Exponentially improved efficient machine learning for quantum many-body states with provable guarantees

Yanming Che [1,2,*] Clemens Gneiting [2,3,†] and Franco Nori [2,3,1,‡]

[1]*Department of Physics, University of Michigan, Ann Arbor, Michigan 48109-1040, USA*
[2]*Theoretical Quantum Physics Laboratory, Cluster for Pioneering Research, RIKEN, Wako-shi, Saitama 351-0198, Japan*
[3]*Center for Quantum Computing, RIKEN, Wako-shi, Saitama 351-0198, Japan*

Solving the ground state and the ground-state properties of quantum many-body systems is generically a hard task for classical algorithms. For a family of Hamiltonians defined on an m-dimensional space of physical parameters, the ground state and its properties at an arbitrary parameter configuration can be predicted via a machine learning protocol up to a prescribed prediction error ε, provided that a sample set (of size N) of the states can be efficiently prepared and measured. In a recent work [Huang *et al.*, Science **377**, eabk3333 (2022)], a rigorous guarantee for such a generalization was proved. Unfortunately, an exponential scaling for the provable sample complexity, $N = m^{\mathcal{O}(\frac{1}{\varepsilon})}$, was found to be universal for generic gapped Hamiltonians. This result applies to the situation where the dimension of the parameter space is large, while the scaling with the accuracy is not an urgent factor, not entering the realm of more precise learning and prediction. In this work, we consider an alternative relevant scenario, where the effective dimension m is a finite, not necessarily large constant, while the scaling with the prediction error becomes the central concern. By jointly preserving the fundamental properties of density matrices in the learning protocol and utilizing the continuity of quantum states in the parameter range of interest, we rigorously obtain a polynomial sample complexity for predicting quantum many-body states and their properties, with respect to the *uniform* prediction error ε and the number of qubits, n, with $N = \text{poly}(\varepsilon^{-1}, n, \ln \frac{1}{\delta})$, where poly denotes a polynomial function, and $(1 - \delta)$ is the probability of success. Moreover, if restricted to learning local quantum-state properties, the number of samples can be further reduced to $N = \text{poly}(\varepsilon^{-1}, \ln \frac{n}{\delta})$. Numerical demonstrations confirm our findings, and an alternative approach utilizing statistical learning theory with reproducing kernel Hilbert space achieves consistent results. The mere continuity assumption indicates that our results are not restricted to gapped Hamiltonian systems and properties within the same phase.

I. INTRODUCTION

The prediction of quantum many-body ground states and their properties lies at the heart of a multitude of branches of modern quantum science [1–4], ranging from condensed-matter physics, quantum computation and simulation, to quantum chemistry. However, finding the exact ground states of quantum many-body Hamiltonians is generically beyond the reach of efficient classical algorithms, under a widely believed conjecture in computational complexity theory that the polynomial hierarchy does not collapse [5]. Such complexity arises from the rich structure of the quantum many-body state space as well as the principles of quantum mechanics. Recently, it was proved [6,7] that for a particular family of gapped Hamiltonians, evaluating ground-state properties,

such as the expectation values of local operators, can be as hard as solving NP-complete problems in classical settings.

On the other hand, machine learning as an effective approximation scheme whose solution can be efficiently obtained via proper optimization processes has found a wide spectrum of applications in solving complex quantum-physics problems [8–17], including approximating and predicting quantum states [1,7,18–51], as well as identifying phases and phase transitions [52–93]. This raises the hope that machine learning may also help, in some cases, with *efficiently* predicting quantum many-body ground states and their properties; for instance, when the Hamiltonian depends on parameters and a training set of ground states or their properties can be efficiently prepared for some specific parameter choices, in which case learning algorithms can help to generalize to unknown parameters with a lower cost compared to that required in experiments, and to make predictions for the entire parameter space, up to a small prediction error ε.

Such a situation has recently been addressed in [7], where a rigorous guarantee for learning and predicting properties of quantum many-body ground states was presented, provided that a training set of the ground states is efficiently available via classical shadow tomography [94,95]. But, unfortunately, a provably exponential scaling of the sample complexity N with the prediction error ε, $N = m^{\mathcal{O}(\varepsilon^{-1})}$, was found to be

---

*Contact author: yanmingche01@gmail.com
†Contact author: clemens.gneiting@riken.jp
‡Contact author: fnori@riken.jp

universal for generic gapped Hamiltonians, where m is the dimension of the parameter space. This situation is therefore mostly useful when the dimension m is large while the error tolerance is high, and precise learning and prediction are not in focus.

An alternative relevant practical scenario may be that the scaling of N with the accuracy is the main concern while the number of parameters m is a finite constant. This includes, for instance, situations where one is only concerned with the spatially local properties of the system or with the continuous macroscopic quantities in the critical region of phase transitions, where only a small number of terms in the Hamiltonian is relevant.

In this work, we are motivated by the framework of the probably approximately correct (PAC) learnable [96,97], which is a fundamental theory of computational learning and requires the minimal number of samples for the learning task to scale (at most) polynomially in the inverse of the prediction error (ε^{-1}), as well as in the inverse of the failure probability (δ^{-1}). We explore theoretical guarantees for polynomially efficient learning that significantly reduce the provable sample complexity scaling exponentially with ε^{-1} in [7], and find a decrease of many orders of magnitude in it. The polynomial sample complexity is numerically demonstrated for a quantum XY model, which highlights that our results apply to more general quantum systems beyond gapped Hamiltonians [98] and properties of a single phase.

While using the framework of kernel-based learning, we emphasize *physical constraints* on the learning protocol to avoid possible unphysical predictions. These constraints lead us to introduce the concept of positive good kernels (PGKs) [99], which will be elaborated below.

Moreover, kernel-based models for learning and predicting quantum states have the following advantages: First, kernel methods have been widely used in supervised and unsupervised learning tasks such as nonlinear regression and classification [97,100], and also dimensionality reduction [82–84,101,102]. The kernel itself contains rich information such that there is no need to calculate the feature maps in the high-dimensional feature space [103,104], which drastically reduces the computational cost. Second, even supervised learning with neural networks, the currently most popular architecture, is equivalent to the neural tangent kernel (NTK) regression in the limit of infinite width of the network [105]. Third, each positive kernel is uniquely associated with a reproducing kernel Hilbert space (RKHS) [97,100,103,106,107], which can be a universal hypothesis space for the learning problem (in the case that the kernel is a universal kernel). With these properties, kernel-based methods not only are of practical value, but also can provide a unified framework of theoretical analysis for various provable guarantees.

II. SETTING AND MODEL

We start from a continuously parametrized family of n-qubit quantum states $\rho(x)$ with $x \in \mathcal{X}$, where $\mathcal{X} = [-\frac{L}{2}, \frac{L}{2}]^m \subset \mathbb{R}^m$ is an m-dimensional space of physical parameters, with L the length of the intervals in each dimension. Without loss of generality, here we restrict to qubit models. For fermionic models, they can be encoded as qubits through,

for example, the Jordan-Wigner [108] or the Bravyi-Kitaev [109] transformation. The states $\rho(x)$ may be, but are not restricted to, the ground states of a parametrized Hamiltonian $H(x)$. Additionally, we assume that the parameter space is equipped with a probability measure μ, according to which a sample x_i is drawn from $\mathcal{X}$, and that a training set of quantum states, $\mathcal{S} = \{x_i, \rho(x_i)\}_{i=1}^N$, can be efficiently prepared classically or quantumly.

While the tomography for accessing $\rho(x_i)$ experimentally is, in general, expensive for multiqubit systems, various strategies have been proposed to reduce the cost of quantum state tomography, e.g., via compressed sensing [110,111], or with neural-network quantum state tomography [21,39–44]. In particular, in classical shadow tomography [7,94,95], the sample state is approximated via randomized Pauli measurements $\rho(x_i) \approx \sigma_T(x_i)$, with $\sigma_T(x_i)$ the classical shadow state over T copies of measurements at the parameter x_i. Moreover, with locality assumptions, the approximation via shadow states has been proved [7,94] to be very efficient, requiring only $T = \mathcal{O}(\ln n/\epsilon^2)$ copies of measurements, where ϵ is the precision of the tomography in trace norm. For more details about the classical shadow tomography, see [7,94].

Without loss of generality, the target functions to be learned, such as the real and imaginary parts of the density-matrix entries, or the order parameters of quantum phases are assumed to be *continuous* real functions $f : \mathcal{X} \to \mathbb{R}$ [112]. Furthermore, periodic boundary conditions can be assumed for $\mathcal{X}$, as in [7], such that the function f is defined on a circle of length L in each dimension. Finally, the predicted density matrix at an arbitrary $x \in \mathcal{X}$ is given by the kernel generalization from the training set $\mathcal{S}$,

$$\sigma_N(x) = \frac{1}{N} \sum_{i=1}^N K_\Lambda(x - x_i)\rho(x_i), \tag{1}$$

where we assume a kernel with translational symmetry, i.e.,

$$K_\Lambda(x, x_i) = K_\Lambda(x - x_i), \tag{2}$$

with positive integer index Λ.

Note that the estimated density matrix $\sigma_N(x)$ in (1) should preserve the fundamental properties of density operators for all $x \in \mathcal{X}$, i.e., it should be Hermitian, positive semidefinite, and have unit trace. The first two conditions require $K_\Lambda(x) \geqslant 0$ to be real, while the last one, due to approximation and statistical errors, can be approximately satisfied in a practical learning process, i.e.,

$$\mathrm{Tr}\sigma_N(x) = \frac{1}{N} \sum_{i=1}^N K_\Lambda(x - x_i) \approx 1. \tag{3}$$

III. POSITIVE GOOD KERNELS (PGKs)

In order to facilitate an efficient and accurate learning of the quantum state (1) satisfying the physical constraints for density matrices, we characterize the following conditions for PGKs (e.g., see [99,113]):

(I) Positivity and boundness:

$$0 \leqslant K_\Lambda(x) \leqslant \mathcal{O}(\Lambda^\tau) \; (\forall x \in \mathcal{X}), \tag{4}$$

with τ some positive integer.

(II) Normalization:

$$\int_{x \in \mathcal{X}} K_\Lambda(x) d\mu(x) = 1. \tag{5}$$

(III) η-convergence: For all $0 < \eta \leqslant L$,

$$\int_{x \in \mathcal{X}, \, \|x\|_2 \geqslant \eta} |K_\Lambda(x)| d\mu(x) \leqslant \mathcal{O}(\Lambda^{-1}). \tag{6}$$

The normalization (II) of the kernel guarantees that $\mathrm{Tr}\sigma_N(x) \approx 1$ with high probability with respect to $\mathcal{S}$. The conditions (I, II, III) allow us to predict the continuous quantum state within a prescribed uniform precision and in an efficient manner (see more details and examples of PGKs in Appendix B). If the uniform distribution on $\mathcal{X}$ is used, the probability measure will be given by $d\mu(x) = \frac{1}{L^m} dx$. As an example, we present a rectangular Fejér kernel, which is a PGK (see Appendix B for proof details) given by [99,114]

$$F_\Lambda(x) = \frac{1}{\Lambda^m} \prod_{i=1}^{m} \frac{\sin^2\left(\frac{\Lambda\pi}{L} x_i\right)}{\sin^2\left(\frac{\pi x_i}{L}\right)}, \tag{7}$$

where we have $0 \leqslant F_\Lambda(x) \leqslant \Lambda^m$.

In [7], an ℓ_2-Dirichlet kernel, $D_\Lambda(x)$, has been shown to perform well in approximating smooth functions; in particular, the average of local observables with respect to the ground state, with a bounded derivative of the first order and through a truncated Fourier series. However, we find that in the task of kernel learning of the density matrix, the Dirichlet kernel may fail to be a PGK, in the sense that it violates the positivity and the L_1-boundedness [99]. For example, taking the dimension $m = 1$, the entries of $D_\Lambda(x)$ can be negative for some values of x, and also $\frac{1}{2\pi} \int_{-\pi}^{\pi} |D_\Lambda(x)| dx \geqslant c \ln \Lambda$, with some constant $c > 0$.

