arXiv:2312.17019v3 [quant-ph] 11 Jan 2025

Efficient Learning of Long-Range and Equivariant Quantum Systems

Štěpán Šmíd and Roberto Bondesan

Department of Computing, Imperial College London, London SW7 2AZ, United Kingdom

In this work, we consider a fundamental task in quantum many-body physics – finding and learning ground states of quantum Hamiltonians and their properties. Recent works have studied the task of predicting the ground state expectation value of sums of geometrically local observables by learning from data. For short-range gapped Hamiltonians, a sample complexity that is logarithmic in the number of qubits and quasipolynomial in the error was obtained. Here we extend these results beyond the local requirements on both Hamiltonians and observables, motivated by the relevance of long-range interactions in molecular and atomic systems. For interactions decaying as a power law with exponent greater than twice the dimension of the system, we recover the same efficient logarithmic scaling with respect to the number of qubits, but the dependence on the error worsens to exponential. Further, we show that learning algorithms equivariant under the automorphism group of the interaction hypergraph achieve a sample complexity reduction, leading in particular to a constant number of samples for learning sums of local observables in systems with periodic boundary conditions. We demonstrate the efficient scaling in practice by learning from DMRG simulations of 1D long-range and disordered systems with up to 128 qubits. Finally, we provide an analysis of the concentration of expectation values of global observables stemming from the central limit theorem, resulting in increased prediction accuracy.

1 Introduction

The simulation of quantum systems underpins our understanding of nature as well as the discovery of novel critical technologies. However, this task is notoriously hard. While a plethora of clever classical algorithms has been developed, such as density functional theory [1], quantum Monte Carlo [2], and tensor networks [3], these methods are still too slow for many practical tasks, such as the discovery of novel catalyst for renewable energy storage [4]. The simulation of quantum systems is also one of the main use cases for quantum computers, and several quantum algorithms with provable advantages have been developed [5]. However, despite tremendous technological advances in recent years, the current quantum computers are too small and noisy to provide a clear advantage [6].

Machine learning (ML) algorithms, such as large language models like Chat-GPT, have become ubiquitous in everyday life, and recent years have also witnessed a wide adoption of ML in science, with successes such as AlphaFold [7]. Many researchers have also applied ML to accelerate the simulation of quantum systems, e.g. [8, 9, 10, 11, 12]. One important difference between applying ML to image or text data and to quantum data is the scale at which data is available. In fact, in contrast to the abundance of images and text on the internet that has fuelled the wide scale application of deep learning to solve tasks in image and natural language processing, data from quantum experiments or simulation is hard to obtain. Experiments can be very expensive and difficult to realise, and the challenges of simulation is precisely the reason to turn to ML in a first place. Thus, when we try to apply ML to predict properties of quantum systems we cannot simply rely on the abundance of data, and we instead need to focus on a theoretical understanding of what tasks can be learned efficiently and what models we should use.

One recent strand of research at the intersection of quantum information and statistical learning theory aims at studying the learning complexity of quantum tasks as a new fundamental concept in quantum information [13]. A particularly important task is learning to predict the expectation value of observables at a new point in parameters space given a dataset of measurements of observables at other values of the parameters. This is in fact the setting of predicting the force field at a new position of the nuclei given the energy at a set of previous coordinates. It is also a relevant setting in the context of variational quantum algorithms, where the learning problem is to predict the expectation value of an observable at new value of the variational parameters, which can save potentially costly experiments, as for example is used in Bayesian optimisation approaches [14].

The recent works [15, 16, 17] have studied the problem of predicting the expectation value of observables within a phase of gapped short range Hamiltonians. They derived efficient ML algorithms that provably achieve a sample complexity logarithmic in the number of qubits for observables that are given by the sum of geometrically local terms. This is remarkable since computing properties of gapped phases include problems that are NP-hard, proving an advantage of algorithms that learn from data compared to algorithms that do not learn from data [18]. Further, in [19] they have considered a related task of predicting arbitrary quantum states smoothly parameterised by a small constant number of parameters. Here they focused primarily on the scaling with respect to the prediction error, and achieved a polynomial sample complexity in the inverse error and the system size. Although the dependence on the prediction error is an important aspect of these learning algorithms, their methods are not applicable to systems with an extensive number of parameters like those ones considered in quantum chemistry.

However, short range interactions are often an approximation to the long range interactions that most systems in nature experience [20]. Important examples of long range interacting systems are: electronic systems with Coulomb forces of relevance to quantum chemistry, dipolar molecules [21], Rydberg atom arrays used to implement quantum gates [22], quantum simulators based on trapped ion crystals [23], and spin glasses [24].

Motivated by the recent advances in understanding ML methods for short range interactions and by the importance of long range interactions in applications, in this work we study provably efficient ML algorithms for predicting properties of long range quantum Hamiltonians. More specifically, the relevance of this questions for the broad quantum computing field stems from the fact that quantum chemistry, which is described by long-range interacting quantum systems, is likely to be one of the earliest and most impactful applications of quantum computers. Downstream tasks in drug design however require performing an error-corrected quantum computation a billion times [25]. ML models that predict properties of long-range interacting systems are thus critical to enable impactful applications of quantum computers to drug design. In this work, we make progress along this line by studying extensions of the generalisation bounds previously derived in [15, 16] to long-range interacting systems. We note that [18] already sketched the extension of their results to long range interacting systems with linear light cone – a technical condition which will be discussed at length below – suggesting ML algorithms that use $N = m^{\mathcal{O}(1/\epsilon)}$ data samples, with m being the number of parameters of the Hamiltonian and ϵ the accuracy. In this work, we improve on this scaling and also introduce novel rigorous generalisation bounds for equivariant quantum systems. Our definition of equivariance encompasses and extends the notion of invariance, and is particularly powerful for disordered lattice systems with periodic boundary conditions. This setup is ubiquitous in quantum and classical condensed matter physics, where disorder models impurities, defects or complex environment interactions [26]. In the companion paper [27] we build on the results for equivariant models we derive here to predict observables within a whole topological phase from data for a single value of the parameters.

We note that the concept of equivariance has been developed extensively in the machine learning literature to deal with symmetries in the data. However, all the works that apply equivariant ML models to quantum physics problems are heuristic in nature. A prominent example of those is the use of equivariant graph neural networks to predict force fields for molecular dynamics, where the symmetry is the invariance of the electronic energy under roto-translations of the nuclei [28, 29]. The heuristic nature of these methods means that in practice they can suffer from uncontrollable failure modes, such as failure to generalise across different regions of the conformational space [30].

In summary, we make the following novel contributions:

- We prove that a *number of samples logarithmic in the system size* suffices to predict properties of gapped quantum systems with interactions that decay exponentially or as a power law with exponent α greater than twice the system's dimension. Our analysis also allows for predicting p-local observables which are not necessarily geometrically local and can depend on the system's parameters – for example the ground state energy.
- We show that ML models that are equivariant under the automorphism group of the interaction hypergraph (i.e. the hypergraph with vertices the qubits and hyperedges connecting qubits in the support of Hamiltonian terms) achieve a *sample complexity reduction*. This implies a *constant sample complexity* w.r.t. the system size for learning expectations of sums of local observables in systems with periodic boundaries. This was the first such known learning algorithm with constant sample complexity, though later [31] has also achieved this for short-range Hamiltonians even for open boundary conditions, but they suggest that their algorithm is not extendable to the long-range case.
- We lift the requirements on the observable having a norm bounded by a constant using a simple modification of the ML algorithm, which allows predictions up until the standard deviation of the expectation values is bounded by a constant. This is substantiated by an analysis of the expectation values of global observables using the central limit theorem.
- We provide extensive numerical simulations of one-dimensional systems with up to 128 qubits, demonstrating the efficient scaling in practice. The corresponding code for these simulations is available at [32].

In the next section we give a technical summary of our main results.

1.1 Summary of main results

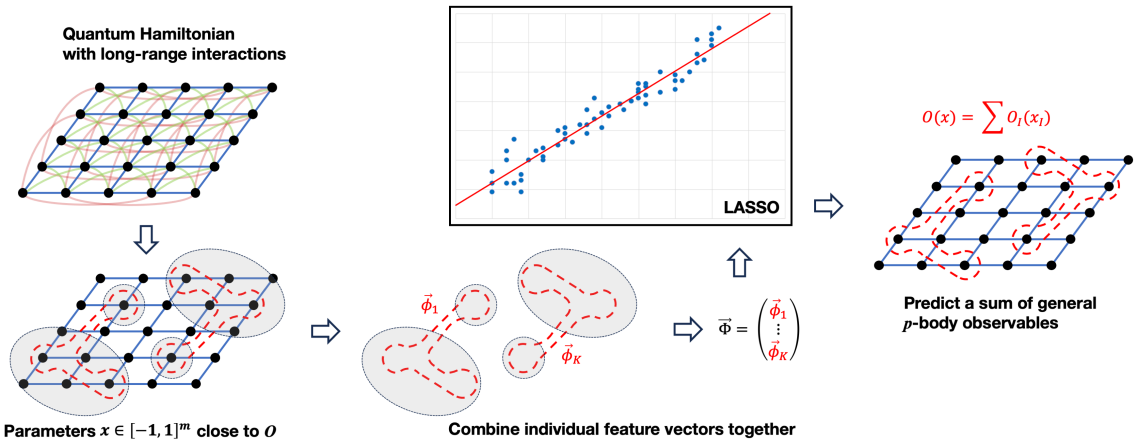


Figure 1: **Overview of the efficient machine learning algorithm.** Given a vector $x \in [-1,1]^m$ parameterising a quantum Hamiltonian with long-range interactions, it gets separated into neighbourhoods of the summands in the observable O, which are then mapped to features $\vec{\phi}_i$, and concatenated into one full feature vector. This vector is then used as an input of the LASSO model, which is trained to predict O for a new value of x.

1.1.1 Machine learning algorithm

The ML algorithm we study is shown in Figure 1. It is an extension of the algorithm introduced in [15] to the setting of long range interactions.

The input to the ML algorithm is a set of data pairs $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$ where $x \in [-1,1]^m$ is the set of parameters of the Hamiltonian and y is the expectation of an observable O in the ground state of the system with Hamiltonian described by x. We assume that the Hamiltonian has interactions $h_I(x_I)$ which are supported on a set of qubits $I = (i_1, \ldots, i_\ell)$, $\ell \leq k$, with $k = \mathcal{O}(1)$,

and that interactions decay exponentially with the distance between two qubits or as a power law with exponent $\alpha > 2D$. We also consider observables $O = \sum_{I \in \mathcal{S}} O_I$ with $\mathcal{S}$ containing sets with at most $p = \mathcal{O}(1)$ qubits, not necessarily geometrically local. We shall normalise O such that the ℓ_1 norm of the coefficients of the expansion in the Pauli basis is $\mathcal{O}(1)$, which implies that $\|O\| = \mathcal{O}(1)$. The data $y^{(i)}$ can be either obtained by direct measurement of O in a quantum experiment or a classical simulation, or by the classical shadow formalism [33] which relies on randomised measurements of N ground states and postprocessing.

The ML model is built out of features defined for each operator O_I as follows. We select the components x_J of x such that J is within distance δ to I, and the diameter of J is also not larger than δ. We call $Z_I(x) \in [-1,1]^{m_{I,\delta}}$ this subset of x. Next we construct a feature map $\vec{\phi}_I(x)$ out of $Z_I(x)$. As in [15], the theoretical analysis uses a feature map based on a discretisation, while the experiments use random Fourier features with input $Z_I(x)$. The feature map based on discretisation is a one-hot vector of dimension equal to the number of points in a grid of mesh size δ_2 in $[-1,1]^{m_{I,\delta}}$. We set $\phi_{I,x'}(x) = 1$ for the point x' in this grid that is closest to $Z_I(x)$.

These features are then concatenated over I to form the feature vector $\vec{\Phi}$ (bottom center of Figure 1). A linear model with those features and a ℓ_1 norm penalisation (LASSO) is learned from the data, and can be used to predict the ground state expectation value of this observable at new values of x within the same gapped phase of the training data.

1.1.2 Theoretical guarantees

We prove the following main theorems about the ML model. The setting and notation is as in the previous paragraph. Specifically, we shall consider k-local Hamiltonians with interaction terms h_I supported on sets of qubits I decaying either exponentially or as a power law with exponent α with respect to the diameters of I. We then want to predict a p-local observable O constituting of $|\mathcal{S}|$ few-body terms.

Theorem 1.1 (Theorem 3.1 informal)**.** *Choose*

$$\delta \geq \begin{cases} \mathcal{O}(\log^2(1/\epsilon)) & \text{exp decay,} \\ \mathcal{O}(\epsilon^{-1/(\nu-D)}) & \text{power law,} \end{cases} \qquad \delta_2 = \mathcal{O}(\epsilon/\sqrt{m_{I,\delta}})\,,$$

with ν the function of α and D given in Proposition A.6, where the exponent $1/(\nu - D)$ diverges as $\alpha \to 2D$ from above. Then the ML model with truncation parameter δ and a mesh size δ_2 achieves generalisation error at most ϵ with probability at least $1-\gamma$ if we choose the number of samples as

$$N = \mathcal{O}(\log(|\mathcal{S}|/\gamma)) \times \begin{cases} 2^{\mathcal{O}(\text{polylog}(1/\epsilon))} & \text{exp decay,} \\ 2^{\mathcal{O}(\epsilon^{-\omega}\log(1/\epsilon))} & \text{power law,} \end{cases} \qquad \omega = kD/(\nu - D)\,.$$

We remark that if $|\mathcal{S}| = \mathcal{O}(1)$, the number of samples does not depend on n, while if $|\mathcal{S}| = n^p$, with $p = \mathcal{O}(1)$, the number of samples grows only logarithmically with n. The dependence on the error ϵ is quasipolynomial in the case of exponentially decaying interactions, which as expected coincides with the behavior for short range interactions [15, 16]. This dependence becomes instead exponential for power law interactions with $\alpha > 2D$. The exponent $1/(\nu - D) \to \infty$ as $\alpha \to 2D$ and our results break down for $\alpha \leq 2D$.

Proof strategy. There are two main parts of proving this theorem.

The first part requires to show that ground state expectation values of few-body non-geometrically-local observables can be accurately approximated only with the parameters in the neighbourhood of their support, where the size of this neighbourhood does not depend on the system size. As we will show, this property will hold true for Hamiltonians with interactions decaying at least as a power law with exponent greater than twice the system's dimension. This property will further hold true even for observables that depend themselves on the parameters of the system. To prove this, one needs to consider the quasi-adiabatic continuation operator [34], which quantifies the change of the expectation values with respect to the parameters within a topological phase. The action of this operator can then be bounded using Lieb-Robinson bounds [35]. Hence one can bound the difference between the expectation values of the full set of parameters and the truncated set of

parameters on the neighbourhood to be arbitrarily small for appropriately large neighbourhood, whose size does not depend on the number of qubits.

The second part then consists of showing that an expectation value of a p-local observable constituting of few-body terms obeying the previously proven property can be accurately expressed across the parameter space as a simple linear function, where the unknown coefficients are to be learned, while the features are based on a discretisation grid of the parameter space. The in-built local geometric bias in these features combined with the previously proven property is what allows this model to be accurate. Finally, we would use well-understood sample complexity theory for linear models [36] to determine how many training samples we would need to learn the unknown coefficients and obtain a constant generalisation error. □

In Theorem 3.7 we also show that we can learn all observables that are sums of at most p-body terms and have bounded ℓ_1 norm in the Pauli basis with a sample complexity as in Theorem 1.1, where $|\mathcal{S}|$ is replaced by the cardinality of the set of Pauli operators with weight at most p. In this case we need to prepare N ground states, one per parameter $x^{(i)}$, and then perform $\mathcal{O}(\log(n/\delta)/\epsilon^2)$ randomised measurements to compute the classical shadow of the density matrix [33], which allows one to produce with accuracy ϵ a dataset $\{x^{(i)}, y^{(i),P}\}_{i=1}^N$ for each Pauli P. This data is then used to learn a ML model for each observable using Theorem 1.1.

Next we present our results on equivariance. We define the interaction hypergraph as the hypergraph with vertex set the vertices of the lattice and one hyperedge per interaction. Its automorphism group G is a subgroup of the permutation group of the qubits that fixes the hyperedge set. It relates the predictions for the observable O_I to that of O_{gI}, with $g \in G$, and allows one to reduce the sample complexity by using an equivariant ML model. The precise definition of equivariance and equivariant weights are in Propositions 3.5 and 3.6.

Corollary 1.1.1 (Corollary 3.5.1)**.** *A G-equivariant ML model achieves a sample complexity reduction that amounts to replace $|\mathcal{S}|$ with $|\mathcal{S}/G|$ in Theorem 1.1.*

This result is particularly powerful if $|\mathcal{S}| \approx |G|$, in which case the complexity becomes independent of n. This is the case for example for the frequently encountered case of models defined on a cubic lattice with periodic boundary conditions and observables with $|\mathcal{S}| = \mathcal{O}(n)$. Note that equivariance reduction applies independently of the details of the Hamiltonian and in particular does not require the Hamiltonian to be translation invariant in the case of periodic boundary conditions. A similar sample complexity reduction from equivariance is obtained in Corollary B.3.1 for predicting expectation values of all observables that are sums of p-body terms.

Physically, we thus expect our results on equivariant models to be most relevant for quantum disordered systems on a lattice, such as the paradigmatic disordered Heisenberg and Ising models [37], spin glasses [38], and atomic chains in the presence of impurities, such as Rydberg atoms [39]; some of which are discussed further in the section on experiments.

While we assume here exact equivariance of the Hamiltonian, we expect our results to be relevant in the case of approximate equivariance as well. That is, if the Hamiltonian under consideration is close to that of an equivariant system, in the limit of large number of qubits the equivariant ML models should provide a good description of the observables. This applies for example to the case where the lattice has open boundary conditions, as we expect that in the thermodynamic limit the system has the same bulk observables as the same system with periodic boundaries.

1.1.3 Experiments

Another contribution of this paper is to verify numerically the analytical results. For the first time, we provide numerical evidence for the logarithmic scaling of the sample complexity with the number of qubits. The data is produced by DMRG simulations of disordered short range Heisenberg chains, disordered Ising chains with spin-spin interactions decaying as a power law with exponent α, and Rydberg atom chains with position displacements with interactions decaying as the 6^{th} power. We predict the expectation value of the Hamiltonian in the ground state across a range of couplings. We point out that due to the short range correlations in the systems, the central limit theorem applies. A naïve normalisation of the observable has an expectation value that concentrates around

the average over disorder in the large n limit, leading to an x-independent target for the machine learning model and a trivial sample complexity – a constant model with a single parameter can achieve zero test error in the $n \to \infty$ limit. We show that scaling and zero-centering the observable so that the Gaussian fluctuations in the large n limit are not suppressed, leads to a non-trivial ML problem also as $n \to \infty$, and this allows us to verify the non-trivial scaling of the sample complexity, see Figures 5a, 6a, and 8. We also verify that using an equivariant ML model allows us to reduce the sample complexity from $\mathcal{O}(\log(n))$ to $\mathcal{O}(1)$ in Figure 5b for the disordered Heisenberg model on a closed chain. Finally, in Figure 6b we show that we can apply successfully the ML derived for $\alpha > 2D$ to a case with $\alpha \leq 2D$. While the numerical data suggest worse, seemingly linear, scaling, we can not definitely conclude the sample complexity to not be logarithmic in the number of qubits. We leave it as an important open problem to derive theoretical guarantees for $\alpha \leq 2D$.

1.2 Outline of the paper

The paper is structured as follows: Section 2 presents the results and comments on the proofs that the expectation values of observables O_I depend only on Hamiltonian parameters that are geometrically close to I; while Appendix A presents the detailed proof. This is shown for gapped Hamiltonians with exponentially decaying interactions and power law decaying interactions with exponent $\alpha > 2D$. These sections constitute the bulk of the novel technical contributions of this paper. Section 3 then builds on these results and the literature on generalisation bounds to derive rigorous guarantees for the sample complexity of the ML models; together with details in Appendix B about predicting many observables via classical shadows. In Section 3 we also introduce the notion of equivariance and show how it allows one to reduce the sample complexity. Finally, Section 4 shows numerical experiments that validate the theoretical findings. We discuss the normalisation of observables and the details of the experimental setup. We also apply the ML model to Hamiltonians with $\alpha \leq 2D$. Appendix C contains combinatorial identities and bounds used to derive results in Section 2, Appendix D discusses Lieb-Robinson bounds for power law interactions, and Appendix E the DMRG implementations of the considered models.

2 Quantum information bounds

2.1 Setup and notation

We consider a D-dimensional lattice with vertex set Λ of cardinality n. We denote by $\mathcal{P}(\Lambda)$ the set of subsets of Λ, and by $\mathcal{P}_k(\Lambda)$ those $I \in \mathcal{P}(\Lambda)$ with cardinality $|I| \leq k$. We denote by $d(i,j)$ the distance between $i, j \in \Lambda$ and for $I, J \in \mathcal{P}(\Lambda)$ we define their distance and diameter as

$$d(I,J) = \min_{i\in I, j\in J} d(i,j)\,, \quad \operatorname{diam}(I) = 1 + \max_{i,j\in I} d(i,j)\,.$$

Now we associate a system of qubits to the vertices in Λ and a Hamiltonian with k-body long range interactions which depend on a set of parameters x as follows:

$$H(x) = \sum_{I\in\mathcal{P}_k(\Lambda)} h_I(x_I)\,,$$

where $h_I(x_I)$ is supported only on the qubits in I and $x_I \in [-1,1]^{q_I}$. We denote the dimensionality of x by

$$m = \sum_{I\in\mathcal{P}_k(\Lambda)} q_I \leq q_* |\mathcal{P}_k(\Lambda)| = q_* \sum_{\ell=1}^{k} \binom{n}{\ell}\,, \quad q_* = \max_{I\in\mathcal{P}_k(\Lambda)} q_I\,.$$

For example, when $q_I = 1$ and $k = 2$, we have the family of Hamiltonians

$$H(x) = \sum_{ij} h_{ij}(x_{ij}) + \sum_i h_i(x_i)\,.$$

The case of interactions with short range r can be seen as a special case where $h_I = 0$ if $\mathrm{diam}(I) > r$.

We are going to study the following machine learning problem. Let us denote the ground state of $H(x)$ by $\rho(x)$, which is defined as

$$\rho(x) = \lim_{\beta\to\infty} \frac{\mathrm{e}^{-\beta H(x)}}{\mathrm{Tr}\left(\mathrm{e}^{-\beta H(x)}\right)}\,.$$

We are given a dataset

$$S = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}\,, \quad y^{(i)} = \mathrm{Tr}\Big(\rho(x^{(i)})O\Big) + \epsilon\,,$$

where $\epsilon \geq 0$ accounts for errors in measuring the observable. We then aim at constructing a predictor for $\mathrm{Tr}(\rho(x_*)O)$ at a test point x_*, and we ask how many samples do we need to solve this problem accurately.

Unless explicitly stated, we will always consider the operator norm $\|\cdot\|$ over Hermitian operators, defined as the absolute value of the largest eigenvalue. If $A = (A_1, \ldots, A_p)$ is a vector of operators, then we denote $\|A\| = \sqrt{\sum_{i=1}^{p} \|A_i\|^2}$.

We are going to solve this ML problem assuming that:

- k, q_I do not depend on n.
- There is a gap γ above the ground state that is uniform over all $x \in [-1,1]^m$.
- $\|h_I(x_I)\|$ and $\|\partial_{x_i} h_I(x_I)\|$ decay exponentially or as a power law with $\mathrm{diam}(I)$ for all x_I. More precise bounds will be discussed below.
- The observable O is such that, with $p = \mathcal{O}(1)$:

$$O = \sum_{I\in\mathcal{P}_p(\Lambda)} O_I\,.$$

This is similar to the setup of [15, 16, 17]; however, differently from those works, we do not assume that neither h_I nor O_I are geometrically-local. In the following we will denote by $\mathbf{1}(\cdot)$ the indicator function and by $\delta(\cdot,\cdot)$ the Kronecker delta.

2.2 Dependency of observables on Hamiltonian parameters

The goal of this section is to prove that under the assumptions of Section 2.1 the function

$$f(O_I, x) = \mathrm{Tr}(O_I \rho(x))\,, \quad I \in \mathcal{P}_p(\Lambda)\,,$$

depends only on the x_J's in a neighborhood of I. This neighborhood is defined as the x_J's that are within a distance $\delta > 0$ from I and are such that $\mathrm{diam}(J) \leq \delta$:

$$S_{I,\delta} = \{J \in \mathcal{P}_k(\Lambda) \,|\, d(I,J) \leq \delta \wedge \mathrm{diam}(J) \leq \delta\}\,. \tag{1}$$

Figure 2 visualizes $S_{I,\delta}$ for a case in $D = 2$. Note that the complementary set is

$$(S_{I,\delta})^c = \{J \in \mathcal{P}_k(\Lambda) \,|\, (d(I,J) \leq \delta \wedge \mathrm{diam}(J) > \delta) \vee d(I,J) > \delta\}\,.$$

We now define x' such that

$$x'_J = \begin{cases} x_J & \text{if } J \in S_{I,\delta}\,, \\ 0 & \text{if } J \notin S_{I,\delta}\,. \end{cases}$$

There is nothing special about setting x' to 0 outside $S_{I,\delta}$, and any other fixed value could have been chosen for the results below to follow. We shall also denote by $\chi_S(x)$ the vector x that is set to zero for $J \in S$. Now we will study under what conditions $f(O_I, x)$ can be approximated by

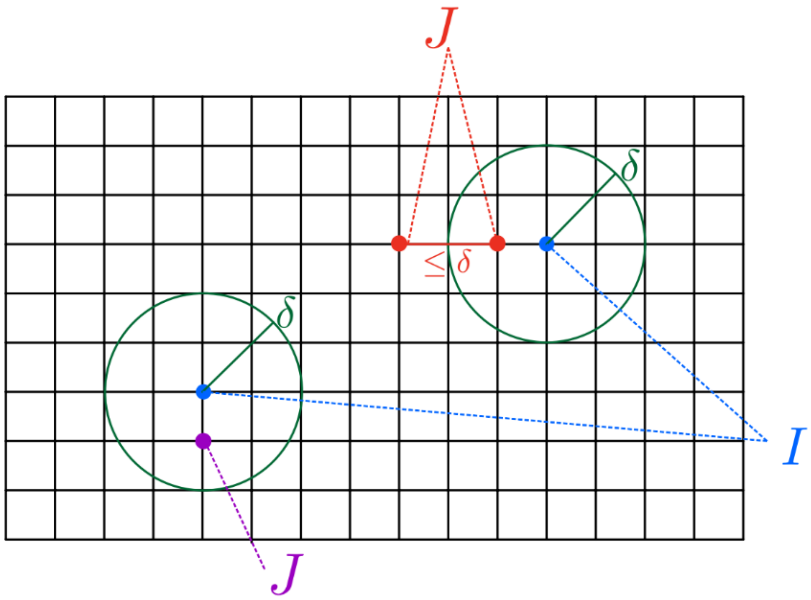


Figure 2: $S_{I,\delta}$ for two choices of J, once with cardinality 1 (purple) and once with 2 (red). Here $D = |I| = 2$.

$f(O_I, x')$. First, following [15, 16] we can rewrite this problem as that of bounding the gradient of f. Let $x(s) = x's + x(1-s)$, then:

$$|f(O_I, x') - f(O_I, x)| = \left| \int_0^1 \mathrm{d}s \partial_s f(O_I, x(s)) \right| \leq \int_0^1 \mathrm{d}s |\partial_s f(O_I, x(s))|$$
$$\leq \int_0^1 \mathrm{d}s \sum_{J \in \mathcal{P}_k(\Lambda)} |(x'_J - x_J) \cdot \nabla_J f(O_I, x(s))| \leq q_* \int_0^1 \mathrm{d}s \sum_{J \not\in S_{I,\delta}} \|\nabla_J f(O_I, x(s))\|$$
$$\leq q_* \int_0^1 \mathrm{d}s \sum_{J \in \mathcal{P}_k(\Lambda)} \left(\mathbf{1}\left(d(I,J) \leq \delta \wedge \mathrm{diam}(J) > \delta\right) + \mathbf{1}\left(d(I,J) > \delta\right) \right) \|\nabla_J f(O_I, x(s))\| ,$$

where we used the Cauchy–Schwarz inequality and the penultimate inequality follows from $\|x_J\| \leq q_J \leq q_*$.

Then recall the quasiadiabatic operator or spectral flow.

