SCIENCE CHINA
Information Sciences

• **Supplementary File** •

A hybrid quantum-classical Hamiltonian learning algorithm

Youle WANG[1,2*], Guangxi LI[1,2*] & Xin WANG[2*]

[1]*Center for Quantum Software and Information, University of Technology Sydney, NSW 2007, Australia;*
[2]*Institute for Quantum Computing, Baidu Research, Beijing 100193, China*

1 Introduction

Hamiltonian learning is a central task in studying quantum physics systems and the experimental realization of quantum computers. For instance, it can predict the quantum system's locality to describe the effective interactions between particles, which plays a crucial role in quantum technology, such as quantum lattice models [68], quantum simulation [53], and adiabatic quantum computation [2]. Moreover, with recent experimental advances in tools for studying complex interacting quantum systems [5], it is becoming more and more essential to learn the dynamics of complicated physical systems, which can predict the evolution of any initial state governed by the Hamiltonian. Another critical utility is relevant to the verification of quantum devices and simulators towards building fault-tolerant quantum computers [40] since certifying that the engineered Hamiltonian matches the theoretically predicted models will always be an indispensable step in developing high-fidelity quantum gates [58].

Hamiltonian of many-body physics is often characterized by some parameters, which describe the interactions between the particles. Technically, a many-body Hamiltonian is composed of polynomially many local Pauli operators, i.e.,

$$H = \sum_{\ell=1}^{m} \mu_\ell E_\ell, \tag{1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m) \in [-1, 1]^m$, and $\{E_\ell\}_{\ell=1}^m$ are n-qubit local Pauli operators, with $m = O(poly(n))$. Despite the number of these parameters $\boldsymbol{\mu}$ in general scales polynomially in the system's size, it is pretty challenging to learn these parameters. Classically characterizing the system's Hamiltonian via tomography would require resources that exponentially scale in the system's size [24]. Other than tomography, there are methods [23, 59, 62–64] that cost polynomially many resources while requiring the ability to simulate the dynamics of the system, which is classically intractable. In particular, it is difficult to perform quantum simulation as a large amount of low-decoherence and fully-connected qubits are required, which are not available on NISQ devices [50]. Moreover, recent work [6] proposes inferring the spatially local Hamiltonian given a single state commuting with the Hamiltonian, which requires solving a linear system relevant to the covariance matrix.

The major goal of this paper is to learn the many-body Hamiltonians using a trusted NISQ device. For this purpose, we exploit the variational quantum algorithms (VQAs) that have been gaining popularity in many areas [10, 13, 29, 41, 47, 49, 54, 60, 61, 66]. VQAs are a class of hybrid quantum-classical algorithms that are expected to be implementable on NISQ devices. The main process is to optimize a certain loss function via parameterized quantum circuits (PQCs). In particular, the loss function depending on parameters of the circuit is evaluated on quantum devices, and then the parameters are updated using gradient-based methods classically. As for Hamiltonian learning, we take advantage of the strategy proposed recently in [3], which allows recovering parameters $\boldsymbol{\mu}$ from the measurement results of a quantum

* Corresponding author (email: youle.wang@student.uts.edu.au, guangxi.li@student.uts.edu.au, wangxin73@baidu.com)

Gibbs state $\rho_\beta = e^{-\beta H}/\operatorname{Tr}(e^{-\beta H})$, i.e., $e_\ell = \operatorname{Tr}(\rho_\beta E_\ell)$ for all $\ell = 1, \ldots, m$. It has been shown that solving the optimization problem below suffices to complete the Hamiltonian learning task.

$$\boldsymbol{\mu} = \operatorname{argmin}_{\boldsymbol{\nu}} \log Z_\beta(\boldsymbol{\nu}) + \beta \sum_{\ell=1}^{m} \nu_\ell e_\ell. \tag{2}$$

Here, $Z_\beta(\boldsymbol{\nu}) = \operatorname{Tr}(e^{-\beta \sum_{\ell=1}^{m} \nu_\ell E_\ell})$ denotes the partition function, parameterized by $\boldsymbol{\nu} = (\nu_1, ..., \nu_m) \in [-1, 1]^m$, and β denotes the inverse temperature of the system.

In this paper, we propose a hybrid quantum-classical algorithm to perform the Hamiltonian learning task, whose aim is to recover the interaction coefficients $\boldsymbol{\mu}$ from the measurement results $\{e_\ell\}_{\ell=1}^m$. The main idea is to solve the optimization problem in Eq. (2) by a gradient-descent method and compute the corresponding gradients utilizing variational quantum algorithms. The challenge of our approach is to compute the log-partition function $\log Z_\beta(\boldsymbol{\nu})$ and its gradient since computing partition function is #P-hard [22, 44].

To overcome this challenge, we accordingly develop a method based on the relation between the log-partition function and the system's free energy. In general, suppose the state of the system is ρ, then the free energy is given by $F(\rho) = \operatorname{Tr}(H\rho) - \beta^{-1} S(\rho)$, where $S(\rho)$ is the von Neumann entropy. The relation states that the global minimum of $F(\rho)$ is proportional to the log-partition function, i.e.,

$$\log \operatorname{Tr}(e^{-\beta H}) = -\beta \min_{\rho} F(\rho). \tag{3}$$

To establish the results, our method for minimizing the free energy depends on two critical steps. First, we choose a suitable PQC with enough expressiveness and train it to learn the eigenvectors of the Hamiltonian and output the corresponding eigenvalues. Second, we combine the post-training PQC with the classical methods for convex optimization to find the global minimum of the free energy. Next, we utilize the post-training PQC and the optimizer of the convex optimization to compute the gradients. Furthermore, we theoretically analyze the estimation precision of the gradients. We also show the efficiency of loss evaluation and gradients estimation by the importance sampling technique when the underlying Hamiltonian is large.

As the proof of principle, we study the effectiveness of our algorithm for Hamiltonian learning by conducting numerical experiments for randomly generated Hamiltonians and several many-body Hamiltonians. To generate random Hamiltonians, we choose Pauli tensor products E_ℓ from the set $\{X, Y, Z, I\}^{\otimes n}$ at random, with n ranging from 3 to 5. The target interaction coefficients are chosen via a uniform distribution over $[-1, 1]$. The tested many-body Hamiltonians consist of Ising, XY-spin, and Heisenberg models, where size also varies from 3 qubits to 5. For these Hamiltonians, we test our algorithm for different parameters β and $\boldsymbol{\mu}$ with different lengths. As a result, the numerical results show that the target interaction coefficients can be estimated with high precision. In these experiments, our algorithm learns all eigenvalues of Hamiltonians. Moreover, we show the effectiveness by partially learning few smallest eigenvalues of Ising Hamiltonians. In particular, the circuit depth of used PQC could be significantly reduced. Finally, we also generalize the experiments to larger Ising Hamiltonians with 6/7 qubits.

Next, we summarize the contribution of this paper and all mentioned results above.

1. We propose a hybrid quantum-classical Hamiltonian learning framework based on the fundamental properties of free energy, which mainly consists of the following two subroutines: log-partition function estimation and stochastic variational quantum eigensolver (SVQE).
2. The main subroutine is the log-partition function estimation algorithm, which combines the SVQE with the classical convex optimization to minimize the free energy.
3. We also propose a feasible scheme for learning the spectrum of the many-body Hamiltonian by integrating variational quantum algorithms with the importance sampling technique.
4. We demonstrate our algorithm's validity by numerical simulations on several random Hamiltonians and many-body Hamiltonians (e.g., Ising model, XY model, and Heisenberg model).

Organization. The remaining paper proceeds as follows. In Sec. 2, we formally define the problems we studied in this work; In Sec. 3, we present the main results, including the Hamiltonian learning algorithm, and its main subroutines log-partition function estimation, stochastic variational quantum eigensolver, and gradient estimation; In Sec. 4, we describe the experimental settings and provide numerical results to demonstrate the efficacy of our algorithm; Lastly, we conclude the paper in Sec. 5. Detailed proofs are presented in the supplementary file.