IV. RESULTS

We are now prepared to present our main findings exposed by the following two theorems.

Theorem 1 [Efficient learning of quantum-state representations with positive good kernels (PGKs)]. *Given:* A family of n-qubit unknown quantum states $\rho(x)$ with continuously [112] parametrized density-matrix entries defined on the parameter space $\mathcal{X} \subset \mathbb{R}^m$, where m is a constant, and a training set $\mathcal{S} = \{x_i, \rho(x_i)\}_{i=1}^{N}$ with x_i drawn from $\mathcal{X}$ according to a probability density associated with the probability measure μ.

Goal: Output an estimator $\sigma_N(x)$ for the density matrix which preserves the positivity and the unit trace through (1) via the PGKs, with an as small as possible number of samples N, and which is close to the true density matrix $\rho(x)$ for an arbitrary $x \in \mathcal{X}$, in terms of the uniform prediction error,

$$\sup_{x \in \mathcal{X}} \|\sigma_N(x) - \rho(x)\|_{\ell_\infty} \leqslant \varepsilon \tag{8}$$

and

$$|\mathrm{Tr}\sigma_N(x) - 1| \leqslant \varepsilon \tag{9}$$

at the same time, with probability at least $1 - \delta$ ($0 < \delta < 1$).

Result (Sample complexity): The above goal can be achieved with only

$$N = \mathrm{poly}\left(\varepsilon^{-1}, \, n, \, \ln\frac{1}{\delta}\right). \tag{10}$$

Note that the matrix ℓ_∞-norm here (different from the operator norm) is defined as $\|\rho\|_{\ell_\infty} = \max_{i,j} |\rho_{ij}|$, with ρ_{ij} the entries of the matrix ρ, and poly denotes a polynomial function. (*This completes Theorem 1*.)

While the computational time can generically depend exponentially on the number of qubits because the number of entries of the density matrix is of $\mathcal{O}(4^n)$, this may be reduced to poly(n) for sparse states. We emphasize that ε represents the uniform prediction error for the whole parameter space. Theorem 1 guarantees that efficient learning and prediction are performed physically by (approximately) preserving the fundamental properties of the density matrix. For more discussions about the continuity of the density matrix, see Appendix C.

Motivated by the importance as well as the hardness of computing the ground-state properties for generic quantum many-body Hamiltonians, we next relax to the target of learning a state property,

$$f_O(x) = \mathrm{Tr}[O\rho(x)] = \sum_{i=1}^{M} f_i(x), \tag{11}$$

a continuous function on $\mathcal{X}$ given by the expectation value of a q-local operator $O = \sum_{i=1}^{M} O_i$, with respect to the quantum many-body state $\rho(x)$. The ith-partite local state property reads

$$f_i(x) = \mathrm{Tr}[O_i\rho(x)], \tag{12}$$

with $i \in [M] = \{1, 2, \ldots, M\}$, where M is a positive integer scaling polynomially in n, and the local function is bounded by $|f_i(x)| \leqslant \mathcal{O}(1)$.

Theorem 2 [Efficient learning of quantum-state properties with positive good kernels (PGKs)]. *Given:* A training set of size N efficiently prepared as $\mathcal{S} = \{x_i, f_O(x_i)\}_{i=1}^{N}$.

Goal: Predict the quantum-state property,

$$\hat{f}_O(x) = \mathrm{Tr}[O\sigma_N(x)] = \sum_{i=1}^{M} \hat{f}_i(x), \tag{13}$$

via the positive $\sigma_N(x)$ with unit trace in (1), with an as small as possible number of samples N, and such that it is close to the true values of $f_O(x)$ in terms of the uniform prediction error,

$$\sup_{x \in \mathcal{X}} |\hat{f}_O(x) - f_O(x)| \leqslant \varepsilon, \tag{14}$$

with probability at least $1 - \delta$ ($0 < \delta < 1$).

Result (Sample complexity): The above goal can be achieved with only

$$N = \mathrm{poly}\left(\varepsilon^{-1}, \, n, \, \ln\frac{1}{\delta}\right). \tag{15}$$

Moreover, if locality is assumed, where the learning only requires $\sup_{x \in \mathcal{X}} |\hat{f}_i(x) - f_i(x)| \leqslant \varepsilon$ ($\forall i \in [M]$) with probability at least $1 - \delta$, then N can be further reduced to

$$N = \mathrm{poly}\left(\varepsilon^{-1}, \, \ln\frac{n}{\delta}\right). \tag{16}$$

Now the computational time scales only (at most) polynomially in the number of qubits n. (*This completes Theorem 2*.)

For a summary of the main ideas and the detailed proofs of the two theorems, see Appendix B.
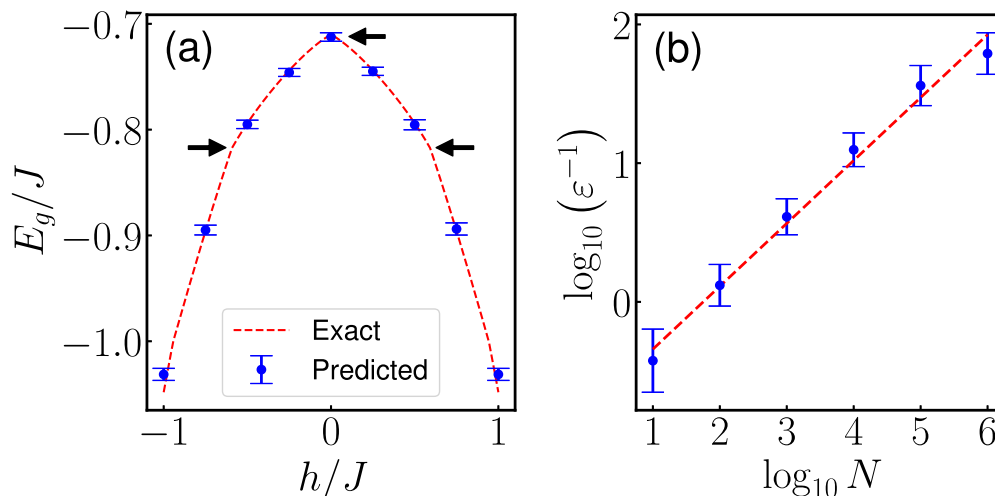
FIG. 1. (a) Comparison between the exact (red dashed curve) ground-state energies per qubit and the positive good kernel (PGK) predicted (blue dots) ones, for a finite-size one-dimensional quantum XY model with sample size $N = 10^6$. The black arrows indicate the locations where the energy curve is *continuous but not smooth* due to the vacua competition [115,116]. Note that there are two additional but not obvious points in the vicinity of $h/J = \pm 1$, where the energy curve is also continuous but not smooth due to the vacua competition, which are not indicated in this figure. (b) Double-logarithmic plot of ε^{-1} vs N (blue dots), and a linear regression (red dashed line) with a slope ≈ 0.45 and an R-squared score ≈ 0.99, confirming a polynomial scaling of the sample complexity. We use the rectangular Fejér kernel with $\Lambda = 50$, $n = 5$ qubits, $\gamma = 1/3$, and samples from a uniform distribution in both (a) and (b). The error ε in (b) is obtained as the maximal error between the predicted and true energies for a set of fixed values of h/J as shown in (a). The mean values and error bars in both panels are calculated with 30 independent runs.

V. COMPARISON WITH EXISTING RESULTS

We compare the provable polynomial sample complexity here with that scaling exponentially with respect to the prediction accuracy, while being restricted to learning the ground-state properties $f_O(x)$ in [7]. With a unified error metric and the same notation, the provable sample complexity in [7] is of the order of

$$N_0 = \frac{B^2}{\varepsilon^2}(2m+1)^{\frac{1}{\varepsilon^2}}. \tag{17}$$

For instance, taking $m = 2$, the number of qubits, $n = 100$, a q-local operator O such that $M = \mathcal{O}(n^q)$ (taking $q = 1$), and $\varepsilon = 0.1$, we then have $N/N_0 \approx 10^{-48}$ with a Fejér kernel, and $N/N_0 \approx 10^{-61}$ with a Gaussian kernel, respectively, where N denotes the number of samples in (15) (see more details in Appendix B 6). This demonstrates that the rigorous guarantees provided here are more powerful when m is not large and the prediction precision is a central concern.

VI. NUMERICAL DEMONSTRATION

While our rigorous results target experimentally obtained training sets, here we demonstrate the above findings regarding learning the ground-state properties with a one-dimensional (1D) quantum XY model [115–117]. The Hamiltonian is given by

$$H = -J\sum_{i=1}^{n}\frac{1+\gamma}{2}\sigma_i^x\sigma_{i+1}^x + \frac{1-\gamma}{2}\sigma_i^y\sigma_{i+1}^y + \frac{h}{J}\sigma_i^z, \tag{18}$$

where $\sigma_i^{x,y,z}$ are the Pauli matrices of the ith qubit, $J > 0$ and h parametrize the nearest-neighbor interaction and the transverse field, respectively, and $0 \leqslant \gamma \leqslant 1$ is the anisotropic

parameter. We assume periodic boundary conditions for the qubits. This model is important for studying quantum many-body physics and is exactly solvable in the dual fermion picture [108,109]. The latter allows us to test our theoretical results, for example, when $f_O(x)$ is the ground-state energy or the correlation function.

An interesting feature of the 1D quantum XY model for finite sizes is that there are two fermion number parity sectors in the Jordan-Wigner fermion picture [115,116], corresponding to periodic and antiperiodic boundary conditions of fermions, respectively. The ground state of the model is the winner of the competition between the vacuum states of the two parity sectors. Therefore, there can be gapless points where the ground-state energy curve is *continuous but not smooth* within the parameter range $|h/J| \leqslant 1$ [e.g., see the locations indicated by black arrows in Fig. 1(a)]. Note that this situation is not covered by the results in [7].

Figure 1(a) shows a comparison between the PGK predicted (blue dots) and exact (red dashed curve) energies E_g/J; the double-logarithmic plot in Fig. 1(b) depicts the scaling of ε^{-1} with N (blue dots), where a good linear behavior (red dashed line) indicates a polynomial sample complexity. Note that with a sufficiently large fixed value of Λ, the generalization error is dominant, then $N \approx \mathcal{O}(\varepsilon^{-2})$ is observed, as expected from our results (see Appendix B 3). We use $n = 5$ qubits and $\gamma = 1/3$ to sharpen the nonsmooth behavior of the energy curve. The excellent agreement between exact and predicted ground-state energies confirms that our provably rigorous guarantee is not limited to smoothly parametrized or gapped Hamiltonian systems as in [7], but can apply to more general quantum systems.

