,i_{\ell-1}) \in [p]^{\ell-1}$ and $(j_{\ell+1},\ldots,j_d) \in [p]^{d-\ell}$ to

$$\sum_{\ell'=1}^{\ell-1} w(i_{\ell'}) = \lambda^{<\ell} \quad \text{and} \quad \sum_{\ell'=\ell+1}^{d} w(i_{\ell'}) = \lambda^{>\ell}, \tag{20}$$

respectively. Note that if we allowed w to take negative values, the maximum block sizes and number of blocks would become very large.

To obtain a *low-rank* block-sparse TT, we enforce the block-sparse structure and limit the block sizes to some maximum value ρ.

4.1. Example 1: Monomial dictionary
Consider the monomial dictionary

$$\Psi_i(x) = x^{i-1}, \quad i \in [p] \tag{21}$$

and the polynomial degree function $w : i \mapsto i - 1$, so that $\text{Image}(w) = \{0,\ldots,p-1\}$. For physical index i at the ℓ-th mode, there are $\lambda - i + 2$ solutions $\lambda^{<\ell}$ and $\lambda^{>\ell}$ to equation (19), so the matrix $(C_\ell)_i$ has $\lambda - i + 2$ non-zero blocks. Assuming that $p > \lambda^{<\ell}, \lambda^{>\ell}$, a combinatorial argument shows that the number of solutions to equation (20) is

$$s_\ell = \begin{pmatrix} \lambda^{<\ell} + \ell - 2 \\ \ell - 2 \end{pmatrix}, \quad t_\ell = \begin{pmatrix} \lambda^{>\ell} + d - \ell - 1 \\ d - \ell - 1 \end{pmatrix}, \tag{22}$$

which gives us the maximum block sizes. For concreteness, choose $\lambda = 3$ and $p = 3$. Now the block-sparse structure becomes

$$(C_\ell)_1 = \begin{pmatrix} * & 0 & 0 & 0 \\ 0 & * & 0 & 0 \\ 0 & 0 & * & 0 \\ 0 & 0 & 0 & * \end{pmatrix}, \quad (C_\ell)_2 = \begin{pmatrix} 0 & * & 0 & 0 \\ 0 & 0 & * & 0 \\ 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$(C_\ell)_3 = \begin{pmatrix} 0 & 0 & * & 0 \\ 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{23}$$

where $*$ indicates the non-zero blocks. Here the ith row corresponds to $\lambda^{<\ell} = i - 1$ and the jth column to $\lambda^{>\ell} = 4 - j$.

4.2. Example 2: Trigonometric dictionary
Consider the trigonometric dictionary

$$\{1, \sin x, \cos x\} \tag{24}$$

and $w(1) = 0, w(2) = 1, w(3) = 1$, which counts the number of sines and cosines. Choose $\lambda = 3$ and $\ell \in \{4,\ldots,d-4\}$. The block-sparse structure now becomes

$$(C_\ell)_1 = \begin{pmatrix} * & 0 & 0 & 0 \\ 0 & * & 0 & 0 \\ 0 & 0 & * & 0 \\ 0 & 0 & 0 & * \end{pmatrix}, \quad (C_\ell)_2 = \begin{pmatrix} 0 & * & 0 & 0 \\ 0 & 0 & * & 0 \\ 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$(C_\ell)_3 = \begin{pmatrix} 0 & * & 0 & 0 \\ 0 & 0 & * & 0 \\ 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{25}$$

where the ith row corresponds to $\lambda^{<\ell} = i - 1$ and the jth column to $\lambda^{>\ell} = 4 - j$.

4.3. Limiting the degree

Theorem 1 states that *fixed degree functions* admit a *block-sparse TT representation*. An often more natural ansatz class are functions with bounded degree. Fortunately, such functions also admit a block-sparse TT, as can be seen from the following argument.

Suppose a function $f : \mathbb{R}^d$ with a bounded degree λ. Hence, it can be written as a linear combination

$$f = \sum_j f_j, \tag{26}$$

where each f_j has a fixed degree $\leqslant \lambda$ and, hence, admits a block-sparse decomposition. The number of terms $n(\lambda)$ in this sum is bounded by the number of options for a degree $\leqslant \lambda$. Suppose that each f_j has a block-sparse TT decomposition $\{C_\ell^{(j)}\}_{\ell \in [d]}$. Now we can represent the function f by a TT with tensor cores $\{C_\ell\}_{\ell \in [d]}$, where for each $\ell \in \{2, \ldots, d-1\}$ and a value $i \in [p]$ of the physical index, the corresponding matrix $(C_\ell)_i$ is block-diagonal with blocks $(C_\ell^{(j)})_i$, the tensor core C_1 is given by concatenating $C_1^{(j)}$ and the tensor core C_d is

$$\tag{27}$$

where $\tilde{C}_d^{(j)}$ has $C_d^{(j)}$ as its jth block, with all other blocks zero. We sum over the blocks corresponding to each f_j by contracting the right index of C_d with a vector of all ones $1_{n(\lambda)}$. Hence, the TT representation of f becomes

$$\tag{28}$$

which is a block-sparse TT. Note that this is equivalent to the TT addition algorithm of [Ose11].

5. Efficient TT representation of dynamical laws

In this section we will show how generic properties of dynamical systems imply efficient TT representations of their dynamical laws. We write the function $f : \mathcal{S} \to \mathbb{R}^d$ in equation (1) using the decomposition (4) as

$$f_k(x) = \sum_{i_1, \ldots, i_d} \Theta_{k; i_1, \ldots, i_d} \Phi_{i_1, \ldots, i_d}(x). \tag{29}$$

We call the TT decomposition of Θ_k the TT representation of f_k. Combining the tensors for $k \in [d]$, we obtain the tensor $\Theta \in \mathbb{R}^{d \times p^d}$, which can be written as a single TT via

$$\tag{30}$$

providing a TT representation of the dynamical law.

In order to show that a function admits an efficient representation, we will bound its *separation rank*, defined as follows.

Definition 2 (Separation rank). Suppose a multivariate function $h(x) : \mathbb{R}^d \to \mathbb{R}$. We say that $h(x)$ has *separation rank s with respect to a bipartition* $\mathcal{P}_k := (\{x_1, \ldots, x_k\}, \{x_{k+1}, \ldots, x_d\}) := (\mathcal{P}_k^{left}, \mathcal{P}_k^{right})$, if the smallest set of functions $\{g_\ell^{left}(\mathcal{P}_k^{left}), g_\ell^{right}(\mathcal{P}_k^{right})\}_{\ell \in [\bar{s}]}$, such that

$$h(x) = \sum_{\ell \in [\bar{s}]} g_\ell^{left}\left(\mathcal{P}_k^{left}\right) \otimes g_\ell^{right}\left(\mathcal{P}_k^{right}\right), \tag{31}$$

has $\tilde{s} = s$.

It can be shown that the minimal ranks r_k of a TT representation of a function $h(x)$ are equal to the separation ranks with respect to $\mathcal{P}_k$. For a formal proof of this statement, see [HRS12].

The first generic property of dynamical systems that has been used in [GGR+20] to show efficient TT representations of dynamical laws is *locality* in one-dimension.

Definition 3 (One dimensional interacting system, definition 4 in [GGR+20]). A dynamical system, governed by the dynamical law $f : \mathcal{S} \subset \mathbb{R}^d$, is *one-dimensional with interaction length L and separation rank N*, if there exists a function set $\{g_i(x)\}_{i=1}^{\tilde{p}}$ and for each $k \in [d]$ an index set $\mathcal{I}_k \subset [\tilde{p}]^{2L+1}$ with $|\mathcal{I}_k| \leqslant N$, such that

$$f_k(x) = \sum_{i_{k-L},\dots,i_{k+L} \in \mathcal{I}_k} g_{i_{k-L}}(x_{k-L}) \dots g_{i_{k+L}}(x_{k+L}), \tag{32}$$

where we set $g_i = 1$ for $i \leqslant 0$ and $i \geqslant d+1$.

By bounding the separation ranks of the dynamical laws of one dimensional interacting systems with respect to bipartitions $\mathcal{P}_k$, the authors prove the following theorem.

Theorem 4 (Efficient TT decomposition of one-dimensional interacting systems, theorem 5 in [GGR+20]). *Suppose a one dimensional interacting system with interaction length L and separation rank N, governed by the dynamical law $f(x) : \mathcal{S} \subset \mathbb{R}^{p^d} \to \mathbb{R}^d$. Then we have the following.*

(a) Each $f_k(x)$ admits a TT representation with rank $r \leqslant N$ and
(b) the function f admits a TT representation with ranks $r_k \leqslant k - L + 1 + 2NL$.

For many systems of interest exact locality as in definition 3 is too strong of an assumption. We would like to be able to use TTs also for systems that do not have a sharp bound on the interaction length, but where instead the interactions decay with distance. This is formalized by the following definition.

Definition 5 (One-dimensional systems with algebraically decaying interactions). A dynamical system, governed by the dynamical law $f : [0,1)^d \to \mathbb{R}^d$, is *one-dimensional with (χ, g)-algebraically decaying interactions and separation rank N*, if there exists a function set $\{g_{k,L}\}_{k,L \in [d]}$, where for all $k, L \in [d]$ the function $g_{k,L}$ depends non-trivially only on x_i for $i \in [k-L, k+L]$, satisfies $\|g_{k,L}\|_2 \leqslant g$ and has separation rank bounded by N with respect to bipartitions $\mathcal{P}_\ell = \{\{x_{k-L}, \dots, x_\ell\}, \{x_{\ell+1}, \dots, x_{k+L}\}\}$ for all $\ell \in \{k-L, \dots, d+L\}$, such that

$$f_k(x) = \sum_{L \in [d]} L^{-\chi} g_{k,L}(x). \tag{33}$$

Note that in this definition we demand that the state space is $\mathcal{S} = [0,1)^d$. This is important so that $\|g_L\|_2$ does not change with d. Physically, in definition 5, χ is the decay parameter of the interactions, where terms depending on modes at most L apart have a contribution that is modulated by $L^{-\chi}$ relative to a common force scale of the dynamical law g. For one-dimensional systems with algebraically decaying interactions we can show the following theorem.