2 Problem Statement

In this paper, the goal of Hamiltonian learning is to learn the interaction coefficients $\boldsymbol{\mu}$ from the measurement results of a quantum Gibbs state. We assume that the Hamiltonian to be learned H is composed of local Pauli operators $\{E_\ell\}_{\ell=1}^m$, and the measurements corresponding to $\{E_\ell\}_{\ell=1}^m$ are performed on the Gibbs state $\rho_\beta = e^{-\beta H}/\operatorname{Tr}(e^{-\beta H})$ at an inverse temperature β. The measurement results are denoted by $\{e_\ell\}_{\ell=1}^m$, given by

$$e_\ell = \operatorname{Tr}(\rho_\beta E_\ell), \quad \forall \ell \in [m]. \tag{4}$$

Recently, there are many methods proposed to efficiently obtain measurement results $\{e_\ell\}_{\ell=1}^m$ [8,17,30]. We, therefore, assume the measurement results $\{e_\ell\}_{\ell=1}^m$ have been given previously and focus on learning interaction coefficients from them. Formally, we define the Hamiltonian learning problem (HLP) as follows:

Definition 1 (HLP)**.** Consider a many-body Hamiltonian with a decomposition given in Eq. (1), where $|\mu_\ell| \leqslant 1$ for all $\ell = 1, ..., m$. Suppose we are given measurement results $\{e_\ell\}_{\ell=1}^m$ of the quantum Gibbs state ρ_β, then the goal is to find an estimate $\widehat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ such that

$$\| \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} \|_\infty \leqslant \epsilon, \tag{5}$$

where $\| \cdot \|_\infty$ means the maximum norm.

To solve the HLP, we adopt a strategy that is proposed recently in Ref. [3], which transforms HLP into an optimization problem by using the Jaynes' principle (or maximal entropy principle) [32]. This strategy is to find a quantum state with the maximal entropy from all states whose measurement results under $\{E_\ell\}_{\ell=1}^m$ match $\{e_\ell\}_{\ell=1}^m$.

$$\begin{aligned} \max_{\rho} \quad & S(\rho) \\ s.t. \quad & \operatorname{Tr}(\rho E_\ell) = e_\ell, \ \forall \ell = 1, ..., m \\ & \rho > 0, \ \operatorname{Tr}(\rho) = 1. \end{aligned} \tag{6}$$

It has been shown in [32] that the optimal state is of the following form:

$$\sigma = \frac{\exp(-\beta \sum_{\ell=1}^m \mu_\ell^* E_\ell)}{\operatorname{Tr}(\exp(-\beta \sum_{\ell=1}^m \mu_\ell^* E_\ell))}. \tag{7}$$

Here, state σ is a quantum Gibbs state of a Hamiltonian with interaction coefficients $\boldsymbol{\mu}^* = (\mu_1^*, ..., \mu_m^*)$. As a result, Ref. [3] shows that coefficients of σ is the target interaction coefficients, i.e., $\boldsymbol{\mu}^* = \boldsymbol{\mu}$. Moreover, Ref. [3] also points out an approach for obtaining $\boldsymbol{\mu}^*$ that is to solve the dual optimization problem (cf. Eq. (2)). More specifically, the dual problem in Eq. (2) is a result of applying the Lagrange multiplier method to problem in Eq. (6), and coefficients $\boldsymbol{\nu}$ are the corresponding Lagrange multipliers.

In Ref. [3], it has shown that the loss function is strongly convex. Hence, a gradient descent method can steadily find the desired solution and recover the unknown Hamiltonian. To this end, we develop a gradient descent method tailored to quantum computers to solve the problem in Eq. (2). A flowchart for illustration is shown in Figure 1. Clearly, the main obstacle is to compute the corresponding gradients of the objective function, which involves computing the partition function. Then, we formalize the gradient estimation problem below.

Definition 2 (Gradient estimation)**.** Given a Hamiltonian parameterized by coefficients $\boldsymbol{\nu}$, i.e., $H(\boldsymbol{\nu}) = \sum_{\ell=1}^m \nu_\ell E_\ell$, let $L(\boldsymbol{\nu})$ be the objective function

$$L(\boldsymbol{\nu}) = \log Z_\beta(\boldsymbol{\nu}) + \beta \sum_{\ell=1}^m \nu_\ell e_\ell, \tag{8}$$

where $Z_\beta(\boldsymbol{\nu}) = \operatorname{Tr}(e^{-\beta H(\boldsymbol{\nu})})$. Then the goal is to estimate the gradient $\nabla L(\boldsymbol{\nu})$ with respect to $\boldsymbol{\nu}$.

According to Eq. (3), one way to compute the loss function $L(\boldsymbol{\nu})$ and the gradient requires the ability to prepare the Gibbs states. In the literature, there are many proposals [19, 31, 61, 65–67] of Gibbs state preparation to this end. In this paper, we propose a method that does not demand the Gibbs state preparation. The following sections are devoted to solving HLP and the Gradient estimation problem.

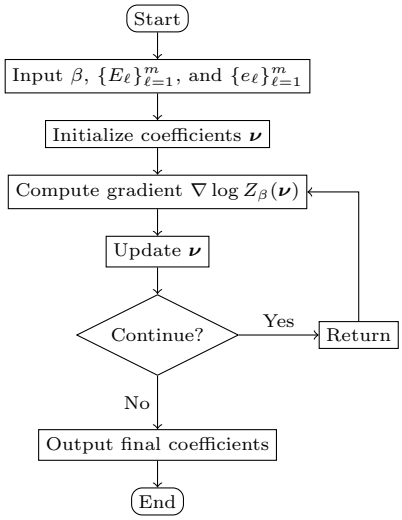

Figure 1 Flowchart of the gradient-descent method for Hamiltonian learning.

3 Main results

This section presents the main results of this paper. Specifically, we first discuss the core idea and outline the framework for computing the log-partition function in Sec. 3.1. In Sec. 3.2, we provide a variational quantum algorithm for learning the eigenvectors of Hamiltonians. Based on the results in Sec 3.1-3.2, we then proceed to give the gradient estimation procedure in Sec. 3.3. Last, Sec. 3.4 provides the main algorithm, the hybrid quantum-classical Hamiltonian learning algorithm (HQHL).

3.1 Log-partition function estimation

Here, we consider computing the log-partition function $\log Z_\beta(\boldsymbol{\nu})$. Motivating our method is the relationship between the log-partition function and free energy. Recall that free energy of the system being state ρ is given by $F(\rho) = \mathrm{Tr}(H(\boldsymbol{\nu})\rho) - \beta^{-1}S(\rho)$, assuming the parameterized Hamiltonian is $H(\boldsymbol{\nu}) = \sum_{\ell=1}^m \nu_\ell E_\ell$. Then the relation states that

$$\log Z_\beta(\boldsymbol{\nu}) = -\beta \min_\rho F(\rho). \tag{9}$$

As shown in Eq. (9), it is natural to minimize the free energy to obtain the value of $\log Z_\beta(\boldsymbol{\nu})$. However, it is infeasible to directly minimize the free energy on NISQ devices since performing entropy estimation with even shallow circuits is difficult [21]. To deal with this issue, we choose an alternate version of Eq. (9):

$$\log Z_\beta(\boldsymbol{\nu}) = -\beta \min_{\mathbf{p}} \sum_{j=1}^N p_j \cdot \lambda_j + \beta^{-1} \sum_{j=1}^N p_j \log p_j, \tag{10}$$

where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_N)$ is the vector of eigenvalues of $H(\boldsymbol{\nu})$, and $\mathbf{p} = (p_1, ..., p_N)$ represents an N-dimensional probability distribution, with N the Hamiltonian's dimension. Please note that proofs for Eqs. (9)-(10) are provided in the supplementary file. Thus, optimizing the R.H.S of Eq. (10) could obtain the desired quantity and avoid the von Neumann entropy estimation simultaneously, assuming eigenvalues of the Hamiltonian $H(\boldsymbol{\nu})$ is given previously. As a result, our task is reduced to solve the following optimization program based on the equality in Eq. (10):

$$\begin{aligned} \min_{\mathbf{p}} \quad & C(\mathbf{p}) \\ \text{s.t.} \quad & \sum_{j=1}^N p_j = 1 \\ & p_j \geqslant 0, \forall j = 1, \ldots, N \end{aligned} \tag{11}$$

where

$$C(\mathbf{p}) = \sum_{j=1}^{N} p_j \cdot \lambda_j + \beta^{-1} \sum_{j=1}^{N} p_j \log p_j. \tag{12}$$

The optimization program in Eq. (11) is a typical convex optimization program. In the context of convex optimization, there are many classical algorithms to solve the optimization program, such as the interior-point method [37], ellipsoid method [25], cutting-plane method [38], and random walks [35], etc. For example, we consider using the cutting plane method [33, 43], which requires the membership and evaluation procedures [42]. Concerning the program in Eq. (11), the membership procedure determines whether a point belongs to the set of probability distributions, and the evaluation procedure takes in a probability distribution $\mathbf{p}$ and returns the value $C(\mathbf{p})$ with high accuracy. Clearly, it is easy to determine whether the given point is a probability distribution while challenging to efficiently evaluate the function value. Thus, we provide a procedure to solve the convex optimization problem as well as overcome this challenge at the same time in Algorithm 1.