As a second example, we provide results for predicting the long-range spin-spin correlation functions of the quantum
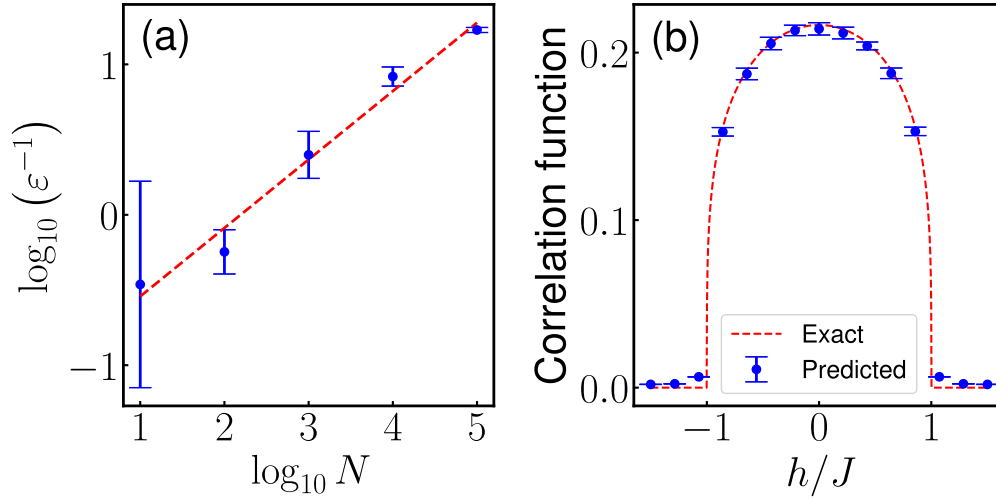
FIG. 2. Predicting the ground-state long-range order (spin-spin correlation function) $\lim_{r\to\infty}(-1)^r\langle S_0^x S_r^x\rangle$ of the quantum XY model in the thermodynamic limit. (a) Double-logarithmic plot of the positive good kernel (PGK) prediction accuracy ε^{-1} vs N (blue dots) within the two-dimensional parameter space $-3/2 \leqslant h/J \leqslant 3/2$ and $0 \leqslant \gamma \leqslant 1$. A linear regression (red dashed line) with a slope ≈ 0.45 and an R-squared score ≈ 0.98 confirms a polynomial scaling of the sample complexity. (b) Comparison between the exact (red dashed curve) long-range spin-spin correlation function and the PGK predicted (blue dots) one with $N = 10^5$, for a quantum XY chain with one-dimensional parameter space $-3/2 \leqslant h/J \leqslant 3/2$ and a fixed value of $\gamma = 1/3$. We use uniform distributions, a Fejér kernel with $\Lambda = 50$, and 30 independent runs as in Fig. 1.

XY model with one- and two-dimensional parameter spaces, respectively. The target is to learn the ground-state two-point function [118],

$$\lim_{r\to\infty}(-1)^r\langle S_0^x S_r^x\rangle, \tag{19}$$

where the spin operators are given by $S_i^{x,y,z} = \sigma_i^{x,y,z}/2$ (set $\hbar = 1$), and the training set is sampled from the exact solution in the thermodynamic limit [118].

In Fig. 2(a), we present a double-logarithmic plot of the inverse uniform error ε^{-1} versus the number of samples, N (blue dots), obtained within the two-dimensional parameter space (h, γ). The linear scaling (red dashed line) indicates a polynomial sample complexity. In Fig. 2(b), the comparison between the exact (red dashed curve) and the predicted (blue dots) correlation functions is plotted, with the parameter h varying and a fixed value of γ. More details can be found in the caption of Fig. 2. We note that in both Figs. 2(a) and 2(b), the range of parameters includes values where the correlation function has a divergent or ill-defined gradient (i.e., it is not smooth), e.g., at $|h/J| = \pm 1$ in Fig. 2(b). This is another strong evidence that our method is applicable to more general cases beyond gapped Hamiltonians, as well as properties of the system within the same phase.

VII. LEARNING QUANTUM-STATE PROPERTIES IN REPRODUCING KERNEL HILBERT SPACE (RKHS)

Now we further elaborate the kernel model and exploit the theory of generalization in machine learning [97], to explicate the result in Theorem 2. We recall that in [100,107], each symmetric and positive-definite kernel $K_\Lambda(x, x')$ defined on $\mathcal{X} \times \mathcal{X}$ is associated to a unique RKHS $\mathcal{H}_{K_\Lambda}$ (see a brief review in Appendix D), with the reproducing property

$$f(x) = \langle f, K_\Lambda(x,)\rangle_{\mathcal{H}_{K_\Lambda}}, \tag{20}$$

for all $f \in \mathcal{H}_{K_\Lambda}$. It is straightforward to verify that the estimator for the state property $\hat{f}_O(x)$ given by (1) lies in the RKHS $\mathcal{H}_{K_\Lambda}$. For universal kernels, the corresponding RKHS is a suitable ansatz space for learning the quantum-state properties.

Provided a training set $\mathcal{S} = \{x_i, f_O(x_i)\}_{i=1}^N$, the optimal function $\hat{f} \in \mathcal{H}_{K_\Lambda}$ can be found by minimizing the empirical error,

$$\mathcal{E}_t = \frac{1}{N}\sum_{i=1}^N |\hat{f}_O(x_i) - f_O(x_i)|, \tag{21}$$

with a bounded functional norm $\|\hat{f}_O\|_{\mathcal{H}_{K_\Lambda}} \leqslant \lambda_f$. The representer theorem [97,100] states that the optimal solution admits the form of

$$\hat{f}_O(x) = \sum_{i=1}^N \alpha_i K_\Lambda(x - x_i), \tag{22}$$

where the real coefficients $\{\alpha_i\}$ are dual variables determined by minimizing $\mathcal{E}_t$.

To obtain a theoretical guarantee with the PGK estimator in (1), the dual variables can be set to be $\alpha_i \approx f_O(x_i)/N$. Moreover, according to the generalization theory of statistical learning, the expected error in terms of the L_1-norm,

$$\mathcal{E}_p = \mathbb{E}_{x\sim\mu} |\hat{f}_O(x) - f_O(x)|, \tag{23}$$

is upper bounded by (with probability at least $1 - \delta$) (see more details in Appendix E)

$$\mathcal{E}_p \leqslant \mathcal{E}_t + \frac{8\lambda_f R}{\sqrt{N}}\sqrt{\frac{\ln\frac{2}{\delta}}{2}}, \tag{24}$$

provided that $K_\Lambda(x, x) \leqslant R^2$ for all $x \in \mathcal{X}$.

With a PGK estimator, the empirical error $\mathcal{E}_t$ can be upper bounded at $\mathcal{O}(\varepsilon)$ with large probability, given a polynomial N

as in (15) (see Appendix E for details). Moreover, the second term on the right-hand side of (24) can also be consistently bounded by $\mathcal{O}(\varepsilon)$. We then obtain a consistent but relaxed result with respect to Theorem 2, i.e., the expected prediction error can be of $\mathcal{O}(\varepsilon)$, *requiring only a polynomially efficient sample complexity*.

VIII. CONCLUSION

We have provided a theoretical guarantee for *polynomially* efficient machine learning quantum many-body states and their properties, provided that efficient access to the training set is available (e.g., via classical shadow tomography) and that the dimension of the parameter space is not necessarily a large constant.

The key idea underlying our finding is to jointly utilize the continuity of quantum states in the parameter space and to preserve the fundamental properties of the density operator in the learning protocol. Our provably rigorous results reveal that a polynomial sample complexity in terms of the *uniform* prediction error and the qubit number is possible, allowing us to predict the quantum states and their properties efficiently and accurately. When restricted to learning local quantum-state properties, the dependence of the number of samples on the number of qubits can be further reduced to poly$(\ln n)$.

We emphasize again that our results are not restricted to learning smooth ground states of gapped Hamiltonians as in [7]. Because our mere continuity assumption of quantum states is more general, it also applies to gapless Hamiltonian systems or continuously parametrized mixed states and their properties (e.g., quantum states smoothly parametrized by Kraus operators [119], and quantum metrological states with finite amounts of quantum Fisher information [120–125]; see Appendix C for more details). An alternative approach from the standard generalization theory of statistical learning, based on the complexity of the RKHS, also achieves consistent results.

Note added. Recently, two related preprints [126,127] appeared, which obtained a quasipolynomial sample complexity, while focusing on learning the average of observables under the assumption of geometric locality, accordingly with more restricted and specific learning models.

ACKNOWLEDGMENTS

APPENDIX A: LIST OF ABBREVIATIONS AND SYMBOLS

For the readers' convenience, here in Table I we list the abbreviations and symbols used most often in this work.
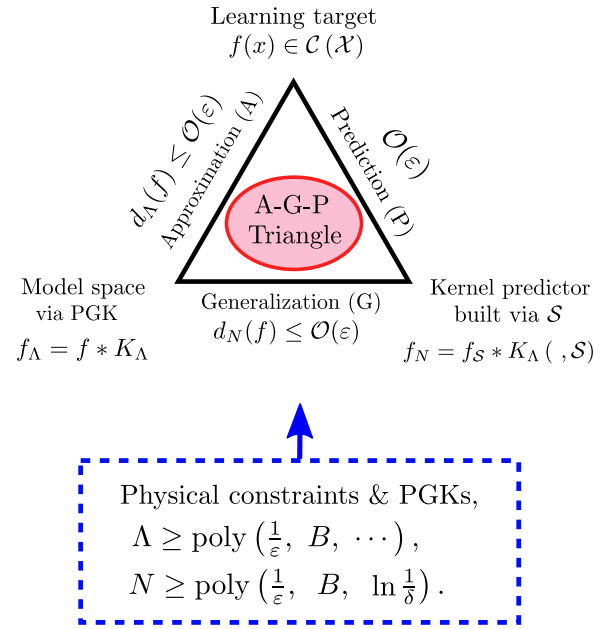

FIG. 3. Schematics of bounding the prediction error of positive good kernel (PGK) predictors, with polynomial number of samples. The target function to be learned (e.g., entries of the density matrix, expectation values of operators with respect to a target quantum state, etc.) is a continuous function defined on the compact parameter space $\mathcal{X}$, the space of which is denoted by $\mathcal{C}(\mathcal{X})$. In the bottom left of the approximation-generalization-prediction (A-G-P) triangle, the model space is obtained via the convolution (denoted by a $*$) between the function itself and a PGK K_Λ. In the bottom right, the kernel predictor f_N is defined as a discrete convolution over the sample set $\mathcal{S}$ of size N. With the physical constraints and proper PGKs, if both the approximation error $d_\Lambda(f)$ and the generalization error $d_N(f)$ can be bounded from above up to a small error bound of $\mathcal{O}(\varepsilon)$, in a consistent manner and with polynomial number of samples, N, then the prediction error can be bounded at $\mathcal{O}(\varepsilon)$, according to the triangle inequality for distances and norms. Note that B is twice the upper bound of the target function $f(x)$ for $x \in \mathcal{X}$ (related to the number of qubits, n) and δ is the probability of success with respect to $\mathcal{S}$.

APPENDIX B: A UNIFIED PROOF FOR THEOREMS 1 AND 2 IN THE MAIN TEXT

1. Setting and summary of main ideas

Summary of proof ideas. Because the target functions to be learned in Theorems 1 and 2 in the main text, such as the entries of the density matrix or the average of the local operators, are continuous functions defined on the parameter space $\mathcal{X}$, the proofs for the two theorems can be given in a unified framework.

The key idea relies on the concept of good kernels [99]. (We added the positivity condition for good kernels to predict density matrices and their properties. Then the L_1-boundness condition for good kernels is automatically satisfied due to the normalization of the kernel.) The convolution between a sequence of good kernels and a continuous target function on a closed interval can approximate the function itself up to an arbitrary *uniform* error at $\mathcal{O}(\varepsilon)$ (also known as the "approximation to the identity") [99]. On the other hand, the

TABLE I. List of abbreviations and symbols.