Theorem 6 (Approximate locality). *Suppose a one-dimensional system with (χ, g)-algebraically decaying interactions and separation rank N, governed by the dynamical law $f : [0,1)^d \to \mathbb{R}^d$, that can be written as*

$$f_k(x) = \sum_{L \in [d]} L^{-\chi} g_{k,L}(x), \tag{34}$$

where $\{g_{k,L}\}_{k,L \in [d]}$ satisfies the assumptions of definition 5. If $\chi > 1$, then for any $\tilde{L} \in [d]$ there exists a one-dimensional interacting system with interaction length $\tilde{L}$ and separation rank $N\tilde{L}$, such that

$$\|f_k(x) - \tilde{f}_k^{\tilde{L}}(x)\|_2 \leqslant c_1(\chi, \tilde{L}) g, \tag{35}$$

where

$$c_1(\chi, \tilde{L}) = (\tilde{L}+1)^{-\chi} \frac{\tilde{L}+\chi}{\chi - 1} = \mathcal{O}(\tilde{L}^{1-\chi}). \tag{36}$$

The proof is given in appendix A.

Theorem 6 allows us to approximate systems with algebraically decaying interactions by strictly local systems with bounded error, which is independent of d.[6] This is formalized in the following corollary.

---

[6] Note that if we were instead looking at the L_2 error of the full function $\|f - \tilde{f}\|_2$, we would get an additional factor of d.

Corollary 7 (Low rank TTs for algebraically decaying interactions). *Suppose a system with (χ, g)-algebraically decaying interactions with separation rank N governed by the dynamical law $f(x) : \mathcal{S} \subset \mathbb{R}^d \to \mathbb{R}^d$. Furthermore, suppose that $\chi > 1$. Then there exists an ε-approximate TT representation of each $f_k(x)$ with rank*

$$r \leqslant N \left[\left(\frac{\chi}{\chi - 1} \frac{g}{\varepsilon} \right)^{\frac{1}{\chi - 1}} - 1 \right]. \tag{37}$$

The details of the proof are given in appendix B.

Many systems of physical interest are *K-mode interacting systems*, formalized by the following definition.

Definition 8 (K-mode interacting systems). A dynamical system, governed by the dynamical law $f : \mathcal{S} \subset \mathbb{R}^d$, is *K-mode interacting with separation rank N*, if there exists a function set $\{g_i(x)\}_{i=1}^{\tilde{p}}$, for each $k \in [d]$ a constant $\mathcal{K}_k$, such that for each $\ell \in [\mathcal{K}_k]$ there are distinct subsets $\mathcal{J}_\ell^{(k)} = \{j_1, \ldots, j_K\} \subset [d]$ with $\mathcal{J}_\ell^{(k)} \ni k$ and subsets $\mathcal{I}_\ell^{(k)} \subset [\tilde{p}]^K$ with $|\mathcal{I}_\ell^{(k)}| \leqslant N$, such that

$$f_k(x) = \sum_{\ell \in [\mathcal{K}_k]} \sum_{(i_{j_1}, \ldots, i_{j_K}) \in \mathcal{I}_\ell^{(k)}} g_{i_{j_1}}(x_{j_1}) \ldots g_{i_{j_K}}(x_{j_K}). \tag{38}$$

An example of a 2-mode interacting system is, e.g. a collection of gravitationally interacting particles, where the total force on each particle is given by the sum of pair-wise forces with respect to the remaining particles. We will now show that if a system is *K*-mode interacting with $K \lesssim 5$, it admits an efficient TT representation.

Theorem 9 (Efficient TT representation of K-mode interacting systems). *Suppose a K-mode interacting system with separation rank N, governed by a dynamical law $f(x) : \mathcal{S} \subset \mathbb{R}^{p^d} \to \mathbb{R}^d$. Then*

(1) *each f_k admits a TT representation with rank $r \leqslant N \begin{pmatrix} d-1 \\ K-1 \end{pmatrix} = N\mathcal{O}(d^{K-1})$ and*

(2) *the function f admits a TT representation with ranks*

$$r_k \leqslant c_2(N, d, k) + k \begin{pmatrix} k-1 \\ K-1 \end{pmatrix} + 1, \tag{39}$$

where $c_2(N, d, k) = \mathcal{O}(Nd^K)$ is a combinatorial factor defined in appendix C.

To prove theorem 9, we adapt the techniques of the proof of theorem 4 from [GGR+20], see appendix C for details.

Already for $K \gtrsim 5$, although $c_2(N, d, k)$ is polynomial in d, the polynomial degree makes working with such TTs prohibitively expensive even for modest d. However, many systems of physical interest are known to have $K = 2, 3, 4$ and hence admit an efficient TT representation. It is important to note that this result does not rely on the underlying systems being one-dimensional. Finally, note that if a dynamical system has *at most* K-mode interactions, it still admits an efficient TT decomposition, since the rank of a TT is sub-additive.

General conditions for the approximability of multivariate functions by a low rank TT in terms of tail control of the singular value spectrum of the matrix unfoldings of their coefficient tensors are derived in [BSU16]. As we discuss in appendix E, these results are in a precise sense analog to the control of MPS approximations of quantum states based on entropy scaling conditions derived in the quantum many-body literature [SWVC08, VC06].

6. ALS-GMWS

6.1. Self-similarity

Additional structure in the system, known prior to learning, can cause certain modes to play the same role in dynamical laws for multiple modes, implying that we would like the corresponding tensor cores to be equal. We call this *self-similarity*. The different roles each mode can play are referred to as *activation types*. For example, in the case of a one-dimensional dynamical system with interaction length L, definition 3, the jth mode plays the same role in all functions f_k with $k < j - L$, namely that the mode is to the left and outside the interaction range. Similarly, the role of the mode is the same for all $k > j + L$. Hence, such systems have $2L + 3$ activation types.

Self-similar systems with α activation types can be described by a set of $d\alpha$ tensor cores $\{C_\ell^{(j)}\}_{\ell\in[d],j\in[\alpha]}$. The recipe to build the corresponding tensor train representations of the dynamical laws $f_k(x)$ can be encoded in a selection table $S \in [\alpha]^{d\times d}$ as in

$$
\begin{array}{c}
C_1^{(S_{k,1})} \quad\quad C_d^{(S_{k,d})} \\
\bullet \quad \cdots \quad \stackrel{\bullet}{\underset{\mathbf{1}_{n(\lambda)}}{\bullet}} \stackrel{}{\underset{}{\bullet}} \mathrm{Id}_{n(\lambda)}
\end{array} \;,
\tag{40}
$$

where block-sparse representation of functions with bounded degree equation (28) is used. We call such systems S-self-similar.

In the case of one-dimensional interacting systems with interaction length L, the selection table takes the form

$$
S_{ij}^L = \begin{cases} 1 & j < i - L, \\ 2L+3 & j > i + L, \\ j - i + L + 2 & \text{otherwise.} \end{cases}
\tag{41}
$$

6.2. ALS optimization of self-similar systems

In learning dynamical laws, given data $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^d$ for $i \in [M]$ with $y_i \approx f(x_i)\ \forall i$, we would like to identify an element $\hat{f}$ of some ansatz class that minimizes the empirical loss

$$
\mathcal{L}_{\mathrm{emp}}\left(\hat{f}\right) := \sum_{i\in[M]} \left\| y_i - \hat{f}(x_i) \right\|_2^2.
\tag{42}
$$

In the previous sections we have shown that a natural choice of an ansatz class for many dynamical systems of interest are block-sparse low rank TTs with self-similarity given by the selection table equation (40).

Previously, in [GGR+20], *alternating least squares* (ALS) optimization (and a rank-adaptive variant [GK19]) has been used to minimize $\mathcal{L}_{\mathrm{emp}}$ over low rank TTs. In the ALS procedure, the tensor cores are iterated over in sweeps, at each step solving a linear least squares problem to minimize the empirical loss as a function of the given core, until convergence. The ALS algorithm can be adapted to block-sparse TT by restricting each contraction in the algorithm to indices labeling elements that are non-zero in the block-sparse structure [GST21].

However, optimization over systems with self-similarity requires more care. In [GGR+20] a selection tensor approach has been taken, where the dynamical law is written as

$$
\theta_k = \begin{array}{c} k \\ \mathcal{T} \\ C_1 \;\; C_2 \;\; \cdots \;\; C_d \end{array} \;,
\tag{43}
$$

where θ_k is the TT representation of f_k and $\mathcal{T}$ is the selection tensor that, given $k \in [d]$, picks the correct activation type. In this form, block-sparse ALS can be directly applied. However, this method leads to numerical instabilities in the optimization, limiting its scalability to $d \lesssim 20$.

The core problem is that the unitary gauge freedom is not properly taken care of. To see this, suppose that the ALS sweeps are performed from left to right and that the tensor core $C_\ell^{(j)}$ corresponding to the mode $\ell \in [d]$ and activation type $j \in [\alpha]$ is being optimized. The left neighbour of this tensor core can have a different activation type in different dynamical law components k. Since each of the left neighbour activation types has been optimized in a separate ALS step, they are each written in a different gauge. Using the form equation (43) does not allow the gauge differences to be corrected for, and therefore we rely on the gauges to converge as more ALS sweeps are performed, which leads to numerical instabilities.

We will now introduce a new algorithm for optimization of block-sparse low rank TTs with self-similarity, called ALS-GMWS, that allows the left neighbour gauge to be adjusted in each step. It will be useful to define the *restricted empirical cost*, which, given a subset $E \subset [d]$, is given by

$$
\mathcal{L}_{\mathrm{emp}}^E\left(\hat{f}\right) := \sum_{i\in[M]} \left\| (y_i)_E - \left(\hat{f}(x_i)\right)_E \right\|_2^2,
\tag{44}
$$

where the notation $(v)_E$ for $v \in \mathbb{R}^d$ denotes the restriction of v to the subspace defined by the set of index values E.