Lemma 2.1 (Corollary 2.8 of [40])**.** *Let $H(s)$ be a Hamiltonian such that 1) $\|H'(s)\|$ is uniformly bounded in $s \in [0,1]$ and 2) the spectrum of $H(s)$ can be decomposed in two parts separated by a spectral gap γ. Then if $P(s)$ is the projector onto the lower part, it satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}s} P(s) = i[D(s), P(s)] , \quad D(s) = \int_{-\infty}^{+\infty} \mathrm{d}t W_\gamma(t) \, \mathrm{e}^{itH(s)} \, H'(s) \, \mathrm{e}^{-itH(s)} ,$$

where the function $W_\gamma(t)$ satisfies $W_\gamma(t) = -W_\gamma(-t)$ and $\sup_t W_\gamma(t) = \frac{1}{2}$.

Assuming O_I does not depend on x – we shall discuss in Section 2.3.1 the extension to the case of O_I depending on x – we have, denoted $\tau_t(A) = \mathrm{e}^{itH(x)} A \, \mathrm{e}^{-itH(x)}$,

$$\|\nabla_J f(O_I, x)\| = \|\operatorname{Tr}(O_I [D_J(x), \rho(x)])\| = \|\operatorname{Tr}([O_I, D_J(x)]\rho(x))\|$$
$$\leq \int_{-\infty}^{+\infty} \mathrm{d}t |W_\gamma(t)| \, \|\operatorname{Tr}([\tau_t(\nabla_J H(x)), O_I]\rho(x))\| \leq \int_{-\infty}^{+\infty} \mathrm{d}t |W_\gamma(t)| \, \|[\tau_t(\nabla_J h_J(x_J)), O_I]\| ,$$

where the last inequality follows from Von Neumann's trace inequality: if A, B have eigenvalues α_i, β_i, then $\operatorname{Tr}(AB) \leq \sum_i \alpha_i \beta_i \leq \max_i \alpha_i \sum_j \beta_j$.

As next step we are going to bound the quantity $\|[\tau_t(\nabla_J h_J(x_J)), O_I]\|$. Let us assume that $R = d(I,J) > 0$. We introduce an invertible function of R, $t_* = t_*(R)$, that we shall determine

later, depending on the exact decay of the interactions. We distinguish two regimes: one for $|t| < t_*$, where the commutator is small since the operator $\tau_t(\nabla_J h_J(x_J))$ has not spread enough in the region I. In this regime we can use a Lieb-Robinson bound [34]. The second regime is $|t| > t_*$, where the Lieb-Robinson bound is vacuous and we can use instead the trivial bound:

$$\|[\tau_t(\nabla_J h_J(x_J)), O_I]\| \leq 2\|\tau_t(\nabla_J h_J(x_J))\| \, \|O_I\| = 2\|\nabla_J h_J(x_J)\| \, \|O_I\| \, .$$

Then we break down the quantity to be estimated as the sum of two terms

$$\int_{-t_*}^{+t_*} \mathrm{d}t |W_\gamma(t)| \, \|[\tau_t(\nabla_J h_J(x_J)), O_I]\| \leq \|\nabla_J h_J(x_J)\| \, \|O_I\| \underbrace{\frac{\int_0^{t_*} \mathrm{d}t \|[\tau_t(\nabla_J h_J(x_J)), O_I]\|}{\|\nabla_J h_J(x_J)\| \, \|O_I\|}}_{\mathcal{I}_1} ,$$

$$\left(\int_{-\infty}^{-t_*} + \int_{t_*}^{+\infty} \right) \mathrm{d}t |W_\gamma(t)| \, \|[\tau_t(\nabla_J h_J(x_J)), O_I]\| \leq \|\nabla_J h_J(x_J)\| \, \|O_I\| \underbrace{4 \int_{t_*}^{+\infty} \mathrm{d}t |W_\gamma(t)|}_{\mathcal{I}_2} ,$$

so that

$$\|\nabla_J f(O_I, x)\| \leq \|\nabla_J h_J(x_J)\| \, \|O_I\| (\mathcal{I}_1 + \mathcal{I}_2) \, . \tag{2}$$

We can bound $\mathcal{I}_2$ using the following result on the function W_γ:

Lemma 2.2 (Lemma 2.6 (iv) of [40])**.** *For $t > 0$, let*

$$I_\gamma(t) = \int_t^\infty \mathrm{d}\xi |W_\gamma(\xi)| \, .$$

Then $I_\gamma(t) \leq G(\gamma|t|)$, where

$$G(\xi) = \frac{1}{\gamma} \begin{cases} \frac{K}{2} & 0 \leq \xi \leq \xi_* \, , \\ 130 \, \mathrm{e}^2 \, \xi^{10} u_{2/7}(\xi) & \xi > \xi_* \, , \end{cases} \qquad u_a(\xi) = \exp\left(-a \tfrac{\xi}{\log^2(\xi)} \right) ,$$

with $K \sim 14708$ and $36057 < \xi_* < 36058$.

Thus $\mathcal{I}_2 \leq 4G(\gamma t_*)$. If we assume that

$$\gamma t_*(d(I, J)) > \xi_* \Rightarrow d(I, J) > t_*^{-1}(\gamma^{-1} \xi_*) \, ,$$

then

$$\mathcal{I}_2 \leq \frac{4}{\gamma} 130 \, \mathrm{e}^2 (\gamma t_*)^{10} u_{2/7}(\gamma t_*) \, . \tag{3}$$

Next, we shall discuss different decays of the long range interactions which lead to different choices of t_* and different bounds for $\mathcal{I}_1$ and thus of $|f(O_I, x') - f(O_I, x)|$. We start with exponential decay in Section 2.2.1 and then move to power law in Sections 2.2.2 – 2.2.3.

2.2.1 Exponential decay

In this section we shall present the results and comment on their proofs for Hamiltonians with interactions decaying exponentially with the distance, while the full derivation is presented in A.1.1. We wish to prove the following proposition:

Proposition 2.3 (Corollary A.3.1)**.** *Assume the Hamiltonian has interactions decaying as*

$$\|h_I\| \leq \mathrm{e}^{-\nu \operatorname{diam}(I)} \, P(\operatorname{diam}(I)) \, , \quad \|\nabla_I h_I\| \leq \mathrm{e}^{-\nu \operatorname{diam}(I)} \, ,$$

where ν is a positive constant and P a positive polynomial. Then we have

$$|f(O_I, x') - f(O_I, x)| \leq \epsilon \|O_I\|,$$

where $x' = \chi_{S_{I,\delta}}(x)$ is the truncated vector of parameters, if $0 < \epsilon \leq e^{-1}$ and $\delta \geq c \log^2(\tilde{c}/\epsilon)$, with c and $\tilde{c}$ being constants specified in Appendix A.1.1.

This means that we can approximate $f(O_I, x)$ with a function $f(O_I, \chi_{S_{I,\delta}}(x))$ that is supported only on $\chi_{S_{I,\delta}}(x)$ if we choose δ as above.

The proof of this result is similar to the case of short range interactions [15, 16] and relies on Lieb-Robinson bounds for exponentially decaying interactions [34] to bound $\mathcal{I}_1$ appearing in (2). There are however two main differences with respect to the case of short range interactions discussed previously in the literature. First, we consider a general observable supported on the qubits I, while previous works consider only a geometrically local observable. Second, as discussed in Section 2.2, bounding $|f(O_I, x') - f(O_I, x)|$ requires to bound contributions from Hamiltonian terms J such that $d(I, J) \leq \delta$ and $\mathrm{diam}(J) > \delta$, which are not present in the short range case. The bulk of the proof develops combinatorial arguments to make sure that these terms do not change the asymptotic behavior of the short range case.

2.2.2 Power law with $\alpha \in (2D, 2D+1)$

In this section we shall present the results and comment on their proofs for Hamiltonians with interactions decaying polynomially with the distance, while the full derivation is presented in A.1.2. We wish to prove the following proposition:

Proposition 2.4 (Corollary A.6.1)**.** *Let $\alpha \in (2D, 2D+1)$ and assume that the Hamiltonian has interactions decaying as follows: for all I such that $|I| > 1$,*

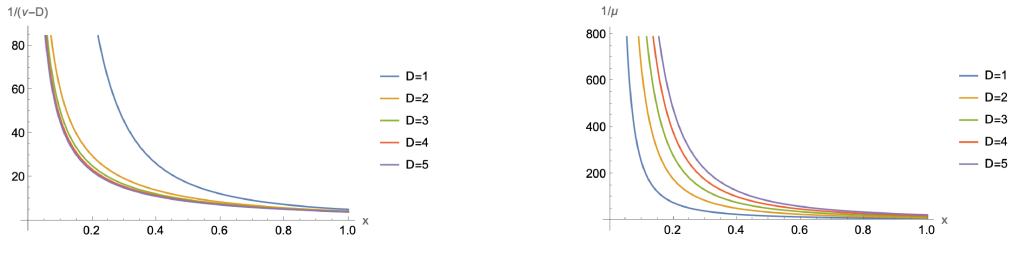
$$\|h_I\| \leq g_{|I|}\,\mathrm{diam}(I)^{-\gamma(|I|)}\,, \quad \|\nabla_I h_I\| \leq \mathrm{diam}(I)^{-\gamma(|I|)}\,, \quad \gamma(\ell) = (\ell-2)D + \alpha\,,$$

for some positive constants g_ℓ. Then we have

$$|f(O_I, x') - f(O_I, x)| \leq \epsilon \|O_I\|,$$

where $x' = \chi_{S_{I,\delta}}(x)$, with δ now being bounded like $\delta \geq \tilde{c}\max(\epsilon^{-1/(\nu-D)}, \log^{2/\mu}(3C_2/\epsilon))$, with the constants being specified in Appendix A.1.2.

Power law interactions with $\alpha \in (2D, 2D+1)$ introduce two new features in the scaling of δ with ϵ compared to exponential decay: a polynomial scaling with $1/\epsilon$ with exponent $1/(\nu - D)$ and the logarithmic dependency $\log^2(1/\epsilon)$ gets replaced with $\log^{2/\mu}(1/\epsilon)$. We show in Figure 3 that both $1/(\nu - D)$ and $1/\mu$ are monotonically decreasing with α, and as $\alpha \to 2D$ both exponents diverge.



(a) (b)

Figure 3: **(a)** $1/(\nu - D)$ as a function of $x = \alpha - 2D$ for $D = 1, \ldots, 5$. **(b)** $1/\mu$ as a function of $x = \alpha - 2D$ for $D = 1, \ldots, 5$.

The proof of this result is similar to that of the case of exponentially decaying interactions and relies on previously derived Lieb Robinson bounds for systems with interactions decaying as a power law [35]. The light cone, which is linear in the case of exponentially decaying or short range interactions, is algebraic for power law interactions. The core of the proof generalises the bounds to series involving exponentially decaying functions of Appendix A.1.1 to the case of power law decay. The summability conditions of the series result in the restrictions to the power law exponent.

2.2.3 Power law with $\alpha > 2D + 1$

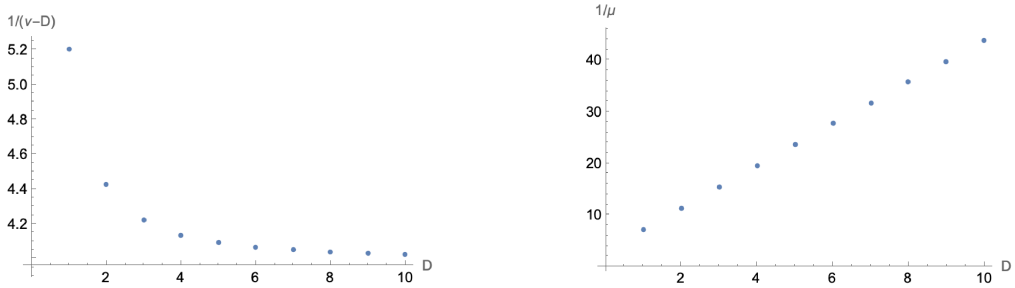
We now note that the results of Section 2.2.2 can be applied to Hamiltonians with $\alpha > 2D + 1$ as well. Indeed if the interactions in the Hamiltonian decay as $1/r^\alpha$ with $\alpha > 2D+1$ we can use that for $r \geq 1$, $1/r^\alpha < 1/r^{2D+1-\zeta}$ for any $\zeta > 0$ and apply Proposition A.6 and Corollary A.6.1 with $\alpha = 2D + 1 - \zeta$, $\zeta \in (0, 1)$:

$$\frac{|f(O_I, x') - f(O_I, x)|}{\|O_I\|} \leq \epsilon \,,$$

if

$$\begin{aligned} \delta &\geq \tilde{c} \max(\epsilon^{-1/(\nu-D)}, \log^{2/\mu}(3C_2/\epsilon))|_{\alpha=2D+1-\zeta} \,, \\ \tilde{c} &= \max((3C_1)^{1/(\nu-D)}, (c/b)^{1/\mu})|_{\alpha=2D+1-\zeta} \,. \end{aligned}$$

In Figure 4 we plot the exponents $1/(\nu - D)$ and $1/\mu$ at $\alpha = 2D + 1 - \zeta$ with $\zeta = 10^{-12}$.



(a) (b)

Figure 4: **(a)** $1/(\nu - D)$ as a function of D at $\alpha = 2D + 1 - \zeta$ with $\zeta = 10^{-12}$. **(b)** $1/\mu$ as a function of D at $\alpha = 2D + 1 - \zeta$ with $\zeta = 10^{-12}$.

So we can extend the results derived for $\alpha \in (2D, 2D + 1)$ to all $\alpha > 2D$. In the following, when we use Corollary A.6.1 for all $\alpha > 2D$, we will understand that if $\alpha > 2D + 1$ we should replace ν, μ with their value at $2D + 1 - \zeta$. A more refined analysis for $\alpha > 2D + 1$ with improved scaling is possible if we use the stronger Lieb-Robinson bounds in this regime [41]. In this case, the dependency on α is such that we recover the short range result as $\alpha \to \infty$.

2.3 Approximation of general p-body observables

Now we consider general observables of the form

$$O = \sum_{I \in \mathcal{P}_p(\lambda)} O_I \,,$$

where O_I is supported on the qubits I as before, and we assume $p = \mathcal{O}(1)$. Then, if we choose δ as in corollary A.3.1 and A.6.1 respectively for exponential and power law decay interactions with exponent $\alpha > 2D$, we have

$$\begin{aligned} &|\operatorname{Tr}(O\rho(x)) - \sum_{I \in \mathcal{P}_p(\Lambda)} f(O_I, \chi_{S_{I,\delta}}(x))| \leq r(O)\epsilon \,, \\ &r(O) = \sum_{I \in \mathcal{P}_p(\Lambda)} \|O_I\| \,. \end{aligned} \tag{4}$$

So we can approximate general p-body observables O provided $r(O)$ is small. If we decompose O as a sum over Paulis, $O = \sum_P \alpha_P P$, then we can bound the difference with $\sum_P \alpha_P f(P, \chi_{\mathrm{dom}(P),\delta})$ by $\sum_P |\alpha_P|\epsilon$. In [15] it was shown that if α_P is non-zero only for geometrically local P, then one can bound $\sum_P |\alpha_P| \leq C\|O\|$. In the following, we are going to assume that we know explicitly

$\|O_I\|$, or a bound of it, which is typically the case in practice where we know what observable we want to measure, so that we can renormalise $O' = O/r(O)$ so that $r(O') = 1$ and we achieve the bound:

$$|\operatorname{Tr}(O'\rho(x)) - \sum_{I\in\mathcal{P}_p(\Lambda)} f(O'_I, \chi_{S_{I,\delta}}(x))| \leq \epsilon\,.$$

2.3.1 Approximation of the energy

In the previous sections we assumed that the observable O_I does not depend on x. This leaves out the important task of predicting the energy of the system, where $O_I = h_I(x_I)$. For this we need to modify the derivations above as follows. We go back to the computation of the gradient of $f(h_I(x_I), x)$:

$$\|\nabla_J f(h_I, x)\| = \|\operatorname{Tr}(\nabla_J h_I \rho(x))\| + \|\operatorname{Tr}(h_I[D_J(x), \rho(x)])\|\,.$$

We need to bound the first term:

$$\|\operatorname{Tr}(\nabla_J h_I \rho(x))\| \leq \|\nabla_J h_I\| = \delta_{I,J}\|\nabla_I h_I\|$$

so that the difference $|f(h_I, x) - f(h_I, x')|$ has the following extra term w.r.t. that of Section 2.2, with $R = d(I, J)$:

$$\left[\sum_{J\in\mathcal{P}_k(\Lambda)} \mathbf{1}\left(R \leq \delta \wedge \operatorname{diam}(J) > \delta\right) + \sum_{J\in\mathcal{P}_k(\Lambda)} \mathbf{1}\left(R > \delta\right)\right] \delta_{I,J}\|\nabla_I h_I\| = \mathbf{1}\left(\operatorname{diam}(I) > \delta\right)\|\nabla_I h_I\|\,.$$

For exponentially decaying interactions, we can bound this by $\mathrm{e}^{-\nu\delta}$, which is subleading compared to the terms in (15), and thus does not alter the scaling of δ with ϵ to ensure that $|f(O_I, \chi_{S_{I,\delta}}(x)) - f(O_I, x)| \leq \epsilon\|O_I\|$ also in the case of $O_I = h_I(x_I)$.

In the case of power law interactions with $\alpha > 2D$, we have from Proposition A.6 that

$$\mathbf{1}\left(\operatorname{diam}(I) > \delta\right)\|\nabla_I h_I\| \leq \mathbf{1}\left(\operatorname{diam}(I) > \delta\right)\operatorname{diam}(I)^{-(|I|-2)D-\alpha} \leq \delta^{-(|I|-2)D-\alpha} \leq \delta^{-\alpha}\,.$$

We assumed $|I| > 1$ since for $|I| = 1$ the condition $\operatorname{diam}(I) > \delta$ cannot be satisfied. After some algebra we have, with $x = \alpha - 2D$,

$$(\nu - D - \alpha)16(2D + 2x - 1)\left(D + x^2 + x\right)^2 = -32D^3(2D-1) - 72D^2(3D-1)x - (80D+39)x^5$$

$$- 2(D(32D+91)+7)x^4 - 8(D(37D+8)-1)x^3 - 4D(D(32D+51)-12)x^2 - 22x^6$$

and $\nu - D - \alpha \leq 0$ since it is the ratio of a polynomial with negative coefficients and one with positive coefficients for all $D = 1, 2, \ldots$ So $-\alpha \leq -\nu + D$ and $\delta^{-\alpha}$ is subleading in the bound of Proposition A.6 and so does not alter the scaling of δ with ϵ to ensure that $|f(O_I, \chi_{S_{I,\delta}}(x)) - f(O_I, x)| \leq \epsilon\|O_I\|$ for $O_I = h_I(x_I)$ also in the case of power law interactions.

For illustration we now compute $r(H)$, with $r(O)$ as in (4), for the case of power law decay with exponent α and $k = 2$. Let us assume that we have a cubic system of side L, so that $n = L^D$. We have, assuming also that the $k = 1$ terms to have norm 1:

$$\begin{aligned} r(H) &= \sum_{ij}\|h_{ij}\| + \sum_i \|h_i\| \leq \sum_{ij}(1 + d(i,j))^{-\alpha} + \sum_i 1 \\ &= \sum_i \sum_j (1 + d(i,j))^{-\alpha} + n \leq \sum_i \int_{\mathbb{R}^D} \mathrm{d}^D x (1 + \|x - i\|)^{-\alpha} + n \\ &\leq \Omega_D \sum_i \int_0^\infty \mathrm{d}r (1+r)^{-\alpha} + n \leq \frac{\Omega_D}{1-\alpha}\sum_i 1 + n = \frac{\Omega_D}{1-\alpha} n + n\,. \end{aligned}$$

Here Ω_D is a constant and so $\sum_I \|h_I\| \leq Cn$ and the observable H/n can be approximated by $\sum_I f(h_I(x_I), \chi_{S_{I,\delta}}(x))$. A similar computation can be done for $k > 2$ and for exponentially decaying interactions.

3 Machine learning model and sample complexity bounds

3.1 Generalisation bounds for single observable

We assume we are given data $S = \{x^{(i)}, y^{(i)}\}_{i=1}^N$ with

$$|y^{(i)} - \mathrm{Tr}\Big(O\rho(x^{(i)})\Big)| \leq \epsilon_2\,, \tag{5}$$

for an observable as in (24):

$$O = \sum_{I\in\mathcal{S}} O_I\,, \quad \mathcal{S} \subseteq \mathcal{P}_p(\Lambda)\,.$$

The machine learning algorithm learns the weights w of

$$h(x) = w\cdot\phi(x) = \sum_{I\in\mathcal{S}}\sum_{x'\in X_{I,\delta}} (w)_{I,x'}\phi(x)_{I,x'}\,, \quad \phi(x)_{I,x'} = \mathbf{1}(x\in T_{x',I})\,,$$

with $X_{I,\delta}$ and $T_{x',I}$ as in (21) and (22). We are going to study the generalisation properties of the predictor obtained by minimising the training error

$$R_S(h) = \frac{1}{N}\sum_{i=1}^N |h(x^{(i)}) - y^{(i)}|^2\,.$$

Given the bound on the norm of w in Lemma (A.11), we consider the optimisation problem

$$\min_{w\in\mathcal{C}_B} R_S(h_w)\,, \quad \mathcal{C}_B = \{w\in\mathbb{R}^m \;|\; \|w\|_1 \leq B\}\,. \tag{6}$$

This algorithm is called LASSO in the statistics literature and has been extensively studied [36].

We have the following result which adapts Theorem 1 of [15] to our setting of long range interactions and observables.

Theorem 3.1. *Consider an observable O*

$$O = \sum_{I\in\mathcal{S}} O_I\,, \quad \mathcal{S} \subseteq \mathcal{P}_p(\Lambda)\,.$$

and a dataset $S = \{x^{(i)}, y^{(i)}\}_{i=1}^N$ with

$$|y^{(i)} - \mathrm{Tr}\Big(O\rho(x^{(i)})\Big)| \leq \epsilon_2\,.$$

Choose δ as

$$\delta > \max(b^{-1}\xi_*, b^{-1}\eta_*, c\log^2(1/\epsilon_1))$$

for exponentially decaying interactions, where the constants are as in Proposition A.3 and Corollary A.3.1 and as

$$\delta > \max((b^{-1}\xi_*)^{1/\mu}, (b^{-1}\eta_*)^{1/\mu}, \tilde{c}\log^{2/\mu}(1/\epsilon_1), \tilde{c}\epsilon_1^{-1/(\nu-D)})\,,$$

for power law decay with $\alpha > 2D$, where the constants are as in Proposition A.6.1, Corollary A.6.1 , and for $\alpha > 2D+1$, ν, μ are evaluated at $\alpha = 2D+1-\zeta$. Then form the linear predictor $h_*(x) = w_*\cdot\phi(x)$ with w_* a solution of the optimisation problem (6) such that

$$R_S(h_*) \leq \frac{\epsilon_3}{2} + \min_{w\in\mathcal{C}_B} R_S(h_w)\,.$$

If the number of samples is

$$N = r(O)^4\epsilon_3^{-2}\mathcal{N}(\epsilon_1)\mathcal{O}(\log(|\mathcal{S}|/\delta))\,, \tag{7}$$

with

$$\omega = \frac{kD}{\nu - D}\,, \quad \mathcal{N}(\epsilon) = \begin{cases} 2^{\mathcal{O}(\text{polylog}(1/\epsilon))} & \text{exp decay,} \\ 2^{\mathcal{O}(\epsilon^{-\omega}\log(1/\epsilon))} & \text{power law } \alpha > 2D\,, \end{cases}$$

the following generalisation bound holds with probability at least $1-\gamma$*:*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}|h_*(x) - y|^2 \leq (\epsilon_1 + \epsilon_2)^2 + \epsilon_3\,.$$

The sample complexity depends polynomially on $r(O)$ and logarithmically on $|\mathcal{S}|$. If we normalise O such that $r(O) = \mathcal{O}(1)$, then the dependency on n will be $\mathcal{O}(\log(n))$ as long as O is the sum of k-body interactions with $|\mathcal{S}| = \mathcal{O}(n^k)$ with $k = \mathcal{O}(1)$. This is the same for both cases of exponentially decaying interactions and power law interactions with $\alpha > 2D$. However, the dependency on $1/\epsilon$ goes from quasi-polynomial to exponential as we go from exponential to power law interactions. Now we will prove this theorem.

Proof. The proof relies on the generalisation bound for LASSO.

Lemma 3.2 ([36])**.** *Denote the input space* $\mathcal{X} \subseteq \mathbb{R}^A$ *and the output space* $\mathcal{Y}$*. We consider a class of linear predictors* $h(x) = w \cdot x$ *with* $\|w\| \leq B$*. If* $\mathcal{D}$ *is the distribution over* $\mathcal{X}\times\mathcal{Y}$ *from which the data* S *of size* N *is drawn, then with probability at least* $1-\gamma$ *we have*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}|h(x) - y|^2 \leq R_S(h) + 2r_\infty BM\sqrt{\frac{2\log(2A)}{N}} + M^2\sqrt{\frac{\log(1/\gamma)}{2N}}$$

with $\|x\|_\infty \leq r_\infty$*,* $|h(x) - y| \leq M$ *for all* $(x,y) \in \mathcal{X}\times\mathcal{Y}$*, and* R_S *is the training error.*

To apply this, first we need to bound the training error.

Lemma 3.3 (Lemma 15 of [15])**.** *Let* w_* *be a solution of the optimisation problem* (6) *such that*

$$R_S(h_*) \leq \frac{\epsilon_3}{2} + \min_{w\in\mathcal{C}_B} R_S(h_w)\,, \quad h_*(x) = w_*\cdot\phi(x)\,.$$

Also assume that

$$|f(O,x) - g(O,x)| \leq \epsilon_1\,.$$

Then, with ϵ_2 *as in* (5)*,*

$$R_S(h_*) \leq \frac{\epsilon_3}{2} + (\epsilon_1+\epsilon_2)^2\,.$$

Next we use Lemma 3.2 with $\mathcal{X} = \{0,1\}^{m_\phi}$, so that $r_\infty = 1$ and

$$A = m_\phi \leq |\mathcal{S}|\mathcal{N}(\epsilon)\,.$$

Since $\text{Tr}(O\rho) - \epsilon_2 \leq y \leq \text{Tr}(O\rho) + \epsilon_2$, we have $|y| \leq |\,\text{Tr}(O\rho)| + |\epsilon_2|$, which can be verified by checking the two cases $y > 0$ and $y < 0$. Then from Lemma A.11,

$$|h(x) - y| \leq |w\cdot\phi(x)| + |y| \leq \|w\|_1\|\phi(x)\|_\infty + \|O\| + \epsilon_2 \leq r(O)\mathcal{N}(\epsilon_1) + r(O) + \epsilon_2 \leq r(O)\mathcal{N}(\epsilon_1)\,,$$

where we used that $\epsilon_2 = \mathcal{O}(1)$ since we consider asymptotics in $1/\epsilon_2$, and so $M = B = r(O)\mathcal{N}(\epsilon_1)$. Now we want to find N such that the generalisation bound is small

$$G = 2r_\infty BM\sqrt{\frac{2\log(2A)}{N}} + M^2\sqrt{\frac{\log(1/\gamma)}{2N}} \leq \frac{\epsilon_3}{2}\,.$$

We have, since $\mathcal{N}(\epsilon_1)^2 = \mathcal{N}(\epsilon_1)$, $2\mathcal{N}(\epsilon_1) = \mathcal{N}(\epsilon_1)$:

$$\begin{aligned} G &\leq r(O)^2\mathcal{N}(\epsilon_1)\frac{1}{\sqrt{N}}\left(\sqrt{2}\sqrt{\log(|\mathcal{S}|\mathcal{N}(\epsilon_1))} + \frac{1}{\sqrt{2}}\sqrt{\log(1/\gamma)}\right) \\ &\leq r(O)^2\mathcal{N}(\epsilon_1)\frac{1}{\sqrt{N}}\left(\sqrt{\log(|\mathcal{S}|\mathcal{N}(\epsilon_1))} + \sqrt{\log(1/\gamma)}\right)\,. \end{aligned}$$

So we need

$$N = \frac{4}{\epsilon_3^2} r(O)^4 \mathcal{N}(\epsilon_1) \left(\sqrt{\log(|\mathcal{S}|) + \log \mathcal{N}(\epsilon_1)} + \sqrt{\log(1/\gamma)} \right)^2$$
$$= r(O)^4 \epsilon_3^{-2} \mathcal{N}(\epsilon_1) \mathcal{O}(\log(|\mathcal{S}|/\gamma)) \,.$$

In the second equality we used that $(a+b)^2 \leq 2(a^2+b^2)$, which follows from $0 \leq (a-b)^2 = a^2 + b^2 - 2ab$, and that $\mathcal{N}(\epsilon_1) \log \mathcal{N}(\epsilon_1) = \mathcal{N}(\epsilon_1)$. □

Assuming $r(O) = \mathcal{O}(1), \gamma = \mathcal{O}(1)$ in the definition of the data size N, as remarked in [15] the training time of LASSO is $\mathcal{O}(m_\phi \log(m_\phi)/\epsilon_3^2) = \mathcal{O}(|\mathcal{S}|N)$.