Algorithm 1 Log-partition function estimation

Require: Parameterized quantum circuit $U(\boldsymbol{\theta})$, Hamiltonian $H(\boldsymbol{\nu})$, constant β;
Ensure: An estimate for $\log Z_\beta(\boldsymbol{\nu})$;
1: `# Evaluation procedure construction`
2: Take probability distribution $\mathbf{p}$ as input;
3: Set integer T and D;
4: Sample TD integers $t_1^1, ..., t_T^1, ..., t_1^D, ..., t_T^D$ according to $\mathbf{p}$;
5: Prepare computational states $|\psi_{t_1^1}\rangle$, ..., $|\psi_{t_T^1}\rangle$, ..., $|\psi_{t_1^D}\rangle$, ..., $|\psi_{t_T^D}\rangle$;
6: Compute approximate eigenvalues: $\lambda_{t_j^s} = \langle\psi_{t_j^s}|U^\dagger(\boldsymbol{\theta})H(\boldsymbol{\nu})U(\boldsymbol{\theta})|\psi_{t_j^s}\rangle$ for all $j = 1, \ldots, T$ and $s = 1, \ldots, D$;
7: Compute averages: $ave_s = \frac{1}{T}\sum_{j=1}^{T} \lambda_{t_j^s}$ for all $s = 1, ..., D$;
8: Take the median value $C(\mathbf{p}) \leftarrow \text{median}(ave_1, ..., ave_D) + \beta^{-1}\sum_{j=1}^{N} p_j \log p_j$;
9: `# Membership procedure construction`
10: Construct a membership procedure;
11: `# Convex optimization solution`
12: Compute the function's global minimum value $C(\mathbf{p}^*)$ and the optimal point $\mathbf{p}^*$ via the cutting plane method.
13: **return** value $-\beta C(\mathbf{p}^*)$ and the final point $\mathbf{p}^*$.

In Algorithm 1, we compute the log-partition function using a classical convex optimization method. For this purpose, we first show the construction process of evaluation procedure. That is, given a point $\mathbf{p}$, find an estimate for $C(\mathbf{p})$. We assume we are given a parameterized quantum circuit $U(\boldsymbol{\theta})$ that can learn eigenvectors of the Hamiltonian $H(\boldsymbol{\nu})$. In our approach, the $U(\boldsymbol{\theta})$ is combined with the importance sampling technique (cf. lines 3-8) to deal with the large-sized Hamiltonians. Specifically, i) we sample TD indices according to the distribution $\mathbf{p}$ (cf. line 4); ii) we evaluate the eigenvalues associated with the sampled indices (cf. lines 5-6); iii) we take the average over T (cf. line 7) and the median over D (cf. line 8) to evaluate the function value $C(\mathbf{p})$ with high accuracy and success probability. Eventually, with the evaluation procedure and the membership procedure, the global minimum of $C(\mathbf{p})$ could be obtained via the cutting plane method [33, 42, 43]. Finally, based on the relationship between $\log Z_\beta(\boldsymbol{\nu})$ and $C(\mathbf{p}^*)$ (cf. Eq. (10)), we could derive the log-partition function value. Here $\mathbf{p}^*$ denotes the optimal distribution of the optimization in Eq. (10)).

Remark 1. Notice that a crucial gadget in Algorithm 1 is the PQC $U(\boldsymbol{\theta})$, which we have assumed to be accessible. To complement the assumption, we provide a procedure for extracting eigenvalues in the next section, *Stochastic variational quantum eigensolver*. In particular, we will prsent a variational quantum algorithm for learning the eigenvectors of the parameterized Hamiltonians.

Now we discuss the cost of applying Algorithm 1. As the efficiency of Algorithm 1 mainly relies on the cost of the evaluation procedure, we only discuss it here. Suppose we have access to Hamiltonian $H(\boldsymbol{\nu})$'s eigenvalues $\boldsymbol{\lambda}$, then the objective function $C(\mathbf{p})$ can be effectively evaluated. Recall that $C(\mathbf{p})$ contains two parts $\sum_{j=1}^{N} p_j \cdot \lambda_j$ and $\beta^{-1}\sum_{j=1}^{N} p_j \log p_j$. On the one hand, the latter value can be computed immediately since $\mathbf{p}$ is stored on classical devices. On the other hand, value $\sum_{j=1}^{N} p_j \cdot \lambda_j$ can be regarded as an expectation of the probability $\mathbf{p}$, where value λ_j is sampled with probability p_j. Notably, the total cost for estimating $C(\mathbf{p})$ is dominated by the number of samples. Then we analyze the number of required samples for loss evaluation in Proposition 1.

Proposition 1. For any constant $\beta > 0$ and parameterized Hamiltonian $H(\boldsymbol{\nu}) = \sum_{\ell=1}^{m} \nu_\ell E_\ell$ with $E_\ell \in \{X, Y, Z, I\}^{\otimes n}$ and $\boldsymbol{\nu} \in \mathbb{R}^m$, suppose we are given access to a parameterized quantum circuit $U(\boldsymbol{\theta})$ that can prepare $H(\boldsymbol{\nu})$'s eigenvectors, then the objective function $C(\mathbf{p})$ can be computed up to precision ϵ with probability larger than 2/3 by taking $T = O(m\|\boldsymbol{\nu}\|_2^2/\epsilon^2)$ samples. Furthermore, the probability can be improved to $1-\eta$ costing an additional multiplicative factor of $D = O(\log(1/\eta))$.

Sketch of Proof. In general, the expectation can be approximated by the sample mean according to Chebyshev's inequality. Specifically speaking, the expectation can be estimated up to precision ϵ with high probability (e.g., larger than 2/3) by taking $O(\mathbf{Var}/\epsilon^2)$ samples, where **Var** denotes the variance of the distribution. Here, the number of samples is $T = O(m\|\boldsymbol{\nu}\|_2^2/\epsilon^2)$, since the variance is bounded by the squared spectral norm of $H(\boldsymbol{\nu})$, which is less than $\sqrt{m}\|\boldsymbol{\nu}\|_2$. Furthermore, Chernoff bounds allow improving success probability to $1-\eta$ at an additional cost of a multiplicative factor of $D = O(\log(1/\eta))$.

As shown in Proposition 1, our evaluation method is computationally efficient, since the number of samples scales polynomially with the number of qubits. Hence Algorithm 1 could be applied to compute the partition function of the parameterized Hamiltonian, given the suitable PQC $U(\boldsymbol{\theta})$.

3.2 Stochastic variational quantum eigensolver

This section discusses learning the eigenvectors of the parameterized Hamiltonian $H(\boldsymbol{\nu})$ using variational quantum algorithms and the importance sampling technique. First, we outline the algorithm in Algorithm 2 and then discuss the fundamental theory. Second, we circumvent the cost for coping with large-scaled Hamiltonians by the importance sampling technique. We also analyze the cost of loss evaluation in the algorithm.

Algorithm 2 Stochastic variational quantum eigensolver (SVQE)

Require: Parameterized quantum circuit $U(\boldsymbol{\theta})$, Hamiltonian $H(\boldsymbol{\nu})$, and weights $\mathbf{q}$;
Ensure: Optimal PQC $U(\boldsymbol{\theta})$;
Set number of iterations I and $l = 1$;
Set integers T and D;
Set learning rate r_θ;
Set probability distribution $\mathbf{q}$;
Sample TD integers $k_1^1,\ldots,k_T^1,\ldots,k_1^D,\ldots,k_T^D$ according to $\mathbf{q}$;
Prepare computational states $|\psi_{k_1^1}\rangle, \ldots, |\psi_{k_T^1}\rangle, \ldots, |\psi_{k_1^D}\rangle, \ldots, |\psi_{k_T^D}\rangle$;
while $l \leqslant I$ **do**
  Compute value $\langle\psi_{k_j^s}|U^\dagger(\boldsymbol{\theta})H(\boldsymbol{\nu})U(\boldsymbol{\theta})|\psi_{k_j^s}\rangle$ for all $j = 1,\ldots,T$ and $s = 1,\ldots,D$;
  Compute averages: $ave_s = \frac{1}{T}\sum_{j=1}^{T}\langle\psi_{k_j^s}|U^\dagger(\boldsymbol{\theta})H(\boldsymbol{\nu})U(\boldsymbol{\theta})|\psi_{k_j^s}\rangle$ for all $s = 1, ..., D$;
  Let $M(\boldsymbol{\theta}) \leftarrow \text{median}(ave_1, ..., ave_D)$;
  Use $M(\boldsymbol{\theta})$ to compute the gradient ∇ by parameter shift rules [46];
  Update parameters $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - r_\theta\nabla$;
  Set $l \leftarrow l + 1$;
end while
return the final $U(\boldsymbol{\theta})$.