Assume without loss of generality that the ALS sweeps are performed left to right. Suppose we want to optimize the core $C_\ell^{(j)}$, for $\ell \in [d]$ and $j \in [\alpha]$. Let $\mathcal{E} \subset [d]$ be the set of indices of functions where this core is used, as determined by the selection table S. If $\ell = 1$, we do not need to worry about the gauge at all, since we are optimizing left to right. We simply perform a block-sparse ALS step to find $C_\ell^{(j)}$ that minimizes $\mathcal{L}_{\text{emp}}^{\mathcal{E}}(\hat{f})$ with all the other tensor cores fixed.

Now consider $\ell \geqslant 2$. Among the functions labeled by $e \in \mathcal{E}$, we are only sure that the gauge is the same in the equivalence classes defined by the values of $S_{e,\ell-1}$, or, in other words, only if the left neighbour, the tensor core of the $\ell - 1$th mode, has the same activation type. Hence, we divide $\mathcal{E}$ into disjoint sets $\{E_a \subset \mathcal{E}\}_{a \in [\alpha]}$ labeled by the activation type of $\ell - 1$th mode, such that, for each $a \in [\alpha]$, $S_{e,\ell-1} = a$ for all $e \in E_a$. In order to use the most information available, we find $\tilde{a} \in [\alpha]$, such that $|E_{\tilde{a}}| \geqslant |E_a| \; \forall a \in [\alpha]$, and perform the block-sparse ALS step to find the $C_\ell^{(j)}$ that minimizes $\mathcal{L}_{\text{emp}}^{E_{\tilde{a}}}(\hat{f})$, with all the other tensor cores fixed.

Now that we found the optimal $C_\ell^{(j)}$ with respect to $E_{\tilde{a}}$, we optimize the gauge of $C_{\ell-1}^{(a)}$ for $a \neq \tilde{a}$, so that using the newly found tensor core components in all the corresponding dynamical law components is justified. Hence, for each $a \in [\alpha]$, such that $a \neq \tilde{a}$ and $|E_a| \neq 0$, we want to find a gauge fixing unitary U_a that minimizes $\mathcal{L}_{\text{emp}}^{E_a}(\hat{f})$ under the transformation

$$
\begin{array}{ccc}
C_{\ell-1}^{(a)} & & C_{\ell-1}^{(a)} \, U_a \\
\bullet & \mapsto & \bullet \!-\! \bullet
\end{array}
\quad , \tag{45}
$$

with all other tensor cores fixed. In order to preserve the block-sparse structure, U_a needs to be block-diagonal with block sizes given by the block-column widths of $C_{\ell-1}^{(a)}$ and block-row heights of $C_\ell^{(j)}$. In practice, we optimize over all block-diagonal matrices, although optimization over block-diagonal unitaries is possible and could lead to improvements.

The ALS-GMWS algorithm is summarized in algorithm 1.

Algorithm 1. ALS optimization with gauge mediated weight sharing.

---

input: Data pairs $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^d$, $i = 1, \ldots, M$, number of activation types α, selection table S, maximum block size ρ
output: $d\alpha$ block-sparse tensor train components.
For $\ell = [d]$, $j = [\alpha]$ randomly initialize block-sparse $C_\ell^{(j)}$ with maximum block size ρ
While not converged
  for $\ell = 1, \ldots, d$
    for $k = 1, \ldots, \alpha$
      Let $\mathcal{E} = \{e \in [d] : S_{e,\ell} = j\}$
      if $\ell = 1$ or $|\mathcal{E}| = 1$ **then**
        Find $C_\ell^{(k)}$ that minimizes $\mathcal{L}_{\text{emp}}^{\mathcal{E}}$ with all other tensor cores fixed;
      end
      else
        For each $a \in [\alpha]$ let $E_a = \{e \in \mathcal{E} : S_{e,\ell-1} = a\}$;
        Choose $\tilde{a} \in [\alpha]$, s.t. $|E_{\tilde{a}}| \geqslant |E_a| \; \forall a \in [\alpha]$;
        Find $C_\ell^{(k)}$ that minimizes $\mathcal{L}_{\text{emp}}^{E_{\tilde{a}}}$ with all other tensor cores fixed;
        for $a \in [\alpha] : a \neq \tilde{a}$, $|E_a| \neq 0$ **do**
          Find a block-diagonal unitary U that minimizes $\mathcal{L}_{\text{emp}}^{E_a}$ after the transformation equation (45) with all the
          other tensor cores fixed   Set $C_{\ell-1}^{(a)} \leftarrow C_{\ell-1}^{(a)} U$
        end
      end
    end
  end
end
return C_k^ℓ, $k = 1, \ldots, d$, $\ell = 1, \ldots, \alpha$.

---

7. Numerical experiments

We will demonstrate our method on three example dynamical systems: The Fermi–Pasta–Ulam–Tsingou (FPUT) system, one-dimensional chain of rotating magnetic dipoles and a chain of atoms interacting via a modified Lennard–Jones interaction. The experiments have been performed using the codebase developed by M Götte and P Trunschke [MP22].

7.1. FPUT system

The FPUT system is a chain of non-linear springs with spring constants κ_ℓ. The dynamical law is given by

$$
\begin{aligned}
f_k(x) =& \kappa_{k+1}(x_{k+1} - x_k) - \kappa_k(x_k - x_{k-1}) \\
& + \beta_{k+1}(x_{k+1} - x_k)^3 - \beta_k(x_k - x_{k-1})^3 \\
& k = 1, \ldots, d,
\end{aligned}
\tag{46}
$$

which is a one-dimensional system with interaction length $L = 1$ and, using the monomial dictionary equation (21), separation rank $N = 4$, such that each f_k can be represented by TTs with rank bounded by $r = 4$. Furthermore, the polynomial degree of the equations is bounded by 3 and, hence, we can use TTs (28) with block-sparse structure given by equation (23) for $\lambda = 3$ and block sizes bounded by 4, to represent the system exactly. We study two instances of this system. First, we study a translationally invariant system, setting $\kappa_k = 1, \beta_k = 0.7 \, \forall k \in [d]$. Second, to demonstrate the ability of our method to recover a disordered system, we use an instance where each κ_k, β_k is drawn i.i.d. randomly from the uniform distribution on $[0, 2]$ and $[0, 1.4]$ respectively.

7.2. Rotating magnetic dipoles

Here, we have a chain of magnetic dipoles at positions X_ℓ with magnetic dipole moments M_ℓ and moments of inertia I_ℓ. They are free to rotate in the plane perpendicular to the chain and their angles of rotation are $x_\ell \in [0, 2\pi)$. The dynamical law is given by

$$
f_k(x) = I_k M_k \sum_{\ell \neq k} \frac{M_\ell}{|X_k - X_\ell|^3} \sin(x_k - x_\ell).
\tag{47}
$$

The positions are chosen so that $X_1 < X_2 < \ldots < X_d$. In fact, we set $M_\ell = I_\ell = 1$, $X_\ell = \ell - 1 \, \forall \ell \in [d]$. According to definition 8, this is a 2-mode interacting system with separation rank $N = 2$, using the trigonometric dictionary (24). Hence, it suffices to use rank $r = 2(d - 1)$ TTs. Since the degree given by $w(1) = 0, w(1) = w(2) = 1$ is bounded by 2, we can use the block-sparse structure

$$
(C_\ell))_1 = \begin{pmatrix} * & 0 & 0 \\ 0 & * & 0 \\ 0 & 0 & * \end{pmatrix}, (C_\ell)_2 = \begin{pmatrix} 0 & * & 0 \\ 0 & 0 & * \\ 0 & 0 & 0 \end{pmatrix},
$$

$$
(C_\ell)_3 = \begin{pmatrix} 0 & * & 0 \\ 0 & 0 & * \\ 0 & 0 & 0 \end{pmatrix}
\tag{48}
$$

with block sizes bounded by $2(d - 1)$ and represent this dynamical law exactly. However, this is also a system with $(3, \frac{1}{2\sqrt{2\pi}})$-algebraically decaying interactions (after rescaling $x_\ell \mapsto (2\pi)^{-1} x_\ell$, so that $x_\ell \in [0, 1)$) and separation rank 2 according to definition 5. Hence, corollary 7 allows us to limit the block sizes to a constant (in d) and get an approximation of the dynamical law with bounded error.

7.3. Lennard–Jones chain

The final example is a chain of particles of masses m_ℓ that interact via a modified Lennard–Jones potential. This is the hardest example to learn of the three. The dynamical law for the positions x_ℓ of the particles along the chain is given by

$$
f_k(x) = 6m_k \sum_{\ell \neq k} \text{sign}(x_k - x_\ell) \frac{\varepsilon_{k,\ell}}{R_{k,\ell}} \left(2 \left(\frac{R_{k,\ell}}{|x_k - x_\ell|} \right)^{2q+1} - \left(\frac{R_{k,\ell}}{|x_k - x_\ell|} \right)^{q+1} \right),
\tag{49}
$$

where $\varepsilon_{k,\ell}$, $R_{k,\ell}$ are parameters of the interaction between the modes k, ℓ. We set $m_k = \varepsilon_{k,\ell} = R_{k,\ell} = 1 \, \forall k, \ell \in [d]$ and $q = 2$. Since it is hard to approximate inverse functions with polynomial dictionaries, we learn

$$
g_k(x) := (x_k - x_{k-1})^{2q+1}(x_k - x_{k+1})^{2q+1} f_k(x),
\tag{50}
$$

instead of equation (49) from the accordingly transformed data $\{x_i, \tilde{y}_i\}_{i \in [M]}$.

This is a 2-mode interacting system with separation rank $N = p^2$ with respect to polynomial dictionaries, truncating the polynomial expansions of inverse functions to p terms. By theorem 9, we can represent the dynamical law by TTs with rank bounded by $r = p^2(d - 1)$. For good approximations of the inverse function,

we require large p, so for practical use-cases we would like to limit the rank more. This is justified by corollary 7, since this is also a system with $(2,g)$-algebraically decaying interactions and separation rank p^2, if there is a finite amount of energy in the system and the initial conditions are chosen so that we can ensure that the state space is $\mathcal{S} \subset \bigtimes_{\ell \in [d]} \mathcal{I}_\ell$, where $\mathcal{I}_\ell$ are finite intervals for all $\ell \in [d]$.