3.2 Sample complexity reduction from equivariance

3.2.1 Equivariance of observables

We first recall given a map $f : \mathcal{X} \to \mathcal{Y}$ and a group G acting on $\mathcal{X}$, $\mathcal{Y}$, f is called G-equivariant if $f(g \cdot x) = g \cdot f(x)$ for all $g \in G$ and x in the input space $\mathcal{X}$, where $g\cdot$ denotes the group action.

Given a Hamiltonian

$$H(x) = \sum_{I \in \mathcal{E}} h_I(x_I) \,,$$

with $\mathcal{E} \subseteq \mathcal{P}(\Lambda)$, we define the interaction hypergraph as the triple $\mathcal{I} = (\Lambda, \mathcal{E}, h)$, where h is a function that assigns to each $I \in \mathcal{E}$ the interaction term $h_I(x_I)$. The automorphism group $G = \mathrm{Aut}(\mathcal{I})$ of the interaction hypergraph is the set of permutations of vertices that permute the interactions terms; specifically, $g \in G$ if

$$\hat{g} h_I(x_I) \hat{g}^{-1} = h_{gI}(x_I) \,,$$

where $\hat{g}$ is the unitary representation of a permutation g on the Hilbert space of the quantum theory and gI is the permuted set of vertices, namely if $I = (i_1, \ldots, i_\ell)$, $gI = (gi_1, \ldots, gi_\ell)$. For example, for a system with periodic boundary conditions and pairwise interactions $h_{ij}(x_{ij}) = x_{ij} f(d(i,j)) H_{ij}$ with $d(i,j)$ the Euclidean distance and f an arbitrary function, G is the Euclidean group of symmetries of the lattice under which $d(i,j)$ is invariant. This setting includes nearest neighbor and long-range interactions.

With this definition, we have the following result:

Lemma 3.4. *Consider a Hamiltonian with interaction hypergraph $\mathcal{I}$. Then the function $f(O_I, x) = \mathrm{Tr}(O_I \rho(x))$ is Aut($\mathcal{I}$)-equivariant:*

$$f(O_{gI}, x) = f(O_I, g \cdot x) \,,$$

for all x and $g \in$ Aut($\mathcal{I}$), where $(g \cdot x)_J = x_{gJ}$.

Proof. To prove the Lemma, we first we show that the Hamiltonian, seen as a map from x to an operator on the qubits, is $\mathrm{Aut}(\mathcal{I})$-equivariant. If $g \in \mathrm{Aut}(\mathcal{I})$ is a permutation of the vertices whose action on vertex i is denoted by gi, the action on the hyperedge $I = \{i_1, \ldots, i_\ell\}$ is $gI = \{gi_1, \ldots, gi_\ell\}$ and by definition it fixes $\mathcal{E} = \{I_0, \ldots, I_{m-1}\}$, i.e. permutes the sets $I_0, \ldots, I_{m-1}$:

$$g\mathcal{E} = \{gI_0, \ldots, gI_{m-1}\} = \mathcal{E} \,.$$

Then we have, denoted by $\hat{g}$ the unitary representation of g on the space of the qubits,

$$\hat{g} H(x) \hat{g}^{-1} = \sum_{I \in \mathcal{E}} \hat{g} h_I(x_I) \hat{g}^{-1} = \sum_{I \in \mathcal{E}} h_{gI}(x_I) = \sum_{J \in \mathcal{E}} h_J(x_{g^{-1}J}) = H(g^{-1} \cdot x) \,,$$

where we set $J = gI$ and used that $g\mathcal{E} = \mathcal{E}$ to move the action from h_I to its argument. Further,

$$Z_\beta(x) = \mathrm{Tr}\Big(\mathrm{e}^{-\beta H(x)}\Big) = \mathrm{Tr}\Big(\hat{g}\, \mathrm{e}^{-\beta H(x)}\, \hat{g}^{-1}\Big) = \mathrm{Tr}\Big(\mathrm{e}^{-\beta H(g^{-1} \cdot x)}\Big) = Z_\beta(g^{-1} \cdot x) \,,$$

$$\hat{g} \rho(x) \hat{g}^{-1} = \lim_{\beta \to \infty} \hat{g} \frac{\mathrm{e}^{-\beta H(x)}}{Z_\beta(x)} \hat{g}^{-1} = \lim_{\beta \to \infty} \frac{\mathrm{e}^{-\beta H(g^{-1} x)}}{Z_\beta(g^{-1} \cdot x)} = \rho(g^{-1} \cdot x) \,, \tag{8}$$

so we have the equivariance property of the function f for any $g \in \mathrm{Aut}(\mathcal{I})$:

$$f(O_{gI}, x) = \mathrm{Tr}(O_{gI}\rho(x)) = \mathrm{Tr}\big(\hat{g}O_I\hat{g}^{-1}\rho(x)\big) = \mathrm{Tr}(O_I\rho(g\cdot x)) = f(O_I, g\cdot x)\,.$$

□

We stress that this notion of equivariance is not the same as invariance of the Hamiltonian, which is a more restrictive concept and of more limited use. Also, note that for the case of interactions over edges of a graph, $\mathrm{Aut}(\mathcal{I})$ reduces to the automorphism group of the graph. For example, if H is the Hamiltonian for qubits on a closed chain with nearest neighbor interactions,

$$H(x) = \sum_{i=0}^{n-1} h_{i,i+1}(x_{i,i+1})\,,$$

with indices understood modulo n, $\mathrm{Aut}(\mathcal{I})$ contains the cyclic group of order n corresponding to translations of the chain, whose generator g acts as $gi = i+1 \mod n$ and cyclically permutes the entries of x as:

$$g\cdot(x_{0,1}, x_{1,2}, \ldots, x_{n-1,0}) = (x_{1,2}, x_{2,3}, \ldots, x_{0,1})\,.$$

3.2.2 Equivariance of the ML model

A ML model for the observable $f(O_I, x)$ should also satisfy the equivariance constraint of Lemma 3.4. In fact we show here that, provided that the model satisfies the equivariance constraints, we can reduce the sample complexity.

Proposition 3.5. *The model*

$$h_w^{O_I}(x) = \sum_{x'\in X_{I,\delta}} w_{x'}\phi(x)_{I,x'}\,, \quad \phi(x)_{I,x'} = \mathbf{1}(x \in T_{x',I})\,,$$

satisfies the equivariance constraint $h_{\tilde{w}}^{O_{gI}}(x) = h_w^{O_I}(g\cdot x)$ *of Lemma 3.4 if the weights transform as*

$$\tilde{w}_{g^{-1}\cdot x'} = w_{x'}\,,$$

for any x' *and* $g \in Aut(\mathcal{I})$.

Proof. We want to relate

$$h_{\tilde{w}}^{O_{gI}}(x) = \sum_{x'\in X_{gI,\delta}} \tilde{w}_{x'}\phi(x)_{gI,x'}\,, \quad \phi(x)_{gI,x'} = \mathbf{1}(x \in T_{x',gI})\,,$$

to $h_w^{O_I}(g\cdot x)$. So we first discuss the transformation of the various sets involved. We have

$$\begin{aligned} S_{gI,\delta} &= \{J \in \mathcal{P}_k(\Lambda)\,|\,d(gI,J) \le \delta \wedge \mathrm{diam}(J) \le \delta\} \\ &= \{J \in \mathcal{P}_k(\Lambda)\,|\,d(I,g^{-1}J) \le \delta \wedge \mathrm{diam}(g^{-1}J) \le \delta\} \\ &= \{gJ' \in \mathcal{P}_k(\Lambda)\,|\,d(I,J') \le \delta \wedge \mathrm{diam}(J') \le \delta\} = gS_{I,\delta}\,. \end{aligned} \tag{9}$$

In the second line we use invariance of the distance under permutation, $d(I,J) = d(gI,gJ)$, and that as a consequence the diameter is also invariant under permutations. In the third line we relabelled $g^{-1}J = J'$ and used that $\mathcal{P}_k(\Lambda) = g\mathcal{P}_k(\Lambda)$. Thus

$$\begin{aligned} X_{gI,\delta} &= \{x \in [-1,1]^m\,|\,x_J = 0 \text{ if } J \notin gS_{I,\delta}\,, \quad x_J \in W_{\delta_2}^{|J|} \text{ if } J \in gS_{I,\delta}\} \\ &= \{x \in [-1,1]^m\,|\,x_{gJ'} = 0 \text{ if } J' \notin S_{I,\delta}\,, \quad x_{gJ'} \in W_{\delta_2}^{|J|} \text{ if } J' \in S_{I,\delta}\} \\ &= \{g^{-1}\cdot x' \in [-1,1]^m\,|\,x'_{J'} = 0 \text{ if } J' \notin S_{I,\delta}\,, \quad x'_{J'} \in W_{\delta_2}^{|J|} \text{ if } J' \in S_{I,\delta}\} \\ &= g^{-1}\cdot X_{I,\delta}\,, \end{aligned}$$

where in the second line we defined $J = gJ'$ and used that $gJ \in gS$ if and only if $J \in S$. In the third line we used that if $x' = g \cdot x$, then $x'_J = x_{gJ}$. Similarly,

$$\begin{aligned} T_{x,gI} &= \{x' \in [-1,1]^m \mid -\frac{\delta_2}{2} < x_J - x'_J < +\frac{\delta_2}{2}\,, \quad J \in gS_{I,\delta}\} \\ &= \{x' \in [-1,1]^m \mid -\frac{\delta_2}{2} < x_{gJ'} - x'_{gJ'} < +\frac{\delta_2}{2}\,, \quad J' \in S_{I,\delta}\} \\ &= \{g^{-1} \cdot z' \in [-1,1]^m \mid -\frac{\delta_2}{2} < (g \cdot x)_{J'} - z'_{J'} < +\frac{\delta_2}{2}\,, \quad J' \in S_{I,\delta}\} \\ &= g^{-1} \cdot T_{g\cdot x,I}\,. \end{aligned}$$

Here we have used the same manipulations as in $X_{gI,\delta}$ and defined $z' = g \cdot x'$. Then

$$\phi(x)_{gI,x'} = \mathbf{1}(x \in T_{x',gI}) = \mathbf{1}(x \in g^{-1} \cdot T_{g\cdot x',I}) = \mathbf{1}(g \cdot x \in T_{g\cdot x',I}) = \phi(g \cdot x)_{I,g\cdot x'}$$

and we can prove the result:

$$\begin{aligned} h_{\tilde{w}}^{O_{gI}}(x) &= \sum_{x' \in X_{gI,\delta}} \tilde{w}_{x'} \phi(x)_{gI,x'} = \sum_{x' \in g^{-1}\cdot X_{I,\delta}} \tilde{w}_{x'} \phi(g \cdot x)_{I,g\cdot x'} \\ &= \sum_{z' \in X_{I,\delta}} \tilde{w}_{g^{-1}\cdot z'} \phi(g \cdot x)_{I,z'} = h_w^{O_I}(g \cdot x)\,, \quad w_{z'} = \tilde{w}_{g^{-1}\cdot z'}\,, \end{aligned}$$

where we called $z' = g \cdot x'$. □

This means that if we have a model

$$h(x) = w \cdot \phi(x) = \sum_{I \in \mathcal{S}} \sum_{x' \in X_{I,\delta}} (w)_{I,x'} \phi(x)_{I,x'}\,, \quad \phi(x)_{I,x'} = \mathbf{1}(x \in T_{x',I})\,,$$

for predicting $\mathrm{Tr}(O\rho(x))$, where

$$O = \sum_{I \in \mathcal{S}} O_I\,, \quad \mathcal{S} \subseteq \mathcal{P}_k(\Lambda)\,,$$

we can take weights that satisfy the equivariance condition of the Proposition 3.5,

$$w_{I,x'} = w_{gI,g^{-1}\cdot x'}\,,$$

for all $g \in \mathrm{Aut}(\mathcal{I})$. This is indeed the equivariance of the true weights $(w')_{I,x'}$ of (23). This leads to a reduction in the number of learnable parameters and thus a sample complexity reduction. More precisely, we consider the quotient space $\mathcal{S}/\mathrm{Aut}(\mathcal{I})$ where two sets $I, J \in \mathcal{S}$ are equivalent if there is a $g \in \mathrm{Aut}(\mathcal{I})$ such that $I = gJ$. Then equivariance relates all the weights related to sets within an equivalence class of $\mathcal{S}/\mathrm{Aut}(\mathcal{I})$, and leaves a number of independent weights equal to the cardinality of the space $\mathcal{S}/\mathrm{Aut}(\mathcal{I})$. We summarise this result in the following corollary.

Corollary 3.5.1. *The sample complexity of Theorem 3.1 is reduced by using an equivariant model to*

$$N = r(O)^4 \epsilon_3^{-2} \mathcal{N}(\epsilon_1) \mathcal{O}(\log(|\mathcal{S}/Aut(\mathcal{I})|/\gamma))\,.$$

Continuing our example of the $D = 1$ periodic chain of Section 3.2.1, if we consider an observable that is a sum of local terms such that

$$O = \sum_{i=0}^{n-1} O_i\,, \quad O_i = O_{g^i 0}\,,$$

with g the translation by one site, $gi = i+1 \mod n$, we have $\mathcal{S} = \{0, \dots, n-1\}$ and $|\mathcal{S}/\mathrm{Aut}(\mathcal{I})| = 1$, leading to a $\mathcal{O}(1)$ complexity of learning to predict this observable, using the following ML model

$$\begin{aligned} h(x) &= \sum_{i=0}^{n-1} \sum_{x'} (w)_{i,x'} \phi(x)_{i,x'} = \sum_{i=0}^{n-1} \sum_{x'} (w)_{0,g^i\cdot x'} \phi(x)_{i,x'} = \sum_{i=0}^{n-1} \sum_{x''} (w)_{0,x''} \phi(x)_{i,g^{-i}\cdot x''} \\ &= \sum_{i=0}^{n-1} (w_0 * \Phi(x))_i\,, \end{aligned}$$

with $*$ the convolution with filter w_0 and $\Phi(x)$ is obtained by concatenating $\{\phi(x)_{i,g^{-i}\cdot x''}\}_{x''}$ over i.

3.2.3 Equivariance of the random feature model

We discuss here equivariance of the random feature model that is used in the experiments Section 4. For a given δ, we define the set of parameters that influence $f(O_I, x)$ according to the results of Section 2.2:

$$Z_I(x) = (x_J \,|\, J \in S_{I,\delta}) \in \bigoplus_{J\in S_{I,\delta}} [-1,1]^{q_J}\,.$$

The random feature model is defined as

$$F^{O_I}(x) = \sum_{i=1}^{R} a_i \sigma(b_i^T Z_I(x))\,,$$

where $a \in \mathbb{R}^R$ are the learnable parameters and b_i's are random vectors of the same size of Z_I. σ is a non-linearity and thus F^{O_I} can be thought as a two-layer neural network with random weights in the first layer [42].

Proposition 3.6. *The random feature model satisfies the equivariance constraint of Lemma 3.4,*

$$F^{O_{gI}}(x) = F^{O_I}(g\cdot x)$$

for all values of a, b.

Proof. From (9), we get, with $gJ' = J$

$$Z_{gI}(x) = (x_J \,|\, J \in S_{gI,\delta}) = (x_J \,|\, J \in gS_{I,\delta}) = (x_{gJ'} \,|\, J' \in S_{I,\delta}) = Z_I(g\cdot x)\,.$$

Thus for any values of the parameters a, b:

$$F^{O_{gI}}(x) = \sum_{i=1}^{R} a_i \sigma(b_i^T Z_{gI}(x)) = \sum_{i=1}^{R} a_i \sigma(b_i^T Z_I(g\cdot x)) = F^{O_I}(g\cdot x)\,.$$

□

As in the case of the feature map $\phi(x)$, this means that we can reduce the number of parameters when predicting an observable by associating a set of a, b's to each equivalence class in $\mathcal{S}/\mathrm{Aut}(\mathcal{I})$. For example, for the case of a $D = 1$ periodic chain and observable

$$O = \sum_{i=0}^{n-1} O_i\,, \quad O_i = O_{g^i 0}$$

we can use the random feature model with $\mathcal{O}(1)$ parameters:

$$F(x) = \sum_{i=0}^{n-1} F^{O_{g^i 0}}(x) = \sum_{i=0}^{n-1} F^{O_0}(g^i\cdot x) = \sum_{i=0}^{n-1}\sum_{j=1}^{R} a_j \sigma(b_j * x)_i = \sum_{i=0}^{n-1} (a * \Phi(x))_i\,.$$

Here we have first noted that

$$(b_j * x)_i = b_j^T Z_0(g^i\cdot x) = \sum_{k=1}^{|S_{0,\delta}|} b_{j,k} x_{k+i \bmod n}$$

is the i-th output of the 1D convolution with filter b_j and input x, with δ determining the filter size. We can think of j as indexing the channel dimension of size R and i the space dimension of size n. The output of the second layer can be also interpreted as a convolution with filter a of size 1 in the space dimension and R in the channel dimension. $\Phi(x)$ is the concatenation of $\{\sigma(b_j * x)_i\}_{j=1}^R$ over i. So the model is a two-layer convolutional neural network, with final layer a global pooling over space.

3.3 Classical shadows

A classical shadow of the density matrix ρ is obtained by averaging repeated random measurements [33]. One selects uniformly at random whether to measure X, Y, Z for each qubit and stores the associated measurement results as classical data $s_i^t \in \{0_X, 1_X, 0_Y, 1_Y, 0_Z, 1_Z\}$ for the measurement outcome of qubit i at time t. Here we denote by $0_A, 1_A$ the possible states after measurement of the Pauli A. After T measurements, the classical shadow is

$$\sigma_T(\rho) = \frac{1}{T}\sum_{t=1}^{T} \left(3\left|s_1^t\right\rangle\left\langle s_1^t\right| - \mathbf{1}_2\right) \otimes \cdots \otimes \left(3\left|s_n^t\right\rangle\left\langle s_n^t\right| - \mathbf{1}_2\right) .$$

The power of classical shadows is that they can ϵ-approximate the reduced density matrices of an $\mathcal{O}(1)$ subsystem with $T = \mathcal{O}(\log(n)/\epsilon^2)$. Our first result is the extension of [15, Corollary 5] to our setting.

Theorem 3.7. *Suppose we have data* $\{x^{(i)}, \sigma_T(\rho(x^{(i)}))\}_{i=1}^N$ *with*

$$T = \mathcal{O}(\log(nN/(\gamma/2))/(\epsilon_2')^2)\,, \quad \epsilon_2' = \epsilon_2/C\,,$$
$$N = (\epsilon_3/C^2)^{-2}\mathcal{N}(\epsilon_1/C)\mathcal{O}(\log(n/\gamma))\,,$$

with $\epsilon_1, \epsilon_2, \epsilon_2, C > 0$ *and* $\gamma \in (0,1)$*. Then we can learn a predictor* $\hat{\rho}(x)$ *that achieves*

$$\mathbb{E}_{x\sim\mathcal{D}} |\operatorname{Tr}(O\rho(x)) - \operatorname{Tr}(O\hat{\rho}(x))|^2 \leq (\epsilon_1 + \epsilon_2)^2 + \epsilon_3\,,$$

with probability at least $1-\gamma$ *for all observables* O *such that*

$$O = \sum_{P\in P_k} \alpha_P P\,, \quad \sum_{P\in P_k} |\alpha_P| \leq C\,, \tag{10}$$

where P_k *is the set of Pauli strings with weight at most* k*.*

This shows that given classical shadow data we can predict all observables of the form $\sum_{I\in\mathcal{P}_k} O_I$ with a sample complexity similar to that of predicting a single one, presented in Theorem 3.1. The proof adapts [15, Corollary 5] to our setting and is presented in Appendix B.1.

We also extend the results for equivariant Hamiltonians to data obtained from classical shadows. Recall the definition of $\mathrm{Aut}(\mathcal{I})$ from 3.2.1. We have the following result:

Corollary 3.7.1. *Under the hypothesis of Theorem 3.7 and assuming further that* $\mathcal{D}$ *is a probability distribution over* x *that is invariant under* $Aut(\mathcal{I})$*, we can predict all observables of the form (10) with*

$$N = (\epsilon_3/C^2)^{-2}\mathcal{N}(\epsilon_1/C)\mathcal{O}(\log(|P_k/Aut(\mathcal{I})|/\gamma))\,,$$

where $Aut(\mathcal{I})$ *is the automorphism group of the interaction hypergraph.*

This result is most powerful if $\mathcal{O}(|P_k/\mathrm{Aut}(\mathcal{I})|) = \mathcal{O}(1)$, since it reduces the complexity from $\mathcal{O}(\log(n))$ to $\mathcal{O}(1)$. As remarked before, this is the case for example if we consider observables that are sum of geometrically local terms (which is the setting of [15, 16, 17]) on a lattice with periodic boundary conditions.

The core of the proof of this result is to show that if $h_*^P(x)$ is an equivariant model as in Proposition 3.5, then

$$\mathbb{E}_{x\sim\mathcal{D}}\left(|h_*^P(x) - y^P(x)|^2\right) = \mathbb{E}_{x\sim\mathcal{D}}\left(|h_*^{P'}(x) - y^{P'}(x)|^2\right)\,, \quad P' = \hat{g}P\hat{g}^{-1}\,,$$

where $y^{(P)}(x)$ is the expectation value of P in the classical shadow, $y^{(P)}(x) = \operatorname{Tr}(P\sigma_T(\rho(x)))$, and $g \in \mathrm{Aut}(\mathcal{I})$. In turn this follows by showing the equivariance of the measure on $y^{(P)}(x)$ induced by the classical shadow measurement protocol. See Appendix B.2 for details.

4 Experiments

In this section, we will cover the details and the setup for the numerical experiments carried out and present the results affirming the theoretically guaranteed bounds. We will discuss implementation of the simulations for different systems and of the machine learning model. For most of the experiments, our main focus will be on demonstrating the predicted efficient scaling of the algorithm. We will also cover an important phenomenon observed for practical usability of the algorithm regarding normalisation of the observables and concentration of expectation values in Section 4.2.

4.1 Setup

The simulations of the quantum systems were typically obtained using exact diagonalisation for up to 20 sites, while larger systems were simulated using tensor networks and the Density Matrix Renormalisation Group (DMRG) method [43]. The tensor network calculations were carried out using the Python library TeNPy [44]. Note that DMRG is a variational method, and so the resulting state is just an approximation of the true ground state, hence there can be possible issues with convergence and small numerical errors. The practical applicability of DMRG is the reason why we focus on one dimensional systems for the numerical experiments; either with open or periodic boundary conditions. For DMRG, the SVD cut-off was set to 10^{-9} while the convergence criterion for the ground energy was a relative change of 10^{-6}. The maximal number of sweeps used was typically set between 100 and 200, and the maximal bond dimension of the MPS was typically gradually increased up to 100 or 200.

To determine whether or not the Hamiltonians remain gapped across the chosen range of parameters, we would also at first measure the spectral gaps for many random choices of parameters as the system sizes increase to see if they aren't closing. To do this for larger systems, the DMRG would be run on a Hamiltonian with projection onto the orthogonal space to the ground state to obtain the first excited state.

Following Appendix D of [15], the machine learning part is based on LASSO, an l_1-regularised linear regression model, with feature mapping with built in local-geometric bias. Instead of using the discretisation as was considered in the theoretical guarantees, we will use random Fourier features, as discussed in Section 3.2.3, which are commonly used in practice [45].

Given an observable $O = \sum_{I \in \mathcal{S}} O_I$, for any $I \in \mathcal{S}$ with $|I| = g$ occurring non-trivially in O, one would consider all the parameters x_J's in its δ-neighbourhood for all J such that $d(I, J) \leq \delta$ and $\mathrm{diam}(J) \leq \delta$ to create a vector

$$Z_I^{(\delta)} = (x_J \,|\, J \in S_{I,\delta})\,.$$

This vector would then be mapped to its randomised Fourier features via

$$\phi : Z \mapsto \begin{pmatrix} \cos\left(\frac{\gamma}{\sqrt{l}}(\omega_1 \cdot Z)\right) \\ \sin\left(\frac{\gamma}{\sqrt{l}}(\omega_1 \cdot Z)\right) \\ \vdots \\ \cos\left(\frac{\gamma}{\sqrt{l}}(\omega_R \cdot Z)\right) \\ \sin\left(\frac{\gamma}{\sqrt{l}}(\omega_R \cdot Z)\right) \end{pmatrix},$$

where l is the length of the vector Z, and γ and R are hyperparameters of the model, to create vectors

$$\phi_I^{(\delta)} = \phi\left(Z_I^{(\delta)}\right). \tag{11}$$

The ω_i's are vectors of length l of independent standard normal random variables, which are the same for all I with $|I| = g$ for a specific g, and all of which are set for a given LASSO model and used for all its training and testing samples. These vectors would then be concatenated over all such I into a g-body interaction feature vector

$$\Phi^{(g,\delta)} = \left(\phi_I^{(\delta)} \,|\, I \in \mathcal{S} \wedge |I| = g\right), \tag{12}$$

which would finally be concatenated into the full feature vector

$$\Phi^{(\delta)} = \left(\Phi^{(g,\delta)} \,|\, 1 \leq g \leq p\right), \tag{13}$$

used as the input for the LASSO model. This means that the full feature dimension for an observable O as defined above would be at most

$$m_\Phi = 2R \cdot \sum_{g=1}^{p} \binom{n}{g} = 2R \cdot |\mathcal{P}_p(\Lambda)| = \mathcal{O}(n^p)\,.$$

Note that in the experiments that will be presented shortly, we won't actually meet an observable for which we would use parameters acting on different numbers of sites (due to the specific implementation of the Ising chain), and hence the last concatenation step won't be used.

For tuning the hyperparameters R, γ, and also the l_1-regularisation parameter α, appearing in the LASSO minimisation problem

$$\min_{\substack{w\in\mathbb{R}^{m_\Phi}\\ y_0\in\mathbb{R}}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left| y^{(i)} - y_0 - w \cdot [\Phi^{(\delta)}]^{(i)} \right|^2 + \alpha \cdot \|w\|_1 \right\},$$

we used 5-fold cross-validation. As per [15], the possible options for these parameters were taken to be

$$\alpha \in \left\{2^{-8}, 2^{-7}, 2^{-6}, 2^{-5}\right\},$$
$$\gamma \in \{0.4, 0.5, 0.6, 0.65, 0.7, 0.75\},$$
$$R \in \{5, 10, 20, 40\}.$$

Note that for a given model, the hyperparameters used are the same everywhere; in particular, when creating the vectors $\Phi^{(g,\delta)}$, they are the same for all I and g. The LASSO model and the cross-validation were implemented using the Python library SCIKIT-LEARN [46].

Having implemented the machine learning algorithm, we wanted to demonstrate the scaling of the number of training samples with respect to the system size. Hence, for any given set of testing samples, we've run the algorithm repeatedly with increasingly more training samples up until a fixed additive root mean squared (RMS) error was achieved. The plots of training samples needed for distinct numbers of qubits will then be used to show the scaling of this algorithm. The corresponding code for these simulations is available at [32].

4.2 Scaling of observables, self-averaging of expectation values, and the central limit theorem

In certain scenarios, for example when evaluating the performance of this algorithm as in the following sections, one needs to be careful when normalising global observables. To illustrate this point, let's consider the nearest-neighbour Heisenberg model on one-dimensional chain with open boundary conditions, as described in Section 4.3. The Hamiltonian for this model is

$$H = \sum_{\langle ij \rangle} J_{ij}(X_iX_j + Y_iY_j + Z_iZ_j),$$

and we would like to measure correlations $C_{ij} = \frac{1}{3}(X_iX_j + Y_iY_j + Z_iZ_j)$ between nearest neighbours.

When one considers these correlations individually, then for any given chain of length n with the coupling coefficients J_{ij} sampled uniformly at random from $[0, 2]$, their expectation values range throughout the whole possible interval $[-1, 0]$, being determined accurately by the local parameters J_{ij} in their neighbourhood as per the theoretical guarantees discussed in Section 2.2.

But now, let's consider the observable to be the average of these correlations,

$$O = \frac{1}{n-1} \sum_{i=0}^{n-2} C_{i,i+1} = \frac{1}{n-1} \sum_{i=0}^{n-2} \frac{1}{3}(X_iX_{i+1} + Y_iY_{i+1} + Z_iZ_{i+1}).$$

This is clearly correctly normalised to $r(O) = \mathcal{O}(1)$ as per the assumptions of Theorems 3.1, 3.7. But we know from Appendix A.3 that in the large n limit, the expectation values will concentrate around their average over the distribution of the parameters, provided that their variance tends to 0 as $\frac{1}{n}$, which we will check numerically.