To incorporate variational quantum algorithms, we utilize the variational principle of Hamiltonian's eigenvalues. That is, Hamiltonian's eigenvalues majorize the diagonal elements, and the dot function with an increasingly ordered vector is Schur concave [52]. A similar idea has already been discussed in [47]. In contrast, our method learns the full spectrum of the Hamiltonian. We define a function $M(\boldsymbol{\theta})$ over all parameters $\boldsymbol{\theta}$ of the circuit.

$$M(\boldsymbol{\theta}) = \sum_{j=1}^{N} q_j \cdot \langle\psi_j|U^\dagger(\boldsymbol{\theta})H(\boldsymbol{\nu})U(\boldsymbol{\theta})|\psi_j\rangle, \tag{13}$$

where $\mathbf{q} = (q_1, ..., q_N)$ is a probability distribution such that $q_1 < q_2 < ... < q_N$, and notations $|\psi_1\rangle, \ldots, |\psi_N\rangle$ denote the computational basis. Suppose that PQC $U(\boldsymbol{\theta})$ has enough expressiveness, then $U(\boldsymbol{\theta})|\psi_j\rangle$ could learn the j-th eigenvector of the Hamiltonian $H(\boldsymbol{\nu})$ with suitable parameters. Particularly, $M(\boldsymbol{\theta})$ will reach the global minimum when all eigenvectors are learned. In other words, we use the PQC $U(\boldsymbol{\theta})$ to learn eigenvectors via finding the global minimum of $M(\boldsymbol{\theta})$ over all parameters $\boldsymbol{\theta}$.

Remark 2. Choosing a suitable $U(\boldsymbol{\theta})$ is critical to many variational quantum algorithms as well as our Algorithm 2. With enough expressibility, training the PQC $U(\boldsymbol{\theta})$ would allow us to exactly or

approximately learn the solution to the certain problem. The expressibility of PQCs has been recently studied in [56]. Note that PQCs with high expressive power generally suffer from the barren plateaus [11, 12, 14, 26] and there exhibits a trade-off between their trainability and expressivity [15, 20, 28]. We also note that we can dynamically design a problem-specific ansatz [7, 18, 51, 69].

Remark 3. In the learning process, we employ a gradient-based method to update the parameters $\boldsymbol{\theta}$ iteratively. In each iteration, the corresponding gradients are computed via the parameter shift rule [46], which outsources the gradient estimation to the loss evaluation. As this is similar to other variational quantum algorithms, we omit the details of gradient computation. For details of gradient derivation, please refer to the proof of Proposition 3 in [60].

Notice that for large Hamiltonians, the loss $M(\boldsymbol{\theta})$ may consist of exponentially many terms, which would be a huge burden to the loss evaluation. However, we could employ the importance sampling technique to circumvent this issue. To this end, $M(\boldsymbol{\theta})$ is taken as an expectation of the distribution $\mathbf{q}$. Hence, $M(\boldsymbol{\theta})$ is to be estimated by the sample mean. Notably, the cost of loss evaluation is dominated by the number of samples, which is why we call our method stochastic variational quantum eigensolver (SVQE). Our algorithm with importance sampling for minimizing $M(\boldsymbol{\theta})$ is depicted in Algorithm 2. In the following, we analyze the sample complexity in the loss evaluation.

Proposition 2. Consider a Hamiltonian $H(\boldsymbol{\nu}) = \sum_{\ell=1}^{m} \nu_\ell E_\ell$ with Pauli operators $E_\ell \in \{X, Y, Z, I\}^{\otimes n}$ and constants $\nu_\ell \in [-1, 1]$. Given any constants $\epsilon > 0$, $\eta \in (0, 1)$, $\beta > 0$, the objective function $M(\boldsymbol{\theta})$ in SVQE can be estimated up to precision ϵ with probability at least $1 - \eta$, costing TD samples with $T = O(m\|\boldsymbol{\nu}\|_2^2/\epsilon^2)$ and $D = O(\log(1/\eta))$. Besides, the total number of measurements is given below:

$$O\left(\frac{mTD\|\boldsymbol{\nu}\|_1^2(n + \log(m/\eta))}{\epsilon^2}\right). \tag{14}$$

Sketch of Proof. The number of samples is determined by the accuracy ϵ and Hamiltonian $H(\boldsymbol{\nu})$. By Chebyshev's inequality, estimating $M(\boldsymbol{\theta})$ up to precision ϵ with high probability requires $T = O(m\|\boldsymbol{\nu}\|_2^2/\epsilon^2)$ samples, since the variance is bounded by the spectral norm, which is less than $\sqrt{m}\|\boldsymbol{\nu}\|_2$. Meanwhile, the expectation value $\langle\psi_j|U^\dagger(\boldsymbol{\theta})H(\boldsymbol{\nu})U(\boldsymbol{\theta})|\psi_j\rangle$ is evaluated by measurements. We compute the expectation value of the observable $H(\boldsymbol{\nu})$ by measuring each Pauli operator E_ℓ separately, since there are only $m = O(poly(n))$ Pauli operators.

Remark 4. Other methods for computing expectation value of Hamiltonians can be found in Ref. [4,57], where importance sampling is employed to sample Pauli operator E_l of the Hamiltonian. Moreover, a technique called classical shadow [30] could also be exploited to this end. Particularly, it can save the resources of quantum states to estimate the expectation of Hamiltonian that consists of many terms of Pauli strings.

Remark 5. In the context of quantum algorithms, there are many proposed methods for learning the low-lying eigenvectors of the Hamiltonian and diagonalizing Hamiltonian. Some known quantum algorithms for Hamiltonian diagonalization are based on quantum fast Fourier transform [1], which may be too costly for NISQ computers and thus not suitable for our purpose. Recently, there have already been some works on finding ground and excited eigenstates of the Hamiltonian with NISQ devices, i.e., variational quantum eigensolvers [16,27,34,36,45,47,49,60]. They may be employed to learn eigenvectors in the Hamiltonian learning framework.

3.3 Gradient estimation

Recall that we employ a gradient-based method to do the optimization in the Hamiltonian learning (cf. Figure 1). We use the tools developed in Sec. 3.1-3.2 to derive the gradient estimation procedure.

Usually, with the estimated gradient, parameters are updated in the following way:

$$\boldsymbol{\nu} \leftarrow \boldsymbol{\nu} - r\nabla L(\boldsymbol{\nu}), \tag{15}$$

where r is the learning rate. The expression of the gradient is given below.

$$\nabla L(\boldsymbol{\nu}) = \left(\frac{\partial L(\boldsymbol{\nu})}{\partial \nu_1}, ..., \frac{\partial L(\boldsymbol{\nu})}{\partial \nu_m}\right). \tag{16}$$

Furthermore, the explicit formula of each partial derivative is given in [3]:

$$\frac{\partial L(\boldsymbol{\nu})}{\partial \nu_\ell} = \frac{\partial}{\partial \nu_\ell} \log Z_\beta(\boldsymbol{\nu}) + \beta e_\ell = -\beta \operatorname{Tr}(\rho_\beta(\boldsymbol{\nu}) E_\ell) + \beta e_\ell, \tag{17}$$

where $\rho_\beta(\boldsymbol{\nu}) = e^{-\beta H(\boldsymbol{\nu})}/Z_\beta(\boldsymbol{\nu})$ represents the Gibbs state associated with the parameterized Hamiltonian $H(\boldsymbol{\nu})$.

Algorithm 3 Gradient estimation

Require: Post-training circuit $U(\boldsymbol{\theta})$, Pauli operators $\{E_\ell\}_{\ell=1}^m$, optimal $\widehat{\mathbf{p}}^*$, and constants β and $\{e_\ell\}_{\ell=1}^m$;
Ensure: Gradient estimate $\nabla L(\boldsymbol{\nu})$;
1: Set $\ell = 1$;
2: Set integer K and D;
3: Sample K integers $l_1^1, ..., l_K^1, ..., l_1^D, ..., l_K^D$, according to $\widehat{\mathbf{p}}^*$;
4: Prepare computational states $|\psi_{l_1^1}\rangle, \ldots, |\psi_{l_K^1}\rangle, \ldots, |\psi_{l_1^D}\rangle, \ldots, |\psi_{l_K^D}\rangle$;
5: **while** $\ell \leqslant m$ **do**
6: Compute value $\langle\psi_{l_j^s}|U^\dagger(\boldsymbol{\theta})E_\ell U(\boldsymbol{\theta})|\psi_{l_j^s}\rangle$ for $j = 1, .., K$ and $s = 1, ..., D$;
7: Calculate averages: $ave_s = \frac{1}{K}\sum_{j=1}^K \langle\psi_{l_j^s}|U^\dagger(\boldsymbol{\theta})E_\ell U(\boldsymbol{\theta})|\psi_{l_j^s}\rangle$ for all $s = 1, ..., D$;
8: Take the median value: $s_\ell = -\beta \cdot \text{median}(ave_1, \ldots, ave_D) + \beta e_\ell$;
9: Set $\ell \leftarrow \ell + 1$;
10: **end while**
11: **return** vector $(s_1, ..., s_m)$.