The polynomial degree of equation (49) is bounded by $2p$, such that we can use block-sparsity, where each core $(C_\ell)_i$ has non-zero blocks only on the ith diagonal.

For completeness, we include the formula for the total energy in the system

$$E = \sum_{k < \ell} \varepsilon_{k,\ell} \left[\left(\frac{R_{k,\ell}}{|x_k(0) - x_\ell(0)|} \right)^{2q} - \left(\frac{R_{k,\ell}}{|x_k(0) - x_\ell(0)|} \right)^q \right] + \sum_k \frac{1}{2} m_k \dot{x}_k^2(0). \tag{51}$$

7.4. Results

For each of the three example systems, we randomly draw data $\mathcal{D} = \{x_i\}_{i \in [M]}$ from the corresponding state space $\mathcal{S}$ and compute $y_i = f(x_i)$ or $y_i = f(x_i) + \eta_i^\sigma$, where the elements of η_i^σ are drawn from the Gaussian distribution with standard deviation σ. Unless stated otherwise, we use noiseless data with $\sigma = 0$. Therefore, the data that we use for learning does not come from trajectories of the dynamical systems, but instead they are random (potentially noisy) evaluations of the dynamical law $f(x)$. This somewhat simplifies the setting, especially since we do not have to approximate $\dot{x}(t)$ or $\ddot{x}(t)$ in equation (1) from the trajectory, using e.g. finite differences. However, since we can hope to learn the dynamical laws only from trajectories that sufficiently explore the state space $\mathcal{S}$, sampling $\mathcal{S}$ at random is not too different from using such trajectories.

For each dynamical system we set the parameters as described in the previous section. We initialize the block sparse TT cores randomly, drawing the non-zero components from the normal distribution with zero mean and unit standard deviation. Since we observe that our algorithm converges on the first try, we only perform the recovery once on each training set, except for the Lennard-Jones system, where we initialize 6 times.

Given $\mathcal{D}$, we use ALS-GMWS to find an S_L-self-similar low rank block-sparse TT representation of an estimate $\hat{f}$ of the dynamical law. We benchmark the quality of the estimate with respect to the true dynamical law f via the residuum

$$\text{res}\left(\hat{f}, f\right) := \sqrt{\frac{\sum_{i \in [M']} \left\| \hat{f}(x_i') - f(x_i') \right\|_2^2}{\sum_{j \in [M']} \| f\left(x_j'\right) \|_2^2}}, \tag{52}$$

where $\{x_i'\}_{i \in [M']}$ are random samples from $\mathcal{S}$, which are different to the samples used for training. In particular, we use $M' = 2 \times 10^4$.
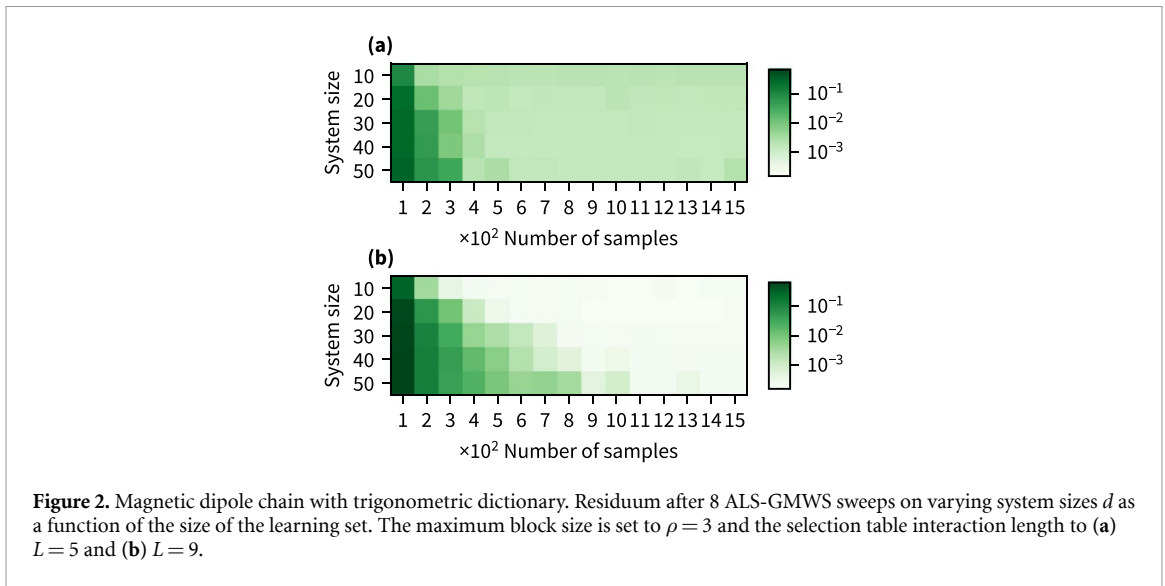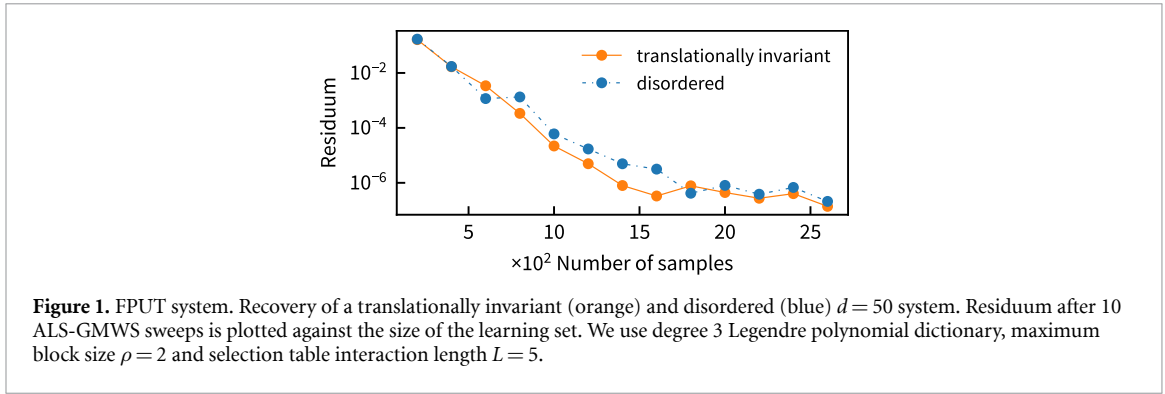
All experiments have been conducted on consumer grade hardware and the code has not been optimized for speed.

The block-sparsity used for polynomial dictionaries is that non-zero blocks of $(C_\ell)_i$ are on the i-th block-diagonal, while for the trigonometric dictionary equation (24) it is such that $(C_\ell)_0$ is block-diagonal, while $(C_\ell)_1$ and $(C_\ell)_2$ have non-zero blocks on the first block-diagonal. In both cases, if we bound the degree by λ, the number of block-rows and block-columns is $\lambda + 1$. The maximum block size is ρ. For all experiments we have used self-similarity given by the selection table S_L defined in equation (40).

7.4.1. FPUT system

To recover FPUT systems, equation (47), with size $d = 50$, we use degree 3 Legendre polynomial dictionary. In figure 1 we show the recovery of the translationally invariant FPUT system and FPUT system with randomly sampled spring constants, to demonstrate applicability of our method to disordered systems. We plot the residuum achieved after 10 ALS-GMWS sweeps, using varying numbers M of training samples. Both systems are successfully recovered using 1800 training samples.

We will now compare these results to those of [GGR+20] and demonstrate the improvement provided by ALS-GMWS with block sparsity. Figure 5 in [GGR+20] shows the recovery of the FPUT system by two methods—ALS and SALSA. The former algorithm achieves a recovery rate of about 90% (with 10^{-6} error threshold) with 4500 samples on a $d = 18$ system, while SALSA has only been demonstrated on up to $d = 12$ systems and shows a 100% recovery rate with about 3000 samples. The authors mention that SALSA becomes intractable for larger systems on a desktop hardware. In contrast, ALS-GMWS presented here is able to achieve a 100% recovery rate already with about 1800 samples on a $d = 50$ system, almost tripling the attainable system size. Furthermore, it only requires 10 ALS-GMWS sweeps, compared to 15 ALS sweeps and 60 SALSA iterations required to obtain the quoted results, keeping the computational cost well suited for desktop hardware.

Figure 1. FPUT system. Recovery of a translationally invariant (orange) and disordered (blue) $d = 50$ system. Residuum after 10 ALS-GMWS sweeps is plotted against the size of the learning set. We use degree 3 Legendre polynomial dictionary, maximum block size $\rho = 2$ and selection table interaction length $L = 5$.

Figure 2. Magnetic dipole chain with trigonometric dictionary. Residuum after 8 ALS-GMWS sweeps on varying system sizes d as a function of the size of the learning set. The maximum block size is set to $\rho = 3$ and the selection table interaction length to (**a**) $L = 5$ and (**b**) $L = 9$.

7.4.2. Rotating magnetic dipoles

We perform three experiments on the rotating magnetic dipole chain.

First, we use the trigonometric dictionary to recover chains with $10 \leqslant d \leqslant 50$. In figure 2 we plot the residua achieved using varying numbers M of training samples. We show the results for $L = 5, 9$. For $d = 50$, successful recovery requires around 4×10^2 training samples for $L = 5$ and 9×10^2 training samples for $L = 9$, which however achieves around 10 times smaller residuum.

Second, we use degree 9 Legendre polynomial dictionary to recover chains with $d = 10, 20, 30$. The residua for varying numbers M of training samples are shown in figure 3. For $d = 30$, we require around 1.7×10^4 samples. This demonstrates the importance of choosing an appropriate dictionary, when compared with the previous experiment.
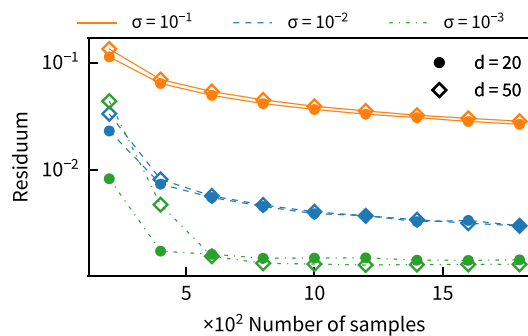
Finally, we use trigonometric dictionary to recover chains with $d = 20, 50$ from noisy data with varying levels σ of noise. We plot the residua achieved using varying numbers M of training samples in figure 4. The results show recovery of the system down to the noise level, demonstrating noise robustness of the proposed method.