Table 1 presents the data collected from simulations of Heisenberg chains of different lengths, each being based on 100 samples, when measuring the observable O as defined previously. Even though each individual correlation in the chain ranges from -1 to 0, it takes only $n = 8$ sites to get the possible range of O down to $[-0.642, -0.468]$, which continues to shrink, and gets down to only $[-0.557, -0.528]$ for $n = 128$ sites. Fitting the results for $\frac{\text{Var}}{\text{Mean}^2}$ from 16 qubits onwards with a function of the form $a + \frac{b}{n}$ yields values $a = -0.000039$ and $b = 0.024$, confirming the applicability of the central limit Theorem as given by Proposition A.13 to this case.

System size n	Range of $\text{Tr}(O\rho(x))$	Size of range	Standard deviation	$\sqrt{n}\cdot$SD
4	$[-0.718, -0.433]$	0.285	0.0662	0.1324
8	$[-0.642, -0.468]$	0.174	0.0444	0.1256
16	$[-0.605, -0.500]$	0.105	0.0206	0.0824
32	$[-0.571, -0.494]$	0.078	0.0162	0.0916
64	$[-0.563, -0.526]$	0.038	0.0079	0.0632
128	$[-0.557, -0.528]$	0.029	0.0063	0.0713

Table 1: Demonstration of self-averaging of expectation values for the average correlation $O = \frac{1}{n-1}\sum_i C_{i,i+1}$ in the Heisenberg model obtained from 100 samples for each system size. Note that the scaled version of standard deviation is actually asymptotically constant, even though it appears here to be oscillating - this is just a result of the intrinsic randomness of the systems.

This behaviour means that if one would like to predict the observable to a given absolute additive RMS error ϵ, it will be actually much easier to do so for larger systems; so much so, that for any sensible choice of the error for smaller systems, there will be a system size whose range of possible expectation values is smaller than this error, so any single sample of this size would be trivially within this error, making the sample complexity at this size be just 1. In other words, we expect that as $n \to \infty$, the expectation values y become independent of the particular parameter choices x, and so one can trivially predict y_* for a new parameter choice x_* by simply outputting the value y at a single training data point. Note that this requires using LASSO with a possibly non-zero y-intercept y_0, as otherwise predicting a constant becomes a non-trivial task.

To remedy this behaviour when demonstrating the scaling complexity of this algorithm, one can utilise few different approaches. One possibility would be based on classical shadows [33], similarly to as what was done in [18, 15], to just consider and predict each of the local terms separately, calculate the RMS error for each of them separately, and finally average over all of these errors. This approach indeed avoids the global effects and gives much more information about the whole chain, but is a priori less accurate when predicting observables such as the ground state energy, as its guaranteed prediction error is a constant ϵ, while the size of the range of possible expectation values will decrease with n below this error bound, and hence training on a single sample of the entire observable eventually becomes trivially more accurate. We used a similar approach when considering an equivariant system in Section 4.3.2, where we trained a single ML model on a single geometrically local observable, but then used it for predicting individual local observables over the whole Heisenberg chain to demonstrate a constant sample complexity.

A different approach, which we used for non-equivariant systems, and which is advantageous when predicting global observables such as the ground state energy directly, is to introduce a scaling of the observable, which is chosen such that the size of the range of sampled expectation values and their standard deviation don't scale with the system size. This would mean considering a new observable $O' \coloneqq \sqrt{n} \cdot O$, where O is a correctly normalised observable $\|O\| = \mathcal{O}(1)$ which represents some type of an average behaviour over the whole system.

The factor of $\sqrt{n}$ can be argued to be appropriate in the following way: Assuming that O is averaging over a sufficiently large disordered sample, such as the whole system, one could observe the physical phenomenon of self-averaging. Looking at the expectation values of the individual terms O_I's making up the whole observable $O = \sum O_I$, assuming them to be weakly correlated,

and hence that we can apply generalised central limit theorem as per the discussion in Appendix A.3, we would find that $y = f(O, x) = \mathrm{Tr}(O\rho(x))$ is distributed according to a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, denoted as $y \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, where $\mu = \mathcal{O}(1)$ and $\sigma = \mathcal{O}(1)$ do not scale with n. Hence, for $O' \coloneqq \sqrt{n} \cdot O$, we would get that $y' \sim \mathcal{N}(\sqrt{n}\mu, \sigma^2)$ is indeed a random variable whose standard deviation σ does not scale with n as we wanted, though its mean $\sqrt{n}\mu$ does grow with n.

For example, in the cases of the Heisenberg and Ising model considered in the following sections, when measuring the ground state energy, this scaling means that we have looked at $O' = \frac{H}{\sqrt{n}}$ instead of $O = \frac{H}{n}$. Note that this means $\|O'\| = \mathcal{O}(\sqrt{n}) > \mathcal{O}(1)$, which implies that the predicted complexity scaling isn't naïvely applicable. But we can further show that the rescaling by $\sqrt{n}$ does not affect the sample complexity for those suitable observables O which exhibit the self-averaging behaviour. Indeed, when looking at the observable $O'_\mu \coloneqq \sqrt{n}(O - \mu)$, we have that the corresponding expectation value y'_μ is distributed as $y'_\mu \sim \mathcal{N}(0, \sigma^2)$, which does not depend on n. This means that this observable is correctly normalised and hence the sample complexity scaling is indeed applicable in this case. But this observable differs from O' only by a shift of $\sqrt{n}\mu$, which for any given n represents only a constant translation of the outcomes, which does not affect the sample complexity of the machine learning model nor its prediction error, as we can simply determine the y-intercept y_0 as the average value of the outputs, and the remaining weights of the model would then be the same as for a new data set $\left\{x^{(i)}, y^{(i)} - y_0\right\}$, which has zero mean. But that makes the predicted efficient sample complexity also applicable for the case of O', even though it is not correctly normalised.

This behaviour can also be reworded in terms of dependency of the sample complexity on the error, as predicting O with error ϵ is analogous to predicting O' with error $\epsilon' = \sqrt{n}\epsilon$. But since the predicted sample complexity with error dependence given by Equation (7) is applicable to O' with error $\epsilon' = \sqrt{n}\epsilon$, this provides a significant sample complexity reduction for observables O exhibiting the self-averaging behaviour, as the dependence of their sample complexity on ϵ gets scaled by $\sqrt{n}$.

4.3 Heisenberg model

The Heisenberg model considered is given by the Hamiltonian

$$H = \sum_{\langle ij \rangle} J_{ij}(X_iX_j + Y_iY_j + Z_iZ_j),$$

where $\langle ij \rangle$ represents nearest neighbours on the lattice. Here we will use only a 1D chain with either open or periodic boundary conditions.

The coupling parameters J_{ij} were sampled uniformly at random from $[0, 2]$ as in [15]. In this section, we consider a local observable $C_{ij} = \frac{1}{3}(X_iX_j + Y_iY_j + Z_iZ_j)$; depending on the situation, either at some specific edge (i.e. C_{01}), or taking some sort of average over all of the nearest neighbours, as will be described in the separate subsections.

4.3.1 Open boundary conditions

For the following simulations, we've looked at a fixed RMS error $\epsilon = 0.55$ when predicting the ground state energy $O = \frac{H}{\sqrt{n}} = \frac{3}{\sqrt{n}} \sum_{i=0}^{n-2} J_{i,i+1}C_{i,i+1}$, where the normalisation was chosen such that the standard deviations in this regime would not scale with the system size, as was discussed in Section 4.2.

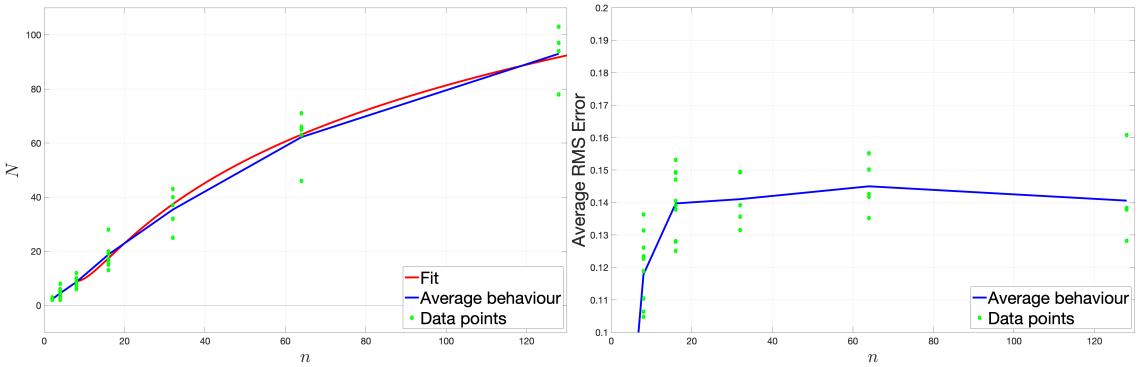
The presented data on Figure 5a, obtained for the number of training samples needed for a given system size, are illustratively fitted with a function of the form $N = a \cdot \log(n) + b + \frac{c}{n}$ (for $n \geq 8$), showing a good correspondence to the predicted asymptotic sample complexity as given by Theorem 3.1. Each shown data point corresponds to a single trial with $N_{\text{test}} = 40$ test samples, using $\delta = 4$. Note that the variance in these results is caused primarily by our testing method rather than the algorithm itself, as each set of N_{test} samples can have noticeably different standard deviation, even though the standard deviation over all parameter choices is normalised. Hence increasing N_{test} would decrease the variance of our results significantly.

4.3.2 Periodic boundary conditions

To demonstrate the absence of scaling in this case, as discussed in Section 3.2, we look into predicting each $C_{i,i+1}$ individually (where the addition is understood modulo n), but we train the model only on C_{01} and then use the same model for all i, simply by cycling through the input parameters.

Note that when training the model for C_{01} with just $\mathcal{O}(1)$ samples, the set $\mathcal{S}$ of non-trivial interactions of O contains a single element, $I = \{0, 1\}$, and so the concatenated feature vectors as defined by equations (13) and (12) are the same as the one local feature vector given by Equation (11) with $I = \{0, 1\}$.

Here we fix a number of training samples at $N = 40$, and we measure the RMS error over all of the testing samples for each $C_{i,i+1}$ individually, which we subsequently average over all i's. And given the sample complexity reduction, we would expect the RMS error to be asymptotically constant. These results are presented on Figure 5b, which indeed shows that the error did not grow between the $n = 16$ and $n = 128$ site cases. Note that saying that the error does not grow with the system size for a fixed number of training samples is tantamount to saying that the number of training samples needed to obtain a fixed error does not grow with the system size, affirming the predicted sample complexity reduction.



(a) Open boundary conditions

(b) Periodic boundary conditions

Figure 5: **Results for Heisenberg model.** Each shown data point corresponds to a single trial with $N_{\text{test}} = 40$ randomly generated test samples. **(a)** Plot of the number of training samples needed to obtain a fixed additive RMS error of $\epsilon = 0.55$ when measuring $\frac{H}{\sqrt{n}}$ in the Heisenberg chain with open boundary conditions, using $\delta = 4$. The logarithmic fit is obtained from all of the data points starting from 8 qubits. **(b)** Plot of the average RMS error for a fixed number of training samples $N = 40$ when measuring all $C_{i,i+1} = \frac{1}{3}(X_iX_{i+1} + Y_iY_{i+1} + Z_iZ_{i+1})$ individually, but training only on C_{01}, in the Heisenberg chain with periodic boundary conditions, using $\delta = 4$.

4.4 Long-range Ising model

The long-range Ising model considered here is given by the Hamiltonian

$$H = \sum_{i<j} \frac{1 + J_i J_j}{d(i,j)^\alpha} Z_i Z_j + \sum_i h_i X_i,$$

where $d(i, j)$ is the distance between sites i and j. Here we will only consider a 1D chain with open boundary conditions, so that $d(i, j) = |i - j|$. This model was studied theoretically in [47]. It can be realised experimentally, especially in the case of dipolar interactions at $\alpha = 3$ studied below, in trapped ions [48] and polar molecules [21].

For the following simulations, the parameters J_i were sampled uniformly at random from $[0, 2]$, while all the h_i's were taken to be $h_i = \mathrm{e}$ for simplicity while also maintaining a spectral gap of a reasonable size. The observable was taken to be $O = \frac{H}{\sqrt{n}}$, where the normalisation was again chosen such that the standard deviations in this regime would not scale with the system size, based on the computation of the norm in Section 2.3.1 and the discussion in Section 4.2.

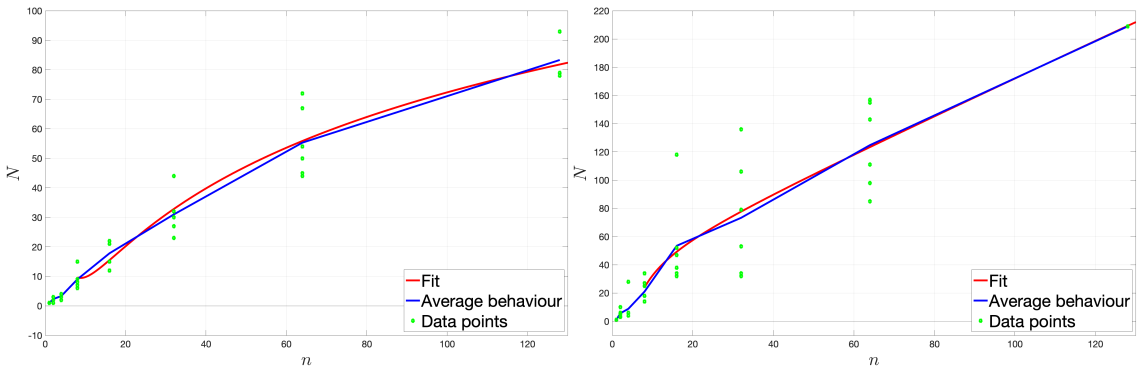
4.4.1 Open boundary conditions

For the following simulations, we have looked at a fixed RMS error $\epsilon = 0.3$. Data presented on Figure 6a are again illustratively fitted with a function of the form $N = a \cdot \log(n) + b + \frac{c}{n}$ (for $n \geq 8$) to demonstrate the expected asymptotic sample complexity of the algorithm given by Theorem 3.1. Each shown data point again corresponds to a single trial with $N_{\text{test}} = 40$ test samples, using $\delta = 4$.

4.4.2 The case of $\alpha \leq 2D$

In this section, we considered the Ising chain with power-law decay with exponent $\alpha = 1.5$. For this regime, we expect the efficient learning guarantees to not hold up as per the discussion in Appendix A.1.3.

Indeed the scaling we observe, plotted on Figure 6b, seems to be noticeably worse than in the previous section with $\alpha = 3$, which is in the correct regime $\alpha > 2D$. As the average behaviour appears to scale seemingly linearly, we illustratively fit the data with a function of the form $N = a \cdot n + b + \frac{c}{n}$. But because of high variance in the obtained data, this fit is not significantly better than the logarithmic one, even though the average behaviour seems rather linear. Further, given the finite sample size and the fact that the number of exponentials needed to fit $x^{-1.5}$ for DMRG simulations is not appreciably larger than to fit x^{-3} to the same precision at this size, we can not definitely conclude that this scaling is not logarithmic.



(a) $\alpha = 3$ (b) $\alpha = 1.5$

Figure 6: **Results for Ising model.** Each shown data point corresponds to a single trial with $N_{\text{test}} = 40$ randomly generated test samples. The fitting curves are obtained from all of the data points starting from 8 qubits. **(a)** Plot of the number of training samples needed to obtain a fixed additive RMS error of $\epsilon = 0.3$ when measuring $O = \frac{H}{\sqrt{n}}$ in the Ising chain with $\alpha = 3$, using $\delta = 4$. **(b)** Plot of the number of training samples needed to obtain a fixed additive RMS error of $\epsilon = 0.15$ when measuring $O = \frac{H}{\sqrt{n}}$ in the Ising chain with $\alpha = 1.5$, using $\delta = 4$.

4.5 Rydberg atom chain with position disorder

The Hamiltonian representing 1D chain with open boundary conditions of interacting Rydberg atoms at displaced positions $i + \delta_i$, i.e. with the metric $d(i,j) = |i + \delta_i - j - \delta_j|$, is given by

$$H = \sum_{i<j} \frac{V}{|i + \delta_i - j - \delta_j|^6} N_i N_j + \sum_i \left(\frac{\Omega}{2} X_i + \Delta N_i \right),$$

where $N = \frac{I-Z}{2} = |1\rangle\langle 1|$ is the Rydberg number operator. Such a system has been previously studied in e.g. [39]. This disordered geometry is illustrated on Figure 7.

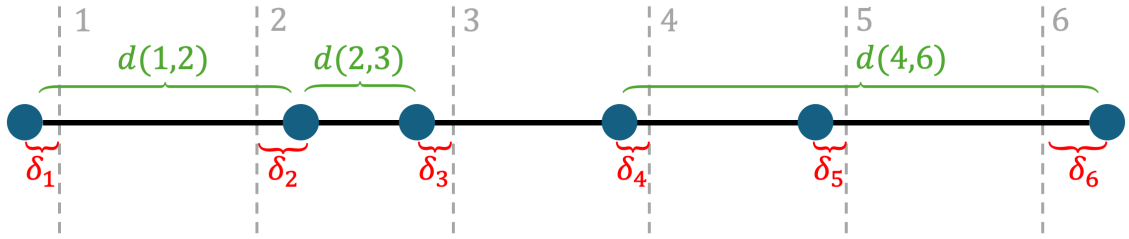


Figure 7: Illustration of the position disorder in the 1D Rydberg chain with 6 sites.

For these simulations, the atom displacements δ_i were sampled uniformly at random from $[-0.25, 0.25]$. Note that if we were to allow displacements such that $d(i,j) \to 0$, then the magnitude of the energy of such configurations wouldn't be bounded. The repulsive coupling coefficient was taken to be $V = -1$, the detuning $\Delta = 2$, and the Rabi frequency $\Omega = 1$. Data presented on Figure 8 were obtained for a fixed error $\epsilon = 2$, and are again illustratively fitted with a function of the form $N = a \cdot \log(n) + b + \frac{c}{n}$ (for $n \geq 8$). Each shown data point again corresponds to a single trial with $N_{\text{test}} = 40$ test samples, using truncation $\delta = 2$.

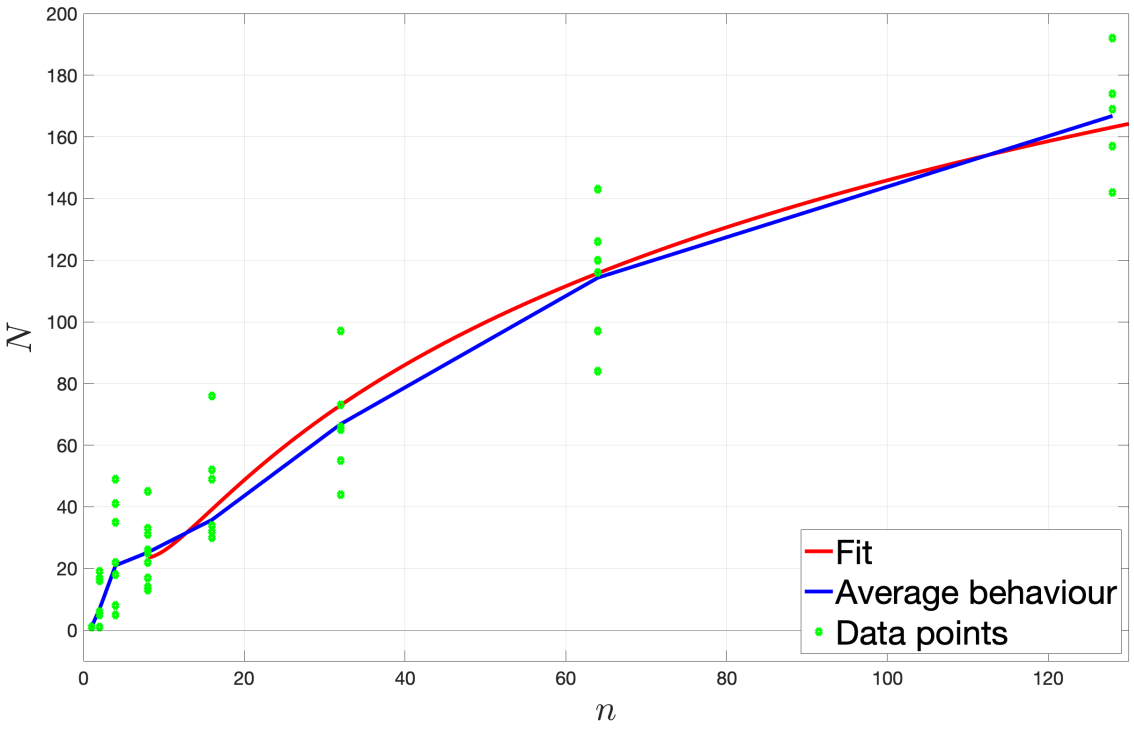


Figure 8: **Results for Rydberg model.** Each shown data point corresponds to a single trial with $N_{\text{test}} = 40$ randomly generated test samples. The fitting curves are obtained from all of the data points starting from 8 qubits. Plotting the number of training samples needed to obtain a fixed additive RMS error of $\epsilon = 2$ when measuring $O = \frac{H}{\sqrt{n}}$ in the Rydberg chain, using $\delta = 2$.

5 Conclusions and outlook

We have shown rigorous guarantees when using a classical machine learning algorithm for predicting ground state properties of Hamiltonians with long range interactions decaying at least polynomially in separation of sites, with the decay exponent α being greater than twice the dimension of the system. Albeit relevant to many systems of interest, this still leaves out important examples with smaller α's, such as the Coulomb interaction. Though our bounds are not applicable for this case, and our simulations do not suggest the efficient logarithmic scaling, it still remains unknown if one

can come up with some theoretical guarantees for these stronger interactions, and if the logarithmic scaling is achievable.

There are also still many other questions of interest when one leaves ground states of gapped Hamiltonians. Is it possible to extend this theory to thermal states? Can one get similar guarantees when working with gapless Hamiltonians, training across different topological phases? We expect that the results implying dependence of the expectation value of a p-local observable on only the parameters x close to it fail if we drop the assumption of a gapped phase. Indeed if it were true, then we could efficiently compute ground state properties of NP-hard Hamiltonians, such as the 3SAT one considered in [18, App. H.1], by reparametrising the Hamiltonian in such a way that there exists a parameter choice that would separate the Hamiltonian into k-local subsystems, allowing for efficient simulation.

In our work, we have explored the importance of self-averaging effects on the concentration of measure for the expectation values of global observables. We have observed this phenomenon for every correctly normalised global observable we have considered. As this is a property of the systems and observables themselves, it will affect any algorithm learning from those data, not only the specific one we have considered. Is this a universal property for global observables within gapped phases, or would some specific distributions of the parameters with long-range correlations not exhibit this behaviour?

A Properties of observables and expectation values

A.1 Dependency of observables on Hamiltonian parameters

A.1.1 Exponential decay

We shall use the following result:

Lemma A.1 ([34])**.** *Assume that the interactions are such that*

$$\sum_{I \ni i,j} \|h_I\| \leq \lambda \, \mathrm{e}^{-\mu d(i,j)} \tag{14}$$

for some positive constants λ, μ. Then, if $d(I,J) > 0$, we have the Lieb-Robinson bound, with $s > 0$:

$$\|[\tau_t(A_J), B_I]\| \leq 2\|A_J\| \, \|B_I\| \, |J| \, \mathrm{e}^{2s|t| - \mu d(I,J)} \, .$$

We start by noting that the assumptions of this Theorem are satisfied if the interactions have exponential decay times a polynomial of the diameter.

Lemma A.2. *Let $\nu = \mu + \epsilon, \epsilon > 0$. If*

$$\|h_I\| \leq \mathrm{e}^{-\nu \operatorname{diam}(I)} \, P(\operatorname{diam}(I)) \, ,$$

with P a positive polynomial, then (14) holds.

Proof. We first rewrite:

$$\begin{aligned}
\sum_{I \ni i,j} \|h_I\| &\leq \sum_{R \geq 1} \mathrm{e}^{-\nu R} \, P(R) \sum_{\ell=2}^{k} \widetilde{M}_\ell(ij, R) \, , \\
\widetilde{M}_\ell(ij, R) &= \sum_{i_1, \ldots, i_\ell} \delta\left(\max_{1 \leq a < b \leq \ell} (d_{ab}), R - 1 \right) \mathbf{1}((i_1, \ldots, i_\ell) \ni ij) \\
&\leq C_\ell \left(\delta(d(i,j), R-1) + [\tfrac{1}{2}\ell(\ell-1) - 1] \, \mathbf{1}(d(i,j) \leq R-1) R^{-1} \right) R^{(\ell-2)D} \, ,
\end{aligned}$$

where the inequality is proved in Lemma C.2 for some $C_\ell > 0$. Defining

$$C_{\max} = \sum_{\ell=2}^{k} C_\ell \max\left(1, \tfrac{1}{2}\ell(\ell-1) - 1\right) \, ,$$

we have

$$\sum_{\ell=2}^{k} \widetilde{M}_\ell(ij,R) \leq R^{(k-2)D} C_{\max} \left(\delta(d(i,j),R-1) + \mathbf{1}(d(i,j) \leq R-1)\right) ,$$

and get

$$\begin{aligned}\sum_{I \ni i,j} \|h_I\| &\leq \sum_{R\geq 1} \mathrm{e}^{-\nu R} P'(R) \left(\delta(d(i,j),R-1) + \mathbf{1}(d(i,j)+1 \leq R)\right) \\ &= \mathrm{e}^{-\nu(d(i,j)+1)} \left(P'(d(i,j)+1) + \sum_{R \geq 1+d(i,j)} \mathrm{e}^{-\nu(R-d(i,j)-1)} P'(R) \right) ,\end{aligned}$$

for a new positive polynomial $P'(R)$. The second sum can be performed by changing variables to $r = R-1-d(i,j)$, to give another positive polynomial $P''(d(i,j)+1)$. The Lemma follows by noting that there exists a constant C_P such that $\mathrm{e}^{-\epsilon(d(i,j)+1)} P(d(i,j)+1) \leq C_P$, so that

$$\sum_{I \ni i,j} \|h_I\| \leq \mathrm{e}^{-\mu d(i,j)} (C_{P'} + C_{P''}) .$$

□

We are going to prove the following result:

Proposition A.3. *Assume that the hypothesis of Lemma A.2 is satisfied, and that*

$$\|\nabla_I h_I\| \leq \mathrm{e}^{-\nu \operatorname{diam}(I)} ,$$

and let ν, μ, s be as in Lemmas A.2 and A.1. Then for $0 < a < \frac{1}{2s}(\mu - \log k)$, define

$$b = \gamma a , \quad \mu' = \mu - 2sa .$$

Also, assume that

$$\delta > \max(b^{-1}\xi_*, b^{-1}\eta_*)$$

with ξ_* as in Lemma 2.2 and η_* such that the function $x^{D+9} u_{2/7}(x)$ is decreasing for $x > \eta_*$ and satisfies the hypothesis of Lemma C.7 for $s = D+9$. Then, there exist positive constants c_0, c_1', c_2' so that:

$$\frac{|f(O_I,x') - f(O_I,x)|}{\|O_I\|} \leq c_0 \delta^{Dk+9} \mathrm{e}^{-\nu\delta} + c_1' \delta^{D-1} \mathrm{e}^{-\mu'\delta} + c_2' (b\delta)^{2D+20} u_{2/7}(b\delta) . \tag{15}$$

Proof. Define

$$t_* = ad(I,J) ,$$

and note that the Lieb-Robinson bound holds for $|J| \mathrm{e}^{2s|t| - \mu d(I,J)} < 1$, which holds for $|t| < t_*$ for all I, J with $d(I,J) > 0$ since we chose

$$a < \min_{I,J} \frac{1}{2s} \left(\mu - \frac{\log |J|}{d(I,J)} \right) = \frac{1}{2s}(\mu - \log k) .$$

Then we can compute $\mathcal{I}_1$ appearing in (2) as:

$$\mathcal{I}_1 \leq 2|J| \mathrm{e}^{-\mu d(I,J)} \int_0^{t_*} \mathrm{e}^{2st} \, \mathrm{d}t = s^{-1} |J| \mathrm{e}^{-\mu d(I,J)} (\mathrm{e}^{2st_*} - 1) \leq s^{-1} |J| \mathrm{e}^{-\mu' d(I,J)} .$$