Here we provide a procedure for gradient estimation without preparing the Gibbs state $\rho_\beta(\boldsymbol{\nu})$ in Algorithm 3. We use the post-training PQC $U(\boldsymbol{\theta})$ and the optimal distribution $\widehat{\mathbf{p}}^*$ (cf. Algorithm 1) from Sec. 3.1-3.2, respectively. And the component of the gradient can be computed in the sense that

$$\frac{\partial L(\boldsymbol{\nu})}{\partial \nu_\ell} \approx -\beta \sum_{j=1}^N \widehat{p}_j^* \cdot \langle\psi_j|U^\dagger(\boldsymbol{\theta})E_\ell U(\boldsymbol{\theta})|\psi_j\rangle + \beta e_\ell. \tag{18}$$

The validity of the relation in Eq. (18) is proved in Proposition 3.

Proposition 3 (Correctness)**.** Consider a parameterized Hamiltonian $H(\boldsymbol{\nu})$ and its Gibbs state $\rho_\beta(\boldsymbol{\nu})$. Suppose the $U(\boldsymbol{\theta})$ from SVQE (cf. Algorithm 2) and $\widehat{\mathbf{p}}^*$ from log-partition function estimation procedure (cf. Algorithm 1) are optimal. Define a density operator ρ_β^* as follows:

$$\rho_\beta^* = \sum_{j=1}^N \widehat{p}_j^* \cdot U(\boldsymbol{\theta})\,|\psi_j\rangle\langle\psi_j|\,U^\dagger(\boldsymbol{\theta}), \tag{19}$$

where $\{|\psi\rangle_j\}$ denote the computational basis. Denote the estimated eigenvalues by $\widehat{\boldsymbol{\lambda}}$, where $\widehat{\lambda}_j = \langle\psi_j|U^\dagger(\boldsymbol{\theta})H(\boldsymbol{\nu})U(\boldsymbol{\theta})|\psi_j\rangle$ for all $j = 1, \ldots, N$. Then, ρ_β^* is an approximation of $\rho_\beta(\boldsymbol{\nu})$ in the sense that

$$D(\rho_\beta^*, \rho_\beta(\boldsymbol{\nu})) \leqslant \sqrt{2\beta \max\left\{\mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|], \mathbf{E}_{\mathbf{p}^*}[|\widehat{\lambda} - \lambda|]\right\}}. \tag{20}$$

where $D(\cdot,\cdot)$ denotes the trace distance, $\boldsymbol{\lambda}$ represent $H(\boldsymbol{\nu})$'s true eigenvalues, $\mathbf{p}^*$ is the distribution corresponding to $\boldsymbol{\lambda}$, i.e., $\lambda_j = e^{-\beta\lambda_j}/\sum_l e^{-\beta\lambda_l}$, and

$$\mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|] = \sum_{j=1}^N \widehat{p}_j^*|\widehat{\lambda}_j - \lambda_j|, \quad \mathbf{E}_{\mathbf{p}^*}[|\widehat{\lambda} - \lambda|] = \sum_{j=1}^N p_j^*|\widehat{\lambda}_j - \lambda_j|. \tag{21}$$

Note that the quantity in Eq. (18) contains an expectation of distribution $\widehat{\mathbf{p}}^*$, then the partial derivative $\frac{\partial L(\boldsymbol{\nu})}{\partial \nu_\ell}$ is estimated by the sample mean. Specifically, we first randomly select the computational basis vectors $|\psi_j\rangle$ complying with distribution $\widehat{\mathbf{p}}^*$ and then compute the associated eigenvalues via $U(\boldsymbol{\theta})$. The detailed procedure of sampling and estimate computation is laid out in Algorithm 3. The number of required samples is analyzed in Proposition 4.

Proposition 4 (Sample complexity)**.** Given $\epsilon > 0$ and $\eta \in (0, 1)$, Algorithm 3 can compute an estimate for the gradient $\nabla L(\boldsymbol{\nu})$ up to precision ϵ with probability larger than $1 - \eta$. Particularly, the overall number of samples is $KD = O(\beta^2 \log(2m/\eta)/\epsilon^2)$ with $K = O(\beta^2/\epsilon^2)$ and $D = O(\log(2m/\eta))$. Besides, the total number of measurements is $O(KD \cdot m\beta^2(n + \log(m/\eta))/\epsilon^2)$.

The proofs for Propositions 3-4 are deferred to the supplementary file.

To validate the gradient estimation, we show that the average of the overall errors determines the accuracy of the gradient estimation. For this purpose, Proposition 3 shows that matrix ρ_β^* is an approximation of the desired density matrix $\rho_\beta(\boldsymbol{\nu})$. Specifically, the trance distance between ρ_β^* and $\rho_\beta(\boldsymbol{\nu})$ is dependent on the averaged errors $\mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda}-\lambda|]$ and $\mathbf{E}_{\mathbf{p}^*}[|\widehat{\lambda}-\lambda|]$. Here, notation $|\widehat{\lambda}-\lambda|$ denotes the difference between estimated eigenvalue and the associated real eigenvalue. $\widehat{\mathbf{p}}^*$ and $\mathbf{p}^*$ are probability distributions, corresponding to $\widehat{\boldsymbol{\lambda}}$ and $\boldsymbol{\lambda}$, respectively. In particular, the distributions $\mathbf{p}^*$ and $\hat{\mathbf{p}}^*$ contain components that are decreasing exponentially and hence $\mathbf{E}_{\mathbf{p}^*}[|\hat{\lambda}-\lambda|]$ and $\mathbf{E}_{\hat{\mathbf{p}}^*}[|\hat{\lambda}-\lambda|]$ can be determined with high accuracy by several components of $|\hat{\lambda}-\lambda|$. As a result, it implies that learning several low-lying eigenvectors with high accuracy may lead to a high precision estimate of the gradient. We numerically verify this feature in Sec. 4.3.

Moreover, Proposition 4 shows the feasibility of our approach as the number of measurements scales polynomially in parameters n, $1/\epsilon$, and β.

Remark 6. Note that the convergence rate may slow down if the estimated gradient has an error. However, we can always set the error as small as possible to alleviate this issue. At the same time, despite the estimation error, the effect on the convergence rate can be suppressed due to the strong convexity of the loss function [9]. For example, experimental results in Sec. 4.3 show that even though the estimated gradients are implicit, the loss function could still converge in a reasonable time.

3.4 Hamiltonian learning algorithm

Eventually, we present our hybrid quantum-classical algorithm for Hamiltonian learning (HQHL) in Algorithm 4. The main idea of HQHL is to find the target interaction coefficients by a gradient-descent method (cf. Figure 1). Thus, HQHL's main process is to compute the gradient of the objective function. Specifically, we take Pauli operators $\{E_\ell\}_{\ell=1}^m$, $\{e_\ell\}_{\ell=1}^m$, and β as input. Then we initialize the coefficients by choosing $\boldsymbol{\nu}$ from $[-1,1]^m$ uniformly at random. Next, we compute the gradient of the objective function $L(\boldsymbol{\nu})$ by Algorithm 3. Then update the coefficients by choosing a suitable learning rate r and using the estimated gradient. In consequence, after repeating the training process sufficiently many times, the final coefficients are supposed to approximate the target coefficients $\boldsymbol{\nu}$.

Algorithm 4 Hybrid quantum-classical Hamiltonian learning algorithm (HQHL)

Require: Pauli operators $\{E_\ell\}_{\ell=1}^m$, constants $\{e_\ell\}_{\ell=1}^m$, and β;
Ensure: An estimate for target coefficients $\boldsymbol{\nu}$;
1: Initialize coefficients $\{\nu_\ell\}_{\ell=1}^m$;
2: Set number of iterations I and $l=1$;
3: Set parameterized quantum circuit $U(\boldsymbol{\theta})$;
4: Set learning rate r;
5: **while** $l \leqslant I$ **do**
6: Set Hamiltonian $H(\boldsymbol{\nu})=\sum_{\ell=1}^m \nu_\ell E_\ell$;
7: Train $U(\boldsymbol{\theta})$ by SVQE with $H(\boldsymbol{\nu})$;
8: Derive a probability $\widehat{\mathbf{p}}^*$ by performing log-partition function estimation with $U(\boldsymbol{\theta})$ and β;
9: Compute gradient $\nabla L(\boldsymbol{\nu})$ by gradient estimation with $U(\boldsymbol{\theta})$, $\widehat{\mathbf{p}}^*$, and β;
10: Update coefficients $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu} - r\nabla L(\boldsymbol{\nu})$;
11: Set $l \leftarrow l+1$;
12: **end while**
13: **return** the final coefficients $\boldsymbol{\nu}$.