7.4.3. Lennard-Jones chain

We recover the Lennard-Jones chain equation (49) for $d = 10$, using degree 8 Legendre polynomial dictionary. We set $L = 5$. Figure 5 shows the residua after 8 ALS-GMWS sweeps for various maximum block sizes, as a function of the training set size. This is clearly the hardest example, requiring around 6×10^3 samples for successful recovery of even such a small system. Furthermore, the plot shows three initialization instances when the algorithm has not converged well. In practice it is therefore sometimes beneficial to run the algorithm multiple times with different initializations.
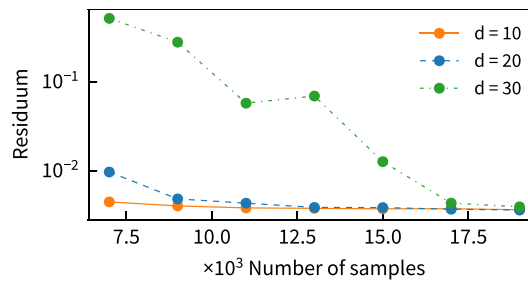
Figure 3. Magnetic dipole chain with Legendre polynomial dictionary. Residuum after 8 ALS-GMWS sweeps on system sizes $d = 10, 20, 30$ is plotted as a function of the learning set size. The maximum block size is set to $\rho = 3$, dictionary is truncated at degree 9 and selection table interaction length to $L = 5$.

Figure 4. Magnetic dipole chain from noisy data. Residuum after 8 ALS-GMWS sweeps using noisy data with varying levels of noise σ for system sizes $d = 20$ (dots) and $d = 50$ (diamonds). We use the trigonometric dictionary, maximum block size $\rho = 3$ and selection table interaction length $L = 5$.
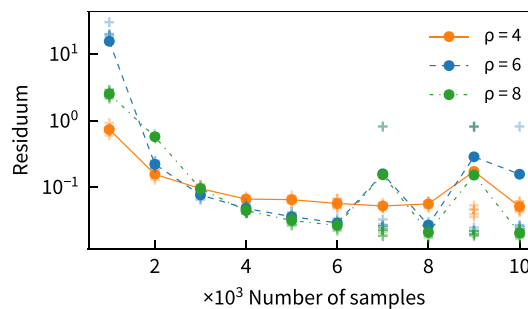
Figure 5. Lennard-Jones chain. The residuum after 8 sweeps of the ALS-GMWS algorithm is plotted for system size $d = 10$ and maximum block sizes $\rho = 4, 6, 8$. The mean (dots) over 6 runs of the algorithm is plotted, as well as the result of the individual instances (plus signs). We use degree 8 Legendre polynomial dictionary and set selection table interaction length to $L = 5$.

8. Conclusions

Learning dynamical laws from data is a key task that has been lately moving into the focus of attention. The well-known and immensely popular SINDy approach introduces a data-driven algorithm for obtaining dynamical systems from data. This approach allows for a reliable recovery of small systems, but is not scalable to systems with a large number of degrees of freedom: For this, a meaningful, physically motivated restriction of the hypothesis class is necessary. In this work, we have overcome this obstacle. We have shown that block-sparse TT with self-similarity provide a suitable efficient ansatz class for learning dynamical laws from data in many contexts of practical interest. In particular, we have proven that these include local one-dimensional systems, one-dimensional systems with algebraically decaying interactions and systems with K-body interactions in any number of spatial dimensions. For learning dynamical laws within this class, we have developed a new variant of the ALS algorithm for block-sparse TTs, which we refer to as ALS-GMWS, which is suitable for self-similar systems. The method has been successfully demonstrated on

three physically relevant one-dimensional dynamical systems and robustness to Gaussian additive noise in the data has been demonstrated.

Data availability statement

Acknowledgments

Appendix A. Proof of theorem 6

Theorem 6 bounds for $\chi > 1$ the error that we obtain by representing the dynamical law of a system with (χ, g)-algebraically decaying interactions and separation rank N by a low rank TT. The original system is governed by the dynamical law

$$f_k(x) = \sum_{L \in [d]} L^{-\chi} g_{k,L}(x). \tag{53}$$

Suppose the approximation of this system

$$\tilde{f}_k^{(\tilde{L})}(x) = \sum_{L \in [\tilde{L}]} L^{-\chi} g_{k,L}(x). \tag{54}$$

The approximation error is bounded by

$$\|f_k(x) - \tilde{f}_k^{(\tilde{L})}(x)\|_2 \leqslant \sum_{L=\tilde{L}+1}^{d} L^{-\chi} g \tag{55}$$

$$\leqslant \sum_{L=\tilde{L}+1}^{\infty} L^{-\chi} g \tag{56}$$

$$\leqslant c_1(\chi, \tilde{L}) g \tag{57}$$

$$= \mathcal{O}(L^{1-\chi}) g, \tag{58}$$

where we use L_2 norm sub-additivity to get the first inequality and the last inequality comes from using the upper bound $\sum_{j=N}^{\infty} f(j) \leqslant f(N) + \int_N^{\infty} f(x) \mathrm{d}x$ for $f(x)$ that decreases monotonously for all $x \in [N, \infty)$. Evaluating the integral for the present case $N = \tilde{L} + 1, f(x) = x^{-\chi}$ with $\chi > 1$, we get

$$\sum_{L=\tilde{L}+1}^{\infty} L^{-\chi} \leqslant (\tilde{L}+1)^{-\chi} \frac{\tilde{L}+\chi}{\chi - 1} \tag{59}$$

$$:= c_1(\chi, \tilde{L}) \tag{60}$$

$$= \mathcal{O}(\tilde{L}^{1-\chi}). \tag{61}$$

It remains to be shown that the truncated dynamical system corresponding to the dynamical law $\tilde{f}^{(\tilde{L})}$ is one-dimensional with interaction length $\tilde{L}$ and separation rank $N\tilde{L}$. Since all the terms in the sum equation (54) depend trivially on x_i with $i \notin [k - \tilde{L}, k + \tilde{L}]$, the system has interaction length $\tilde{L}$ by definition. Moreover, since it is a sum of $\tilde{L}$ functions with separation rank at most N with respect to any bipartition $\mathcal{P}_k$, the total separation rank is bounded by $N\tilde{L}$. $\qquad\square$

Appendix B. Proof of corollary 7

Corollary 7 shows that there exists ε-approximate low rank TT representations of dynamical systems with (χ, g)-algebraically decaying interactions and separation rank N, if $\chi > 1$. To prove this, use theorem 6 to find that there exists a one-dimensional dynamical system with interaction length $\tilde{L}$ and separation rank $N\tilde{L}$, which is governed by a dynamical law $\tilde{f}$, such that, for all $k \in [d]$, $\tilde{f}_k$ is an ε-approximation of f_k in the L_2 norm, if

$$\left(\tilde{L}+1\right)^{-\chi} \frac{\tilde{L}+\chi}{\chi-1} \leqslant \frac{\varepsilon}{g} \tag{62}$$

$$\left(\tilde{L}+1\right)^{1-\chi} + (\chi-1)\left(\tilde{L}+1\right)^{-\chi} \leqslant (\chi-1)\frac{\varepsilon}{g}. \tag{63}$$

This will certainly be satisfied if

$$\left(\tilde{L}+1\right)^{1-\chi} \leqslant \frac{\chi-1}{\chi}\frac{\varepsilon}{g}. \tag{64}$$

Rearranging, we obtain

$$\tilde{L} \geqslant \left(\frac{\chi}{\chi-1}\frac{g}{\varepsilon}\right)^{\frac{1}{\chi-1}} - 1. \tag{65}$$

From theorem 4 we know that in this approximation, each $\tilde{f}_k$ admits a TT representation with rank $r \geqslant N\tilde{L}$, so

$$r \geqslant N\left[\left(\frac{\chi-1}{\chi}\frac{\varepsilon}{g}\right)^{\frac{1}{1-\chi}} - 1\right]. \tag{66}$$

$\square$

Appendix C. Proof of theorem 9

Theorem 9 bounds the TT rank of the TT representation of the dynamical law $f(x)$. In order to do this, we need to bound the separation rank with respect to bipartitions $\mathcal{P}_k = (\{x_1, \ldots, x_k\}, \{x_{k+1}, \ldots, k_d\})$. To prove 1., first we need to show that

$$\mathcal{K}_k \leqslant \binom{d-1}{K-1}. \tag{67}$$

To see this note that we require each $\mathcal{J}_\ell^{(k)}$ to be distinct and to contain k. Hence, the upper bound is the number of ways of selecting the remaining $K-1$ elements of $\mathcal{J}_\ell^{(k)}$. Now, rewrite the decomposition of the dynamical law of a K-mode interacting system with separation rank N as

$$\begin{aligned}
f_k(x) &= \sum_{\ell \in [\mathcal{K}_k]} \sum_{(i_{j_1}, \ldots, i_{j_K}) \in \mathcal{I}_\ell^{(k)}} g_{i_{j_1}}\left(x_{j_1}\right) \ldots g_{i_{j_K}}\left(x_{j_K}\right) \\
&= \sum_{\ell \in [\mathcal{K}_k]} \sum_{(i_{j_1}, \ldots, i_{j_K}) \in \mathcal{I}_\ell^{(k)}} \bigotimes_{m=1}^{d} \begin{cases} g_{i_m}\left(x_m\right) & m \in \mathcal{J}_\ell^{(k)}, \\ \mathrm{Id}\left(x_{j_m}\right) & \text{otherwise}, \end{cases}
\end{aligned} \tag{68}$$

where $\mathrm{Id}(x) = 1$. This is a sum of tensor products of single variable functions of each mode. Hence, the separation rank of f_k with respect to any $\mathcal{P}_k$ is upper bounded by the number of terms in the sum, which is $\mathcal{K}_k|\mathcal{I}_k^{(\ell)}| \leqslant \binom{d-1}{K-1}N$. Since the minimal ranks of the TT representation of f_k are equal to the corresponding separation ranks, we get that

$$r_k \leqslant r = N\binom{d-1}{K-1} = N\mathcal{O}\left(d^{K-1}\right). \tag{69}$$