We have that, denoting $R = d(I,J)$,

$$\|\nabla_J f(O_I,x)\| \leq \|\nabla_J h_J\| \, \|O_I\| \left(c_1/q_* \, \mathrm{e}^{-\mu' R} + 4G(bR) \right) , \quad c_1 = s^{-1}|J| q_* ,$$

with G as in Lemma 2.2. Next, under the assumption (all the calculations below can be easily adopted to the case where we have a decay of $\mathrm{e}^{-\nu \operatorname{diam}(I)}$ times a polynomial of $\operatorname{diam}(I)$):

$$\delta > b^{-1}\xi_* \,, \quad \|\nabla_I h_I\| \leq \mathrm{e}^{-\nu \operatorname{diam}(I)} \,,$$

we have that, with $c_2 = \frac{4}{\gamma} 130\, \mathrm{e}^2\, q_*$,

$$\begin{aligned}
&\frac{|f(O_I, x') - f(O_I, x)|}{\|O_I\|} \leq S_1 + S_2 \,, \\
&S_1 = \sum_{J \in \mathcal{P}_k(\Lambda)} \mathbf{1}\left(d(I,J) \leq \delta \wedge \operatorname{diam}(J) > \delta\right) \mathrm{e}^{-\nu \operatorname{diam}(J)} \left(c_1\, \mathrm{e}^{-\mu' d(I,J)} + 4G(bd(I,J)) \right) \,, \\
&S_2 = \sum_{J \in \mathcal{P}_k(\Lambda)} \mathbf{1}\left(d(I,J) > \delta\right) \mathrm{e}^{-\nu \operatorname{diam}(J)} \left(c_1\, \mathrm{e}^{-\mu' d(I,J)} + c_2 (bd(I,J))^{10} u_{2/7}(bd(I,J)) \right) \,.
\end{aligned}$$

Denoting $R_* = b^{-1}\xi_*$, we can rewrite these as follows:

$$\begin{aligned}
S_1 &= \sum_{r > \delta} \mathrm{e}^{-\nu r} \left(c_1 \sum_{R \leq \delta} \mathrm{e}^{-\mu' R} + \sum_{R \leq R_*} 2 \frac{K q_*}{\gamma} + c_2 \sum_{R_* < R \leq \delta} (bR)^{10} u_{2/7}(bR) \right) M(I, r, R) \,, \\
S_2 &= \sum_{R > \delta} \sum_{r \geq 1} \mathrm{e}^{-\nu r} \left(c_1\, \mathrm{e}^{-\mu' R} + c_2 (bR)^{10} u_{2/7}(bR) \right) M(I, r, R) \,, \\
M(I, r, R) &= \sum_{J \in \mathcal{P}_k(\Lambda)} \delta(r, \operatorname{diam}(J)) \delta(R, d_{IJ}) \leq C_{\mathrm{sum}} r^{(k-1)D-1} R^{D-1} \,,
\end{aligned}$$

where the bound is according to Corollary C.3.1. Putting things together we have

$$S_1 \leq \Phi \left(c_1 \sum_{R \leq \delta} R^{D-1}\, \mathrm{e}^{-\mu' R} + \sum_{R \leq R_*} 2 \frac{K q_*}{\gamma} R^{D-1} + c_2 \sum_{R_* < R \leq \delta} R^{D-1} (bR)^{10} u_{2/7}(bR) \right) \,.$$

Setting $d = D(k-1) - 1$, Φ can be upper bounded by

$$\Phi / C_{\mathrm{sum}} = \sum_{r > \delta} r^d\, \mathrm{e}^{-\nu r} = \mathrm{e}^{-\nu\delta} \sum_{y > 0} (y + \delta)^d\, \mathrm{e}^{-\nu y} = \mathrm{e}^{-\nu\delta} \sum_{\ell=0}^{d} \delta^\ell \binom{d}{\ell} \sum_{y > 0} y^{d-\ell}\, \mathrm{e}^{-\nu y} \leq C'\, \mathrm{e}^{-\nu\delta}\, \delta^d \quad (16)$$

Now we bound each term in parenthesis in S_1. We use that $\mathrm{e}^{-\mu' R}, u_{2/7}(bR) \leq 1$, to see that the term in parenthesis is bounded by

$$c_1 \delta^D + c_1' + c_2 b^{10} \delta^{D+10} \leq \max(c_1, c_1', c_2 b^{10}) \delta^{D+10} \,,$$

for a constant $c_1' > 0$. Thus we get:

$$S_1 \leq c_0 \delta^{D(k-1)-1+D+10}\, \mathrm{e}^{-\nu\delta} = c_0 \delta^{Dk+9}\, \mathrm{e}^{-\nu\delta} \,, \quad c_0 = C_{\mathrm{sum}} C' \max(c_1, c_1', c_2 b^{10}) \,.$$

Note that the case $R = 0$, which was excluded when deriving the Lieb-Robinson bound is included in the sum, and just contributes as an additive constant.

Now we get back to S_2. Defining

$$g(R) = R^{D-1} \left(c_1\, \mathrm{e}^{-\mu' R} + c_2 (bR)^{10} u_{2/7}(bR) \right) \,,$$

it is equal to

$$S_2 = C_{\mathrm{sum}} \sum_{r \geq 0} r^{D(k-1)-1}\, \mathrm{e}^{-\nu r} \sum_{R > \delta} g(R) \,.$$

The sum over r gives the constant

$$\Phi = \sum_{r \geq 1} r^{D(k-1)-1}\, \mathrm{e}^{-\nu r} \,.$$

The first sum over R can be bounded as in (16):

$$\sum_{R>\delta} R^{D-1}\,\mathrm{e}^{-\mu' r} \le C'\delta^{D-1}\,\mathrm{e}^{-\mu'\delta}\ .$$

The remaining sum is

$$\sum_{R>\delta}(bR)^{D+9}u_{2/7}(bR)\,.$$

We are going to use Lemma C.7. Since $b\delta > \eta_*$, we can bound the sum as

$$\begin{aligned}\sum_{R>\delta}(bR)^{D+9}u_{2/7}(bR) &\le \int_\delta^\infty \mathrm{d}x(bx)^{D+9}u_{2/7}(bx) = b^{-1}\int_{b\delta}^\infty \mathrm{d}\eta\eta^{D+9}u_{2/7}(\eta)\\ &\le \frac{7}{2b}(2D+21)(b\delta)^{2D+20}u_{2/7}(b\delta)\,.\end{aligned}$$

Putting things together we have

$$\begin{aligned}&S_2 \le c_1'\delta^{D-1}\,\mathrm{e}^{-\mu'\delta} + c_2'(b\delta)^{2D+20}u_{2/7}(b\delta)\,,\\ &c_1' = C_{\mathrm{sum}}\Phi C' c_1\,,\quad c_2' = C_{\mathrm{sum}}\Phi\frac{7}{2b}(2D+21)c_2\,,\end{aligned}$$

which proves the result. □

We have the following corollary:

Corollary A.3.1. *Under the same assumptions of Proposition A.3, we have*

$$|f(O_I,x') - f(O_I,x)| \le \epsilon\|O_I\|$$

if $0<\epsilon\le \mathrm{e}^{-1}$ *and*

$$\delta \ge c\log^2(\tilde{c}/\epsilon)\,,\quad \tilde{c} = 3\max(c_0,c_1',c_2')\,. \tag{17}$$

The proof follows from Lemmas C.4 and C.5 to bound each term in (15) by $\epsilon/3$.

A.1.2 Power law with $\alpha \in (2D, 2D+1)$

We start to discuss the case of power law interactions with exponent $\alpha \in (2D, 2D+1)$ and then show how the results obtained in this case can be extended to $\alpha \ge 2D+1$. We shall use the following Lieb-Robinson bound:

Lemma A.4. *Assume that*

$$H = \sum_{I\in\mathcal{P}_k(\Lambda)} h_I\,,\quad \sum_{I\ni i,j}\|h_I\| \le g d(i,j)^{-\alpha}\,, \tag{18}$$

for a positive constant g. For any $\alpha\in(2D,2D+1)$, let

$$\epsilon_* = \frac{(\alpha-2D)^2}{(\alpha-2D)^2+\alpha-D}\,,\quad \alpha' = \alpha-2D-\epsilon\,,\quad \beta = \frac{\alpha-D}{\alpha-2D}-\frac{\epsilon}{2}\,,$$

with $\epsilon\in(0,\epsilon_*)$. Then there exist constants $c, C_1, C_2 \ge 0$ s.t. for all $0\le t\le cr^{\alpha'}$ and $|I|,|J| = \mathcal{O}(1)$

$$\|[A_I(t),B_J]\| \le \|A_I\|\|B_J\| f(t,d(I,J))\,,\quad f(t,r) = C_1\left(\frac{t}{r^{\alpha'}}\right)^\beta + C_2\frac{t}{r^{\alpha-d}}\,.$$

This is proven in Corollary D.1.1 based on results in the literature. The proof relies on a conjecture that was made in [35], which applies to the case at hand of a Hamiltonian given by a sum of terms with at most k-body interactions, while only the case $k=2$ was proved in [35].

The assumptions of the Lieb-Robinson bound are satisfied if the interactions decay as in the following lemma:

Lemma A.5. *If for all I such that $|I| > 1$,*

$$\|h_I\| \leq g_{|I|} \operatorname{diam}(I)^{-\gamma(|I|)}\,, \quad \gamma(\ell) = (\ell-2)D + \alpha\,,$$

for some constant $g_\ell > 0$, then (18) holds.

Proof. Let us fix $|I| = \ell$ and consider the sum over only those sets:

$$\begin{aligned} f_\ell(i,j) &\coloneqq \sum_{I,|I|=\ell} \|h_I\|\, \mathbf{1}(i,j\in I) \leq g_\ell \sum_{I,|I|=\ell} \operatorname{diam}(I)^{-\gamma(\ell)}\, \mathbf{1}(i,j\in I) \\ &= g_\ell \sum_{R\geq 1} R^{-\gamma(\ell)} \widetilde{M}_\ell(ij,R)\,, \quad \widetilde{M}_\ell(ij,R) = \sum_{i_1,\dots,i_\ell} \delta\left(\max_{1\leq a<b\leq \ell}(d_{ab}), R-1\right) \mathbf{1}(i,j\in I)\,. \end{aligned}$$

Here we denoted $d_{ab} = d(i_a, i_b)$. From Lemma C.2, we know

$$\widetilde{M}_\ell(ij,R) \leq C_\ell \left(\delta(d(i,j),R-1) R^{(\ell-2)D} + [\tfrac{1}{2}\ell(\ell-1)-1]\, \mathbf{1}(d(i,j)\leq R-1) R^{(\ell-2)D-1}\right)\,,$$

and, denoting $C'_\ell = g_\ell C_\ell$,

$$f_\ell(i,j) = C'_\ell \sum_{R\geq 1} R^{-\gamma(\ell)+(\ell-2)D} \left(\delta(d(i,j),R-1) + [\tfrac{1}{2}\ell(\ell-1)-1]\, \mathbf{1}(d(i,j)\leq R-1) R^{-1}\right)\,.$$

Now if we plug in the value $\gamma(\ell) = (\ell-2)D + \alpha$, we get for the first term

$$\sum_{R\geq 1} R^{-\alpha} \delta(d(i,j),R-1) = (1+d(i,j))^{-\alpha}$$

and the second term

$$\sum_{R\geq 1} R^{-1-\alpha}\, \mathbf{1}(d(i,j)\leq R-1) = \sum_{R\geq 1+d(i,j)} R^{-1-\alpha} \leq \int_{d(i,j)}^{\infty} x^{-1-\alpha} \mathrm{d}x = \frac{1}{\alpha} d(i,j)^{-\alpha}\,.$$

The Lemma follows since this holds for all ℓ. □

Proposition A.6. *Let $\alpha \in (2D, 2D+1)$ and assume that the hypothesis of Lemma A.5 are satisfied and that for all I such that $|I| > 1$,*

$$\|\nabla_I h_I\| \leq \operatorname{diam}(I)^{-\gamma(|I|)}\,, \quad \gamma(\ell) = (\ell-2)D + \alpha\,.$$

With the definitions of Lemma A.4, let

$$b = ac\gamma\,, \quad \eta = \frac{1}{2}\frac{\alpha-2D}{2\alpha-2D-1}\,, \quad \epsilon = \frac{\epsilon_*}{2}\,, \quad \mu = \alpha'\eta\,, \quad \nu = \alpha'\beta - \mu(\beta+1)\,.$$

Also, assume that

$$\delta > \left(b^{-1}\max(\xi_*,\eta_*)\right)^{1/\mu}\,,$$

with ξ_* as in Lemma 2.2 and η_* is such that the function $t^{9+\frac{D}{\mu}} u_{2/7}(t)$ is monotonically decreasing for $t > \eta_*$ and the hypothesis of Lemma C.7 are satisfied. Then, there exist positive constants C_1, C_2 such that:

$$\frac{|f(O_I,x') - f(O_I,x)|}{\|O_I\|} \leq C_1 \delta^{-\nu+D} + C_2 (b\delta^\mu)^{21+2\frac{D}{\mu}} u_{2/7}(b\delta^\mu)\,.$$

Proof. We first note that by assumption we can use the Lieb-Robinson bound of Lemma A.4, and discuss the bounds on the various exponents used there. For $\alpha \in (2D, 2D+1)$ we have

$$0 < \epsilon_* < \frac{1}{D+2}$$

since ϵ_* is an increasing function of α with $\lim_{\alpha\to 2D}\epsilon_* = 0$, $\lim_{\alpha\to 2D+1}\epsilon_* = 1/(D+2)$. Next, note that α' is a decreasing function of ϵ, so we can bound its range by first using the bounds on ϵ and then on α:

$$\alpha' > \alpha'|_{\epsilon=\epsilon_*} = \alpha - 2D - \epsilon_* > \alpha - 2D > 0\,, \quad \alpha' < \alpha'|_{\epsilon=0} = \alpha - 2D < 1\,.$$

Similarly for β:

$$\beta > \beta|_{\epsilon=\epsilon_*} > \lim_{\alpha\to 2D+1}\beta|_{\epsilon=\epsilon_*} = D + 1 - \frac{1}{2(D+1)}\,, \quad \beta < \beta|_{\epsilon=0} = \frac{\alpha - D}{\alpha - 2D} < \infty\,.$$

where we used that $\beta|_{\epsilon=\epsilon_*}$ is the sum of two decreasing functions of α, so their minimum is at $2D+1$. We can do the same analysis for $\alpha'\beta$ using bounds for each:

$$\alpha'\beta > \alpha'\beta|_{\epsilon=\epsilon_*} > \lim_{\alpha\to 2D}\alpha'\beta|_{\epsilon=\epsilon_*} = D\,, \quad \alpha'\beta < \alpha'\beta|_{\epsilon=0} = \alpha - D < D+1\,,$$

where to obtain the lower bound we noticed that $\alpha'\beta|_{\epsilon=\epsilon_*}$ is an increasing function of α.

Now we evaluate the integrals $\mathcal{I}_1, \mathcal{I}_2$ of (2). The Lieb-Robinson bound holds for $|t| \le cR^{\alpha'}$, with $R = d(I,J)$. So we define

$$t_* = acR^{\alpha'\eta}\,, \quad a \le 1\,, \quad 0 \le \eta \le 1\,,$$

so that the bound holds for $|t| \le t_*$. Then,

$$\begin{aligned}\mathcal{I}_1 &\le \int_0^{t_*} \mathrm{d}t \left(C_1 t^\beta R^{-\alpha'\beta} + C_2 t R^{D-\alpha}\right) = C_1 R^{-\alpha'\beta}\frac{t_*^{\beta+1}}{\beta+1} + C_2 R^{D-\alpha}\frac{t_*^2}{2} \\ &= C_1' R^{-\alpha'\beta+(\beta+1)\alpha'\eta} + C_2' R^{D-\alpha+2\alpha'\eta}\,, \quad C_1' = C_1\frac{(ac)^{\beta+1}}{\beta+1}\,, \quad C_2' = C_2\frac{(ac)^2}{2}\,.\end{aligned}$$

Now from the bounds on $\alpha'\beta, \beta$ discussed above, we know that

$$-\alpha'\beta > D - \alpha\,, \quad \beta + 1 > D + 2 - \frac{1}{2(D+2)} > 2\,,$$

so that the first power law in $\mathcal{I}_1$ dominates the sum:

$$\mathcal{I}_1 \le CR^{-\nu}\,, \quad \nu = \alpha'(\beta - (\beta+1)\eta)\,, \quad C = C_1' + C_2'\,.$$

Next we want so show that we can choose η such that $\nu > D$ for all $\alpha \in (2D, 2D+1)$. This is required for the sum in $|f(O_I, x) - f(O_I, x')|$ to converge. We set $\epsilon = \epsilon_*/2$ to simplify the discussion. We now set

$$\eta = \frac{1}{2}\frac{\alpha - 2D}{2\alpha - 2D - 1} \in \left(0, \frac{1}{2}\frac{1}{2D+1}\right)\,.$$

The range of η can be computed by noting that it is an increasing function of α and $\lim_{\alpha\to 2D}\eta = 0$, $\lim_{\alpha\to 2D+1}\eta = 1/(2(2D+1))$. Now we claim that with this choice we have $\nu > D$. Indeed after some algebra we have that $\nu - D$ is the ratio of two polynomials of $x = \alpha - 2D$ with positive coefficients for $D = 1, 2, \ldots,$

$$\nu = D + \frac{8(D-1)D^2x + 4D(5D-4)x^2 + \left(24D^2 - 8\right)x^3 + (42D - 14)x^4 + (16D+9)x^5 + 10x^6}{16(2D - 1 + 2x)\left(D + x^2 + x\right)^2}$$

showing that $\nu - D > 0$. $\mu = \alpha'\eta$ is also an increasing function of α with range

$$0 < \mu < \left(1 - \frac{1}{2(D+2)}\right)\frac{1}{2(2D+1)}\,.$$

Thus, from Lemma 2.2 we have

$$\|\nabla_J f(O_I, x)\| \leq \|\nabla_J h_J\| \, \|O_I\| \left(c_1 R^{-\nu} + 4G(bR^\mu)\right) .$$

Note the difference with respect to the case of exponential decay or short range Hamiltonians: the exponential in their first term has been replaced by a power law with exponent ν, while in the second term the distance is shrunk with a power μ. We then have:

$$\begin{aligned}
&\frac{|f(O_I, x') - f(O_I, x)|}{\|O_I\|} \leq S_1 + S_2 \,, \\
&S_1 = \sum_{J \in \mathcal{P}_k(\Lambda)} \mathbf{1}\left(R \leq \delta \wedge \mathrm{diam}(J) > \delta\right) \|\nabla_J h_J\| \left(c_1 R^{-\nu} + 4G(bR^\mu)\right) , \\
&S_2 = \sum_{J \in \mathcal{P}_k(\Lambda)} \mathbf{1}\left(R > \delta\right) \|\nabla_J h_J\| \left(c_1 R^{-\nu} + c_2 (bR^\mu)^{10} u_{2/7}(bR^\mu)\right) .
\end{aligned}$$

Here the constants are $c_1 = q_* C$ and $c_2 = 4/\gamma \times 130\, \mathrm{e}^2\, q_*$. We can rewrite these, since $\|\nabla_J h_J\| \leq \mathrm{diam}(J)^{-\gamma(|J|)}$ and denoting $R_* = (b^{-1}\xi_*)^{1/\mu}$, as follows:

$$\begin{aligned}
S_1 &= \sum_{r>\delta} \sum_{\ell=1}^{k} r^{-\gamma(\ell)} \left(c_1 \sum_{R \leq \delta} R^{-\nu} + \sum_{R \leq R_*} 2\frac{K q_*}{\gamma} + c_2 \sum_{R_* < R \leq \delta} (bR^\mu)^{10} u_{2/7}(bR^\mu) \right) M_\ell(I, r, R) \\
S_2 &= \sum_{r>0} \sum_{R>\delta} \sum_{\ell=1}^{k} r^{-\gamma(\ell)} M_\ell(I, r, R) \left(c_1 R^{-\nu} + c_2 (bR^\mu)^{10} u_{2/7}(bR^\mu)\right) , \\
M_\ell(I, r, R) &= \sum_{J, |J|=\ell} \delta(r, \mathrm{diam}(J)) \delta(R, d_{IJ}) \,.
\end{aligned}$$

Note that the sum over r starts from 1 since the $\mathrm{diam}(I) > 0$. Now use Lemma C.3, which proves that for $\ell > 1$,

$$M_\ell(I, r, R) \leq C_\ell r^{(\ell-1)D-1} R^{D-1} \,, \quad M_1(I, r, R) \leq \delta(r, 1) C R^{D-1} \,,$$

to bound S_1. The sum over r is, using that $\delta > 1$:

$$\begin{aligned}
&\sum_{r>\delta} (C\delta(r,1) + \sum_{\ell=2}^{k} r^{-\gamma(\ell)} C_\ell r^{(\ell-1)D-1}) = \sum_{\ell=2}^{k} C_\ell \sum_{r>\delta} r^{(\ell-1)D-1-(\ell-2)D-\alpha} \\
&= \sum_{\ell=2}^{k} C_\ell \sum_{r>\delta} r^{D-1-\alpha} \leq C_{\mathrm{sum}} \delta^{D-\alpha} \,,
\end{aligned}$$

where in the bound we used that $D - \alpha - 1 < -1$ so that

$$\sum_{r>\delta} r^{D-\alpha-1} \leq \int_\delta^\infty x^{D-\alpha-1} \mathrm{d}x = \frac{1}{D-\alpha} \delta^{D-\alpha} \,.$$

Then we perform the sums over R in S_1:

$$\begin{aligned}
&\left(c_1 \sum_{R \leq \delta} R^{-\nu} + \sum_{R \leq R_*} 2\frac{K q_*}{\gamma} + c_2 \sum_{R_* < R \leq \delta} (bR^\mu)^{10} u_{2/7}(bR^\mu) \right) R^{D-1} \\
&\leq c_1 \delta^{-\nu+D} + c_1' + c_2 \sum_{1 \leq R \leq \delta} (bR^\mu)^{10} u_{2/7}(bR^\mu) \,.
\end{aligned}$$

Then we use the assumption that $b\delta^\mu > \eta_*$, so that the second term in $g(R)$ attains its maximum at a value $\tilde{R} < \delta$:

$$c_2 \sum_{1 \leq R \leq \delta} (bR^\mu)^{D-1+10\mu} u_{2/7}(bR^\mu) \leq c_2 \tilde{C} \delta \,, \quad \tilde{C} = (b\tilde{R}^\mu)^{D-1+10\mu} u_{2/7}(b\tilde{R}^\mu) \,.$$

Putting things together we can bound S_1 as follows, since $\nu > D$,

$$S_1 \leq C_{\text{sum}} \delta^{D-\alpha} (c_1 \delta^{-\nu+D} + c_1' + c_2 \tilde{C} \delta) \leq C' \delta^{D-\alpha+1}\,, \quad C' = \max(c_1, c_1', c_2 \tilde{C})\,.$$

Now we turn to S_2. Defining

$$g(R) = R^{D-1} \left(c_1 R^{-\nu} + c_2 (bR^\mu)^{10} u_{2/7}(bR^\mu) \right)\,,$$

we get

$$S_2 \leq \sum_{R>\delta} g(R) \sum_{r>0} \left(C\delta(r,1) + \sum_{\ell=2}^{k} r^{-(\ell-2)D-\alpha} C_\ell r^{(\ell-1)D-1} \right) = \sum_{R>\delta} g(R) \left(C + C_{\text{sum}} \sum_{r>0} r^{D-\alpha-1} \right)\,.$$

Using

$$\sum_{r>0} r^{D-\alpha-1} = 1 + \sum_{r>1} r^{D-\alpha-1} \leq 1 + \int_1^\infty x^{D-\alpha-1} \mathrm{d}x = 1 + \frac{1}{\alpha - D}\,,$$

we have

$$S_2 \leq C'' \sum_{R>\delta} g(R)\,.$$

To estimate S_2 we use Lemma C.7. We assumed that $b\delta^\mu > \eta_*$, with η_* is such that the function $t^{9+\frac{D}{\mu}} u_{2/7}(t)$ is monotonically decreasing for $t > \eta_*$ and the hypotheses of Lemma C.7 are satisfied. Therefore, we first bound the sum with an integral – recall also that $\nu > D$ so that the first integral converges:

$$\begin{aligned}
\sum_{R>\delta} g(R) &\leq \int_\delta^\infty \mathrm{d}R \left(c_1 R^{-\nu+D-1} + c_2 R^{D-1} (bR^\mu)^{10} u_{2/7}(bR^\mu) \right) \\
&= \frac{c_1}{\nu - D} \delta^{-\nu+D} + c_2 b^{-\frac{D-1}{\mu}} \int_\delta^\infty \mathrm{d}R (bR^\mu)^{10+\frac{D-1}{\mu}} u_{2/7}(bR^\mu)\,.
\end{aligned}$$

Then we change variables to $\eta = bR^\mu$, so that $\mathrm{d}R = (\mu b^{1/\mu})^{-1} \eta^{1/\mu - 1} \mathrm{d}\eta$ and

$$\int_\delta^\infty \mathrm{d}R (bR^\mu)^{10+\frac{D-1}{\mu}} u_{2/7}(bR^\mu) = (\mu b^{1/\mu})^{-1} \int_{b\delta^\mu}^\infty \mathrm{d}\eta\, \eta^{9+\frac{D}{\mu}} u_{2/7}(\eta)\,.$$

Finally we apply Lemma C.7 to get

$$S_2 \leq \frac{c_1 C''}{\nu - D} \delta^{-\nu+D} + C_2' (b\delta^\mu)^{21+2\frac{D}{\mu}} u_{2/7}(b\delta^\mu)\,, \quad C_2' = C''(\mu b^{1/\mu})^{-1} c_2 b^{-\frac{D-1}{\mu}}\,.$$

Putting things together, we get

$$\frac{|f(O_I, x') - f(O_I, x)|}{\|O_I\|} \leq C' \delta^{-\alpha+D+1} + \frac{c_1 C''}{\nu - D} \delta^{-\nu+D} + C_2' (b\delta^\mu)^{21+2\frac{D}{\mu}} u_{2/7}(b\delta^\mu)\,.$$

Finally we claim that

$$-\alpha + 1 \leq -\nu\,.$$

Indeed, after some algebra we have that, if $x = \alpha - 2D$:

$$\begin{aligned}
(\nu - \alpha + 1) 16(2D - 1 + 2x) \left(D + x^2 + x \right)^2 &= -16(D-1)D^2(2D-1) - (48D+7)x^5 \\
&- (2D(16D+51) - 34)x^4 - 8(D(21D-8)-1)x^3 - 4(D+2)(D(16D-13)+2)x^2 \\
&- 8D(3D-1)(5D-4)x - 22x^6\,.
\end{aligned}$$

This shows that $\nu - \alpha + 1 \leq 0$ since it is the ratio of a polynomial of x with negative coefficients for all $D = 1, 2, \ldots$ and a polynomial with positive coefficients. Then, since $\delta > 1$ we have that $\delta^{-\alpha+D+1} \leq \delta^{-\nu+D}$, and we get the result of the Lemma:

$$\frac{|f(O_I, x') - f(O_I, x)|}{\|O_I\|} \leq C_1' \delta^{-\nu+D} + C_2'(b\delta^\mu)^{21+2\frac{D}{\mu}} u_{2/7}(b\delta^\mu)\,, \quad C_1' = \max(\tfrac{c_1 C''}{\nu - D}, C')\,.$$

□

We see that compared to the case of exponentially decaying interactions of Proposition A.3, in the second term we have replaced δ with δ^μ and a different power that depends on μ. Also, the first term which was exponentially decaying is now a power law. We have the following Corollary whose proof again follows from Lemma C.5:

Corollary A.6.1. *Under the same assumptions of Proposition A.6, we have*

$$\begin{aligned} &\delta \geq \tilde{c} \max(\epsilon^{-1/(\nu-D)}, \log^{2/\mu}(3C_2/\epsilon))\,, \\ &\tilde{c} = \max((3C_1)^{1/(\nu-D)}, (c/b)^{1/\mu})\,. \end{aligned} \tag{19}$$

A.1.3 Failure of bounding the local approximation for power law decay with $\alpha \leq 2D$

We have seen in Corollary A.6.1 that $\delta \to \infty$ as $\alpha \to 2D$. In this section we study this phenomenon by using a Lieb-Robinson bound that applies for all $\alpha \in (D, 2D)$ and implies a logarithmic light cone for operator spreading. We start by recalling the following Lieb-Robinson bound.