Notably, the learning process is in the "**while**" loop of HQHL. In the loop, the subroutine SVQE (cf. Sec. 3.2) is first called to learn Hamiltonian's eigenvectors and eigenvalues. Here, we choose a suitable parameterized quantum circuit $U(\boldsymbol{\theta})$ and train it to prepare the eigenvectors of the Hamiltonian $H(\boldsymbol{\nu})$. Afterwards, we enter the process of the log-partition function estimation (cf. Sec. 3.1). It first exploits the $U(\boldsymbol{\theta})$ to output the estimated eigenvalues of the parameterized Hamiltonian $H(\boldsymbol{\nu})$ and then computes the objective function $L(\boldsymbol{\nu})$. We would obtain a probability distribution $\widehat{\mathbf{p}}^*$ that consists of eigenvalues of the associated Gibbs state $\rho_\beta(\boldsymbol{\nu})=e^{-\beta H(\boldsymbol{\nu})}/Z_\beta(\boldsymbol{\nu})$. Lastly, we exploit the resultant results (post-training circuit $U(\boldsymbol{\theta})$ and distribution $\widehat{\mathbf{p}}^*$) to compute the gradients following the procedure in Algorithm 3 and update the coefficient $\boldsymbol{\nu}$ accordingly (cf. Eq. (15)).

Remark 7. To improve the scalability of our method, on the one hand, we use the importance sampling technique to circumvent many resource requirements for Hamiltonian diagonalization. On the other hand, we discuss that learning partial eigenvalues could also lead to the target Hamiltonian. In particular, we

Table 1 Hyper-parameters setting. The number of qubits (# qubits) varies from 3 to 5, and the number of μ (# μ) from 3 to 6. β is chosen as 0.3, 1, 3. "LR" denotes learning rate. The values of μ are sampled uniformly in the range of [-1, 1]. The term, likes "[[0 2 1] [2 1 3] [0 3 3]]", indicates there are three E_l's and each has three qubits with the corresponding Pauli tensor product. Here "0,1,2,3" represent "I, X, Y, Z" respectively. For example, for the first sample, the corresponding Hamiltonian is taken as H=0.3408 $\cdot I \otimes Y \otimes X$ -0.6384 $\cdot Y \otimes X \otimes Z$ -0.4988 $\cdot I \otimes Z \otimes Z$.

Three aspects	# qubits	# μ	β	LR	$\boldsymbol{\mu}$	E_l
	3	3	1	1.0	[0.3408 -0.6384 -0.4988]	[[0 2 1] [2 1 3] [0 3 3]]
Vary β	3	3	0.3	8.0	[-0.4966 -0.8575 -0.7902]	[[1 0 0] [3 0 2] [3 1 3]]
			3	0.1	[0.5717 -0.1313 0.2053]	[[1 0 0] [3 3 3] [0 2 3]]
Vary # μ	3	4	1	1.0	[-0.7205 -0.3676 -0.7583 -0.3002]	[[3 2 1] [2 1 3] [0 0 2] [2 0 0]]
		5			[-0.5254 -0.1481 -0.0037 -0.4373 0.7326]	[[1 3 0] [2 1 1] [3 3 2] [2 3 1] [0 2 0]]
		6			[-0.5992 0.7912 0.5307 -0.5422 -0.9239 0.0354]	[[3 2 2] [0 2 1] [1 2 1] [2 2 0] [0 1 2] [3 2 1]]
Vary # qubits	4	3	1	1.0	[0.0858 0.3748 -0.1007]	[[0 2 0 1] [1 0 0 1] [2 0 1 0]]
	5				[-0.0411 0.7882 0.6207]	[[2 2 2 1 2] [2 3 3 2 1] [1 2 0 2 3]]

numerically show that learning several low-lying eigenvectors could help recover the unknown Hamiltonian in Sec. 4.3.

4 Numerical Results

In this section, we conduct numerical experiments to verify the correctness of our algorithm. Specifically, we consider recovering interactions coefficients of several Hamiltonians, including randomly generated Hamiltonians and many-body Hamiltonians. To ensure the performance of the algorithm, we choose a PQC (shown in Fig. 2) and set the circuit with enough expressibility. When testing our algorithm, we first use SVQE to learn the full spectrum of Hamiltonians, where size of the Hamiltonian varies from 3 to 5. In SVQE, weights $\mathbf{q}$ consists of a normalized sequence of arithmetic sequence. For instance, when $n = 3$, $\mathbf{q} = (1, 2, 3, \ldots, 8)/S_3$, where $S_3 = \sum_{l=1}^{8} l$. Furthermore, in order to reduce quantum resources, we also partially learn the few smallest eigenvalues of the selected Ising models and derive estimates for coefficients up to precision 0.05. With fewer eigenvalues to be learned, the depth of the used PQC is significantly reduced.

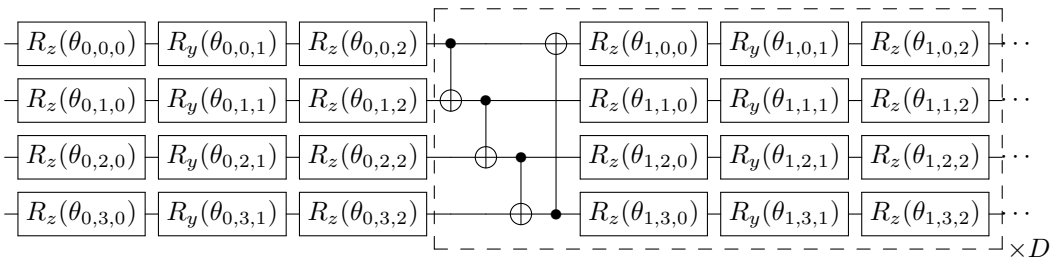

Figure 2 The selected quantum circuit $U(\boldsymbol{\theta})$ for stochastic variational quantum eigensolver (SVQE). Here, D represents circuit depth. Parameters $\boldsymbol{\theta}$ are randomly initialized from a uniform distribution in [0, 2π] and updated via gradient descent method.

4.1 Random Hamiltonian models

This section shows the effectiveness of our algorithm with random Hamiltonians from three aspects: different β, different numbers of μ (# μ) and a different number of qubits (# qubits).

In the experimental setting, we randomly choose Pauli tensor products E_ℓ from $\{X, Y, Z, I\}^{\otimes n}$ and target coefficients $\boldsymbol{\mu}$ by a uniform distribution over $[-1, 1]$. Specifically, we first vary the values of β by fixing the number of μ and the number of qubits to explore our method's sensitivity to temperature. We similarly vary the number of μ and the number of qubits by fixing other hyper-parameters to explore our method's scalability. The actual values of these hyper-parameters sampled/chosen in each trial are concluded in Table 1. In addition, the deep, D, of the PQC $U(\boldsymbol{\theta})$ is set according to the size of Hamiltonian. As number of qubits ranges from $n = 3$ to $n = 5$, the depth D is set to be $10, 20, 40$, respectively.

In Table 1, Hamioltonian is represented by a tuple. Each number $0, 1, 2, 3$ corresponds to matrices I, X, Y, Z, respectively. $\boldsymbol{\mu}$ denotes the interaction coefficients to be learned. For instance, [[0 2 1] [2 1 3] [0 3 3]] means that the Hamiltonian consists of three Pauli operators, where each term represents a Pauli operator, e.g., [0 2 1] means $I \otimes Y \otimes X$. Then, the parameters in the top second row represents the following Hamiltonian.

$$0.3408 I \otimes Y \otimes X - 0.6384 Y \otimes X \otimes Z - 0.4988 I \otimes Z \otimes Z. \tag{22}$$

Other Hamiltonians to be tested are represented in a similar fashion.

The results for these three aspects are illustrated in Fig. 3. We find that all curves converge to the values close to 0 in less than ten iterations, which shows our method is effective. In particular, our method works for different β means that it is robust to temperature. And the results for the different number of μ and qubits reveals our method's scalability to a certain extent.