The proof of 2. proceeds along similar lines. We write the total dynamical law $f(x)$, which we think of in terms of equation (30), as

$$f(x) = \sum_{k' \in [d]} \sum_{\ell \in [\mathcal{K}_{k'}]} \tilde{g}_{k',\ell}(x) \otimes e_{k'}, \tag{70}$$

where $e_{k'} \in \mathbb{R}^d$ is the vector with one at the k'-th element and zeroes elsewhere. Here each $\tilde{g}_{k',\ell}$ represents a function that depends non-trivially only on x_i with $i \in \mathcal{J}_\ell^{(k')}$ and which has a separation rank at most N with respect to any bipartition of $[d]$, due to the assumption that $f(x)$ is a dynamical law of a K-body interacting system with separation rank N. Given a bipartition $\mathcal{P}_k = (\mathcal{P}^{\text{left}}, \mathcal{P}^{\text{right}}) = \{\{x_1, \ldots, x_k\}, \{x_{k+1}, \ldots, x_d\}\}$, we can write

$$f(x) = \text{Id}_{1,k} \otimes \sum_{k' \in [d]} \sum_{\ell \in [\mathcal{K}_{k'}]:\mathcal{J}_\ell^{(k')} \subset \mathcal{P}_k^{\text{right}}} \tilde{g}'_{k',\ell} \otimes e_{k'} \tag{71}$$

$$+ \sum_{k' \in [d]} \sum_{\ell \in [\mathcal{K}_{k'}]:\mathcal{J}_\ell^{(k')} \subset \mathcal{P}_k^{\text{left}}} \tilde{g}''_{k',\ell} \otimes \text{Id}_{k+1,d} \otimes e_{k'} \tag{72}$$

$$+ \sum_{k' \in [d]} \sum_{\ell \in [\mathcal{K}_{k'}]:\mathcal{J}_\ell^{(k')} \not\subseteq \mathcal{P}_k^{\text{left,right}}} \tilde{g}_{k',\ell} \otimes e_{k'}, \tag{73}$$

where the Id_{k_1,k_2} is the one function of $x_{k_1}, \ldots, x_{k_2}$, and we abuse the subset notation in $\mathcal{J}_\ell^{(k')} \subset \mathcal{P}_k^{\text{left,right}}$ and similar to indicate that $x_j \in \mathcal{P}_k^{\text{left,right}}$ for all $j \in \mathcal{J}_\ell^{(k')}$. Furthermore, we denote by $g'_{k',\ell}, g''_{k',\ell}$ the restriction of $g_{k',\ell}$ onto the modes $x_{k+1}, \ldots, x_d$ and $x_1, \ldots, x_k$ respectively. Note that this is well defined since we always use this notation in the cases where g depends trivially on the modes that we throw away.

Each term (71), (72), (73) is now written in such a way that we can read of a bound on its separation rank with respect to $\mathcal{P}_k$. The term (71) has separation rank with respect to $\mathcal{P}_k$ at most 1. The term (72) vanishes if $k' \geqslant k+1$, since then the condition on the second sum cannot be satisfied. For each $k' \leqslant k$, the number of terms in the sum, each of which has separation rank 1 with respect to $\mathcal{P}_k$ is upper bounded by $\binom{k-1}{K-1}$, so the bound on the separation rank of equation (72) is $k\binom{k-1}{K-1}$. Finally, in the last term (73), we know that each $\tilde{g}_{k',\ell}$ has separation rank with respect to any bipartition bounded by N. The number of $\ell \in [\mathcal{K}_{k'}]$ such that $\mathcal{J}_\ell^{(k')}$ satisfies the condition is

$$\begin{aligned} c_2(N, d, k) &:= N\left[d\binom{d-1}{K-1} - k\binom{k-1}{K-1} - (d-k)\binom{d-k-1}{K-1}\right] \\ &= \mathcal{O}\left(Nd^K\right). \end{aligned} \tag{74}$$

Putting all the bounds together, we recover claim 2. of the theorem.   □

Appendix D. Proof of theorem 1

Theorem 1 states that functions with fixed degree given by the degree map w admit a TT representation, such that the left and right interface tensors satisfy the eigenvalue equations

$$\phi^{>\ell} L^{>\ell} = \Lambda^{>\ell} \phi^{>\ell}, \tag{75}$$

$$L^{<\ell+1} \phi^{<\ell+1} = \phi^{<\ell+1} \left(\lambda \text{Id} - \Lambda^{>\ell}\right). \tag{76}$$

We have a TT ϕ with tensor cores $\{C_i\}_{i \in [d]}$, such that it can be written in tensor network notation as

$$\phi = \quad \overset{C_1 \ \ C_2}{\underset{\textstyle |\quad |}{\bullet\!\!-\!\!\bullet}} \ \cdots \ \overset{C_d}{\underset{\textstyle |}{-\!\!\bullet}} \quad . \tag{77}$$

Without loss of generality, we can assume that this TT is in left-canonical form, so that the tensor cores satisfy equation (8), and that the ranks are minimal. If this is not the case, we can always find a gauge transformation that puts the TT into this form. We also assume that ϕ is an eigenvector of L, such that

$$L\phi = \lambda\phi. \tag{78}$$

To prove the theorem, we will inductively gauge transform each tensor core, starting at C_d and proceeding one-by-one towards C_1, such that after transforming C_ℓ, for all $\ell' \geqslant \ell - 1$ the right interface tensor satisfies

$$\phi^{>\ell'} L^{>\ell'} = \Lambda^{>\ell'} \phi^{>\ell'}. \tag{79}$$

Finally, we will show that equation (76) follows from equation (75).

   Base case. First, we will find an appropriate gauge transformation for C_d. We can write the eigenvalue equation in the form

$$L\phi = \lambda\phi = \tag{80}$$

where the double line combines multiple indices into a single edge and $\Omega = \mathrm{diag}(w(1), \ldots, w(d))$. We can now contract the first $d-1$ physical indices with $(\phi^{<d})^*$ and use the assumption that the TT is written in left-canonical form to obtain

$$\lambda \tag{81}$$

which, rearranging, we can write in matrix notation as

$$(\phi^{<d})^\dagger L^{<d} \phi^{<d} C_d = C_d (\lambda \mathrm{Id} - \Omega). \tag{82}$$

Since $(\phi^{<d})^\dagger L^{<d} \phi^{<d}$ is Hermitian, there exists a unitary U_d and a diagonal matrix $\Lambda^{<d}$ with non-decreasing diagonal entries, such that $(\phi^{<d})^\dagger L^{<d} \phi^{<d} = U_d^\dagger \Lambda^{<d} U_d$. Now we can write equation (82) as

$$\Lambda^{<d} \tilde{C}_d = \tilde{C}_d (\lambda \mathrm{Id} - \Omega), \tag{83}$$

where $\tilde{C}_d = U_d C_d$. Since U_d is a unitary, it defines a gauge transformation

$$C_{d-1} \mapsto \hat{C}_{d-1} = C_{d-1} U_{d-1}^\dagger, \quad C_d \mapsto \tilde{C}_d = U_{d-1} C_d, \tag{84}$$

which leaves ϕ invariant and preserves its left-canonical form. Rewriting equation (83) as $\tilde{C}_d \Omega = (\lambda \mathrm{Id} - \Lambda^{<d}) \tilde{C}_d$ and noticing that $\Omega = L^{>d-1}$ and $\tilde{C}_d = \phi^{>d-1}$, we get

$$\phi^{>d-1} L^{>d-1} = \Lambda^{>d-1} \phi^{>d-1}, \tag{85}$$

where we defined $\Lambda^{>d-1} = \lambda \mathrm{Id} - \Lambda^{<d}$, which has non-increasing entries. This is equation (75) for $\ell = d-1$.

   Induction step. Take $\ell \in \{2, \ldots, d-1\}$ and assume that ϕ is in left canonical form with

$$\phi^{>\ell'} L^{>\ell'} = \Lambda^{>\ell'} \phi^{>\ell'} \quad \forall \ell' \geqslant \ell, \tag{86}$$

where $\Lambda^{>\ell}$ is a diagonal matrix with non-increasing entries. We can decompose the LHS of equation (78) to get

$$\lambda\phi = L\phi = \tag{87}$$

Contracting the first $\ell - 1$ physical indices with $(\phi^{<\ell})^*$, using the left-canonical gauge condition and equation (86), when we fix the ℓth physical index to i, we obtain

$$\lambda \quad \begin{array}{c} C_\ell \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet \\ i \end{array} = \begin{array}{c} \phi^{<\ell} \quad C_\ell \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet\!\!-\!\!\bullet \\ L^{<\ell}\;\; i \\ \bullet \\ (\phi^{<\ell})^* \end{array} + \; w(i) \begin{array}{c} C_\ell \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet \\ i \end{array} + \begin{array}{c} \phi^{<\ell} \quad C_\ell \quad \Lambda^{>\ell} \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet \\ i \end{array} , \tag{88}$$

which we can rearrange and write in matrix notation as

$$\left(\phi^{<\ell}\right)^\dagger L^{<\ell}\phi^{<\ell}\left(C_\ell\right)_i = \left(C_\ell\right)_i\left[(\lambda - w(i))\,\mathrm{Id} - \Lambda^{>\ell}\right]. \tag{89}$$