Lemma A.7 ([49])**.** *Assume that*

$$H = \sum_{I \in \mathcal{P}_k(\Lambda)} h_I\,, \quad \sum_{I \ni i,j} \|h_I\| \leq g d(i,j)^{-\alpha}\,,$$

for a positive constant g. For any $\alpha \in (D, 2D)$, we have

$$\|[A_I, B_J]\| \leq C \|A_I\| \, \|B_J\| \, |I| \, |J| \, \frac{\mathrm{e}^{v|t|} - 1}{(1 + d(I,J))^\alpha}$$

with C, v depending only on the Hamiltonian and the metric of the lattice.

We have the following result:

Proposition A.8. *Assume the hypotheses of Lemma A.7 are satisfied. Then the bound of the quantity $|f(O_I, x) - f(O_I, x')|$ introduced in Section 2.2 becomes vacuous in this regime.*

Proof. We use the bound in Lemma A.7 to compute $\mathcal{I}_1$ appearing in (2):

$$\frac{\mathcal{I}_1}{|I||J|} \leq C \frac{1}{(1 + d(I,J))^\alpha} \int_0^{t_*} (\mathrm{e}^{vt} - 1) \leq \frac{C}{v} \frac{\mathrm{e}^{vt_*}}{(1 + d(I,J))^\alpha}\,.$$

We now take

$$t_*(R) = \frac{\beta}{v} \log(1 + R)\,,$$

where we will fix β later. If

$$\delta > \gamma^{-1} t_*^{-1}(\xi_*) = \gamma^{-1}(\mathrm{e}^{\frac{v}{\beta}\xi_*} - 1)$$

with ξ_* as in Lemma 2.2, we have that, from (3) with $R = d(I,J)$ and $b = \beta\gamma v^{-1}$:

$$\|\nabla_J f(O_I, x)\| \leq \|\nabla_J h_J(x_J)\| \, \|O_I\| (c_1 (1 + R)^{\beta - \alpha} + 4G(b \log(1 + R)))\,,$$

with G as in Lemma 2.2. As in the proofs of propositions A.3 and A.6 we can write

$$\frac{|f(O_I, x') - f(O_I, x)|}{\|O_I\|} \leq S_1 + S_2 \,,$$
$$S_1 = \sum_{J \in \mathcal{P}_k(\Lambda)} \mathbf{1}\left(r \leq \delta \wedge \operatorname{diam}(J) > \delta\right) \|\nabla_J h_J\| \left(c_1 R^{\beta-\alpha} + 4G(b\log(1+R))\right)$$
$$S_2 = \sum_{J \in \mathcal{P}_k(\Lambda)} \mathbf{1}\left(R > \delta\right) \|\nabla_J h_J\| \left(c_1 R^{\beta-\alpha} + c_2 (b\log(1+R))^{10} u_{2/7}(b\log(1+R))\right) \,.$$

We will show that $S_2 = \infty$. If $\|\nabla_J h_J\| = h(\operatorname{diam}(J))$:

$$S_2 = \sum_{r>0} h(r) \sum_{R>\delta} \left(c_1 R^{\beta-\alpha} + c_2 (b\log(1+R))^{10} u_{2/7}(b\log(1+R))\right) M(I, r, R) \,,$$
$$M(I, r, R) = \sum_{J \in \mathcal{P}_k(\Lambda)} \delta(r, \operatorname{diam}(J)) \delta(R, d_{IJ}) \,.$$

To simplify the discussion and notation we will consider below the case of Hamiltonians with terms h_J with $|J| = 2$ only. Then Lemma C.3 gives:

$$M(I, r, R) \leq C r^{D-1} R^{D-1} \,.$$

Since for $|J| = 2$ we have

$$\|\nabla_J h_J\| \leq \operatorname{diam}(J)^{-\alpha} \,,$$

defining

$$g(R) = R^{D-1} \left(c_1 R^{\beta-\alpha} + c_2 (b\log(1+R))^{10} u_{2/7}(b\log(1+R))\right) \,,$$

we get the following result by replacing the sum over r with an integral, since the summand is a decreasing function and $\alpha > D$ so $-\alpha + D - 1 < -1$ and the integral converges:

$$S_2 = \sum_{r>0} r^{-\alpha+D-1} \sum_{R>\delta} g(R) \leq C \sum_{R>\delta} g(R) \,.$$

Now we consider the sum:

$$\sum_{R>\delta} \left(c_1 R^{D-1+\beta-\alpha} + c_2 R^{D-1} (b\log(1+R))^{10} u_{2/7}(b\log(1+R))\right) \,.$$

We need $D-1+\beta-\alpha < -1$ for the first term to converge, so $0 < \beta < \alpha - D$ which can be satisfied since $\alpha - D > 0$. For the second term, we have

$$u_{2/7}(b\log(1+R)) = \exp\left(-\frac{2}{7} \frac{b\log(1+R)}{\log^2(b\log(1+R))}\right) = (1+R)^{-a(R)} \,, \quad a(R) = \frac{2}{7} b \log^{-2}(b\log(1+R))$$

For the sum to converge, we want $a(R) > D$ for large R, so, defining $z = \exp\left(\sqrt{\frac{2b}{7D}}\right)$,

$$\log^2(b\log(1+R)) < \frac{2b}{7D} \Rightarrow b\log(1+R) < z \Rightarrow R < \mathrm{e}^{z/b} - 1 \,.$$

This means that for $R > \mathrm{e}^{z/b} - 1$ the summand will decay as a non-summable power of R, leading to a divergent bound: $S_2 = \infty$. □

This shows that for $\alpha \leq 2D$ the observable $f(O_I, x)$ does not only depend on x_J with $J \in S_{I,\delta}$, but also on parameters outside this region.

A.1.4 Bound on the gradient

In Section 2.2 we bounded $\|\nabla_J \operatorname{Tr}(O_I, \rho(x))\|$ and its sum over $S^c_{\delta,I}$. It will be useful for deriving the machine learning model to have a bound on the norm of the gradient itself. We now generalize [18, Lemma 4] to exponentially decaying and power law interactions with $\alpha > 2D$:

Lemma A.9. *The gradient of* $\operatorname{Tr}(O\rho(x))$ *for* $O = \sum_I O_I$ *is bounded by*

$$\|\nabla \operatorname{Tr}(O\rho(x))\| \leq C' \sum_I \|O_I\|$$

with $C' > 0$ *a constant for both exponentially decaying interactions and power law interactions with* $\alpha > 2D$.

Proof. The computation is very similar to the one done above for $\|\nabla f(O_I, \rho(x))\|$. As in [18, Lemma 4], we use that

$$\|\nabla \operatorname{Tr}(O\rho(x))\| = \|u \cdot \nabla \operatorname{Tr}(O\rho(x))\|\,, \quad u = \frac{\nabla \operatorname{Tr}(O\rho(x))}{\|\nabla \operatorname{Tr}(O\rho(x))\|}\,,$$

and bound the quantity $\|\partial_u \operatorname{Tr}(O\rho(x))\|$ for a generic unit vector u. u has the same size as x and we denote its J-th component as u_J. Since $\partial_u = \sum_J u_J \cdot \nabla_J$, we have

$$|\partial_u f(O_I, x)| \leq \sum_J |u_J| |\nabla_J f(O_I, x)| \leq \|O_I\| \sum_J \|\nabla_J h_J\| (\mathcal{I}_1 + \mathcal{I}_2)\,,$$

where in the second inequality we used $|u_J| \leq 1$ (since u is a unit vector) and (2). We now discuss the case of exponentially decaying interactions. From the proof of Proposition A.3 we know that, with G as in Lemma 2.2,

$$\sum_J \|\nabla_J h_J\| (\mathcal{I}_1 + \mathcal{I}_2) \leq C_{\text{sum}} \sum_{R\geq 0} \sum_{r\geq 1} \mathrm{e}^{-\nu r}\, r^{(k-1)D-1} R^{D-1} \left(c_1\, \mathrm{e}^{-\mu' R} + 4G(bR) \right)\,.$$

The sum over r as well as the sum over R of the term multiplying c_1 produce a constant. We then claim that the remaining term, which is proportional to

$$\sum_{R\geq 0} G(bR)\,,$$

is also constant. This is because the sum is convergent, as used in the computation of S_2 in Proposition A.3, and so it gives a finite constant.

Similarly, we can deal with power law interactions with $\alpha > 2D$. We know from Proposition A.6 that

$$\sum_J \|\nabla_J h_J\| (\mathcal{I}_1 + \mathcal{I}_2) \leq g \sum_{R\geq 0} R^{D-1} \left(c_1 R^{-\nu} + 4G(bR^{\mu}) \right)\,,$$

$$g = C + C \sum_{\ell=2}^{k} \sum_{r>0} r^{-\gamma(\ell)+(\ell-1)D-1} = C + C(k-1) \sum_{r>0} r^{-\alpha+D-1} < \infty\,,$$

since $-\alpha + D - 1 < -1$. The term proportional to c_1 is a constant since $-\nu + D - 1 < -1$ because $\nu > D$. The second term proportional to c_2 is also a constant because the sum is convergent, as used in the computation of S_2 in Proposition A.3. □

A.2 Approximation by discretisation

It is straightforward to adapt the discretisation of the function $f(O_I, \chi_{S_{I,\delta}}(x))$ discussed in [15] to our case, as we detail next. For given $\epsilon > 0$ we define the fraction

$$\delta_2 = \left\lceil \sqrt{C'|S_{I,\delta}|}\epsilon^{-1} \right\rceil^{-1} \tag{20}$$

with C' as in Lemma A.9 and $S_{I,\delta}$ as in (1). Next we define the discretised space:

$$X_{I,\delta} = \left\{ x \in [-1,1]^m \;\middle|\; x_J = 0 \text{ if } J \notin S_{I,\delta}\,, \quad x_J \in W_{\delta_2}^{|J|} \text{ if } J \in S_{I,\delta} \right\}, \tag{21}$$
$$W_{\delta_2} = \{0, \pm\delta_2, \pm 2\delta_2, \dots, \pm 1\}\,,$$

and the space of points close to an $x \in X_{I,\delta}$:

$$T_{x,I} = \left\{ x' \in [-1,1]^m \;\middle|\; -\frac{\delta_2}{2} < x_J - x'_J < +\frac{\delta_2}{2}\,, \quad J \in S_{I,\delta} \right\}, \tag{22}$$

where for a vector x, $x > 0$ means that every component is greater than zero. We now adapt [15, Lemma 5] to our notation. The proof relies on the bound on the gradient of Lemma A.9.

Lemma A.10 (Lemma 5 of [15])**.** *Given $\epsilon > 0$ and defining the discretised version of f:*

$$g(O_I, x) = \sum_{x' \in X_{I,\delta}} f(O_I, x')\, \mathbf{1}(x \in T_{x',I})\,,$$

if we choose δ in terms of ϵ as in (17) or (19), with our choice of δ_2 in (20), we have that

$$|f(O_I, x) - g(O_I, x)| < \epsilon \|O_I\|\,.$$

We see that g has the form of a linear function of the parameters

$$(w')_{I,x'} = f(O_I, x')\,. \tag{23}$$

Then for any observable

$$O = \sum_{I \in \mathcal{S}} O_I\,, \quad \mathcal{S} \subseteq \mathcal{P}_p(\Lambda)\,, \tag{24}$$

we get

$$|f(O,x) - g(O,x)| \le \sum_{I \in \mathcal{S}} |f(O_I,x) - g(O_I,x)| \le \epsilon r(O)\,,$$

with $r(O)$ as in (4) and

$$g(O,x) = w' \cdot \phi(x) = \sum_{I \in \mathcal{S}} \sum_{x' \in X_{I,\delta}} (w')_{I,x'} \phi(x)_{I,x'}\,, \quad \phi(x)_{I,x'} = \mathbf{1}(x \in T_{x',I})\,.$$

While the true weights are unknown, we show below they can be learned efficiently. Before moving on, we note the following bounds that will be useful in the derivation of the sample complexity of the machine learning model:

Lemma A.11. *Denote by m_ϕ the size of the weight vector w' of (23) associated to the observable (24). With ν as in Proposition A.6 define*

$$\omega = \frac{kD}{\nu - D}\,, \quad \mathcal{N}(\epsilon) = \begin{cases} 2^{\mathcal{O}(\mathrm{polylog}(1/\epsilon))} & \text{exp decay,} \\ 2^{\mathcal{O}(\epsilon^{-\omega}\log(1/\epsilon))} & \text{power law } \alpha > 2D\,. \end{cases}$$

We have the following bounds:

$$m_\phi \le |\mathcal{S}| \mathcal{N}(\epsilon)\,, \quad \|w'\|_1 \le r(O)\mathcal{N}(\epsilon) \equiv B\,.$$

Proof. We note that the size of weights w and the feature map ϕ is

$$m_\phi = \sum_{I \in \mathcal{S}} |X_{I,\delta}|\,.$$

We can bound $|X_{I,\delta}|$ as

$$|X_{I,\delta}| \leq (2\delta_2^{-1}+1)^{|S_{I,\delta}|} \leq \left(2\frac{\sqrt{C'|S_{I,\delta}|}}{\epsilon}+3\right)^{|S_{I,\delta}|}$$

and

$$|S_{I,\delta}| = \sum_{J\in\mathcal{P}_k(\Lambda)} \mathbf{1}\left(d(I,J)\leq\delta \wedge \operatorname{diam}(J)\leq\delta\right) = \sum_{r,R\leq\delta} M(I,r,R)\,,$$

where M is defined in Corollary C.3.1. Then

$$|S_{I,\delta}| \leq C\sum_{r,R\leq\delta} r^{(k-1)D-1}R^{D-1} \leq C\delta^{kD}\,.$$

Note that compared to [15], we have δ^{kD} instead of δ^D because of the requirement that $\operatorname{diam}(J)\leq\delta$ which is non-trivial for general k-body interactions. Then

$$|X_{I,\delta}| \leq \left(2\sqrt{C'}\delta^{kD/2}\epsilon^{-1}+3\right)^{C\delta^{kD}} = 2^{C\delta^{kD}\log_2(2\sqrt{C'}\delta^{kD/2}\epsilon^{-1}+3)}\,.$$

To proceed, we assume exponentially decaying interactions first. Then we know from Proposition A.3 that $\delta = C_{\max}\log^2(1/\epsilon)$ for small ϵ, which implies

$$|X_{I,\delta}| \leq 2^{\mathcal{O}(\operatorname{polylog}(1/\epsilon))}\,.$$

So the number of features is

$$m_\phi \leq \sum_{I\in\mathcal{S}} 2^{\mathcal{O}(\operatorname{polylog}(1/\epsilon))} = |\mathcal{S}|2^{\mathcal{O}(\operatorname{polylog}(1/\epsilon))}\,.$$

We can also bound the ℓ_1 norm of the true weights:

$$\|w'\|_1 = \sum_{I\in\mathcal{S}}\sum_{x'\in X_{I,\delta}} |f(O_I,x')| \leq \sum_{I\in\mathcal{S}}\sum_{x'\in X_{I,\delta}} \|O_I\| \leq 2^{\mathcal{O}(\operatorname{polylog}(1/\epsilon))}r(O)\,.$$

Then we consider power law decay with $\alpha > 2D$. From Proposition A.6 we have $\delta = C_{\max}\epsilon^{-1/(\nu-D)}$ for a fixed α and small ϵ, which implies

$$|X_{I,\delta}| \leq 2^{\mathcal{O}(\epsilon^{-\omega}\log\left(\epsilon^{-1-\omega/2}\right))} = 2^{\mathcal{O}(\epsilon^{-\omega}\log(1/\epsilon))}\,,\quad \omega = \frac{kD}{\nu-D}\,.$$

So the number of features is

$$m_\phi \leq \sum_{I\in\mathcal{S}} 2^{\mathcal{O}(\epsilon^{-\omega}\log(1/\epsilon))} = |\mathcal{S}|2^{\mathcal{O}(\epsilon^{-\omega}\log(1/\epsilon))}\,.$$

and the ℓ_1 norm of the true weights:

$$\|w'\|_1 = \sum_{I\in\mathcal{S}}\sum_{x'\in X_{I,\delta}} |f(O_I,x')| \leq \sum_{I\in\mathcal{S}}\|O_I\| \sum_{x'\in X_{I,\delta}} 1 \leq r(O)2^{\mathcal{O}(\epsilon^{-\omega}\log(1/\epsilon))}\,.$$

□

Note that if we used a polynomial approximation of $f(O_I,x)$ rather than discretisation, we would have a similar growth of the number of parameters with the error ϵ, as can be derived from classical approximation theory, see e.g. [50, Theorem 4].

A.3 Limit theorem for global observables

We now recall some facts about random fields [51]. A random field is a collection of random variables $X = (X_i)_{i\in S}$, $S \subseteq \mathbb{Z}^D$. We call a random field R-dependent if for any pair $i, j \in S$ such that $d(i,j) > R$, X_i and X_j are independent: $\mathbb{E}(X_iX_j) = \mathbb{E}(X_i)\mathbb{E}(X_j)$. We then have the following central limit theorem.

Lemma A.12 (Prop. B1 in [51] for R-dependent random fields)**.** *Suppose D_n is a strictly increasing sequence of finite subsets of $\mathbb{Z}^d$, X a random field with zero mean that is R-dependent, and define $S_n = \sum_{i\in D_n} X_i$, $\sigma_n^2 = Var(S_n)$. Then $\sigma_n^{-1} S_n \to \mathcal{N}(0,1)$, the standard normal distribution, provided that*

1. *There exists $\delta > 0$ such that $\sup_i \mathbb{E}(|X_i|^{2+\delta}) < \infty$*
2. $\liminf_n |D_n|^{-1}\sigma_n^2 > 0$.

We shall apply this result to

$$X_i = \mathrm{Tr}\big(O_i\rho(\chi_{S_{i,\delta}}(x))\big) - \mu_i\,, \quad \mu_i = \mathbb{E}(\mathrm{Tr}\big(O_i\rho(\chi_{S_{i,\delta}}(x))\big))\,,$$

for $i \in \Lambda$. We recall that the random fields X_i, whose randomness is derived from the random parameters $x \in [-1,1]^m$, are functions of x_J with $d(J,i) \le \delta$ and $\mathrm{diam}(J) \le \delta$, so for any $j_a \in J$, $J \in S_{i,\delta}$:

$$d(j_a, i) \le d(j_b, i) + d(j_b, j_a) \le 2\delta\,,$$

where j_b is the closest point to i. Therefore if $d(i,j) > 4\delta$, X_i, X_j depend on different x variables and thus are independent, if we assume that the x_J's are independent. We thus have the following Proposition as a Corollary of the central limit theorem for random fields above.

Proposition A.13. *Assume the x_J's are independent random variables. Then*

$$X_i = \mathrm{Tr}\big(O_i\rho(\chi_{S_{i,\delta}}(x))\big) - \mu_i$$

are 4δ-dependent, and defining

$$S_n = \sum_{i=1}^n X_i = \sum_{i=1}^n \mathrm{Tr}\big(O_i\rho(\chi_{S_{i,\delta}}(x))\big) - \sum_{i=1}^n \mu_i\,, \quad \sigma_n^2 = Var(S_n)\,,$$

we have $\sigma_n^{-1} S_n \to \mathcal{N}(0,1)$, provided that

1. *There exists $\delta > 0$ such that $\sup_i \mathbb{E}(|X_i|^{2+\delta}) < \infty$*
2. $\liminf_n \frac{\sigma_n^2}{n} > 0$.

Note that

$$|X_i| \le |\,\mathrm{Tr}\big(O_i\rho(\chi_{S_{i,\delta}}(x))\big)| + |\mu_i| \le 2\|O_i\|\,.$$

So the first condition is

$$\mathbb{E}(|X_i|^{2+\delta}) \le (2\|O_i\|)^{2+\delta}\,,$$

and $\sup_i \mathbb{E}(|X_i|^{2+\delta}) < \infty$ as long as $\sup_i \|O_i\|$ is finite, which we assume. We can bound the variance of S_n as

$$\sigma_n^2 = \mathbb{E}\left(\sum_{i=1}^n X_i\right)^2 \le \mathbb{E}\left(\sum_{i=1}^n |X_i|\right)^2 \le 4\left(\sum_{i=1}^n \|O_i\|\right)^2\,.$$

The second condition means that $\sigma_n^2 = \sigma^2 n + o(n)$. Even if we normalise $\|O_i\|$ so that $\sum_{i=1}^n \|O_i\| = \mathcal{O}(\sqrt{n})$, and we have a bound of cn, we cannot however conclude that the second condition is

satisfied. We shall discuss this further in Section 4, where we will check it numerically for specific examples.

We note that under the assumption of Proposition A.13 for the random variable S_n/n we would get $\mathrm{Var}(S_n/n) = \sigma_n^2/n^2 \to 0$ so that S_n/n tends to the average mean and fluctuations are suppressed for large n:

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}\big(O_i\rho(\chi_{S_{\delta,i}}(x))\big) = \frac{1}{n}\sum_{i=1}^{n}\mu_i + \frac{\sigma}{\sqrt{n}}\xi + o(n^{-1/2})\,.$$

Here ξ is a random variable that depends on x and is distributed as a standard normal. If $\|O_i\| = \mathcal{O}(1)$ and want a bounded operator norm, we shall consider S_n/n, since $\|S_n/n\| \leq \sum_i \|O_i\|/n = \mathcal{O}(1)$.

Now we recall that if we choose δ as in (17) for exponential decay and (19) for power law decay with $\alpha > 2D$, we have

$$|\,\mathrm{Tr}(O_i\rho(x)) - \mathrm{Tr}\big(O_i\rho(\chi_{S_{i,\delta}}(x))\big)| \leq \epsilon\|O_i\|\,.$$

Thus,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}(O_i\rho(x)) - \frac{1}{n}\sum_i \mu_i - \frac{\sigma}{\sqrt{n}}\xi + o(n^{-1/2})\right| \leq \epsilon\frac{1}{n}\sum_{i=1}^{n}\|O_i\| \equiv \epsilon'\,.$$

So if $\|O_i\| = \mathcal{O}(1)$, we can approximate $n^{-1}\sum_{i=1}^{n}\mathrm{Tr}(O_i\rho(x))$ by the x-independent number $n^{-1}\sum_i \mu_i$ with constant error ϵ', meaning that for large n the observable concentrates around its mean.

So far we have considered observables of the type O_I with $|I| = 1$. We could not find a central limit theorem for the fields supported on sets rather than single sites in the literature, but we conjecture that the same behavior holds for the more general class of observables of the form $\sum_{I\in\mathcal{P}_p(\Lambda)} O_I$ with $p = \mathcal{O}(1)$.

B Machine learning model for classical shadows

B.1 Classical shadows and prediction of many observables

This section extends [15, Corollary 5] to our setting, and shows how to predict many observables from classical shadows data. We start by recalling that a classical shadow is an approximation of the density matrix ρ obtained by repeated random measurements [33]. For each copy of ρ we select uniformly at random whether to measure X, Y, Z for each qubit and store the associated measurement results as classical data $s_i^t \in \{0_X, 1_X, 0_Y, 1_Y, 0_Z, 1_Z\}$ for the measurement outcome of qubit i at time t. Here we denote by $0_A, 1_A$ the possible states after measurement of the Pauli A. After T measurements, the classical shadow is

$$\sigma_T(\rho) = \frac{1}{T}\sum_{t=1}^{T}\left(3\left|s_1^t\right\rangle\left\langle s_1^t\right| - \mathbf{1}_2\right)\otimes\cdots\otimes\left(3\left|s_n^t\right\rangle\left\langle s_n^t\right| - \mathbf{1}_2\right)\,.$$

Note that $\mathrm{Tr}\left(3\left|s_i^t\right\rangle\left\langle s_i^t\right| - \mathbf{1}_2\right) = 1$, so the reduced density matrix for subsystem $I = \{i_1, \ldots, i_\ell\}$ is simply

$$\mathrm{Tr}_{I^c}(\sigma_T(\rho)) = \frac{1}{T}\sum_{t=1}^{T}\left(3\left|s_{i_1}^t\right\rangle\left\langle s_{i_1}^t\right| - \mathbf{1}_2\right)\otimes\cdots\otimes\left(3\left|s_{i_\ell}^t\right\rangle\left\langle s_{i_\ell}^t\right| - \mathbf{1}_2\right)\,.$$

We are going to use the following result.

Lemma B.1 (Lemma 1 in [18])**.** *Given $\epsilon, \gamma \in (0,1)$, we have with probability at least $1-\delta$,*

$$\|\,\mathrm{Tr}_{I^c}(\rho) - \mathrm{Tr}_{I^c}(\sigma_T(\rho))\|_1 \leq \epsilon$$

for all $I \in \mathcal{P}_k(\Lambda)$, if

$$T = \mathcal{O}(k12^k\log(n/\gamma)/\epsilon^2)\,.$$

Proof. We here sketch some steps of the proof that are going to be useful below. The Lemma follows from the Bernstein matrix inequality, which gives [18]:

$$\Pr(\|\operatorname{Tr}_{I^c}(\rho) - \operatorname{Tr}_{I^c}(\sigma_T(\rho))\|_1 \geq \epsilon) \leq 2^{k+1} \exp\left(-\frac{3T\epsilon^2}{8\times 12^k}\right).$$

Now we want to bound

$$\Pr(\|\operatorname{Tr}_{I^c}(\rho) - \operatorname{Tr}_{I^c}(\sigma_T(\rho))\|_1 \leq \epsilon\,, \forall I \in \mathcal{P}_k(\Lambda)) = \Pr(\bigcap_{I\in\mathcal{P}_k(\Lambda)} \|\operatorname{Tr}_{I^c}(\rho) - \operatorname{Tr}_{I^c}(\sigma_T(\rho))\|_1 \leq \epsilon)\,,$$

and we are going to use that

$$\Pr(A\cap B) = 1 - \Pr(\overline{A\cap B}) = 1 - \Pr(\overline{A}\cup\overline{B})\,,$$

and then compute the r.h.s. with the union bound $\Pr(\cup_i A_i) \leq \sum_i \Pr(A_i)$. We have

$$\begin{aligned}\Pr\left(\bigcup_{I\in\mathcal{P}_k(\Lambda)} \|\operatorname{Tr}_{I^c}(\rho) - \operatorname{Tr}_{I^c}(\sigma_T(\rho))\|_1 \geq \epsilon\right) &\leq \sum_{I\in\mathcal{P}_k(\Lambda)} \Pr\left(\|\operatorname{Tr}_{I^c}(\rho) - \operatorname{Tr}_{I^c}(\sigma_T(\rho))\|_1 \geq \epsilon\right)\\ &\leq C_k n^k 2^{k+1} \exp\left(-\frac{3T\epsilon^2}{8\times 12^k}\right).\end{aligned}$$

Setting this equal to γ, gives the value of T in the Lemma. □

In our case, $k = \mathcal{O}(1)$ and so we can accurately predict the reduced density matrix on any subsystems of size k with only $\mathcal{O}(\log(n/\gamma)/\epsilon^2)$ randomized measurements.

The following result shows that if we have access to classical shadow data we can predict all observables of the form $\sum_{I\in\mathcal{P}_k} O_I$ with the same sample complexity of predicting a single one of Theorem 3.1.