Abbreviation	Full form
PAC-learnable	Probably approximately correct-learnable
PGK	Positive good kernel
NTK	Neural tangent kernel
RKHS	Reproducing kernel Hilbert space
A-G-P	Approximation-generalization-prediction

Symbol	Description
$\mathcal{X}$	Input parameter space, with $\mathcal{X} \subset \mathbb{R}^m$
m	Dimension of the parameter space
L	Length of $\mathcal{X}$ in each dimension
μ	Probability measure
$\mathcal{S}$	Training set
N	Number of samples/size of the training set
n	Number of qubits in the system
ε	Prediction error bound of the machine learning model
δ	Probability of failure ($1 - \delta$ is the success probability)
η	Quantity used in the definition of continuous functions
Λ	Index of the kernel/Scale of cutoff
$\mathcal{K}_\Lambda(x, x')$	Kernel defined on $\mathcal{X} \times \mathcal{X}$
$\mathcal{K}_\Lambda(x - x')$	Kernel with translational symmetry
$\mathcal{M}$	Upper bound of the L_1 norm of the kernel
$\sigma_N(x)$	Kernel density matrix estimator $\forall x \in \mathcal{X}$
Tr	Trace
$\text{poly}(\alpha, \gamma, \dots)$	Polynomial function of α, γ, etc.
$d_\Lambda(f)$	Distance (error) from the convolution approximation of f with PGKs
$d_N(f)$	Distance (error) from the generalization over the training set
O	q-local observable defined as $O = \sum_{i=1}^{M} O_i$, with $M = \mathcal{O}(n^q)$
B	Twice the upper bound of a continuous function $f(x)$ defined on $\mathcal{X}$
C_L	Lipschitz constant
$\mathcal{H}_{K_\Lambda}$	Reproducing kernel Hilbert space (RKHS) associated with kernel K_Λ
α_i	Dual variables for $i = 1, 2, \dots, N$
$\hat{f}_O(x)$	Kernel estimator of the expectation value of O with respect to the quantum state $\rho(x)$
λ_f	Upper bound of the functional norm of $\hat{f}_O(x)$ in the RKHS
$\mathcal{E}_p$	Expected error in terms of L_1 norm
$\mathcal{E}_t$	Corresponding empirical error/training error with respect to $\mathcal{E}_p$
$\mathcal{R}_\mathcal{S}$	Empirical Rademacher complexity
R	Square root of the upper bound of the kernel K_Λ
$D_\mathbf{n}(x)$	Rectangular Dirichlet kernel with the vector index $\mathbf{n}$
$F_\Lambda(x)$	Rectangular Fejér kernel with the cutoff scale Λ
$K_h(x)$	Gaussian kernel parametrized by h

kernel convolution can be approximated as the sample mean over the training set of size N drawn from the parameter space $\mathcal{X}$ and the target $f(x)$ [e.g., see Eq. (1) in the main text], up to a generalization error $\mathcal{O}(\varepsilon)$ and with a probability of success (at least) $1 - \delta$. If the two approximation errors can be consistently controlled (for all $x \in \mathcal{X}$) with a polynomially scaling number of samples (relative to ε^{-1}, the qubit number n, and $\ln \delta^{-1}$), then Theorems 1 and 2 follow naturally. Note that we have used the Heine-Cantor theorem which states that a continuous function defined on a compact metric space is uniformly continuous. Here the input space $\mathcal{X}$ is a compact metric space. See Fig. 3 for a schematic illustration of the main idea. We will elaborate the details of the proof in the following, where we use positive good kernels such as the multidimensional Fejér kernel and the Gaussian kernel, and the uniform proba-

bility distribution. The results can remain unchanged with nonuniform distributions. (This completes the summary of the proof.)

Without loss of generality, we set the input parameter space as $\mathcal{X} = [-\frac{L}{2}, \frac{L}{2}]^m \subset \mathbb{R}^m$, an m-dimensional tensor product of closed intervals of length L. The target to learn is a *continuous* real function $f : \mathcal{X} \to \mathbb{R}$, defined on a circle of length L in each dimension (i.e., with periodic boundary conditions in the parameter space). As in the main text, we restrict ourselves to kernels with translational symmetry. For a sequence of kernels, $\{K_\Lambda(x)\}_{\Lambda \in \mathcal{I}}$, defined on the parameter space $\mathcal{X}$, with $\mathcal{I}$ the index set of integers by default, we first use the uniform distribution for the probability measure on $\mathcal{X}$ (the case with nonuniform distributions will be discussed later), i.e., $d\mu(x) = dx / L^m$. Then, the properties for positive good kernels read as follows:

(I) Positivity and boundedness: $0 \leqslant K_\Lambda(x) \leqslant \mathcal{O}(\Lambda^\tau)$ ($\forall x \in \mathcal{X}$), with τ some positive integer.

(II) Normalization: $\frac{1}{L^m} \int_{x \in \mathcal{X}} K_\Lambda(x) dx = 1$.

(III) η-convergence: For all $0 < \eta \leqslant L$, $\frac{1}{L^m} \int_{x \in \mathcal{X}, \|x\|_2 \geqslant \eta} |K_\Lambda(x)| dx \leqslant \mathcal{O}(\Lambda^{-1})$ as $\Lambda \to \infty$.

We first use a Fejér kernel [99,114] as an example of the proof. The rectangular Fejér kernel, obtained as the mean of a sequence of rectangular Dirichlet kernels, is given by [114]

$$F_\Lambda(x) = \frac{1}{\Lambda^m} \sum_{\mathbf{n} \in V_\Lambda} D_\mathbf{n}(x), \tag{B1}$$

where the kernel index set $V_\Lambda = \{\mathbf{n} = (n_1, n_2, \ldots, n_m) \in \mathbb{Z}^m : 0 \leqslant n_i < \Lambda, \forall i \in [1, 2, \ldots, m]\}$, and the rectangular Dirichlet kernel reads

$$D_\mathbf{n}(x) = \sum_{k_1=-n_1}^{n_1} \sum_{k_2=-n_2}^{n_2} \cdots \sum_{k_m=-n_m}^{n_m} e^{2\pi k \cdot x / L}, \tag{B2}$$

where the wave vector $k = (k_1, k_2, \ldots, k_m)$ takes integer entries. An equivalent but closed expression for the Fejér kernel above is

$$F_\Lambda(x) = \frac{1}{\Lambda^m} \prod_{i=1}^m \frac{\sin^2\left(\frac{\Lambda\pi}{L} x_i\right)}{\sin^2\left(\frac{\pi x_i}{L}\right)}, \tag{B3}$$

where we have $0 \leqslant F_\Lambda(x) \leqslant \Lambda^m$. The closed form (B3) of the rectangular kernel significantly accelerates the numerical convolution compared to the ℓ_2 kernels, as the latter require to first generate wave vectors whose ℓ_2-norm does not exceed Λ, while the cardinality of the wave-vector set can be exponentially large relative to Λ [7].

Lemma 1. The rectangular Fejér kernel is a positive good kernel (PGK) with respect to the uniform distribution.

Proof. The positivity and boundedness, as well as the normalization, can be easily verified through the closed expression of the Fejér kernel. The last property (III) can be found to be true in the following Appendix B 2 while bounding the approximation error in (B6). ∎

Here the target function $f(x)$ to be learned can be either the entries of the density matrix (real or imaginary parts, respectively) or the average of observables with respect to the quantum state, $f_O(x) = \text{Tr}[\hat{O}\rho(x)]$, where $\hat{O}$ is a local bounded operator as in the main text.

Notation and definition. Because $f(x)$ is assumed to be continuous on a compact metric space $\mathcal{X} \subset \mathbb{R}^m$, according to the Heine-Cantor theorem, it is uniformly continuous. Then, $\forall \varepsilon > 0$, there exists $\eta > 0$ such that if $\|x - y\|_2 \leqslant \eta$, then $|f(x) - f(y)| \leqslant \frac{\varepsilon}{4}$ [called $(\frac{\varepsilon}{4}, \eta)$-continuous]. The function is uniformly continuous, so η is independent of x, but can depend on ε. Furthermore, for a rich family of continuous functions, η can be chosen to be of $\mathcal{O}(\varepsilon^k)$, where $k \geqslant 0$. For example, for Lipschitz continuous functions, which satisfy the Lipschitz condition

$$|f(x) - f(y)| \leqslant C_L \|x - y\|_2 \tag{B4}$$

for all $x, y \in \mathcal{X}$, with C_L the Lipschitz constant, we have $\eta = \varepsilon/(4C_L)$. Or, more generally, for Hölder continuous functions satisfying

$$|f(x) - f(y)| \leqslant C_H \|x - y\|_2^{\alpha_H}, \tag{B5}$$

where C_H is a constant and $\alpha_H \in (0, 1]$, we have $k = 1/\alpha_H$. As a consequence, a function that scales polynomially in η^{-1} is also polynomial in ε^{-1}. In addition, because $f(x)$ is continuous on the closed set $\mathcal{X}$, it is guaranteed to be bounded by $|f(x)| \leqslant B/2$ for all $x \in \mathcal{X}$, for some positive constant B.

2. Bounding the approximation error

With positive good kernels, such as the Fejér kernel $F_\Lambda(x)$, the error from the convolutional approximation of the function is given by

$$\begin{aligned}
d_\Lambda(f) &= |f * F_\Lambda(x) - f(x)| \\
&= \frac{1}{L^m} \left| \int_\mathcal{X} [f(x - y) - f(x)] F_\Lambda(y) dy \right| \\
&\leqslant \frac{1}{L^m} \int_\mathcal{X} |f(x - y) - f(x)| \times |F_\Lambda(y)| dy \\
&= \frac{1}{L^m} \int_{y \in \mathcal{X}, \|y\|_2 \leqslant \eta} |f(x - y) - f(x)| \times |F_\Lambda(y)| dy \\
&\quad + \frac{1}{L^m} \int_{y \in \mathcal{X}, \|y\|_2 \geqslant \eta} |f(x - y) - f(x)| \times |F_\Lambda(y)| dy \\
&\leqslant \frac{\varepsilon}{4} + B\left(\frac{C}{\Lambda}\right)^m \leqslant \frac{\varepsilon}{2}, \tag{B6}
\end{aligned}$$

where

$$C = \frac{4mL^2}{\pi^2 \eta^2}. \tag{B7}$$

The second equation comes from the symmetry property of the kernel convolution on $\mathcal{X}$, $f * F_\Lambda(x) = F_\Lambda * f(x)$, as well as the normalization property of the Fejér kernel. In the third equation, the first integration is upper bounded by $\varepsilon/4$ because f is $(\frac{\varepsilon}{4}, \eta)$-continuous and because the kernel is normalized. The bound for the second term in the third equation is obtained based on the fact that the integration over the region with $\|y\|_2 \geqslant \eta$ does not exceed the integration over the region with $\|y\|_\infty \geqslant \eta/\sqrt{m}$, where the latter is a multidimensional integration over a cubic region that can be factorized into the product of the integrations in each dimension. In each factor dimension, the Fejér kernel is bounded from above by

$$\frac{1}{\Lambda \sin^2[\pi\eta/(\sqrt{m}L)]} \leqslant \frac{4mL^2}{\pi^2 \Lambda \eta^2}, \tag{B8}$$

by using $\sin x \geqslant x/2$ while $0 < x < \pi/3$ is a small number. Then the second inequality holds. *This line of argument also serves as a proof for the statement that the multidimensional Fejér kernel satisfies the property* (III) *of good kernels (see above).* Finally, if we set

$$\Lambda^m \geqslant \frac{4BC^m}{\varepsilon}, \tag{B9}$$

it suffices to bound the total approximation error $d_\Lambda(f)$ up to $\varepsilon/2$, as in the last inequality.

3. Bounding the generalization error

The Fejér kernel estimator of the density matrix is given by

$$\sigma_N(x) = \frac{1}{N} \sum_{i=1}^{N} F_\Lambda(x - x_i)\rho(x_i). \tag{B10}$$

By taking the target $f(x)$ to be the entries of the density matrix or the state properties $f_O(x)$, and with the corresponding sample set $\mathcal{S} = \{x_i, f(x_i)\}_{i=1}^{N}$ drawn uniformly from $\mathcal{X}$ and the function space, Eq. (B10) is a sample mean of the kernel convolution $f * F_\Lambda(x)$. Equivalently, denoting a sequence of independent random variables as $\{Z_i(x) = F_\Lambda(x - x_i)f(x_i)\}_{i=1}^{N}$, which is bounded by $|Z_i| \leqslant \Lambda^m B/2$, we have

$$\mathbb{E}_{x_i \sim \mu(\mathcal{X})} \frac{1}{N} \sum_{i=1}^{N} Z_i(x) = \frac{1}{L^m} \int_{\mathcal{X}} F_\Lambda(x - y)f(y)dy$$
$$= f * F_\Lambda(x). \tag{B11}$$

As in Eq. (B10), the Fejér kernel estimator can be written as $\hat{f}_N(x) = \frac{1}{N} \sum_{i=1}^{N} Z_i(x)$. The McDiarmid's inequality [97] leads to

$$\text{Prob}\left[|\hat{f}_N(x) - f * F_\Lambda(x)| \geqslant \frac{\varepsilon}{2}\right] \leqslant 2 \exp\left(-\frac{N\varepsilon^2}{2\Lambda^{2m}B^2}\right). \tag{B12}$$

Therefore, if the number of samples is set to be

$$N \geqslant \frac{2B^2\Lambda^{2m}}{\varepsilon^2} \ln \frac{2}{\delta}, \tag{B13}$$

then we have the generalization error

$$d_N(f) = |\hat{f}_N(x) - f * F_\Lambda(x)| \leqslant \frac{\varepsilon}{2}, \tag{B14}$$

with probability at least $1 - \delta$ [$\delta \in (0, 1)$].