Since $\left(\phi^{<\ell}\right)^\dagger L^{<\ell}\phi^{<\ell}$ is Hermitian, there exists a unitary U_ℓ and a diagonal matrix with non-decreasing diagonal entries $\Lambda^{<\ell}$, such that $\left(\phi^{<\ell}\right)^\dagger L^{<\ell}\phi^{<\ell} = U_\ell^\dagger \Lambda^{<\ell} U_\ell$. Hence, we can write equation (89) as

$$\Lambda^{<\ell}\left(\tilde{C}_\ell\right)_i = \left(\tilde{C}_\ell\right)_i\left[(\lambda - w(i))\,\mathrm{Id} - \Lambda^{>\ell}\right] \quad \forall i \in [p], \tag{90}$$

with $\tilde{C}_\ell = U_\ell C_\ell$. Furthermore, for all $i \in [p]$

$$\begin{array}{c} \tilde{C}_\ell \quad \phi^{>\ell-1} \\ \bullet\!\!-\!\!\bullet \\ L^{>\ell-1} \\ i \end{array} = \begin{array}{c} \tilde{C}_\ell \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet \\ i \quad L^{>\ell} \end{array} + \begin{array}{c} \tilde{C}_\ell \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet \\ \Omega \\ i \end{array}$$
$$= \begin{array}{c} \tilde{C}_\ell \quad \Lambda^{>\ell} \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet\!\!-\!\!\bullet \\ i \end{array} + \; w(i) \begin{array}{c} \tilde{C}_\ell \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet \\ i \end{array} \tag{91}$$
$$= \begin{array}{c} \tilde{C}_\ell \quad \Lambda^{>\ell} + w(i)\,\mathrm{Id} \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet\!\!-\!\!\bullet \\ i \end{array}$$
$$= \begin{array}{c} \lambda\,\mathrm{Id} - \Lambda^{<\ell} \quad \tilde{C}_\ell \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet\!\!-\!\!\bullet \\ i \end{array} ,$$

where in the last equality we have used equation (90). After the gauge transformation $C_\ell \mapsto \tilde{C}_\ell$, $C_{\ell-1} \mapsto C_{\ell-1}U_\ell^\dagger$, we can write equation (91) in matrix notation as

$$\phi^{>\ell-1}L^{>\ell-1} = \Lambda^{>\ell-1}\phi^{>\ell-1}, \tag{92}$$

where $\Lambda^{>\ell-1} = \lambda\,\mathrm{Id} - \Lambda^{<\ell}$ has non-increasing entries. Hence, we are left with a left canonical tensor train, such that the right interface vectors satisfy

$$\phi^{>\ell'}L^{>\ell'} = \Lambda^{>\ell'}\phi^{>\ell'} \quad \forall \ell' \geqslant \ell - 1. \tag{93}$$

This is the inductive hypothesis for $\ell - 1$.

Conclusion. We have found a gauge transformation that puts the tensor train into a form such that equation (75) is satisfied. We will now show that this in fact implies equation (76). For any $\ell \in [d-1]$ we can write the eigenvalue equation as

$$\lambda \begin{array}{c} \phi^{<\ell+1} \quad \phi^{>\ell} \\ \bullet\!\!=\!\!\bullet \end{array} = \begin{array}{c} \phi^{<\ell+1} \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet \\ L^{<\ell+1} \end{array} + \begin{array}{c} \phi^{<\ell+1} \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet \\ L^{>\ell} \end{array}$$
$$= \begin{array}{c} \phi^{<\ell+1} \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet \\ L^{<\ell+1} \end{array} + \begin{array}{c} \phi^{<\ell+1} \quad \Lambda^{>\ell} \quad \phi^{>\ell} \\ \bullet\!\!-\!\!\bullet\!\!-\!\!\bullet \end{array} , \tag{94}$$

which, rearranging, can be written in matrix notation as

$$L^{<\ell+1}\phi^{<\ell+1}\phi^{>\ell} = \phi^{<\ell+1}\left(\lambda\mathrm{Id} - \Lambda^{>\ell}\right)\phi^{>\ell}, \tag{95}$$

which, since $\phi^{>\ell}$ has full row rank by the assumption of minimal ranks, implies

$$L^{<\ell+1}\phi^{<\ell+1} = \phi^{<\ell+1}\left(\lambda\mathrm{Id} - \Lambda^{>\ell}\right), \tag{96}$$

which is equation (15).

$\square$

Appendix E. Conditions on low-rankness

We here connect the conditions on low rank TT approximate representations of multivariate functions [BSU16] with entanglement conditions on low rank TT approximations of quantum states [SWVC08, ECP10], known in this context as MPS.

Multivariate $L^2(\mathbb{R}^d)$ functions can be naturally associated with $\ell^2(\mathbb{N}^d)$ sequences, where the ℓ^2-norm is defined by $\|u\|_{\ell^2} := \sqrt{\sum_{\mathbf{i}\in\mathbb{N}^d} u_{\mathbf{i}}^2}$. Here, given a product basis $\Phi_{i_1,\dots,i_d}(x_1,\dots,x_d) = \Psi_{i_1}(x_1)\dots\Psi_{i_d}(x_d)$, where $\{\Psi_i\}_{i\in\mathbb{N}}$ is an orthonormal basis of $L^2(\mathbb{R})$, a function

$$f(x_1,\dots,x_d) = \sum_{i_1,\dots,i_d\in\mathbb{N}} u_{i_1,\dots,i_d}\Phi_{i_1,\dots,i_d}(x_1,\dots,x_d) \quad \in L^2\left(\mathbb{R}^d\right) \tag{97}$$

is associated with the sequence $(u_{\mathbf{i}})_{\mathbf{i}\in\mathbb{N}^d}$, in the sense that $\|f\|_{L^2} = \|u\|_{\ell^2}$.

Let us introduce weak-ℓ^p-norms of sequences, defined for $(a_n)_{n\in\mathbb{N}}$ by

$$|a|_{w\ell^p} := \sup_{n\in\mathbb{N}} n^{1/p}\tilde{a}_n, \tag{98}$$

where $(\tilde{a})_{n\in\mathbb{N}}$ is a non-increasing reordering of $(|a_n|)_{n\in\mathbb{N}}$. These norms satisfy

$$\|\cdot\|_{\ell^{p'}} \leqslant |\cdot|_{w\ell^p} \leqslant \|\cdot\|_{\ell^p} \tag{99}$$

for $p' < p$.

Suppose a sequence $(u_{\mathbf{i}})_{\mathbf{i}\in\mathbb{N}^d} \in \ell^2(\mathbb{N}^d)$. For each $\eta \in [d-1]$ we define a linear operator $\mathcal{M}_\eta^u : \ell^2(\mathbb{N}^{d-\eta}) \to \ell^2(\mathbb{N}^\eta)$, which acts on $(v_{\mathbf{i}})_{\mathbf{i}\in\mathbb{N}^{d-\eta}}$ as

$$\left(\mathcal{M}_\eta^u(v)\right)_{j_1,\dots,j_\eta} = \sum_{i_1,\dots,i_{d-\eta}\in\mathbb{N}} u_{j_1,\dots,j_\eta,i_1,\dots,i_{d-\eta}} v_{i_1,\dots,i_{d-\eta}}, \tag{100}$$

which can be thought of as a matrification at the interface labeled η of the tensor corresponding to u. The operator $\mathcal{M}_\eta^u$ has a finite Frobenius norm $\|\mathcal{M}_\eta^u\|_F = \|u\|_{\ell^2}$, so its singular values $\sigma_\eta(u) \in \ell^2(\mathbb{N})$. Let us now define the weak-ℓ_*^p-norm for multi-index sequences, which we define for $u \in \ell^p(\mathbb{N}^d)$ (with $d \geqslant 2$) by

$$\|u\|_{w\ell_*^p} := \max_{\eta\in[d-1]} |\sigma_\eta(u)|_{w\ell^p}. \tag{101}$$

By *proposition 5.1* in [BSU16], the weak-ℓ_*^p-norm determines the approximability of functions by low rank TTs. Suppose $f \in L^2(\mathbb{R}^d)$, which is associated, through a given dictionary, with a sequence $u \in \ell^2(\mathbb{N}^d)$. The Proposition states that, for $0 < p < 2$, there exists a sequence $\hat{u} \in \ell^2(\mathbb{N}^d)$, associated to a function $\hat{f}$, such that

$$\|f - \hat{f}\|_2 = \|u - \hat{u}\|_{\ell^2} \leqslant C\sqrt{d}\|u\|_{w\ell_*^p}\left(\max_{\eta\in[d-1]} \mathrm{rank}_\eta(\hat{u})\right)^{-s}, \tag{102}$$

where $C > 0$ is a constant, $s := \frac{1}{p} - \frac{1}{2}$ and $\mathrm{rank}_\eta(\hat{u})$ counts the number of non-zero elements of $\sigma_\eta(\hat{u})$. To see the significance of this result, note that $\max_{\eta\in[d-1]} \mathrm{rank}_\eta(\hat{u})$ is equal to the TT rank of $\hat{u}$, so through equations (102) the weak-ℓ_*^p-norm puts limits on the existence of low rank TT approximations of f. We will now show how this result relates to *entanglement entropy scaling* used in quantum physics as a condition on low-rank MPS approximations of quantum states [SWVC08, ECP10].

Quantum states are normalized elements of a complex Hilbert space $\mathcal{H}$ (or more accurately rays in $\mathcal{H}$) associated to a given system. Many-body systems have a tensor product structure, where $\mathcal{H} = \bigotimes_{k\in[d]} \mathcal{L}_k$, where $\mathcal{L}_k$ is the local Hilbert space of a single constituent sub-system and d is the number of sub-systems.

Suppose $\mathcal{L}_k = \mathcal{L}$ for all $k \in [d]$ and choose an orthonormal basis $(\phi_i)_i$ for $\mathcal{L}$. This allows us to represent many-body quantum states by tensor networks, in an analogous way to how we use them to represent multivariate functions, although for quantum states we need to allow the tensor networks to be complex.