Theorem B.2. *Suppose we have data* $\{x^{(i)}, \sigma_T(\rho(x^{(i)})\}_{i=1}^N$ *with*

$$\begin{aligned}T &= \mathcal{O}(\log(nN/(\gamma/2))/(\epsilon_2')^2)\,, \quad \epsilon_2' = \epsilon_2/C\,,\\ N &= (\epsilon_3/C^2)^{-2}\mathcal{N}(\epsilon_1/C)\mathcal{O}(\log(n/\gamma))\,,\end{aligned}$$

with $\epsilon_1, \epsilon_2, \epsilon_2, C > 0$ *and* $\gamma \in (0,1)$*. Then we can learn a predictor* $\hat{\rho}(x)$ *that achieves*

$$\mathbb{E}_{x\sim\mathcal{D}} |\operatorname{Tr}(O\rho(x)) - \operatorname{Tr}(O\hat{\rho}(x))|^2 \leq (\epsilon_1+\epsilon_2)^2 + \epsilon_3\,,$$

with probability at least $1-\gamma$ *for all observables* O *such that*

$$O = \sum_{P\in P_k} \alpha_P P\,, \quad \sum_{P\in P_k} |\alpha_P| \leq C\,,$$

where P_k *is the set of Pauli strings with weight at most* k*.*

Proof. We start to note that

$$|P_k| = \sum_{\ell=0}^k \binom{n}{\ell} 3^\ell \leq (k+1)\binom{n}{k} 3^k \leq (k+1)(3\,\mathrm{e}\,n/k)^k = \mathcal{O}(n^k)\,.$$

Then by the union bound

$$\begin{aligned}&\Pr\left(\bigcup_{i=1}^N \bigcup_{I\in\mathcal{P}_k(\Lambda)} \|\operatorname{Tr}_{I^c}(\rho(x^{(i)})) - \operatorname{Tr}_{I^c}(\sigma_T(\rho(x^{(i)})))\|_1 \geq \epsilon\right)\\ &\leq \sum_{i=1}^N \sum_{I\in\mathcal{P}_k(\Lambda)} \Pr\left(\|\operatorname{Tr}_{I^c}(\rho(x^{(i)})) - \operatorname{Tr}_{I^c}(\sigma_T(\rho(x^{(i)})))\|_1 \geq \epsilon\right)\\ &\leq NC_k n^k 2^{k+1} \exp\left(-\frac{3T\epsilon^2}{8\times 12^k}\right).\end{aligned}$$

If we set this equal to $\gamma/2$ we get that with T as in the assumptions we can achieve with probability at least $1-\gamma/2$,

$$\|\operatorname{Tr}_{I^c}(\rho(x^{(i)}) - \operatorname{Tr}_{I^c}(\sigma_T(\rho(x^{(i)})))\|_1 \leq \frac{\epsilon_2}{C}\,,$$

for all $i = 1, \ldots, N$ and subsets $I \in \mathcal{P}_k$. Then we can compute for each $P \in P_k$,

$$y^{(P)}(x^{(i)}) = \operatorname{Tr}\Big(P\sigma_T(\rho(x^{(i)}))\Big)\,, \tag{25}$$

which, with probability at least $1-\gamma/2$, satisfies, since $\|P\| = 1$,

$$|y^{(P)}(x^{(i)}) - \operatorname{Tr}\Big(P\rho(x^{(i)})\Big)| \leq \|\operatorname{Tr}_{I^c}(\rho) - \operatorname{Tr}_{I^c}(\sigma_T(\rho))\|_1 \leq \frac{\epsilon_2}{C}\,, \tag{26}$$

thus producing a dataset $\{x^{(i)}, y^{(P)}(x^{(i)})\}_{i=1}^N$ for each $P \in P_k$. Note that we have N data points for each P, but this counts as N samples only since they can all be produced by preparing N different quantum states $\rho(x_\ell)$.

Now we learn a model h_*^P for each of those P using this dataset. Called $\mathcal{D}$ the distribution over x, we know from Theorem 3.1 that if we choose

$$N = (\epsilon_3/C^2)^{-2}\mathcal{N}(\epsilon_1/C)\mathcal{O}(\log(1/\gamma'))\,, \quad \gamma' = \gamma/(2|P_k|)\,,$$

we have, denoting $\epsilon = \frac{1}{C^2}((\epsilon_1+\epsilon_2)^2 + \epsilon_3)$, that

$$\Pr\Big(\mathbb{E}_{x\sim\mathcal{D}}|h_*^P(x) - y^{(P)}(x)| \geq \epsilon\Big) \leq \gamma'\,,$$

and this holds for any $P \in P_k$, since we have used the generalisation bound for each dataset $\{x_\ell, y^{(P)}(x_\ell)\}_{\ell=1}^N$, $P \in P_k$. Then, again by using the union bound:

$$\Pr\left(\bigcup_{P\in P_k} \mathbb{E}_{x\sim\mathcal{D}}|h_*^P(x) - y^{(P)}(x)| \geq \epsilon\right) \leq \sum_{P\in P_k} \Pr\Big(\mathbb{E}_{x\sim\mathcal{D}}|h_*^P(x) - y^{(P)}(x)| \geq \epsilon\Big) \leq |P_k|\gamma' = \frac{\gamma}{2} \tag{27}$$

conditioned on (26) to occur. The probability for $\mathbb{E}_{x\sim\mathcal{D}}|h_*^P(x) - y^{(P)}(x)| \leq \epsilon$ for all $P \in P_k$ and (26) is thus at least $1-\gamma$ again by the union bound.

We can then construct a ground state representation for all observables

$$\hat{\rho}(x) = \frac{1}{2^n}\sum_{P\in P_k} h_*^P(x)P\,,$$

which attains for any O as in (10) the bound:

$$\mathbb{E}_{x\sim\mathcal{D}}|\operatorname{Tr}(O\rho(x)) - \operatorname{Tr}(O\hat{\rho}(x))|^2 = \sum_{P,P'\in P_k} |\alpha_P\|\alpha_{P'}|\mathbb{E}_{x\sim\mathcal{D}}|h_*^P(x) - \operatorname{Tr}(P\rho(x))||h_*^{P'}(x) - \operatorname{Tr}(P'\rho(x))|$$

$$\leq \sum_{P,P'\in P_k} |\alpha_P\|\alpha_{P'}|\sqrt{\mathbb{E}|h_*^P(x) - \operatorname{Tr}(P\rho(x))|^2}\sqrt{\mathbb{E}|h_*^{P'}(x) - \operatorname{Tr}(P'\rho(x))|^2}$$

$$\leq \left(\sum_{P\in P_k}|\alpha_P|\right)^2 \frac{1}{C^2}((\epsilon_1+\epsilon_2)^2+\epsilon_3) = (\epsilon_1+\epsilon_2)^2 + \epsilon_3\,.$$

In the first equality we used $\operatorname{Tr}(PP') = 2^n\delta_{P,P'}$, and in the second inequality we used Hölder's inequality. □

B.2 Equivariant classical shadows

In this section, we'll show how equivariance reduces the number of samples needed for predicting any observable of the form (10). We will repeat the derivation of Theorem 3.7 in the case we have a non-trivial automorphism group of the interaction hypergraph $\mathcal{I}$. Using the notation of Appendix B.1, we first show the following result.

Lemma B.3. *If $\mathcal{D}$ is a probability distribution over x that is invariant under $Aut(\mathcal{I})$, and $h^P_*(x)$ is an equivariant model as in Proposition 3.5, then*

$$\mathbb{E}_{x\sim\mathcal{D}}\left(|h^P_*(x)-y^P(x)|^2\right)=\mathbb{E}_{x\sim\mathcal{D}}\left(|h^{P'}_*(x)-y^{P'}(x)|^2\right)\,,\quad P'=\hat{g}P\hat{g}^{-1}\,,$$

where $y^{(P)}(x)$ is the expectation value of P in the classical shadow as in (25).

Proof. Explicitly, we have

$$\mathbb{E}_{x\sim\mathcal{D}}\left(|h^P_*(x)-y^P(x)|^2\right)=\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\mu^P(\cdot|x)}\left(|h^P_*(x)-y|^2\right)\,,$$

with

$$\mu^P=F^P\#q$$

namely the push forward of the measure of the randomized measurements

$$q(s|x)=\prod_{t=1}^{T}q_1(s^t|\rho(x))\,,\quad q_1(s|\rho)=\frac{1}{3^n}\operatorname{Tr}(|s\rangle\langle s|\,\rho)$$

under the map

$$F^P(s)=\frac{1}{T}\sum_{t=1}^{T}\operatorname{Tr}\left(P\bigotimes_{i=1}^{n}(3\,|s^t_i\rangle\langle s^t_i|-\mathbf{1}_2)\right).$$

Note that the factor 3^n in $q_1(s|\rho)$ ensures the correct normalisation

$$\sum_{s\in\mathcal{S}^n}q_1(s|\rho)=\operatorname{Tr}\left(\left(\frac{1}{3}\sum_{s\in\mathcal{S}}|s\rangle\langle s|\right)^{\otimes n}\rho\right)=\operatorname{Tr}(\rho)=1\,,\quad \mathcal{S}=\{0_X,1_X,0_Y,1_Y,0_Z,1_Z\}\,,$$

which in turns ensures the normalisation of $q(s|x)$ which is a product distribution over t. We can write μ^P explicitly as

$$\mu^P(y|x)=\sum_{s\in\mathcal{S}^{nT}}\delta(y-F^P(s))q(s|x)\,.$$

Now we show how these pieces transform under $P\to\hat{g}P\hat{g}^{-1}$. Since $\hat{g}\bigotimes_{i=1}^n|s_i\rangle=\bigotimes_{i=1}^n|s_{gi}\rangle$, we have

$$\begin{aligned}F^{\hat{g}P\hat{g}^{-1}}(s)&=\frac{1}{T}\sum_{t=1}^{T}\operatorname{Tr}\left(\hat{g}P\hat{g}^{-1}\bigotimes_{i=1}^{n}(3\,|s^t_i\rangle\langle s^t_i|-\mathbf{1}_2)\right)=\frac{1}{T}\sum_{t=1}^{T}\operatorname{Tr}\left(P\bigotimes_{i=1}^{n}(3\,|r^t_i\rangle\langle r^t_i|-\mathbf{1}_2)\right)\\&=F^P(r)\,,\quad r^t_i=(g^{-1}\cdot s)^t_i=s^t_{g^{-1}i}\,.\end{aligned}$$

Also, from (8)

$$q_1(g\cdot s|\rho(x))=\frac{1}{3^n}\operatorname{Tr}\big(\hat{g}\,|s\rangle\langle s|\,\hat{g}^{-1}\rho(x)\big)=q_1(s|\rho(g\cdot x))\,,$$

and so

$$q(g\cdot s|x)=\prod_{t=1}^{T}q_1(g\cdot s^t|\rho(x))=\prod_{t=1}^{T}q_1(s^t|\rho(g\cdot x))=q(s|g\cdot x)\,.$$

Putting these together, with $s'=g^{-1}\cdot s$ and since $g\cdot\mathcal{S}^{nT}=\mathcal{S}^{nT}$,

$$\begin{aligned}\mu^{\hat{g}P\hat{g}^{-1}}(y|x)&=\sum_{s\in\mathcal{S}^{nT}}\delta(y-F^{\hat{g}P\hat{g}^{-1}}(s))q(s|x)=\sum_{s\in\mathcal{S}^{nT}}\delta(y-F^P(g^{-1}\cdot s))q(s|x)\\&=\sum_{s'\in\mathcal{S}^{nT}}\delta(y-F^P(s'))q(g\cdot s'|x)=\sum_{s'\in\mathcal{S}^{nT}}\delta(y-F^P(s'))q(s'|g\cdot x)=\mu^P(y|g\cdot x)\,.\end{aligned}$$

We also know from Proposition 3.5 that an equivariant model has weights such that:

$$h_*^{\hat{g}P\hat{g}^{-1}}(x) = h_*^P(g \cdot x)\,.$$

Then, with $x' = g \cdot x$,

$$\begin{aligned}&\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\mu^{\hat{g}P\hat{g}^{-1}}(\cdot|x)}\left(|h_*^{\hat{g}P\hat{g}^{-1}}(x) - y|^2\right) = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\mu^P(\cdot|g\cdot x)}\left(|h_*^P(g\cdot x) - y|^2\right)\\ &= \mathbb{E}_{x'\sim\mathcal{D}'}\mathbb{E}_{y\sim\mu^P(\cdot|x')}\left(|h_*^P(x') - y|^2\right)\,, \quad \mathcal{D}' = g\#\mathcal{D}\,.\end{aligned}$$

The result follows since we assume $\mathcal{D}' = \mathcal{D}$. □

This gives the following corollary:

Corollary B.3.1. *Under the hypothesis of Theorem 3.7 and Lemma B.3, we can predict all observables of the form (10) with*

$$N = (\epsilon_3/C^2)^{-2}\mathcal{N}(\epsilon_1/C)\mathcal{O}(\log(|P_k/Aut(\mathcal{I})|/\gamma))\,.$$

Proof. If we denote $P_k' = P_k/\mathrm{Aut}(\mathcal{I})$, the proof follows that of Theorem 3.7 and by noting that we can replace (27) with

$$\Pr\left(\bigcup_{P\in P_k}\mathbb{E}_{x\sim\mathcal{D}}|h_*^P(x) - y^{(P)}(x)| \geq \epsilon\right) = \Pr\left(\bigcup_{P\in P_k'}\mathbb{E}_{x\sim\mathcal{D}}|h_*^P(x) - y^{(P)}(x)| \geq \epsilon\right) \leq |P_k'|\gamma'\,,$$

which equals $\gamma/2$ if we choose a $\gamma' = \gamma/(2|P_k'|)$ that leads to the result of the corollary. □

If $\mathcal{O}(|P_k/\mathrm{Aut}(\mathcal{I})|) = \mathcal{O}(1)$, then this reduces the complexity from $\mathcal{O}(\log(n))$ to $\mathcal{O}(1)$. This is the case for example if we consider observables that are sum of geometrically local terms (which is the setting of [15, 16, 17]) on a lattice with periodic boundary conditions, which is a typical setting for numerical experiments, so that $\mathrm{Aut}(\mathcal{I})$ constrains the D-dimensional translation group.

C Technical Lemmas

C.1 Counting Lemmas

Lemma C.1. *Let $\Lambda = \mathbb{Z}^D$ be the D-dimensional lattice. Then if $R > D$, there exists $C > 0$ such that*

$$\sum_{i\in\Lambda}\delta(d(i,j),R) \leq CR^{D-1}\,, \quad \sum_{i\in\Lambda}\mathbf{1}(d(i,j)\leq R) \leq CR^D\,, \quad \forall j\in\Lambda\,.$$

Proof. By translation invariance of the distance and Λ, we can evaluate each term at $j = 0$, and then use the exact formulas for both quantities, see e.g. [52, Prop. 31]:

$$\sum_{i\in\Lambda}\delta(d(i,0),R) = \binom{R-1, D}{D-1, 2}\,, \quad \sum_{i\in\Lambda}\mathbf{1}(d(i,0)\leq R) = \binom{R, D}{D, 2}\,,$$

with

$$\binom{m,n}{k,2} = 2^{n-k}\sum_{i=0}^{m}2^i\binom{m}{i}\binom{n}{k-i}\,.$$

So, assuming $R > D$ and recalling that $\binom{n}{k} = 0$ if $k < 0$ or $n - k < 0$, there exist $C_1, C_2 > 0$ such that:

$$\sum_{i\in\Lambda}\delta(d(i,j),R) \leq 2\sum_{i=0}^{R-1}2^i\binom{R-1}{i}\binom{D}{D-1-i} = 2\sum_{i=0}^{D-1}2^i\binom{R-1}{i}\binom{D}{D-1-i} \leq C_1(R-1)^{D-1}\,,$$

$$\sum_{i\in\Lambda}\delta(d(i,j),R) \leq 2\sum_{i=0}^{R}2^i\binom{R}{i}\binom{D}{D-i} = 2\sum_{i=0}^{D}2^i\binom{R}{i}\binom{D}{D-i} \leq C_2(R-1)^D\,.$$

The Lemma follows by taking $C = \max(C_1, C_2)$ and noting that $(R-1)^d < R^d$ for $R, d > 0$. □

Lemma C.2. *We have*

$$\widetilde{M}_\ell(ij,R) = \sum_{i_1,\dots,i_\ell} \delta\left(\max_{1\le a<b\le \ell}(d_{ab}), R-1\right)\mathbf{1}(ij\in(i_1,\dots,i_\ell))$$
$$\le C_\ell\left(\delta(d(i,j),R-1) + [\tfrac{1}{2}\ell(\ell-1)-1]\,\mathbf{1}(d(i,j)\le R-1)R^{-1}\right)R^{(\ell-2)D}\,,$$

where we denoted $d_{ab} = d(i_a,i_b)$.

Proof. Note that the summand in $\widetilde{M}_\ell(ij,R)$ is symmetric in $i_1,\dots,i_\ell$. Writing $\mathbf{1}(ij\in(i_1,\dots,i_\ell))$ as a sum where ij can be any of the $\ell(\ell-1)/2$ pairs, we then have that each summand is the same, and we can write:

$$\widetilde{M}_\ell(ij,R) = \tfrac{1}{2}\ell(\ell-1)\sum_{i_3,\dots,i_\ell}\delta\left(\max_{1\le a<b\le\ell}(d_{ab}),R-1\right)\,,\quad i_1=i,i_2=j\,.$$

We are further going to write this as a sum over the $\ell(\ell-1)/2$ pairs where the a,b term corresponds to the case when ther maximum is attained at i_a,i_b:

$$\widetilde{M}_\ell(ij,R) = \tfrac{1}{2}\ell(\ell-1)\sum_{1\le a<b\le\ell} M_{\ell,ab}(ij,R)\,,$$
$$\widetilde{M}_{\ell,ab}(ij,R) = \sum_{i_3,\dots,i_\ell}\delta(d_{ab},R-1)\prod_{\substack{1\le c<d\le\ell\\ cd\neq ab}}\mathbf{1}(d_{cd}\le R-1)\,.$$

Let us start by looking at $a,b=1,2$:

$$\widetilde{M}_{\ell,12}(ij,R)\le\delta(d_{12},R-1)\sum_{i_3,\dots,i_\ell}\prod_{2\le c<d\le\ell}\mathbf{1}(d_{cd}\le R-1)$$
$$=\delta(d_{12},R-1)\sum_{i_3,\dots,i_\ell}\prod_{b=3}^{\ell}\mathbf{1}(d_{2,b}\le R-1)\prod_{b=4}^{\ell}\mathbf{1}(d_{3,b}\le R-1)\cdots\mathbf{1}(d_{\ell-1,\ell}\le R-1)\,.$$

In the first inequality we have used $\mathbf{1}(d_{1b}\le R-1)\le 1$. Then we can perform each summation in the order $i_\ell,i_{\ell-1},i_{\ell-2},\dots,i_3$. At step i_ℓ

$$\sum_{i_\ell}\prod_{a=2}^{\ell-1}\mathbf{1}(d_{a,\ell}\le R-1)\le\sum_{i_\ell}\mathbf{1}(d_{\ell-1,\ell}\le R-1)\le CR^D\,,$$

since we can use the bound $\mathbf{1}(d_{a-1,b}\le R-1)\le 1$ and we used the result from Lemma C.1. Then, similarly, for $b=\ell-1,\dots,3$

$$\sum_{i_b}\prod_{a=2}^{b-1}\mathbf{1}(d_{a,b}\le R-1)\le\sum_{i_b}\mathbf{1}(d_{b-1,b}\le R-1)\le CR^D\,. \tag{28}$$

Thus we have shown

$$\widetilde{M}_{\ell,12}(ij,R)\le\delta(d_{12},R-1)(CR^D)^{\ell-2}\,.$$

Now we note that the sum in $\widetilde{M}_\ell(ij,R)$ is symmetric under permutations of $i_3,\dots,i_\ell$. So if $b>a>2,d>c>2$, then $\widetilde{M}_{\ell,ab}=\widetilde{M}_{\ell,cd}$. In fact, we can simply relabel the dummy indices to show this equivalence. So to deal with the cases $b>a>2$, we can set $a=\ell-1,b=\ell$:

$$\widetilde{M}_{\ell,ab}(ij,R)\le\sum_{i_3,\dots,i_\ell}\delta(d_{\ell-1,\ell},R-1)\,\mathbf{1}(d_{12}\le R-1)\prod_{d=2}^{\ell}\mathbf{1}(d_{1d}\le R-1)\prod_{c=2}^{\ell-2}\prod_{d=c+1}^{\ell}\mathbf{1}(d_{cd}\le R-1)$$

Now we perform the sum over i_ℓ:

$$\sum_{i_\ell}\prod_{a=1}^{\ell-2}\mathbf{1}(d_{a,\ell}\le R-1)\delta(d_{\ell-1,\ell},R-1)\le\sum_{i_\ell}\delta(d_{\ell-1,\ell},R-1)\le CR^{D-1}\,,$$

and for $i_{\ell-1}, i_{\ell-2}, \dots, i_3$ we have like in (28). Thus if $b > a > 2$,

$$\widetilde{M}_{\ell,ab}(ij, R) \leq \mathbf{1}(d_{12} \leq R-1) C^{\ell-2} R^{(\ell-2)D-1} .$$

Then, assume that $a = 1, b > 2$. Again, any $b > 2$ will give the same result due to symmetry and we choose $b = \ell$. Then

$$\widetilde{M}_{\ell,1b}(ij, R) \leq \sum_{i_3,\dots,i_\ell} \delta(d_{1,\ell}, R-1)\, \mathbf{1}(d_{12} \leq R-1) \prod_{d=2}^{\ell-1} \mathbf{1}(d_{1d} \leq R-1) \prod_{2\leq c<d\leq \ell} \mathbf{1}(d_{cd} \leq R-1) .$$

Again, we perform the sum over i_ℓ first. We have

$$\sum_{i_\ell} \delta(d_{1,\ell}, R-1) \prod_{a=2}^{\ell-1} \mathbf{1}(d_{a,\ell} \leq R-1) \leq \sum_{i_\ell} \delta(d_{1,\ell}, R-1) \leq C R^{D-1} ,$$

and the other sums follow like in (28). Then for $b > 2$:

$$\widetilde{M}_{\ell,1b}(ij, R) \leq \mathbf{1}(d_{12} \leq R-1) C^{\ell-2} R^{(\ell-2)D-1} .$$

Finally, since i_1, i_2 are symmetric, the same result holds for $a = 2, b > a$. Putting everything together, we have

$$\widetilde{M}_\ell(ij, R) = C_\ell \left(\delta(d_{12}, R-1) R^{(\ell-2)D} + [\tfrac{1}{2}\ell(\ell-1) - 1]\, \mathbf{1}(d_{12} \leq R-1) R^{(\ell-2)D-1} \right) ,$$

where we denoted $C_\ell = \frac{1}{2}\ell(\ell-1)C^{\ell-2}$. □

Lemma C.3. *For $\ell > 1$, there exists a positive constant C_ℓ such that*

$$M_\ell(I, r, R) = \sum_{J, |J|=\ell} \delta(r, \mathrm{diam}(J)) \delta(R, d_{IJ}) \leq C_\ell r^{(\ell-1)D-1} R^{D-1} .$$

Proof. We want to compute, defining $r' = r - 1$ and $d_{ab} \equiv d(j_a, j_b)$, the value of

$$M_\ell(I, r, R) = \sum_{j_1,\dots,j_\ell} \delta\left(r', \max_{1\leq a<b\leq \ell}(d_{ab})\right) \delta\left(R, \min_{i\in I, j\in J} d(i,j)\right) .$$

First of all, we note that

$$\delta\left(R, \min_{i\in I, j\in J} d(i,j)\right) \leq \sum_{a=1}^{|I|} \sum_{b=1}^{|J|} \delta(R, d(i_a, j_b)) .$$

This is because if the δ on the l.h.s. is 1, then the sum on the r.h.s. must be ≥ 1. Then

$$M_\ell(I, r, R) \leq \sum_{j_1,\dots,j_\ell} \sum_{c=1}^{|I|} \sum_{d=1}^{|J|} \delta\left(r', \max_{1\leq a<b\leq \ell}(d_{ab})\right) \delta(R, d(i_c, j_d)) .$$

Now note that the function we are summing over is invariant under permutation of the j's. Therefore, each choice of j_d will give the same result and we can set $d = 1$:

$$M_\ell(I, r, R) \leq \ell \sum_{c=1}^{|I|} \sum_{j_1,\dots,j_\ell} \delta\left(r', \max_{1\leq a<b\leq \ell}(d_{ab})\right) \delta(R, d(i_c, j_1)) .$$

Then we consider each of the cases where the maximum is attained at each of the $\frac{1}{2}\ell(\ell-1)$ pairs j_a, j_b:

$$M_\ell(I, r, R) \leq \ell \sum_{c=1}^{|I|} \sum_{1\leq a<b\leq \ell} M_\ell(a, b, c, r', R)$$

$$M_\ell(a, b, c, r', R) = \sum_{j_1,\dots,j_\ell} \delta(r', d_{ab}) \prod_{\substack{1\leq e<f\leq \ell \\ ef\neq ab}} \mathbf{1}(d_{ef} \leq r') \delta(R, d(i_c, j_1)) .$$

We consider first $a, b = 1, 2$:

$$M_\ell(1,2,c,r',R) = \sum_{j_1,\dots,j_\ell} \delta(r', d_{12}) \prod_{f=3}^{\ell} \mathbf{1}(d_{1f} \le r') \prod_{2\le e<f\le \ell} \mathbf{1}(d_{ef} \le r') \delta(R, d(i_c, j_1)) \,.$$

We now perform the sum over j_2. Using $\mathbf{1}(\cdot) \le 1$ and Lemma C.1:

$$\sum_{j_2} \delta(r', d_{12}) \prod_{f=3}^{\ell} \mathbf{1}(d_{2f} \le r') \le \sum_{j_2} \delta(r', d_{12}) \le C(r')^{D-1} \,.$$

Now summing sequentially over $j_3, j_4, \dots, j_\ell$, gives at a-th step:

$$\sum_{j_a} \prod_{f=a+1}^{\ell} \mathbf{1}(d_{af} \le r') \le \sum_{j_a} \mathbf{1}(d_{a,a+1} \le r') \le C(r')^{D} \,.$$

Finally, we sum over j_1:

$$\sum_{j_1} \prod_{f=3}^{\ell} \mathbf{1}(d_{1f} \le r') \delta(R, d(i_c, j_1)) \le \sum_{j_1} \delta(R, d(i_c, j_1)) \le C R^{D-1} \,,$$

which gives

$$M_\ell(1,2,c,r',R) \le C^\ell R^{D-1} (r')^{(\ell-1)D-1} \,.$$

Now note that $M_\ell(1,2,c,r',R) = M_\ell(1,b,c,r',R)$ for any $b \ge 2$ since the summand is symmetric under exchanging j_b and j_2. Then consider $M_\ell(a,b,c,r',R)$, $b > a > 1$. Again by symmetry these are all equal since we can relabel any a, b with $b > a > 1$ as $\ell - 1, \ell$, and we can consider for definiteness

$$M_\ell(\ell-1,\ell,c,r',R) = \sum_{j_1,\dots,j_\ell} \delta(r', d_{\ell-1,\ell}) \prod_{e=1}^{\ell-2} \prod_{f=e+1}^{\ell} \mathbf{1}(d_{ef} \le r') \delta(R, d(i_c, j_1)) \,.$$

We start with summing over $j_{\ell-1}$:

$$\sum_{j_{\ell-1}} \delta(r', d_{\ell-1,\ell}) \prod_{e=1}^{\ell-2} \mathbf{1}(d_{e,\ell-1} \le r') \le \sum_{j_{\ell-1}} \delta(r', d_{\ell-1,\ell}) \le C(r')^{D-1} \,.$$

Now, any other sum apart from j_1 can be bounded by $C(r')^D$ as done before. The sum over j_1 gives:

$$\sum_{j_1} \prod_{f=2}^{\ell} \mathbf{1}(d_{1f} \le r') \delta(R, d(i_c, j_1)) \le \sum_{j_1} \delta(R, d(i_c, j_1)) \le C R^{D-1} \,,$$

so that

$$M_\ell(\ell-1,\ell,c,r',R) \le C^\ell R^{D-1} (r')^{(\ell-1)D-1} \,,$$

and thus

$$M_\ell(I,r,R) \le \ell C^\ell |I| \tfrac{1}{2} \ell(\ell-1) R^{D-1} (r')^{(\ell-1)D-1}$$

□

We have the following immediate consequence:

Corollary C.3.1. *We have*

$$M(I,r,R) = \sum_{J \in \mathcal{P}_k(\Lambda)} \delta(r, \mathrm{diam}(J)) \delta(R, d_{IJ}) = \sum_{\ell=1}^{k} M_\ell(I,r,R) \le C_{sum} r^{(k-1)D-1} R^{D-1} \,.$$

C.2 Solutions to inequalities

Lemma C.4 (Lemma 6 in [15])**.** *Given $a > 0, s > 1$, $0 < \epsilon < \mathrm{e}^{-1}$, there exists a constant $c > 0$ such that for all $y \geq c \log^2(1/\epsilon)$,*

$$y^s \, \mathrm{e}^{-ay} \leq \epsilon \, .$$

Proof. We can rewrite the condition we want to prove taking $-\log$ of both sides as

$$ay - s \log(y) \geq \log(1/\epsilon) \, .$$

We can then apply [15, Lemma 6] to show that the Lemma holds with

$$c = \frac{(2s + \sqrt{4s^2 + a})^2}{4a^2} \, .$$

□

Lemma C.5 (Adapting Lemma 7 in [15])**.** *Given $a, s > 0$, $0 < \epsilon \leq \mathrm{e}^{-1}$, there exists a constant $c > 0$ such that for all $y \geq c \log^2(1/\epsilon)$,*

$$y^s u_a(y) \leq \epsilon \, .$$

Proof. We assume $y \geq 0$. We can rewrite the condition to prove as

$$f(y) \coloneqq a \frac{y}{\log^2(y)} - s \log(y) \geq x \, , \quad x = \log(1/\epsilon) \, .$$

We compute

$$f'(y) = \frac{a}{\log^3(y)} (\log(y) - 2) - \frac{s}{y} \, ,$$

and find y_0 such that $f'(y) \geq 0$ for $y \geq y_0$. We use $\log(z) \leq 3z^{1/3}$, or $1/\log^3(z) \geq 1/(3^3 z)$ for $z > 0$ to get

$$f'(y) \geq \frac{1}{y}(a3^{-3}(\log(y) - 2) - s)$$

so we need $a3^{-3}(\log(y) - 2) - s \geq 0$ or $\log(y) - 2 - \frac{s}{a}3^3 \geq 0$, namely

$$y \geq y_0 \equiv \exp\Big(2 + \frac{s}{a}3^3\Big) \, .$$

We note that numerically we can see that this bound is quite loose but it will serve our purposes here. So for $y \geq y_0$, f is monotonically increasing. Next we will show that for $x > 1$ ($\epsilon < 1/\mathrm{e}$), there is a $c \geq y_0$ such that $f(cx^2) \geq x$. This will imply that $f(y) \geq x$ for all $y \geq cx^2$ due to the monotonicity of f for $cx^2 \geq y_0$ that is ensured for $x \geq 1$ as long as $c \geq y_0$. To show we can get the desired bound:

$$f(cx^2) = a \frac{cx^2}{\log^2(cx^2)} - s \log\big(cx^2\big) \geq x \, ,$$

we will show the two inequalities

$$\frac{a}{2} \frac{cx^2}{\log^2(cx^2)} \geq x \, , \quad \frac{a}{2} \frac{cx^2}{\log^2(cx^2)} \geq s \log\big(cx^2\big) \, .$$

For the first, we have, using $\log(z) \leq 4z^{1/4}$ for $z \geq 0$,

$$\frac{a}{2} \frac{cx^2}{\log^2(cx^2)} \geq \frac{a}{32} \frac{cx^2}{\sqrt{cx^2}} = \frac{a}{32} \sqrt{c} x \geq x \quad \text{if } c \geq \frac{32^2}{a^2} \, .$$

For the second, since $x \geq 1$, $acx^2 \geq acx^{3/2}$. Using again $\log(z) \leq 4z^{1/4}$ with $z = cx^2$, we have

$$acx^2 \geq ac^{1/4}((cx^2)^{1/4})^3 \geq \frac{ac^{1/4}}{4^3} \log^3(cx^2)\,,$$

so

$$\frac{a}{2}\frac{cx^2}{\log^2(cx^2)} \geq \frac{ac^{1/4}}{2\times 4^3}\log\bigl(cx^2\bigr)\,,$$

and this is greater or equal to $s\log\bigl(cx^2\bigr)$ if

$$\frac{ac^{1/4}}{2\times 4^3} \geq s \Rightarrow c \geq 2^4 4^{12}\frac{s^4}{a^4}\,.$$

Finally, under the assumption

$$c \geq \max\left(y_0, \frac{32^2}{a^2}, 2^4 4^{12}\frac{s^4}{a^4}\right)\,,$$

we can get the desired bound. □

C.3 Integral bounds

We use

Lemma C.6 ([53])**.** *For any $a \geq 2, x > 0$:*

$$\Gamma(a,x) \leq \frac{(x+b_a)^a - x^a}{ab_a}\,\mathrm{e}^{-x}\,, \quad b_a = \Gamma(a+1)^{1/(a-1)}\,.$$

We can simplify this bound for $x \geq 1$, so that $(1+b_a)x = x + b_a x \geq x + b_a$, as follows

$$\Gamma(a,x) \leq \frac{(1+b_a)^a - 1}{ab_a}x^a\,\mathrm{e}^{-x}\,. \tag{29}$$

We are going to use this bound on the incomplete Gamma function to prove an extension of Lemma 2.5 [40] to non-integral powers s.