4. Bounding the supremum-norm prediction error with polynomial sample complexity

Combing Eqs. (B9) and (B13), we can finally obtain the prediction error of the Fejér kernel estimator as (see Fig. 3 for an illustration)

$$|\hat{f}_N(x) - f(x)| \leqslant d_N(f) + d_\Lambda(f) \leqslant \varepsilon, \quad \forall x \in \mathcal{X}, \tag{B15}$$

with probability at least $1 - \delta$, by using a sample set of size

$$N \geqslant \frac{32B^4C^{2m}}{\varepsilon^4} \ln \frac{2}{\delta}$$
$$= \text{poly}\left(\frac{1}{\varepsilon}, B, \ln \frac{1}{\delta}\right), \tag{B16}$$

provided that m is a finite constant, the scaling of which is not a concern. The polynomial scaling of N in ε^{-1} is due to the fact that $C = \mathcal{O}(\eta^{-2}) = \mathcal{O}(\varepsilon^{-2k})$, as in (B7) and in the *Notation and definition* of Appendix B 1.

Now we can apply (B16) to the quantum-state properties learning as in Theorem 2 of the main text, where the target function $f_O(x) = \text{Tr}[O\rho(x)] = \sum_{i=1}^{M} f_i(x)$ is a continuous function on $\mathcal{X}$ given by the average of a local operator $O = \sum_{i=1}^{M} O_i$, as in the main text. The bound of $f_O(x)$ here is

$B = \mathcal{O}(M) = \text{poly}(n)$, provided that $|f_i(x)| \leqslant \mathcal{O}(1)$ as in the main text.

So far we have focused on approximating a single function. Next we consider to simultaneously bound the approximation errors up to ε for a set of continuous functions $\{f_i(x)\}_{i=1}^{M}$ with $x \in \mathcal{X}$, with the Fejér kernel predictor as above. This situation applies to, for example, learning the overall density matrix as in Theorem 1 and learning the quantum-state properties with locality assumptions as in Theorem 2.

Define the event $\mathcal{A}_i$ to be $|\hat{f}_{iN}(x) - f_i * F_\Lambda(x)| \leqslant \frac{\varepsilon}{2}$; then,

$$\text{Prob}\left(|\hat{f}_{iN}(x) - f_i * F_\Lambda(x)| \leqslant \frac{\varepsilon}{2}, \ \forall i \in [M]\right)$$
$$= \text{Prob.}\left(\bigcap_{i=1}^{M} \mathcal{A}_i\right)$$
$$= 1 - \text{Prob.}\left(\bigcup_{i=1}^{M} \mathcal{A}_i^c\right), \tag{B17}$$

where $\mathcal{A}_i^c =: |\hat{f}_{iN}(x) - f_i * F_\Lambda(x)| \geqslant \frac{\varepsilon}{2}$ is the complementary event of $\mathcal{A}_i$. Also, we have the following relation:

$$\text{Prob}\left(\bigcup_{i=1}^{M} \mathcal{A}_i^c\right) \leqslant \sum_{i=1}^{M} \text{Prob}(\mathcal{A}_i^c)$$
$$\leqslant M \max_{i \in [M]} \text{Prob}(\mathcal{A}_i^c)$$
$$\leqslant 2M \max_{i \in [M]} \exp\left(-\frac{N\varepsilon^2}{2\Lambda_i^{2m}B_i^2}\right), \tag{B18}$$

where the first inequality is from the fact that $\text{Prob}(A_1 \cup A_2) = \text{Prob}(A_1) + \text{Prob}(A_2) - \text{Prob}(A_1 \cap A_2) \leqslant \text{Prob}(A_1) + \text{Prob}(A_2)$, and the last inequality is derived from (B12). Lastly, by combining Eqs. (B9), (B17), and (B18), we obtain that with success probability at least $1 - \delta$, all of the functions $f_i(x)$ can be approximated by the kernel predictor up to a maximal error ε, if the kernel parameter and the number of samples are set to be

$$\Lambda^m \geqslant \frac{4B_{\max}C_{\max}^m}{\varepsilon} \tag{B19}$$

and

$$N \geqslant \frac{32B_{\max}^4 C_{\max}^{2m}}{\varepsilon^4} \ln \frac{2M}{\delta}$$
$$= \text{poly}\left(\frac{1}{\varepsilon}, \ B_{\max}, \ \ln \frac{M}{\delta}\right), \tag{B20}$$

respectively, where the respective maximal values of B and C are obtained by taking i running from 1 to M. Note that for the density matrix learning in Theorem 1, $B_{\max} = \|\rho\|_{\ell_\infty} \leqslant 1$ and $M = \mathcal{O}(2^n)$, while for the quantum-state properties learning with strong locality in Theorem 2, $B_{\max} = \mathcal{O}(1)$ and $M = \text{poly}(n)$.

In addition, the normalization property of good kernels is particularly important for the density-matrix learning problem, in keeping the trace of $\sigma_N(x)$ to be (approximately) one for all $x \in \mathcal{X}$. The Hoeffding inequality (or, more generally, the McDiarmid's inequality) [97] ensures that the trace of the predicted density matrix, given by $\text{Tr}\sigma_N(x) = \frac{1}{N} \sum_{i=1}^{N} F_\Lambda(x -$

x_i), lies in the vicinity of one with high probability. Concretely, we have

$$\text{Prob}[|\text{Tr}\sigma_N(x) - 1| \geqslant \varepsilon] \leqslant 2\exp\left(-\frac{2N\varepsilon^2}{\Lambda^{2m}}\right). \quad \text{(B21)}$$

Therefore, by taking N in Eq. (B20), it suffices to set $|\text{Tr}\sigma_N(x) - 1| \leqslant \varepsilon$ with probability at least $1 - \delta$.

So far we have proved the statements in Theorems 1 and 2 in the main text with an example of PGKs, the rectangular Fejér kernel. Alternatively, one can use a Gaussian kernel as another choice of the PGKs, in which case the index set of the kernel is not restricted to positive integers. Then the Fejér kernel in Eq. (B10) can be replaced by a Gaussian kernel, which is given by

$$K_h(x) = C_h \exp\left(-\frac{\|x\|_{\ell_2}^2}{h}\right), \quad \text{(B22)}$$

where $x \in \mathcal{X}$ and C_h is the normalization coefficient. Note that here $1/\sqrt{h}$ plays the same role as Λ in the Fejér kernel. We set $h \ll L/2$; then, $C_h = [L/(\sqrt{\pi h})]^m$. We extend the Gaussian kernel to be defined periodically outside $\mathcal{X}$, i.e., to be defined on a circle of length L in each dimension as in the case of the Fejér kernel. Then, following a similar procedure as above, it can be straightforwardly shown that for $m \geqslant 2$, the required number of samples for $|\hat{f}_N(x) - f(x)| \leqslant d_N(f) + d_\Lambda(f) \leqslant \varepsilon$, with confidence at least $1 - \delta$ [$\delta \in (0,1)$], is

$$N \geqslant \frac{2B^2 C_g^m}{\varepsilon^2} \ln\frac{2}{\delta}$$
$$= \text{poly}\left(\frac{1}{\varepsilon}, B, \ln\frac{1}{\delta}\right), \quad \text{(B23)}$$

where $C_g = \frac{mL^2}{\pi\eta^2}\ln\frac{2mB}{\varepsilon}$. So the result is similar to that with a Fejér kernel in (B16). Other examples of PGKs, which may have better performance, will be left for future work.

5. Nonuniform distributions

We have demonstrated a polynomial sample complexity with a uniform probability density, $d\mu(x) = dx/L^m$. We recall that [97] the framework of the probably approximately correct (PAC) learning also requires that the polynomial sample complexity is distribution free. So it is meaningful to explore other practical probability densities, $\varrho(x)$, with which the sample x_i is drawn from $\mathcal{X}$, denoted by $d\mu(x) = \varrho(x)dx$. Starting from the above rectangular Fejér kernel as an example, one can use a weighted Fejér kernel, given by

$$\tilde{F}_\Lambda(x) = \omega(x)F_\Lambda(x), \quad \text{(B24)}$$

where $\omega(x) = \varrho_0(x)/\varrho(x)$, with $\varrho_0(x) = 1/L^m$ the uniform density distribution. Then the weighted Fejér kernel is a PGK when the probability density $\varrho(x)$ is used, and serves as the kernel used for a similar derivation leading to a polynomial sample complexity as in the above proving procedure, provided that $\tilde{F}_\Lambda(x)$ is also upper bounded by $\mathcal{O}(\Lambda^m)$ for $x \in \mathcal{X}$, as in the case of the Fejér kernel. While this is one of the possible methods to construct PGKs for nonuniform distributions, other approaches may exist.

In summary, by combining the above Appendices B 1–B 5, we complete the unified proof for Theorems 1 and 2 in the main text.

6. Comparison with existing results

Again, we compare the number of samples above with the exponential scaling for learning the ground-state properties $f_O(x)$ in [7], with more mathematical details compared to that in the main text. The target function defined on the compact parameter space (with $L = 2$) in [7] is assumed to have a bounded first-order derivative, which means it satisfies the Lipschitz condition mentioned earlier,

$$|f_O(x) - f_O(y)| \leqslant C_L \|x - y\|_2, \quad \text{(B25)}$$

for all $x, y \in \mathcal{X}$. This is a result of the mean value theorem, where the first-order derivative of the function is upper bounded by C_L here. As a consequence, we have $\eta = \varepsilon/(4C_L)$ (or, equivalently, $k = 1$). Also, note that Ref. [7] uses an error metric of averaged square distance between the predicted and target functions, while here we used the supremum-norm distance bounded by ε. With the same notation, the required number of samples in [7] is of the order (with $C_L = 1$)

$$N_0 = \frac{B^2}{\varepsilon^2}(2m+1)^{\frac{1}{\varepsilon^2}}. \quad \text{(B26)}$$

Note that $|f_O(x)| \leqslant B/2 = \mathcal{O}(M)$ by assuming that $\|O_i\|_\infty = \mathcal{O}(1)$, and $M = \mathcal{O}(n^q)$ for the q-local operator O.

For example, taking $m = 2$, the qubit number $n = 100$, $q = 1$, and $\varepsilon = 0.1$, one can easily obtain

$$\frac{N}{N_0} \approx 10^{-48}, \quad \text{(B27)}$$

with a Fejér kernel, where the required number of samples, N, is given by (B16), and

$$\frac{N}{N_0} \approx 10^{-61}, \quad \text{(B28)}$$

with a Gaussian kernel, where the required number of samples, N, is given by (B23), respectively. *We see that when m is a small constant, very significant reductions in the provable sample complexity can be achieved here with the PGKs, as in the main text.*

APPENDIX C: BOUNDED SMOOTH PARAMETRIZATION IMPLIES CONTINUOUS DENSITY MATRICES

In Theorem 1, we assume that the entries of the density matrix are continuous over the parameter space. In this

Appendix, we show the continuity of the entries of the density matrix, provided that the operator norm of the gradient of the parametrized density operator is uniformly upper bounded. This holds for ground states of gapped Hamiltonians and other smoothly parametrized models.

Lemma 2. For a family of density operators $\rho(x)$ smoothly parametrized by x, if the gradient is upper bounded uniformly in the operator norm, i.e., $\|\partial_x \rho(x)\|_\infty \leqslant C_0$, then each entry of the density matrix is continuous over the parameter space.