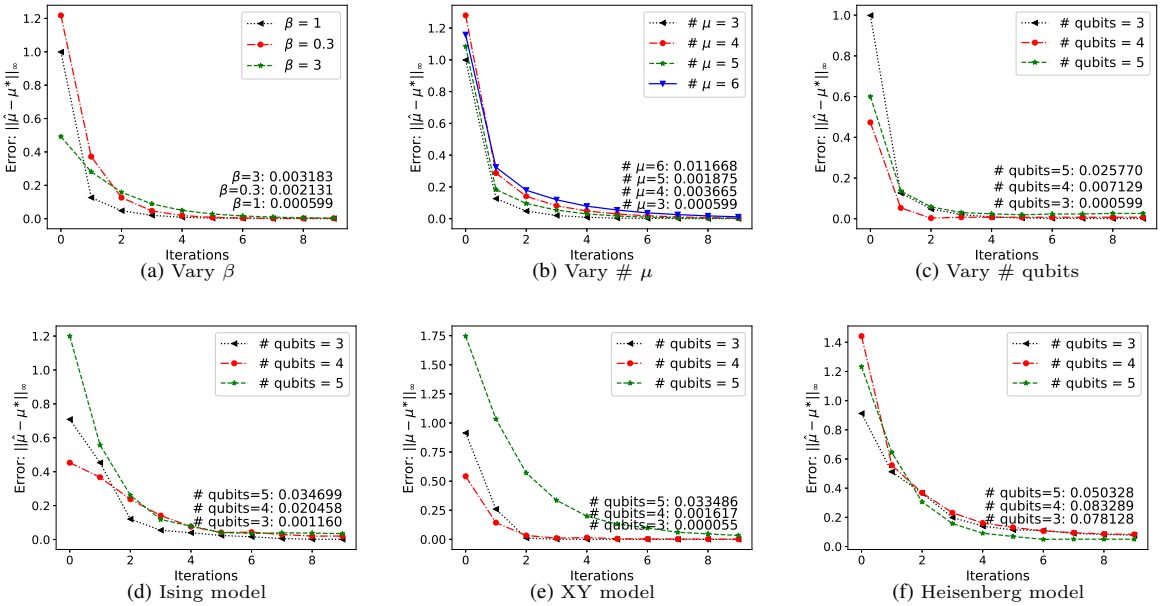

Figure 3 The curves in (a), (b), (c) represent the infinity norm of the error of μ with different β, different number of μ, and different number of qubits, respectively. In (d), (e), (f), the curves represent the infinity norm of the error of μ for different many-body Hamiltonians with the number of qubits varies from 3 to 5. The numbers on the line represent the values of the last iteration. These numbers close to 0 indicate that our algorithm is effective.

4.2 Quantum many-body models

Here, we demonstrate the performance of our algorithm for quantum many-body models. Specifically, we consider the one-dimensional nearest-neighbor Ising model, XY model, and Heisenberg model. These many-body models are described by the Hamiltonians shown below:

$$\text{(Ising model)} \quad H_0 = J_0 \sum_{l=1}^{n} Z^l Z^{l+1} + h_0 \sum_{l=1}^{n} X^l, \tag{23}$$

$$\text{(XY model)} \quad H_1 = J_1 \sum_{l=1}^{n} (X^l X^{l+1} + Y^l Y^{l+1}), \tag{24}$$

$$\text{(Heisenberg model)} \quad H_2 = J_2 \sum_{l=1}^{n} (X^l X^{l+1} + Y^l Y^{l+1} + Z^l Z^{l+1}) + h_2 \sum_{l=1}^{n} Z^l, \tag{25}$$

where periodic boundary conditions are assumed (i.e., $X^{n+1} = X^1$, $Y^{n+1} = Y^1$, and $Z^{n+1} = Z^1$). Coefficient J is the coupling constant for the nearest neighbor interaction, and h represents the external transverse magnetic field. The experimental parameters are concluded in Table 2.

Table 2 Hyper-parameters setting for many-body models. For each Hamiltonian model, the number of qubits varies from 3 to 5, and the number of μ is determined by the number of Pauli operators. "LR" denotes learning rate. The values of μ are sampled uniformly in the range of $[-1, 1]$.

Many-body models	# qubits	# μ	β	LR	$\boldsymbol{\mu}$
Ising model	3	6	1.0	2.0	$[J_0 = 0.1981, \quad h_0 = 0.7544]$
	4	8		1.0	$[J_0 = 0.5296, \quad h_0 = 0.4996]$
	5	10		0.5	$[J_0 = -0.6916, \quad h_0 = 0.4801]$
XY model	3	6	1.0	1.0	$J_1 = -0.0839$
	4	8		1.0	$J_1 = 0.2883$
	5	10		0.6	$J_1 = -0.7773$
Heisenberg	3	12	1.0	1.0	$[J_2 = 0.0346, \quad h_2 = 0.8939]$
	4	16		1.0	$[J_2 = -0.5831, \quad h_2 = -0.0366]$
	5	20		1.0	$[J_2 = 0.2883, \quad h_2 = -0.2385]$

Table 3 Parameters setting for HQHL. The script index means the length of the tuple, e.g., $()_8$ indicates the tuple consists of 8 entries. The notation $0, \ldots$ means the entries following 0 are all zeros as well. Notation $\#\lambda$ means the number of eigenvalues we learned. Please note that we omit the $\beta = 1$ in the table.

# qubits n	weights $\mathbf{q}$	# μ	depth D	LR	$\#\lambda$
3	$(0.1, 0.2, 0.3, 0.4, 0, \ldots)_8$	6	5	0.4	4
4	$(0.1, 0.15, 0.2, 0.25, 0.3, 0, \ldots)_{16}$	8	10	0.55	5
5	$(0.1, 0.15, 0.2, 0.25, 0.3, 0, \ldots)_{32}$	10	20	0.7	5
6	$(1/21, 2/21, 3/21, 4/21, 5/21, 6/21, 0, \ldots)_{64}$	12	30	0.55	6
7	$(1/21, 2/21, 3/21, 4/21, 5/21, 6/21, 0, \ldots)_{128}$	14	40	0.6	6

We consider the models with a different number of qubits, varying from $n = 3$ to $n = 5$. The inverse temperature is set as $\beta = 1$. The coefficients J_0, J_1, J_2 and h_0, h_2 are sampled uniformly from a uniform distribution on [-1,1]. We also employ the parameterized quantum circuit $U(\boldsymbol{\theta})$ in Fig. 2 for the SVQE. And the depth of $U(\boldsymbol{\theta})$ is also set as $D = 10, 20, 40$ for different n. Moreover, the numerical results are shown in Fig. 3, which imply our method is applicable to recover quantum many-body Hamiltonians.

4.3 Numerical results using fewer eigenvalues of Ising Hamiltonians

Notice that we use a PQC $U(\boldsymbol{\theta})$ with deep depths to learn the full spectrum of small-sized Hamiltonians in Secs. 4.1-4.2, which may be beyond the capacity of NISQ devices. However, this section demonstrates the efficacy of HQHL in learning the Ising Hamiltonians using a circuit with reduced depth, where few eigenvalues (instead of the full spectrum) are learned. In particular, only halved circuit depths are needed for Hamiltonians with 3-5 qubits, given in Table 2. Furthermore, the performance on $n = 6$ and $n = 7$-qubit Ising models, given below, is tested as well.

$$H = 0.1981 \sum_{l=1}^{n} Z^l Z^{l+1} + 0.7544 \sum_{l=1}^{n} X^l. \tag{26}$$

To reduce the number of eigenvalues to be learned, we tune the weights $\mathbf{q}$ of the SVQE such that the $U(\boldsymbol{\theta})$ can output several smallest eigenvalues. For instance, five eigenvalues are learned for 4 & 5-qubit Ising Hamiltonians, and four eigenvalues are learned for 3-qubit Ising Hamiltonians. As a result, the circuit depth of the used $U(\boldsymbol{\theta})$ is significantly reduced. For example, we only use depth $D = 20$ to learn the coefficients with precision 0.05 for 5-qubit Ising models. While, in Sec. 4.2, we use the depth $D = 40$. Moreover, we find out that using a circuit with 35 depths suffices to learn well the 6-qubit Ising model, where SVQE only learns six eigenvalues. Using the circuit with 40 depths could also reach a precision of 0.05 for the 7-qubit Ising Hamiltonian. The details of parameters setting (weights, depth, learning rate, etc.) are given in Table 3. Besides, the experimental results are depicted in Figure 4.

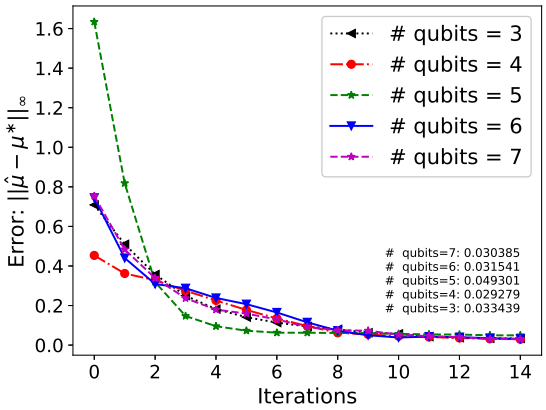

Figure 4 Experimental results by using fewer eigenvalues. Each line corresponds to the results by running HQHL with Ising Hamiltonians of different sizes. Results show that using halved circuit depth, compared to the setting in Sec. 4.2, could learn coefficients up to precision 0.05 for different sized Ising models and a different number of μ.