Lemma C.7. *For all $t > 1, s > 3$ such that $\log^4(t) \leq t$ and*

$$\tau(t) \coloneqq a\frac{t}{\log^2(t)} \geq b_s = \Gamma(s+1)^{1/(s-1)}\,,$$

there exists a t-independent positive number C such that:

$$F(s,t) = \int_t^\infty \mathrm{d}\eta\,\eta^s u_a(\eta) \leq Ct^{2s+3}u_a(t)\,.$$

Proof. We have

$$\frac{\mathrm{d}\eta}{\mathrm{d}\tau(\eta)} = \left(a\frac{1}{\log^2(\eta)} - 2a\frac{1}{\log^3(\eta)}\right)^{-1} = \frac{1}{a}\frac{\log^2(\eta)}{1-\frac{2}{\log(\eta)}} \leq \frac{\eta}{a}\,,$$

and from assumption, $\log^4(\eta) \leq \eta$, so

$$\eta \leq \frac{\eta^2}{\log^4(\eta)} = \frac{\tau^2}{a^2}\,.$$

Then we change the variable from η to τ:

$$F(s,t) \leq \frac{1}{a^{2s+3}}\int_{\tau(t)}^\infty \mathrm{d}\tau\tau^{2(s+1)}\,\mathrm{e}^{-\tau} = \frac{1}{a^{2s+3}}\Gamma(2s+3,\tau(t))\,.$$

Now we use the bound (29) to get

$$F(s,t) \leq C'\tau(t)^{2s+3}u_a(t) \leq a^{2s+3}C't^{2s+3}u_a(t)\,, \quad C' = \frac{(1+b_{2s+3})^{2s+3}-1}{(2s+3)b_{2s+3}}\,.$$

□

D Lieb-Robinson bounds for power law interactions

The following result covers the case of two body interactions.

Theorem D.1 (Thm 1, S1 [35])**.** *Assume that $H = \sum_{ij} H_{ij}$ with $\|H_{ij}\| \leq 1/d(i,j)^\alpha$. Then let O be a norm 1 operator supported at the origin, and define $\mathbb{P}_r O$ the projection of O onto sites that are at least of distance r from the origin. For any $\alpha \in (2D, 2D+1)$, $\epsilon \in (0, \epsilon_*)$, $\epsilon_* = \frac{(\alpha-2D)^2}{(\alpha-2D)^2+\alpha-D}$, there exist constants $c, C_1, C_2 \geq 0$ s.t.*

$$\|\mathbb{P}_r O(t)\| \leq f(t,r)\,, \quad f(t,r) = C_1 \left(\frac{t}{r^{\alpha-2D-\epsilon}}\right)^{\frac{\alpha-D}{\alpha-2D}-\frac{\epsilon}{2}} + C_2 \frac{t}{r^{\alpha-D}}\,,$$

for all $0 \leq t \leq c r^{\alpha-2D-\epsilon}$.

In [35] this result was also conjectured to be valid for the more general class of Hamiltonians

$$H = \sum_{I \in \mathcal{P}(\Lambda)} h_I\,, \quad \sum_{I \ni i,j} \|h_I\| \leq g d(i,j)^{-\alpha}\,,$$

for a constant g, and this is the form that we are going to use. We derive the following corollary.

Corollary D.1.1. *For any $\alpha \in (2D, 2D+1)$, $\epsilon \in (0, \epsilon_*)$, $\epsilon_* = \frac{(\alpha-2D)^2}{(\alpha-2D)^2+\alpha-D}$, there exist constants $c, C_1, C_2 \geq 0$ s.t. for all $0 \leq t \leq c r^{\alpha-2D-\epsilon}$ and $|I|, |J| = \mathcal{O}(1)$*

$$\|[O_I(t), O_J]\| \leq \|O_I\|\,\|O_J\|\, f(t, d(I,J))\,.$$

Proof. We assume unit norm operators since for general operators we can apply the result to the normalized operator. For a set I, let P_I be the projector onto the non-trivial part of a Pauli string on I, i.e. it is zero unless there is a non-trivial operator in I:

$$P_I \sigma_1^{a_1} \cdots \sigma_n^{a_n} = \mathbf{1}(\exists j \in I \,|\, a_j \neq 0)\sigma_1^{a_1} \cdots \sigma_n^{a_n} = \mathbf{1}\Big(\bigvee_{j \in I}(a_j \neq 0)\Big)\sigma_1^{a_1} \cdots \sigma_n^{a_n}\,.$$

Then define $\mathbb{P}_r^I$ to be the projector onto the set $B_r^I = \{j \,|\, d(I,j) \geq r\}$. $\mathbb{P}_r$ of the Theorem D.1 coincides with $\mathbb{P}_r^0$. Since $B_r^I = \bigcap_{i \in I} B_r^i$ and $I(\vee_{i \in I \cap J}(a_i \neq 0)) \leq I(\vee_{i \in I}(a_i \neq 0)) + I(\vee_{i \in J}(a_i \neq 0))$ because there are more configurations of the a_i's that make the r.h.s. non-zero, we have

$$\mathbb{P}_r^I \leq \sum_{i \in I} \mathbb{P}_r^i\,.$$

Then note that $(\mathbf{1} - \mathbb{P}_r^I)O$ for any operator O is supported on sites of distance smaller than r to I. Then if $r = d(I,J)$, we have that $[(\mathbf{1} - \mathbb{P}_r^I)O, O_J] = 0$ for any O since the support of the two operators in the commutator has no overlap. Then, with $r = d(I,J)$, we have

$$\begin{aligned}\|[O_I(t), O_J]\| &= \|[\mathbb{P}_r^I O_I(t) + (\mathbf{1} - \mathbb{P}_r^I)O_I(t), O_J]\| = \|[\mathbb{P}_r^I O_I(t), O_J]\| \leq \sum_{i \in I} \|[\mathbb{P}_r^i O_I(t), O_J]\| \\ &\leq 2 \sum_{i \in I} \|\mathbb{P}_r^i O_I(t)\|\,.\end{aligned}$$

Then we use formula [54, Lemma 4]: if $\|P_J O_i(t)\| \leq f(t, d(i,J))$ for all unit norm operators O_i with $i \in I$, then there exists a positive constant K such that

$$\|P_J O_I(t)\| \leq K \sum_{i \in I} f(t, d(i,J))\,.$$

From Theorem D.1 we know

$$\|\mathbb{P}_r^i O_i(t)\| \leq f(t,r)\,,$$

so we have that

$$\|\mathbb{P}_r^i O_I(t)\| \leq K|I| f(t,r)\,,$$

and putting things together:

$$\|[O_I(t), O_J]\| \leq 2K\,|I|^2 f(t,r)\,.$$

□

E Implementation of Matrix Product Operators

E.1 Heisenberg model

Implementing the nearest neighbour Heisenberg model

$$H = \sum_{\langle ij \rangle} J_{ij}(X_i X_j + Y_i Y_j + Z_i Z_j)$$

with open boundary conditions is straightforward. Following the discussion in Appendix E.4, the Hamiltonian on a periodic chain may be represented as an MPO in the following way:

$$H = W_L \cdot \prod_{i=1}^{n-2} W[i] \cdot W_R,$$

where the grid matrices are given by

$$W[i] = \begin{pmatrix} I & 0 & 0 & 0 & J_{i,i+1}X_i & J_{i,i+1}Y_i & J_{i,i+1}Z_i & 0 \\ 0 & I & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & X_i \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & Y_i \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & Z_i \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}$$

together with

$$W_L = (I \quad J_{0,n-1}X_0 \quad J_{0,n-1}Y_0 \quad J_{0,n-1}Z_0 \quad J_{01}X_0 \quad J_{01}Y_0 \quad J_{01}Z_0 \quad 0),$$

$$W_R = (0 \quad X_{n-1} \quad Y_{n-1} \quad Z_{n-1} \quad X_{n-1} \quad Y_{n-1} \quad Z_{n-1} \quad I)^T.$$

E.2 Ising model

The Hamiltonian

$$H = \sum_{i<j} \frac{1 + J_i J_j}{d(i,j)^\alpha} Z_i Z_j + \sum_i h_i X_i,$$

may be represented as a matrix product operator using matrices of the form

$$W[i] = \begin{pmatrix} I & \lambda_1 J_i Z_i & \lambda_2 J_i Z_i & \dots & \lambda_K J_i Z_i & \lambda_1 Z_i & \dots & \lambda_K Z_i & h_i X_i \\ 0 & \lambda_1 I & 0 & \dots & 0 & 0 & \dots & 0 & a_1 J_i Z_i \\ 0 & 0 & \lambda_2 I & \dots & 0 & 0 & \dots & 0 & a_2 J_i Z_i \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_K I & 0 & \dots & 0 & a_k J_i Z_i \\ 0 & 0 & 0 & \dots & 0 & \lambda_1 I & \dots & 0 & a_1 Z_i \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & \lambda_K I & a_K Z_i \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 & I \end{pmatrix}$$

together with

$$W_L = \begin{pmatrix} I & \lambda_1 J_0 Z_0 & \lambda_2 J_0 Z_0 & \dots & \lambda_K J_0 Z_0 & \lambda_1 Z_0 & \dots & \lambda_K Z_0 & h_0 X_0 \end{pmatrix},$$

$$W_R = \begin{pmatrix} h_{n-1}X_{n-1} & a_1 J_{n-1} Z_{n-1} & a_2 J_{n-1} Z_{n-1} & \dots & a_K J_{n-1} Z_{n-1} & a_1 Z_{n-1} & \dots & a_K Z_{n-1} & I \end{pmatrix}^T$$

like

$$H = W_L \cdot \prod_{i=1}^{n-2} W[i] \cdot W_R,$$

where we used exponential-sum fitting to approximate the decay of interactions as

$$\frac{1}{d(i,j)^\alpha} \approx \sum_{l=1}^{K} a_l \lambda_l^{|i-j|},$$

using an algorithm described in the Appendix of [55]. For periodic boundary conditions, see the approach described in Appendix E.4.

E.3 Rydberg model

The Hamiltonian

$$H = \sum_{i<j} \frac{V}{|i+\delta_i - j - \delta_j|^6} N_i N_j + \sum_i \left(\frac{\Omega}{2} X_i + \Delta N_i \right),$$

may be represented as a matrix product operator using matrices of the form

$$W[i] = \begin{pmatrix} I & V\lambda_1^{1-\delta_i} N_i & V\lambda_2^{1-\delta_i} N_i & \dots & V\lambda_K^{1-\delta_i} N_i & \frac{\Omega}{2} X_i + \Delta N_i \\ 0 & \lambda_1 I & 0 & \dots & 0 & a_1 \lambda_1^{\delta_i} N_i \\ 0 & 0 & \lambda_2 I & \dots & 0 & a_2 \lambda_2^{\delta_i} N_i \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_K I & a_k \lambda_K^{\delta_i} N_i \\ 0 & 0 & 0 & \dots & 0 & I \end{pmatrix}$$

together with

$$W_L = \begin{pmatrix} I & V\lambda_1^{1-\delta_0} N_0 & V\lambda_2^{1-\delta_0} N_0 & \dots & V\lambda_K^{1-\delta_0} N_0 & \frac{\Omega}{2} X_0 + \Delta N_0 \end{pmatrix},$$

$$W_R = \begin{pmatrix} \frac{\Omega}{2} X_{n-1} + \Delta N_{n-1} & a_1 \lambda_1^{\delta_{n-1}} N_{n-1} & a_2 \lambda_2^{\delta_{n-1}} N_{n-1} & \dots & a_k \lambda_K^{\delta_{n-1}} N_{n-1} & I \end{pmatrix}^T$$

like

$$H = W_L \cdot \prod_{i=1}^{n-2} W[i] \cdot W_R,$$

where we again used exponential-sum fitting to approximate the decay of interactions as

$$\frac{1}{|i-j|^6} \approx \sum_{l=1}^{K} a_l \lambda_l^{|i-j|}$$

like previously. This exponential fit further allowed us to separate the site displacements from the overall decay of interactions and hence implement this nontrivial modified geometry using tensor networks.

E.4 Implementation of periodic boundary conditions

All particular details regarding specific implementations for both open and periodic boundary conditions for specific models are discussed in their corresponding sections. In this appendix, we look at some general ideas behind implementing periodic boundary conditions. For practical reasons, the simulations of periodic boundary conditions in larger systems were implemented using open boundary tensor networks.

When considering only a nearest-neighbour coupling on a ring, it is a common practice to use a so-called folded ordering of the MPO, which transforms the model from having only nearest-neighbour interactions with one additional interaction over the length of the whole chain, into a model with nearest-neighbour and next-nearest-neighbour interactions. However, when carrying out the simulations for the periodic Heisenberg model, we found it to be faster to implement this long-range interaction using a non-uniform exponentially decaying coupling, where the base would be equal to 1 (so that there wouldn't actually be any decay), and the only non-zero interaction would be between the first and last site. The implementation for this case is shown in Appendix

E.1, while the general non-uniform exponential coupling is discussed in more detail in Appendix E.2.

When considering general long-range interactions, the periodic boundary conditions were implemented by appropriately changing the metric in the interaction terms. The metric specifying one-dimensional chain with open boundary conditions is simply $d(i,j) = |i-j|$, while the corresponding metric for periodic boundary conditions is $d(i,j) = \min\{|i-j|, n-|i-j|\}$. The newly obtained symmetric decay rate of interactions was then fitted using correspondingly symmetric sum of exponentials

$$\frac{1}{d(i,j)^\alpha} \approx \sum_{l=1}^{K} c_l \cdot \left(b_l^{|i-j|} + b_l^{n-|i-j|} \right).$$

As this task proved to be numerically problematic, we've devised the following approach: denote $\sum_{l=1}^{k} c_l \cdot b_l^x$ by $g(x)$, so that the desired polynomial decay rate is given by $\frac{1}{d(x)^\alpha} \approx g(x) + g(n-x)$, where $x = |i-j|$. If we can find a function $g(x)$ obeying this relation, which is decaying and smooth enough, we could use a robust non-iterative algorithm to fit it with a sum of exponentials [55]. A suitable function for this task was found to be

$$g(x) = \begin{cases} \frac{1}{x^\alpha} - \frac{1/2}{(n-x)^\alpha} & \text{for } x \leq \frac{n}{2}, \\ \frac{1/2}{x^\alpha} & \text{for } x \geq \frac{n}{2}. \end{cases}$$

Hence, after obtaining the fitting parameters $\{c_l, b_l\}_{l=1}^K$, we would create the MPO using $\{a_l, \lambda_l\}_{l=1}^{2K} = \{c_l, b_l\}_{l=1}^K \cup \{c_l \cdot b_l^n, \frac{1}{b_l}\}_{l=1}^K$ as the parameters for the exponential fit to correctly account for the symmetry.

References

[1] R.G. Parr and Y. Weitao. "Density-functional theory of atoms and molecules". International Series of Monographs on Chemistry. Oxford University Press. (1994).

[2] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal. "Quantum Monte Carlo simulations of solids". Rev. Mod. Phys. **73**, 33–83 (2001).

[3] J. Ignacio Cirac, David Pérez-García, Norbert Schuch, and Frank Verstraete. "Matrix product states and projected entangled pair states: Concepts, symmetries, theorems". Rev. Mod. Phys. **93**, 045003 (2021).

[4] C. Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, Muhammed Shuaibi, Anuroop Sriram, Kevin Tran, Brandon Wood, Junwoong Yoon, Devi Parikh, and Zachary Ulissi. "An introduction to electrocatalyst design using machine learning for renewable energy storage" (2020). arXiv:2010.09435.

[5] Alexander M. Dalzell, Sam McArdle, Mario Berta, Przemyslaw Bienias, Chi-Fang Chen, András Gilyén, Connor T. Hann, Michael J. Kastoryano, Emil T. Khabiboulline, Aleksander Kubica, Grant Salton, Samson Wang, and Fernando G. S. L. Brandão. "Quantum algorithms: A survey of applications and end-to-end complexities" (2023). arXiv:2310.03011.

[6] Joseph Tindall, Matt Fishman, Miles Stoudenmire, and Dries Sels. "Efficient tensor network simulation of IBM's Eagle kicked Ising experiment" (2023). arXiv:2306.14887.

[7] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. "Highly accurate protein structure prediction with Alphafold". Nature **596**, 583–589 (2021).

[8] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. "Neural message passing for quantum chemistry" (2017). arXiv:1704.01212.

[9] Giuseppe Carleo and Matthias Troyer. "Solving the quantum many-body problem with artificial neural networks". Science **355**, 602–606 (2017).

[10] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. "Neural-network quantum state tomography". Nature Physics **14**, 447–450 (2018).

[11] Juan Carrasquilla. "Machine learning for quantum matter". Advances in Physics: X **5**, 1797528 (2020).

[12] David Pfau, James S. Spencer, Alexander G. D. G. Matthews, and W. M. C. Foulkes. "Ab initio solution of the many-electron Schrödinger equation with deep neural networks". Physical Review Research **2** (2020).

[13] Anurag Anshu and Srinivasan Arunachalam. "A survey on the complexity of learning quantum states" (2023). arXiv:2305.20069.

[14] Kim Andrea Nicoli, Christopher J. Anders, Lena Funcke, Tobias Hartung, Karl Jansen, Stefan Kuhn, Klaus Robert Muller, Paolo Stornati, Pan Kessel, and Shinichi Nakajima. "Physics-informed Bayesian optimization of variational quantum circuits" (2024). arXiv:2406.06150.

[15] Laura Lewis, Hsin-Yuan Huang, Viet T Tran, Sebastian Lehner, Richard Kueng, and John Preskill. "Improved machine learning algorithm for predicting ground state properties". Nature communications **15**, 895 (2024).

[16] Emilio Onorati, Cambyse Rouzé, Daniel Stilck França, and James D. Watson. "Efficient learning of ground & thermal states within phases of matter" (2023). arXiv:2301.12946.

[17] Emilio Onorati, Cambyse Rouzé, Daniel Stilck França, and James D. Watson. "Provably efficient learning of phases of matter via dissipative evolutions" (2023). arXiv:2311.07506.

[18] Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V. Albert, and John Preskill. "Provably efficient machine learning for quantum many-body problems". Science **377** (2022).

[19] Yanming Che, Clemens Gneiting, and Franco Nori. "Exponentially improved efficient machine learning for quantum many-body states with provable guarantees". Phys. Rev. Res. **6**, 033035 (2024).

[20] Nicolò Defenu, Tobias Donner, Tommaso Macrì, Guido Pagano, Stefano Ruffo, and Andrea Trombettoni. "Long-range interacting quantum systems". Rev. Mod. Phys. **95**, 035002 (2023).

[21] Bo Yan, Steven A Moses, Bryce Gadway, Jacob P Covey, Kaden RA Hazzard, Ana Maria Rey, Deborah S Jin, and Jun Ye. "Observation of dipolar spin-exchange interactions with lattice-confined polar molecules". Nature **501**, 521–525 (2013).

[22] M. Saffman, T. G. Walker, and K. Mølmer. "Quantum information with Rydberg atoms". Reviews of Modern Physics **82**, 2313–2363 (2010).

[23] Joseph W Britton, Brian C Sawyer, Adam C Keith, C-C Joseph Wang, James K Freericks, Hermann Uys, Michael J Biercuk, and John J Bollinger. "Engineered two-dimensional Ising interactions in a trapped-ion quantum simulator with hundreds of spins". Nature **484**, 489–492 (2012).

[24] K. Binder and A. P. Young. "Spin glasses: Experimental facts, theoretical concepts, and open questions". Rev. Mod. Phys. **58**, 801–976 (1986).

[25] Raffaele Santagati, Alan Aspuru-Guzik, Ryan Babbush, Matthias Degroote, Leticia González, Elica Kyoseva, Nikolaj Moll, Markus Oppel, Robert M. Parrish, Nicholas C. Rubin, Michael Streif, Christofer S. Tautermann, Horst Weiss, Nathan Wiebe, and Clemens Utschig-Utschig. "Drug design on quantum computers". Nature Physics **20**, 549–557 (2024).

[26] Thomas Vojta. "Disorder in quantum many-body systems". Annual Review of Condensed Matter Physics **10**, 233–252 (2019).

[27] Štěpán Šmíd and Roberto Bondesan. "Accurate learning of equivariant quantum systems from a single ground state" (2024). arXiv:2405.12309.

[28] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. "E(n) equivariant graph neural networks" (2022). arXiv:2102.09844.

[29] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials". Nature Communications **13** (2022).

[30] Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. "Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations" (2023). arXiv:2210.07237.

[31] Marc Wanner, Laura Lewis, Chiranjib Bhattacharyya, Devdatt Dubhashi, and Alexandru Gheorghiu. "Predicting ground state properties: Constant sample complexity and deep learning algorithms" (2024). arXiv:2405.18489.

[32] Štěpán Šmíd and Roberto Bondesan (2024). code: Quantum-AI-Lab-ICL/Efficient-Learning-Long-Range-QS.

[33] Hsin-Yuan Huang, Richard Kueng, and John Preskill. "Predicting many properties of a quantum system from very few measurements". Nature Physics **16**, 1050–1057 (2020).

[34] M. B. Hastings. "Locality in quantum systems" (2010). arXiv:1008.5137.

[35] Minh C. Tran, Andrew Y. Guo, Christopher L. Baldwin, Adam Ehrenberg, Alexey V. Gorshkov, and Andrew Lucas. "Lieb-Robinson light cone for power-law interactions". Physical Review Letters **127** (2021).

[36] M. Mohri, A. Rostamizadeh, and A. Talwalkar. "Foundations of machine learning, second edition". Adaptive Computation and Machine Learning series. MIT Press. (2018). url: MIT Press.

[37] David J. Luitz, Nicolas Laflorencie, and Fabien Alet. "Many-body localization edge in the random-field Heisenberg chain". Physical Review B **91** (2015).

[38] Hidetoshi Nishimori. "Statistical physics of spin glasses and information processing: An introduction". Oxford University Press. (2001).

[39] Matteo Marcuzzi, Jiří Minář, Daniel Barredo, Sylvain de Léséleuc, Henning Labuhn, Thierry Lahaye, Antoine Browaeys, Emanuele Levi, and Igor Lesanovsky. "Facilitation dynamics and localization phenomena in Rydberg lattice gases with position disorder". Phys. Rev. Lett. **118**, 063606 (2017).

[40] Sven Bachmann, Spyridon Michalakis, Bruno Nachtergaele, and Robert Sims. "Automorphic equivalence within gapped phases of quantum lattice systems". Communications in Mathematical Physics **309**, 835–871 (2011).

[41] Tomotaka Kuwahara and Keiji Saito. "Strictly linear light cones in long-range interacting systems of arbitrary dimensions". Physical Review X **10** (2020).

[42] Weinan E, Chao Ma, and Lei Wu. "Machine learning from a continuous viewpoint, I". Science China Mathematics **63**, 2233–2266 (2020).

[43] Ulrich Schollwöck. "The density-matrix renormalization group in the age of matrix product states". Annals of Physics **326**, 96–192 (2011).

[44] Johannes Hauschild and Frank Pollmann. "Efficient numerical simulations with Tensor Networks: Tensor Network Python (TeNPy)". SciPost Phys. Lect. NotesPage 5 (2018).

[45] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, Advances in Neural Information Processing Systems. Volume 20. Curran Associates, Inc. (2007). url: NeurIPS proceedings.

[46] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. "Scikit-learn: Machine learning in Python". J. Mach. Learn. Res. **12**, 2825–2830 (2011).

[47] R. Juhász, I. A. Kovács, and F. Iglói. "Random transverse-field Ising chain with long-range interactions". Europhysics Letters **107**, 47008 (2014).

[48] Philipp Hauke, Fernando M Cucchietti, Alexander Müller-Hermes, Mari-Carmen Bañuls, J Ignacio Cirac, and Maciej Lewenstein. "Complete devil's staircase and crystal–superfluid transitions in a dipolar XXZ spin chain: a trapped ion quantum simulation". New Journal of Physics **12**, 113037 (2010).

[49] Matthew B. Hastings and Tohru Koma. "Spectral gap and exponential decay of correlations". Communications in Mathematical Physics **265**, 781–804 (2006).

[50] D. J. Newman and H. S. Shapiro. "Jackson's theorem in higher dimensions". Pages 208–219. Springer Basel. Basel (1964).

[51] C. Gaetan and X. Guyon. "Spatial statistics and modeling". Springer Series in Statistics. Springer New York. (2009).

[52] Milan Janjic and Boris Petkovic. "A counting function" (2013). arXiv:1301.4550.

[53] Iosif Pinelis. "Exact lower and upper bounds on the incomplete gamma function" (2020). arXiv:2005.06384.

[54] Minh C. Tran, Chi-Fang Chen, Adam Ehrenberg, Andrew Y. Guo, Abhinav Deshpande, Yifan Hong, Zhe-Xuan Gong, Alexey V. Gorshkov, and Andrew Lucas. "Hierarchy of linear light cones with long-range interactions". Physical Review X **10** (2020).

[55] B Pirvu, V Murg, J I Cirac, and F Verstraete. "Matrix product operator representations". New Journal of Physics **12**, 025012 (2010).

[56] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. "Graph networks as a universal machine learning framework for molecules and crystals". Chemistry of Materials **31**, 3564–3572 (2019).