Proof. Because $\partial_x \rho(x)$ is a Hermitian operator, it has a spectral decomposition given by

$$\partial_x \rho(x) = \sum_k \lambda_k |e_k\rangle\langle e_k|, \tag{C1}$$

where λ_k is the eigenvalue and $|e_k\rangle$ is the corresponding eigenvector (both of which may be x dependent).

As a result, the gradient of the entry of the density matrix is bounded by

$$
\begin{aligned}
|\partial_x \rho_{ij}(x)| &= \left| \sum_k \lambda_k \langle i|e_k\rangle\langle e_k|j\rangle \right| \\
&\leqslant \sum_k |\lambda_k| \, |\langle i|e_k\rangle\langle e_k|j\rangle| \\
&\leqslant \|\partial_x \rho(x)\|_\infty \sum_k |\langle i|e_k\rangle\langle e_k|j\rangle| \\
&\leqslant C_0 \left(\sum_k |\langle i|e_k\rangle|^2 \sum_k |\langle j|e_k\rangle|^2 \right)^{\frac{1}{2}} \\
&\leqslant C_0, \tag{C2}
\end{aligned}
$$

where $|i\rangle$ is a parameter-independent basis of the Hilbert space, and in the second to last inequality, we have used the Cauchy-Schwartz inequality and the normalization condition $\langle i|i\rangle = \sum_k |\langle i|e_k\rangle|^2 = 1 \; \forall i$.

Because the gradient of the entry is uniformly bounded by C_0 as above, it is a Lipschitz continuous function over the parameter space, which can be directly obtained via the mean value theorem. It can be directly generalized to the multiparameter case, provided that Lemma 2 holds in each dimension of the parameter space. ∎

In the following, we list a few examples where the gradient of the density operator can be uniformly bounded.

1. Ground states of gapped Hamiltonians

The gradient of the ground-state density operator for a local Hamiltonian $H(x)$ ($0 \leqslant x \leqslant 1$) with a uniform constant gap $\gamma > 0$ is given by [98]

$$\partial_x \rho(x) = i[\rho(x), D(x)],$$

with $D(x) = \int_{-\infty}^{+\infty} dt \, W_\gamma(t) \, e^{itH(x)} \, \partial_x H(x) \, e^{-itH(x)}, \tag{C3}$

where $W_\gamma(t) \in L^1(\mathbb{R})$ is a decaying weight function satisfying $\|W_\gamma\|_1 \leqslant c_0/\gamma$, for some positive constant c_0 [98].

It is straightforward to show that

$$
\begin{aligned}
\|\partial_x \rho(x)\|_\infty &\leqslant 2\|D(x)\|_\infty \|\rho(x)\|_\infty \\
&\leqslant 2 \int_{-\infty}^{+\infty} dt \, |W_\gamma(t)| \|\partial_x H(x)\|_\infty \\
&\leqslant \frac{2c_0}{\gamma} \|\partial_x H(x)\|_\infty. \tag{C4}
\end{aligned}
$$

Note that we have used that $\|\rho(x)\|_\infty \leqslant 1$ and that the operator norm is invariant under unitary transformations in the second inequality. Therefore, the gradient of the density matrix above can be bounded provided that each component of the local Hamiltonian $H(x) = \sum_j h_j(x)$ has a bounded gradient in the operator norm, i.e., $\|\partial_x h_j(x)\|_\infty \leqslant$ const.

2. States smoothly parametrized by Kraus operators

In the Kraus representation [119], starting from a parameter-independent initial state ρ_0, the parametrized state (which can be a pure or a mixed state) is given by

$$\rho(x) = \sum_i K_i(x)\rho_0 K_i^\dagger(x), \tag{C5}$$

where $K_i(x)$ are the Kraus operators satisfying $\sum_i K_i(x)K_i^\dagger(x) = \mathrm{Id}$ (identity). The gradient of the quantum states satisfies

$$\|\partial_x \rho(x)\|_\infty \leqslant 2 \sum_i \|K_i(x)\|_\infty \|\partial_x K_i(x)\|_\infty, \tag{C6}$$

where we have used the subadditivity and submultiplicity of the operator norm, and $\|\rho_0\|_\infty \leqslant 1$.

Therefore, the gradient of $\rho(x)$ can be uniformly upper bounded, as long as the operator norms of $K_i(x)$ and $\partial_x K_i(x)$ ($\forall i$) for the physical process are uniformly bounded from above.

3. Quantum parameter estimation

In quantum parameter estimation or quantum metrology, quantum Fisher information plays a central role in characterizing the fundamental precision limit (e.g., for a recent review, see [124]; see, also, [120–123]). The quantum Fisher information with respect to the single parameter x for the quantum state $\rho(x)$ reads

$$F = \mathrm{Tr}[\rho(x)\mathcal{L}^2], \tag{C7}$$

where $\mathcal{L}$ is the symmetric logarithmic derivative (SLD) operator given by

$$\partial_x \rho(x) = \tfrac{1}{2}[\rho(x)\mathcal{L} + \mathcal{L}\rho(x)]. \tag{C8}$$

Consequently, we can obtain that $\|\partial_x \rho(x)\|_\infty \leqslant \|\mathcal{L}\|_\infty$. This condition holds in quantum metrological schemes with finite amount of quantum Fisher information (e.g., without quantum phase transitions in the parameter range of interest).

APPENDIX D: REVIEW OF REPRODUCING KERNEL HILBERT SPACE (RKHS)

Here we provide a brief review of the key ideas and main results for the RKHS [100,107].

Definition 1. Let $\mathcal{X}$ be a compact space and $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, where a kernel associated with $\mathcal{H}$ is defined as $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. If $K(\mathcal{X}, \mathcal{X})$ is symmetric and positive definite, and has the following reproducing property, i.e., for all functions $f \in \mathcal{H}$, $f(x) = \langle f, K(x,)\rangle_{\mathcal{H}}$, then $\mathcal{H}$ is called the reproducing kernel Hilbert space (RKHS) of K, denoted by $\mathcal{H}_K$.

Properties of RKHS:

(1) $\forall x \in \mathcal{X}, K(x,) \in \mathcal{H}_K$.

(2) $K(x, x') = \langle K(x,), K(x',)\rangle_{\mathcal{H}_K}$, which naturally follows from (1) and the reproducing property of RKHS.

(3) The linear evaluation operator $\mathcal{E}_x$ is bounded, i.e.,

$$|\mathcal{E}_x f| = |f(x)| = \left| \langle f, K(x,)\rangle_{\mathcal{H}_K} \right|$$
$$\leqslant \sqrt{K(x, x)} \|f\|_{\mathcal{H}_K}, \tag{D1}$$

where the second equation follows from the reproducing property and the last inequality results from the Cauchy-Schwartz inequality.

Theorem 3 (Moore-Aronszajn). Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive-definite kernel. There is a unique RKHS $\mathcal{H}_K$ with the reproducing kernel K.

Universal kernels. Let $C(\mathcal{X})$ be the space of bounded continuous functions on $\mathcal{X}$. A kernel $K(\mathcal{X}, \mathcal{X})$ is called universal if the RKHS $\mathcal{H}_K$ is dense in $C(\mathcal{X})$, i.e., $\forall f \in C(\mathcal{X})$ and $\forall \varepsilon > 0$, and there is a $f^* \in \mathcal{H}_K$ such that $\|f(x) - f^*(x)\|_\infty \leqslant \varepsilon$.

Representer theorem. The optimal function $\hat{f}$ in the RKHS that minimizes the functional-norm regularized empirical loss on the training set $\mathcal{S} = \{x_i, y_i\}_{i=1}^N$,

$$\hat{f} = \arg\min_{f \in \mathcal{H}_K} \frac{1}{N} \sum_{i=1}^N \ell[f(x_i), y_i] + \lambda \|f\|_{\mathcal{H}_K}, \tag{D2}$$

takes the form

$$\hat{f}(x) = \sum_{i=1}^N \alpha_i K(x, x_i), \tag{D3}$$

where the dual variables $\{\alpha_i\} \subset \mathbb{R}$ are determined by minimizing the specific empirical loss function, and $\lambda > 0$ is a regularization parameter to avoid the overfitting. This is particularly useful for reducing the complexity of the problem, by transforming the optimization over an infinite-dimensional Hilbert space into that over a finite set of real variables.

APPENDIX E: LEARNING QUANTUM-STATE PROPERTIES IN REPRODUCING KERNEL HILBERT SPACE AND GENERALIZATION ERROR

Here we present the generalization error bound for ansatz functions from the reproducing kernel Hilbert space (RKHS) [97]. Given spaces $\mathcal{X} \subset \mathbb{R}^m$ and $Y \subset \mathbb{R}$, and a joint distribution $\mathcal{D}$ over the joint space, where $(x, y) \sim \mathcal{D}$ for $x \in \mathcal{X}$ and $y \in Y$, we are aiming at learning the continuous target

function $f_O : \mathcal{X} \to Y$, provided a sample set

$$\mathcal{S} = \{z_i = (x_i, y_i)\}_{i=1}^N = \{z_i = [x_i, f_O(x_i)]\}_{i=1}^N \tag{E1}$$

drawn from the distribution $\mathcal{D}$. An optimal function $\hat{f}_O \in \mathcal{H}_K$ can be obtained by minimizing an empirical loss defined on the sample set; then the estimator $\hat{f}_O$ will be applied to the whole space $\mathcal{X}$ for predictions.

If we take the loss to be

$$\ell[\hat{f}_O(x), y] = |\hat{f}_O(x) - f_O(x)|, \tag{E2}$$

the empirical error is given by

$$\mathcal{E}_t = \frac{1}{N} \sum_{i=1}^N |\hat{f}_O(x_i) - f_O(x_i)|, \tag{E3}$$

and the expected prediction error is

$$\mathcal{E}_p = \mathbb{E}_{(x,y) \sim \mathcal{D}} |\hat{f}(x) - y|$$
$$= \mathbb{E}_x |\hat{f}_O(x) - f_O(x)|. \tag{E4}$$

In statistical learning theory [97,100], the expected prediction error is upper bounded by the empirical error plus a term proportional to $1/\sqrt{N}$. For the model with the RKHS $\mathcal{H}_K$, it can be written as follows:

With probability at least $1 - \delta$ [$\delta \in (0, 1)$],

$$\mathcal{E}_p \leqslant \mathcal{E}_t + \mathcal{R}_{\mathcal{S}}(\mathcal{L} \circ \mathcal{H}_K) + 3\beta \sqrt{\frac{\ln \frac{2}{\delta}}{2N}}, \tag{E5}$$

where

$$\mathcal{R}_{\mathcal{S}}(\mathcal{L} \circ \mathcal{H}_K) \leqslant \mathcal{R}_{\mathcal{S}}(\mathcal{H}_K) = \frac{2\lambda_f \sqrt{\mathrm{tr}K}}{N} \tag{E6}$$

is the empirical Rademacher complexity of the loss space, and provided that $\|\hat{f}_O\|_{\mathcal{H}_K} \leqslant \lambda_f$, and $0 \leqslant \ell[\hat{f}_O(x), f_O(x)] \leqslant \beta$ for all $x \in \mathcal{X}$. Note that we have used the property of the Rademacher complexity that

$$\mathcal{R}_{\mathcal{S}}(\mathcal{L} \circ \mathcal{H}_K) \leqslant C_L \mathcal{R}_{\mathcal{S}}(\mathcal{H}_K), \tag{E7}$$

provided that $\ell(\cdot, y) \in \mathcal{L}$ is a C_L-Lipschitz continuous function ($C_L = 1$ for the loss we use here). The trace of the kernel defined on the training set satisfies $\mathrm{tr}K \leqslant NR^2$ provided that $K(x, x) \leqslant R^2$.