Given a quantum state ψ, the corresponding density matrix ρ is defined as the rank-one projector onto ψ. For each $\eta \in [d-1]$, we define the reduced density matrix $\rho_\eta = \mathrm{tr}_\eta \rho$, where tr_η denotes the partial trace over all $\mathcal{L}_\ell$ with $\ell > \eta$. The Rényi entropy for $\alpha \in [0,\infty]$ is defined by the analytic continuation of

$$S_\alpha(\rho_\eta) := \frac{\log \mathrm{tr}\left(\rho_\eta^\alpha\right)}{1-\alpha}. \tag{103}$$

Suppose the singular value decomposition (SVD) $\mathcal{M}_\eta^\psi = U\Sigma V^\dagger$, where U, V are isometries and $\Sigma = \mathrm{diag}(\sigma_\eta(\psi))$. Hence, we can write $\rho_\eta = U\Sigma^2 U^\dagger$, which shows that the eigenvalues of ρ_η are squares of the singular values $\sigma_\eta(\psi)$. Therefore, we have

$$\mathrm{tr}\left(\rho_\eta^\alpha\right) = \|\sigma_\eta(\psi)\|_{\ell^{2\alpha}}^{2\alpha}. \tag{104}$$

In quantum many-body physics, we are interested in approximations $(\hat{\psi}_d)_d$ of families of states $(\psi_d)_d$ on growing number d of sub-systems, that satisfy $\|\psi_d - \hat{\psi}_d\|_2 \leqslant \delta$ for all d. We call such approximations efficient it the TT rank of $\hat{\psi}_d$ scales at most polynomially with d. In [SWVC08] and [VC06] it is shown that efficient approximations exist if, for some $0 < \alpha < 1$, there exist $c, c' > 0$ such that for all $\eta \in [d-1]$ we have that $S_\alpha(\rho_{d,\eta}) = c \log d + c'$, where $\rho_{d,\eta} = \mathrm{tr}_\eta \rho_d$ and ρ_d is the density matrix corresponding to ψ_d. We will now show that this follows from equation (102).

Equation (102) implies that polynomial bond dimension approximations are guaranteed to exist if

$$\|\psi_d\|_{w\ell_*^p} \leqslant \frac{\delta}{\sqrt{d}} \left(\mathrm{poly}(d)\right)^s \tag{105}$$

for some $0 < p < 2$. Using equation (99) and with $\alpha := p/2$ we find a simpler sufficient condition

$$\max_{\eta \in [d-1]} \|\sigma_\eta(\psi_d)\|_{\ell^{2\alpha}} \leqslant \frac{\delta}{\sqrt{d}} \left(\mathrm{poly}(d)\right)^s. \tag{106}$$

Through equations (103) and (104), we can rewrite this condition, for $0 < \alpha < 1$, as

$$\max_{\eta \in [d-1]} S_\alpha(\rho_{\eta,d}) \leqslant \frac{2\alpha}{1-\alpha} \log \frac{\delta}{\sqrt{d}} \left(\mathrm{poly}(d)\right)^s \tag{107}$$

and hence there exist $c, c' > 0$, such that $S_\alpha(\rho_{\eta,d}) \leqslant c \log d + c'$ for all $\eta \in [d-1]$, which is the result of [SWVC08] and [VC06].

ORCID iDs

J Fuksa ● https://orcid.org/0000-0003-4606-2584
I Roth ● https://orcid.org/0000-0002-1191-7442
J Eisert ● https://orcid.org/0000-0003-3033-1292

References

[BC17] Bridgeman J C and Chubb C T 2017 Hand-waving and interpretive dance: an introductory course on tensor networks *J. Phys. A: Math. Theor.* A **50** 223001

[BGP21] Bachmayr M, Götte M and Pfeffer M 2022 Particle number conservation and block structures in matrix product states *Calcolo* **59** 24

[BPK16] Brunton S L, Proctor J L and Kutz J N 2016 Discovering governing equations from data by sparse identification of nonlinear dynamical systems *Proc. Natl Acad. Sci.* **113** 3932–7

[BSU16] Bachmayr M, Schneider R and Uschmajew A 2016 Tensor networks and hierarchical tensors for the solution of high-dimensional partial differential equations *Found. Comput. Math.* **16** 1423–72

[CDA+21] Cornelio C, Dash S, Austel V, Josephson T R, Goncalves J, Clarkson K L, Megiddo N, El Khadir B and Horesh L 2021 Combining data and theory for derivable scientific discovery with AI-Descartes *Nat. Commun.* **14** 1777

[CLKB19] Champion K, Lusch B, Kutz J N and Brunton S L 2019 Data-driven discovery of coordinates and governing equations *proc. Natl Acad. Sci.* **116** 22445–51

[CPGSV21] Cirac J I, Pérez-García D, Schuch N and Verstraete F 2021 Matrix product states and projected entangled pair states: concepts, symmetries, theorems *Rev. Mod. Phys.* **93** 045003

[CPSW21] Carderera A, Pokutta S, Schütte C and Weiser M 2021 CINDy: conditional gradient-based identification of non-linear dynamics – noise-robust recovery (arXiv:2101.02630)

[CSS15] Cohen N, Sharir O and Shashua A 2015 On the expressive power of deep learning: a tensor analysis *29th Annual Conf. on Learning Theory* vol 49 pp 689–728 (arXiv:1509.05009)

[DHZ+21] Ding M, Huang T-Z, Zhao X-L, Ng M K and Ma T-H 2021 Tensor train rank minimization with nonlocal self-similarity for tensor completion *Inverse Probl. Imaging* **15** 475–98

[dSCQ+20] de Silva B M, Champion K, Quade M, Loiseau J-C, Kutz J N and Brunton S L 2020 PySINDy: a python package for the sparse identification of nonlinear dynamics from data *J. Open Source Softw.* **5** 2104–7

[ECP10] Eisert J, Cramer M and Plenio M B 2010 Colloquium: area laws for the entanglement entropy *Rev. Mod. Phys.* **82** 277–306

[GGR+20] Goeßmann A, Götte M, Roth I, Sweke R, Kutyniok G and Eisert J 2020 Tensor network approaches for learning non-linear dynamical laws (arXiv:2002.12388)

[GK19] Grasedyck L and Krämer S 2019 Stable ALS approximation in the TT-format for rank-adaptive tensor completion *Numer. Math.* **143** 855–904

[GKES19] Gelß P, Klus S, Eisert J and Schütte C 2019 Multidimensional approximation of nonlinear dynamical systems *J. Comput. Nonlinear Dyn.* **14** 061006

[GPA+18] Glasser I, Pancotti N, August M, Rodriguez I D and Cirac J I 2018 Neural-network quantum states, string-bond states and chiral topological states *Phys. Rev. X* **8** 011006

[GSP+19] Glasser I, Sweke R, Pancotti N, Eisert J and Cirac J I Expressive power of tensor-network factorizations for probabilistic modeling, with applications from hidden markov models to quantum machine learning *Advances in Neural Information Processing Systems* **32**

[GST21] Götte M, Schneider R and Trunschke P 2021 A block-sparse tensor train format for sample-efficient high-dimensional polynomial regression *Front. Appl. Math. Stat.* 7

[HLO+16] Haegeman J, Lubich C, Oseledets I V, Vandereycken B and Verstraete F 2016 Unifying time evolution and optimization with matrix product states *Phys. Rev. B* **94** 165116

[HMOV14] Haegeman J, Marien M, Osborne T J and Verstraete F 2014 Geometry of matrix product states: Metric, parallel transport and curvature *J. Math. Phys.* **55** 021902

[HRS12] Holtz S, Rohwedder T and Schneider R 2012b On manifolds of tensors of fixed TT rank *Numer. Math.* **120** 701–31

[IMW+20] Iten R, Metger T, Wilming H, del Rio L and Renner R 2020 Discovering physical concepts with neural networks *Phys. Rev. Lett.* **124** 010508

[KBK22] Kaheman K, Brunton S L and Kutz J N 2022 Automatic differentiation to simultaneously identify nonlinear dynamics and extract noise probability distributions from data *Mach. Learn.: Sci. Technol.* **3** 015031

[KG19] Klus S and Gelß P 2019 Tensor-based algorithms for image classification *Algorithms* **12** 240

[LYCS17] Levine Y, Yakira D, Cohen N, and Shashua A 2017 Deep learning and quantum entanglement: fundamental connections with implications to network design (arXiv:1704.01552 [quant-ph])

[MBPK16] Mangan N M, Brunton S L, Proctor J L and Kutz J N 2016 Inferring biological networks by sparse identification of nonlinear dynamics

[MP22] Michael G and Philipp T 2022 (available at: https://github.com/JonasFuksa/bstt.git)

[Oru14] Orus R 2014 A practical introduction to tensor networks: matrix product states and projected entangled pair states *Ann. Phys.* **349** 117–58

[Ose11] Oseledets I V 2011 Oseledets Tensor-train decomposition *SIAM J. Sci. Comput.* **33** 2295–317

[RCR+20] Riemer M, Cases I, Rosenbaum C, Liu M and Tesauro G 2020 On the role of weight sharing during deep option learning (arXiv:1912.13408 [cs, stat])

[SBK21] Shea D E, Brunton S L and Kutz J N 2021 Sindy-bvp: sparse identification of nonlinear dynamics for boundary value problems *Phys. Rev. Res.* **3** 023255

[SL09] Schmidt M and Lipson H 2009 Distilling free-form natural laws from experimental data *Science* **324** 81–85

[SPV10] Singh S, Pfeifer R N C and Vidal G 2010 Tensor network decompositions in the presence of a global symmetry *Phys. Rev. A* **82** 050301

[SS16] Stoudenmire E and Schwab D J 2016 Supervised learning with tensor networks *Neural Information Processing Systems*

[SWVC08] Schuch N, Wolf M M, Verstraete F and Cirac J I 2008 Entropy scaling and simulability by matrix product states *Phys. Rev. Lett.* **100** 030504

[VC04] Verstraete F and Cirac J I 2004 Renormalization algorithms for quantum-many body systems in two and higher dimensions (arXiv:cond-mat/0407066)

[VC06] Verstraete F and Cirac J I 2006 Matrix product states represent ground states faithfully *Phys. Rev. B* **73** 094423

[VMC08] Verstraete F, Murg V and Cirac J I 2008 Matrix product states, projected entangled pair states and variational renormalization group methods for quantum spin systems *Adv. Phys.* **57** 143–224