With the PGK estimator,

$$\hat{f}_O(x) = \frac{1}{N} \sum_{i=1}^{N} K_\Lambda(x - x_i) f_O(x_i), \qquad (E8)$$

the empirical error can be approximated as (when Λ is a large number, the PGK plays a role of the Dirac δ function)

$$\mathcal{E}_t \approx \frac{1}{N} \sum_{i=1}^{N} \left| \frac{1}{N} \sum_{j=1}^{N} K_\Lambda(x_i - x_j) - 1 \right| \times |f_O(x_i)|, \quad (E9)$$

which can be upper bounded at $O(\varepsilon)$ with large probability, given a polynomial N as in (B20), with which $\mathrm{Tr}\sigma_N(x_i)$ is close to one, up to a $\mathcal{O}(\varepsilon)$ error. Also, note that $f_O(x)$ is bounded. Once $\mathcal{E}_t$ is minimized up to $\mathcal{O}(\varepsilon)$, one needs to bound the second plus the third terms on the right-hand side of (E5) also up to $\mathcal{O}(\varepsilon)$, which will result in $\mathcal{E}_p \leqslant \mathcal{O}(\varepsilon)$, to facilitate a good generalization and prediction. Then we get the minimal number of samples required for this purpose,

$$N \geqslant \frac{1}{\varepsilon^2} \left(8\lambda_f R \sqrt{\frac{\ln\frac{2}{\delta}}{2}} \right)^2$$

$$= \mathrm{poly}\left(\frac{1}{\varepsilon}, \ B, \ \ln\frac{1}{\delta} \right), \qquad (E10)$$

as in the main text, *which indicates an efficient sample complexity*. Note that with the reproducing property for the function $\hat{f}_O(x)$, one obtains that

$$\ell[\hat{f}_O(x), f_O(x)] \leqslant |\hat{f}_O(x)| + |f_O(x)|$$

$$\leqslant \lambda_f R + \frac{B}{2}$$

$$\leqslant 2\lambda_f R =: \beta, \qquad (E11)$$

where the second inequality is obtained from (D1), and the last inequality is because R is a large number such that $\lambda_f R \geqslant B/2$. Furthermore, if the dual variables are set to be

$$\alpha_i \approx \frac{f_O(x_i)}{N}, \qquad (E12)$$

as in the PGK estimator in the main text, we have $\lambda_f = BR/2$. Also, to obtain (E10), we have assumed $\ln\frac{2}{\delta} \geqslant 2$ for a small value of δ. The result in (E10) is consistent with that obtained from (B12). Note that due to the normalization of the kernel, R can be, in general, dependent on m exponentially. This can be further improved by the dimensionality reduction in specific learning tasks.

[1] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).

[2] T. Albash and D. A. Lidar, Adiabatic quantum computation, Rev. Mod. Phys. **90**, 015002 (2018).

[3] I. M. Georgescu, S. Ashhab, and F. Nori, Quantum simulation, Rev. Mod. Phys. **86**, 153 (2014).

[4] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. D. Sawaya, S. Sim, L. Veis, and A. Aspuru-Guzik, Quantum chemistry in the age of quantum computing, Chem. Rev. **119**, 10856 (2019).

[5] S. Arora and B. Barak, *Computational Complexity: A Modern Approach* (Cambridge University Press, Cambridge, 2009).

[6] N. Abrahamsen, Solving degenerate 2D frustration-free spin systems in sub-exponential time, arXiv:2004.02850.

[7] H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, and J. Preskill, Provably efficient machine learning for quantum many-body problems, Science **377**, eabk3333 (2022).

[8] M. Schuld, I. Sinayskiy, and F. Petruccione, An introduction to quantum machine learning, Contemp. Phys. **56**, 172 (2014).

[9] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature (London) **549**, 195 (2017).

[10] V. Dunjko and H. J. Briegel, Machine learning and artificial intelligence in the quantum domain: A review of recent progress, Rep. Prog. Phys. **81**, 074001 (2018).

[11] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[12] S. Das Sarma, D.-L. Deng, and L.-M. Duan, Machine learning meets quantum physics, Phys. Today **72**(3), 48 (2019).

[13] P. Mehta, M. Bukov, C.-H. Wang, A. G.R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, Phys. Rep. **810**, 1 (2019).

[14] E. Rinaldi, X. Han, M. Hassan, Y. Feng, F. Nori, M. McGuigan, and M. Hanada, Matrix-model simulations using quantum computing, deep learning, and lattice Monte Carlo, PRX Quantum **3**, 010324 (2022).

[15] Y. Zeng, Z.-Y. Zhou, E. Rinaldi, C. Gneiting, and F. Nori, Approximate autonomous quantum error correction with reinforcement learning, Phys. Rev. Lett. **131**, 050601 (2023).

[16] E. Smolina, L. Smirnov, D. Leykam, F. Nori, and D. Smirnova, Identifying topology of leaky photonic lattices with machine learning, Nanophotonics **13**, 271 (2024).

[17] Z.-C. Shi, J.-T. Ding, Y.-H. Chen, J. Song, Y. Xia, X.X. Yi, and F. Nori, Supervised learning for robust quantum control in composite-pulse systems, Phys. Rev. Appl. **21**, 044012 (2024).

[18] X. Gao and L.-M. Duan, Efficient representation of quantum many-body states with deep neural networks, Nat. Commun. **8**, 662 (2017).

[19] J. Chen, S. Cheng, H. Xie, L. Wang, and T. Xiang, Equivalence of restricted Boltzmann machines and tensor network states, Phys. Rev. B **97**, 085104 (2018).

[20] D.-L. Deng, X. Li, and S. Das Sarma, Machine learning topological states, Phys. Rev. B **96**, 195145 (2017).

[21] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, Nat. Phys. **14**, 447 (2018).

[22] K. Choo, G. Carleo, N. Regnault, and T. Neupert, Symmetries and many-body excitations with neural-network quantum states, Phys. Rev. Lett. **121**, 167204 (2018).

[23] J. Gao, L.-F. Qiao, Z.-Q. Jiao, Y.-C. Ma, C.-Q. Hu, R.-J. Ren, A.-L. Yang, H. Tang, M.-H. Yung, and X.-M. Jin, Experimental machine learning of quantum states, Phys. Rev. Lett. **120**, 240501 (2018).

[24] F. Noé, S. Olsson, J. Köhler, and H. Wu, Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning, Science **365**, eaaw1147 (2019).

[25] N. Yoshioka and R. Hamazaki, Constructing neural stationary states for open quantum many-body systems, Phys. Rev. B **99**, 214306 (2019).

[26] M. J. Hartmann and G. Carleo, Neural-network approach to dissipative quantum many-body dynamics, Phys. Rev. Lett. **122**, 250502 (2019).

[27] A. Nagy and V. Savona, Variational quantum Monte Carlo method with a neural-network ansatz for open quantum systems, Phys. Rev. Lett. **122**, 250501 (2019).

[28] F. Vicentini, A. Biella, N. Regnault, and C. Ciuti, Variational neural-network ansatz for steady states in open quantum systems, Phys. Rev. Lett. **122**, 250503 (2019).

[29] I. A. Luchnikov, A. Ryzhov, P.-J. Stas, S. N. Filippov, and H. Ouerdane, Variational autoencoder reconstruction of complex many-body physics, Entropy **21**, 1091 (2019).

[30] K. Cranmer, S. Golkar, and D. Pappadopulo, Inferring the quantum density matrix with machine learning, arXiv:1904.05903.

[31] H. Y. Kwon, N. J. Kim, C. K. Lee, and C. Won, Searching magnetic states using an unsupervised machine learning algorithm with the Heisenberg model, Phys. Rev. B **99**, 024423 (2019).

[32] O. Sharir, Y. Levine, N. Wies, G. Carleo, and A. Shashua, Deep autoregressive models for the efficient variational simulation of many-body quantum systems, Phys. Rev. Lett. **124**, 020503 (2020).

[33] A. Glielmo, Y. Rath, G. Csányi, A. De Vita, and G. H. Booth, Gaussian process states: A data-driven representation of quantum many-body physics, Phys. Rev. X **10**, 041026 (2020).

[34] A. Borin and D. A. Abanin, Approximating power of machine-learning ansatz for quantum many-body states, Phys. Rev. B **101**, 195141 (2020).

[35] C.-Y. Park and M. J. Kastoryano, Geometry of learning neural quantum states, Phys. Rev. Res. **2**, 023232 (2020).

[36] T. Ohtsuki and T. Mano, Drawing phase diagrams of random quantum systems by deep learning the wave functions, J. Phys. Soc. Jpn. **89**, 022001 (2020).

[37] R. Zen, L. My, R. Tan, F. Hébert, M. Gattobigio, C. Miniatura, D. Poletti, and S. Bressan, Transfer learning for scalability of neural-network quantum states, Phys. Rev. E **101**, 053301 (2020).

[38] A. Barr, W. Gispen, and A. Lamacraft, Quantum ground states from reinforcement learning, PMLR **107**, 635 (2020).

[39] A. Melkani, C. Gneiting, and F. Nori, Eigenstate extraction with neural-network tomography, Phys. Rev. A **102**, 022412 (2020).

[40] A. M. Palmieri, E. Kovlakov, F. Bianchi, D. Yudin, S. Straupe, J. D. Biamonte, and S. Kulik, Experimental neural network enhanced quantum tomography, npj Quantum Inf. **6**, 20 (2020).

[41] S. Lohani, B. T. Kirby, M. Brodsky, O. Danaci, and R. T. Glasser, Machine learning assisted quantum state estimation, Mach. Learn.: Sci. Technol. **1**, 035007 (2020).

[42] M. Neugebauer, L. Fischer, A. Jäger, S. Czischek, S. Jochim, M. Weidemüller, and M. Gärttner, Neural-network quantum state tomography in a two-qubit experiment, Phys. Rev. A **102**, 042604 (2020).

[43] S. Ahmed, C. Sánchez Muñoz, F. Nori, and A. F. Kockum, Quantum state tomography with conditional generative adversarial networks, Phys. Rev. Lett. **127**, 140502 (2021).

[44] S. Ahmed, C. Sánchez Muñoz, F. Nori, and A. F. Kockum, Classification and reconstruction of optical quantum states with deep neural networks, Phys. Rev. Res. **3**, 033278 (2021).

[45] N. Yoshioka, W. Mizukami, and F. Nori, Solving quasiparticle band spectra of real solids using neural-network quantum states, Commun. Phys. **4**, 106 (2021).

[46] J. Yao, L. Lin, and M. Bukov, Reinforcement learning for many-body ground-state preparation inspired by counterdiabatic driving, Phys. Rev. X **11**, 031070 (2021).

[47] Y. Nomura, N. Yoshioka, and F. Nori, Purifying deep Boltzmann machines for thermal quantum states, Phys. Rev. Lett. **127**, 060601 (2021).

[48] Q. H. Tran and K. Nakajima, Learning temporal quantum tomography, Phys. Rev. Lett. **127**, 260401 (2021).

[49] P.-L. Zheng, S.-J. Du, and Y. Zhang, Ground-state properties via machine learning quantum constraints, Phys. Rev. Res. **4**, L032043 (2022).

[50] Y. Che, C. Gneiting, and F. Nori, Estimating the Euclidean quantum propagator with deep generative modeling of Feynman paths, Phys. Rev. B **105**, 214205 (2022).

[51] A. Anshu and S. Arunachalam, A survey on the complexity of learning quantum states, Nat. Rev. Phys. **6**, 59 (2024).

[52] L. Wang, Discovering phase transitions with unsupervised learning, Phys. Rev. B **94**, 195105 (2016).

[53] T. Ohtsuki and T. Ohtsuki, Deep learning the quantum phase transitions in random two-dimensional electron systems, J. Phys. Soc. Jpn. **85**, 123706 (2016).

[54] S. J. Wetzel, Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders, Phys. Rev. E **96**, 022140 (2017).

[55] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, Machine learning quantum phases of matter beyond the fermion sign problem, Sci. Rep. **7**, 8823 (2017).

[56] W. Hu, R. R. P. Singh, and R. T. Scalettar, Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination, Phys. Rev. E **95**, 062122 (2017).

[57] P. Ponte and R. G. Melko, Kernel methods for interpretable machine learning of order parameters, Phys. Rev. B **96**, 205146 (2017).

[58] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, Nat. Phys. **13**, 431 (2017).

[59] E. P. L. Van Nieuwenburg, Y.-H. Liu, and S. D. Huber, Learning phase transitions by confusion, Nat. Phys. **13**, 435 (2017).

[60] Y. Zhang and E.-A. Kim, Quantum loop topography for machine learning, Phys. Rev. Lett. **118**, 216401 (2017).

[61] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, Machine learning phases of strongly correlated fermions, Phys. Rev. X **7**, 031038 (2017).

[62] K. Ch'ng, N. Vazquez, and E. Khatami, Unsupervised machine learning account of magnetic transitions in the Hubbard model, Phys. Rev. E **97**, 013306 (2018).

[63] M. J. S. Beach, A. Golubeva, and R. G. Melko, Machine learning vortices at the Kosterlitz-Thouless transition, Phys. Rev. B **97**, 045207 (2018).

[64] W.-J. Rao, Z. Li, Q. Zhu, M. Luo, and X. Wan, Identifying product order with restricted Boltzmann machines, Phys. Rev. B **97**, 094207 (2018).

[65] Y.-H. Liu and E. P. L. van Nieuwenburg, Discriminative cooperative networks for detecting phase transitions, Phys. Rev. Lett. **120**, 176401 (2018).

[66] P. Zhang, H. Shen, and H. Zhai, Machine learning topological invariants with neural networks, Phys. Rev. Lett. **120**, 066401 (2018).

[67] N. Sun, J. Yi, P. Zhang, H. Shen, and H. Zhai, Deep learning topological invariants of band insulators, Phys. Rev. B **98**, 085402 (2018).

[68] P. Huembeli, A. Dauphin, and P. Wittek, Identifying quantum phase transitions with adversarial neural networks, Phys. Rev. B **97**, 134109 (2018).

[69] N. Yoshioka, Y. Akagi, and H. Katsura, Learning disordered topological phases by statistical recovery of symmetry, Phys. Rev. B **97**, 205110 (2018).

[70] L. Pilozzi, F. A. Farrelly, G. Marcucci, and C. Conti, Machine learning inverse problem for topological photonics, Commun. Phys. **1**, 57 (2018).

[71] J. Venderley, V. Khemani, and E.-A. Kim, Machine learning out-of-equilibrium phases of matter, Phys. Rev. Lett. **120**, 257204 (2018).

[72] W. Zhang, J. Liu, and T.-C. Wei, Machine learning of phase transitions in the percolation and XY models, Phys. Rev. E **99**, 032142 (2019).

[73] P. Huembeli, A. Dauphin, P. Wittek, and C. Gogolin, Automated discovery of characteristic features of phase transitions in many-body localization, Phys. Rev. B **99**, 104106 (2019).

[74] X.-Y. Dong, F. Pollmann, and X.-F. Zhang, Machine learning of quantum phase transitions, Phys. Rev. B **99**, 121104(R) (2019).

[75] B. S. Rem, N. Käming, M. Tarnowski, L. Asteria, N. Fläschner, C. Becker, K. Sengstock, and C. Weitenberg, Identifying quantum phase transitions using artificial neural networks on experimental data, Nat. Phys. **15**, 917 (2019).

[76] K. Liu, J. Greitemann, and L. Pollet, Learning multiple order parameters with interpretable machines, Phys. Rev. B **99**, 104410 (2019).

[77] C. Giannetti, B. Lucini, and D. Vadacchino, Machine learning as a universal tool for quantitative investigations of phase transitions, Nucl. Phys. B **944**, 114639 (2019).

[78] A. Canabarro, F. F. Fanchini, A. L. Malvezzi, R. Pereira, and R. Chaves, Unveiling phase transitions with machine learning, Phys. Rev. B **100**, 045129 (2019).

[79] F. Schäfer and N. Lörch, Vector field divergence of predictive model output as indication of phase transitions, Phys. Rev. E **99**, 062107 (2019).

[80] S. Durr and S. Chakravarty, Unsupervised learning eigenstate phases of matter, Phys. Rev. B **100**, 075102 (2019).

[81] W. Lian, S.-T. Wang, S. Lu, Y. Huang, F. Wang, X. Yuan, W. Zhang, X. Ouyang, X. Wang, X. Huang, L. He, X. Chang, D.-L. Deng, and L. Duan, Machine learning topological phases with a solid-state quantum simulator, Phys. Rev. Lett. **122**, 210503 (2019).

[82] J. F. Rodriguez-Nieva and M. S. Scheurer, Identifying topological order through unsupervised machine learning, Nat. Phys. **15**, 790 (2019).

[83] Y. Che, C. Gneiting, T. Liu, and F. Nori, Topological quantum phase transitions retrieved through unsupervised machine learning, Phys. Rev. B **102**, 134213 (2020).

[84] M. S. Scheurer and R.-J. Slager, Unsupervised machine learning and band topology, Phys. Rev. Lett. **124**, 226401 (2020).

[85] Y. Long, J. Ren, and H. Chen, Unsupervised manifold clustering of topological phononics, Phys. Rev. Lett. **124**, 185501 (2020).

[86] A. Lidiak and Z. Gong, Unsupervised machine learning of quantum phase transitions using diffusion maps, Phys. Rev. Lett. **125**, 225701 (2020).

[87] O. Balabanov and M. Granath, Unsupervised learning using topological data augmentation, Phys. Rev. Res. **2**, 013354 (2020).

[88] E. Greplova, A. Valenti, G. Boschung, F. Schäfer, N. Lörch, and S. D. Huber, Unsupervised identification of topological phase transitions using predictive models, New J. Phys. **22**, 045003 (2020).

[89] A. Berezutskii, M. Beketov, D. Yudin, Z. Zimborás, and J. D. Biamonte, Probing criticality in quantum spin chains with neural networks, J. Phys. Complex. **1**, 03LT01 (2020).

[90] L.-W. Yu and D.-L. Deng, Unsupervised learning of non-hermitian topological phases, Phys. Rev. Lett. **126**, 240402 (2021).

[91] T. Mendes-Santos, X. Turkeshi, M. Dalmonte, and A. Rodriguez, Unsupervised learning universal critical behavior via the intrinsic dimension, Phys. Rev. X **11**, 011040 (2021).

[92] T. Mendes-Santos, A. Angelone, A. Rodriguez, R. Fazio, and M. Dalmonte, Intrinsic dimension of path integrals: Data-mining quantum criticality and emergent simplicity, PRX Quantum **2**, 030332 (2021).

[93] Y. Long and B. Zhang, Unsupervised data-driven classification of topological gapped systems with symmetries, Phys. Rev. Lett. **130**, 036601 (2023).

[94] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).

[95] H.-Y. Huang, R. Kueng, and J. Preskill, Efficient estimation of Pauli observables by derandomization, Phys. Rev. Lett. **127**, 030503 (2021).

[96] L. G. Valiant, A theory of the learnable, Commun. ACM **27**, 1134 (1984).

[97] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed., Adaptive Computation and Machine Learning (MIT Press, Cambridge, MA, 2018).

[98] S. Bachmann, S. Michalakis, B. Nachtergaele, and R. Sims, Automorphic equivalence within gapped phases of quantum lattice systems, Commun. Math. Phys. **309**, 835 (2012).

[99] E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction* (Princeton University Press, Princeton, NJ, 2003).

[100] I. Steinwart and A. Christmann, *Support Vector Machines* (Springer, New York, 2008).

[101] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, Proc. Natl. Acad. Sci. USA **102**, 7426 (2005).

[102] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators, in *Advances in Neural Information Processing Systems*, edited by Y. Weiss, B. Schölkopf, and J. C. Platt (MIT Press, Cambridge, 2006), pp. 955–962.

[103] R. Chatterjee and T. Yu, Generalized coherent states, reproducing kernels, and quantum support vector machines, Quantum Inf. Commun. **17**, 1292 (2017).

[104] K. Bartkiewicz, C. Gneiting, A. Černoch, K. Jiráková, K. Lemr, and F. Nori, Experimental kernel-based quantum machine learning in finite feature space, Sci. Rep. **10**, 12356 (2020).

[105] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems*, edited by S. Bengio *et al.* (Curran Associates, Inc., 2018), Vol. 31, pp. 8571–8580.

[106] H. Q. Minh, L. Bazzani, and V. Murino, A unifying framework in vector-valued reproducing kernel Hilbert spaces for manifold regularization and co-regularized multi-view learning, J. Mach. Learn. Res. **17**, 1 (2016).

[107] M. Schuld, Supervised quantum machine learning models are kernel methods, arXiv:2101.11020.

[108] P. Jordan and E. Wigner, Über das Paulische Äquivalenzverbot, Z. Phys. **47**, 631 (1928).

[109] S. B. Bravyi and A. Y. Kitaev, Fermionic quantum computation, Ann. Phys. **298**, 210 (2002).

[110] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, Quantum state tomography via compressed sensing, Phys. Rev. Lett. **105**, 150401 (2010).

[111] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators, New J. Phys. **14**, 095022 (2012).

[112] For the definition and more details about continuous functions here, see the paragraph *Notation and definition* in Appendix B 1.

[113] Note that in the usual definition of good kernels [99], there is no requirement for the positivity, but a L_1-boundness condition of the kernel is necessary, i.e., $\int_{x \in \mathcal{X}} |K_\Lambda(x)| d\mu(x) \leqslant \mathcal{M}$ as $\Lambda \to \infty$, with $\mathcal{M}$ a positive constant. Here, for the density-matrix learning, the positivity and normalization conditions together lead to the L_1-boundness, with $\mathcal{M} = 1$. Therefore, we do not write this condition explicitly.

[114] L. Pfister and Y. Bresler, Bounding multivariate trigonometric polynomials, IEEE Trans. Signal Process. **67**, 700 (2019).

[115] A. De Pasquale and P. Facchi, *XY* model on the circle: Diagonalization, spectrum, and forerunners of the quantum phase transition, Phys. Rev. A **80**, 032102 (2009).

[116] M. Okuyama, Y. Yamanaka, H. Nishimori, and M. M. Rams, Anomalous behavior of the energy gap in the one-dimensional quantum *XY* model, Phys. Rev. E **92**, 052116 (2015).

[117] S. Sachdev, *Quantum Phase Transitions* (Cambridge University Press, Cambridge, 2011).

[118] E. Barouch and B. M. McCoy, Statistical mechanics of the *XY* model. II. spin-correlation functions, Phys. Rev. A **3**, 786 (1971).

[119] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2012).

[120] C. W. Helstrom, *Quantum Detection and Estimation Theory* (Academic, New York, 1976).

[121] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam, 1982).

[122] J. Ma, X. Wang, C.P. Sun, and F. Nori, Quantum spin squeezing, Phys. Rep. **509**, 89 (2011).

[123] Y. Che, J. Liu, X.-M. Lu, and X. Wang, Multiqubit matter-wave interferometry under decoherence and the Heisenberg scaling recovery, Phys. Rev. A **99**, 033807 (2019).

[124] J. Liu, H. Yuan, X.-M. Lu, and X. Wang, Quantum Fisher information matrix and multiparameter estimation, J. Phys. A: Math. Theor. **53**, 023001 (2020).

[125] E. Rinaldi, M. González Lastre, S. García Herreros, S. Ahmed, M. Khanahmadi, F. Nori, and C. Sánchez Muñoz, Parameter estimation from quantum-jump data using neural networks, Quantum Sci. Technol. **9**, 035018 (2024).

[126] L. Lewis, H.-Y. Huang, V. T. Tran, S. Lehner, R. Kueng, and J. Preskill, Improved machine learning algorithm for predicting ground state properties, Nat. Commun. **15**, 895 (2024).

[127] E. Onorati, C. Rouzé, D. S. França, and J. D. Watson, Efficient learning of ground and thermal states within phases of matter, arXiv:2301.12946.