5 Conclusion

We have proposed a hybrid quantum-classical Hamiltonian learning algorithm that employs a gradient-descent method to find the desired interaction coefficients. We achieve this purpose by unifying the variational quantum algorithms (VQAs) with the strategy proposed in [3]. To this end, we develop several subroutines: log-partition function estimation, stochastic variational quantum eigensolver (SVQE), and gradient estimation. In SVQE, we propose a method to learn the full/partial spectrum of the Hamiltonian and use the importance sampling to circumvent the resources in the loss evaluation. In the log-partition function, we propose a method that combines the parameterized quantum circuits and convex optimization to find the global minimum of the free energy as well as compute the log-partition function. In gradient estimation, we present a procedure to compute the gradient of the objective function costing polynomially many resources. Finally, we conduct numerical experiments to demonstrate the effectiveness of our approach with randomly generated Hamiltonians and selected many-body Hamiltonians. In consequence, we show that learning the full spectrum of Hamiltonians in the learning process could produce high-precision estimates of the desired interaction coefficients. Moreover, we also show that partially learning several smallest eigenvalues of Ising Hamiltonians could derive estimates up to a precision of 0.05. Overall, this paper develops a concrete near-term quantum algorithm for Hamiltonian learning and demonstrates the effectiveness as well, which has potential applications in quantum device certification, quantum simulation, and quantum machine learning.

We believe our approach would shed lights on near-term quantum applications. For example, SVQE might enrich the VQE family in the fields of molecules and materials. Moreover, as many problems in computer science can be framed as partition function problems (e.g., counting coloring), our method may contribute to these fields as well. Furthermore, it is reasonable to explore our algorithm's applications in quantum machine learning [55], quantum error correction [58], and tomography [39].

Acknowledgements Y. W. and G. L. acknowledge support from the Australian Research Council (Grant No: DP180100691) and the Baidu-UTS AI Meets Quantum project. G. L. acknowledges the financial support from China Scholarship Council (No. 201806070139).

References

1 Abrams, D. S., and Lloyd, S. Quantum algorithm providing exponential speed increase for finding eigenvalues and eigenvectors. *Phys. Rev. Lett. 83* (Dec 1999), 5162–5165.
2 Aharonov, D., van Dam, W., Kempe, J., Landau, Z., Lloyd, S., and Regev, O. Adiabatic Quantum Computation Is Equivalent to Standard Quantum Computation. *SIAM Review 50*, 4 (jan 2008), 755–787.
3 Anshu, A., Arunachalam, S., Kuwahara, T., and Soleimanifar, M. Sample-efficient learning of interacting quantum systems. *Nature Physics 17*, 8 (aug 2021), 931–935.
4 Arrasmith, A., Cincio, L., Somma, R. D., and Coles, P. J. Operator Sampling for Shot-frugal Optimization in Variational Algorithms. *arXiv preprint arXiv:2004.06252* (2020), 1–11.
5 Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., Biswas, R., Boixo, S., Brandao, F. G., Buell, D. A., Burkett, B., Chen, Y., Chen, Z., Chiaro, B., Collins, R., Courtney, W., Dunsworth, A., Farhi, E., Foxen, B., Fowler, A., Gidney, C., Giustina, M., Graff, R., Guerin, K., Habegger, S., Harrigan, M. P., Hartmann, M. J., Ho, A., Hoffmann, M., Huang, T., Humble, T. S., Isakov, S. V., Jeffrey, E., Jiang, Z., Kafri, D., Kechedzhi, K., Kelly, J., Klimov, P. V., Knysh,

S., Korotkov, A., Kostritsa, F., Landhuis, D., Lindmark, M., Lucero, E., Lyakh, D., Mandrà, S., McClean, J. R., McEwen, M., Megrant, A., Mi, X., Michielsen, K., Mohseni, M., Mutus, J., Naaman, O., Neeley, M., Neill, C., Niu, M. Y., Ostby, E., Petukhov, A., Platt, J. C., Quintana, C., Rieffel, E. G., Roushan, P., Rubin, N. C., Sank, D., Satzinger, K. J., Smelyanskiy, V., Sung, K. J., Trevithick, M. D., Vainsencher, A., Villalonga, B., White, T., Yao, Z. J., Yeh, P., Zalcman, A., Neven, H., and Martinis, J. M. Quantum supremacy using a programmable superconducting processor. *Nature 574*, 7779 (2019), 505–510.

6 Bairey, E., Arad, I., and Lindner, N. H. Learning a Local Hamiltonian from Local Measurements. *Physical Review Letters 122*, 2 (2019), 1–8.

7 Bilkis, M., Cerezo, M., Verdon, G., Coles, P. J., and Cincio, L. A semi-agnostic ansatz with variable structure for quantum machine learning. *arXiv preprint arXiv:2103.06712* (2021).

8 Bonet-Monroig, X., Babbush, R., and O'Brien, T. E. Nearly Optimal Measurement Scheduling for Partial Tomography of Quantum States. *arXiv preprint arXiv:1908.05628* (2019), 1–9.

9 Boyd, S., Vandenberghe, L., and Faybusovich, L. Convex optimization. *IEEE Transactions on Automatic Control 11*, 51 (2006), 1859.

10 Bravo-Prieto, C., LaRose, R., Cerezo, M., Subasi, Y., Cincio, L., and Coles, P. J. Variational quantum linear solver: A hybrid algorithm for linear systems. *arXiv preprint arXiv:1909.05820* (2019).

11 Caro, M. C., and Datta, I. Pseudo-dimension of quantum circuits. *Quantum Machine Intelligence 2*, 2 (2020), 1–14.

12 Caro, M. C., Gil-Fuster, E., Meyer, J. J., Eisert, J., and Sweke, R. Encoding-dependent generalization bounds for parametrized quantum circuits. *arXiv preprint arXiv:2106.03880* (2021).

13 Cerezo, M., Sharma, K., Arrasmith, A., and Coles, P. J. Variational Quantum State Eigensolver. *arXiv preprint arXiv:2004.01372*, 1 (apr 2020), 1–14.

14 Cerezo, M., Sone, A., Volkoff, T., Cincio, L., and Coles, P. J. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature communications 12*, 1 (2021), 1–12.

15 Chen, C.-C., Watabe, M., Shiba, K., Sogabe, M., Sakamoto, K., and Sogabe, T. On the expressibility and overfitting of quantum circuit learning. *ACM Transactions on Quantum Computing 2*, 2 (2021), 1–24.

16 Commeau, B., Cerezo, M., Holmes, Z., Cincio, L., Coles, P. J., and Sornborger, A. Variational Hamiltonian Diagonalization for Dynamical Quantum Simulation. *arXiv preprint arXiv:2009.02559* (sep 2020), 1–12.

17 Cotler, J., and Wilczek, F. Quantum Overlapping Tomography. *Physical Review Letters 124*, 10 (2020).

18 Du, Y., Huang, T., You, S., Hsieh, M.-H., and Tao, D. Quantum circuit architecture search: error mitigation and trainability enhancement for variational quantum solvers. *arXiv preprint arXiv:2010.10217* (2020).

19 Du, Y., and Tao, D. On exploring practical potentials of quantum auto-encoder with advantages. *arXiv preprint arXiv:2106.15432* (2021), 1–25.

20 Du, Y., Tu, Z., Yuan, X., and Tao, D. An efficient measure for the expressivity of variational quantum algorithms. *arXiv preprint arXiv:2104.09961* (2021).

21 Gheorghiu, A., and Hoban, M. J. Estimating the entropy of shallow circuit outputs is hard. *arXiv preprint arXiv:2002.12814* (feb 2020), 1–31.

22 Goldberg, L. A., and Jerrum, M. Inapproximability of the tutte polynomial. *Information and Computation 206*, 7 (2008), 908–929.

23 Granade, C. E., Ferrie, C., Wiebe, N., and Cory, D. G. Robust online Hamiltonian learning. *New Journal of Physics 14*, 10 (oct 2012), 103013.

24 Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S., and Eisert, J. Quantum state tomography via compressed sensing. *Physical review letters 105*, 15 (2010), 150401.

25 Grötschel, M., Lovász, L., and Schrijver, A. Geometric algorithms and combinatorial optimization. *Algorithms and Combinatorics* (1993).

26 Gyurik, C., van Vreumingen, D., and Dunjko, V. Structural risk minimization for quantum linear classifiers. *arXiv preprint arXiv:2105.05566* (2021).

27 Higgott, O., Wang, D., and Brierley, S. Variational quantum computation of excited states. *Quantum 3* (2019), 156.

28 Holmes, Z., Sharma, K., Cerezo, M., and Coles, P. J. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *arXiv preprint arXiv:2101.02138* (2021).

29 Huang, H.-Y., Bharti, K., and Rebentrost, P. Near-term quantum algorithms for linear systems of equations. *arXiv preprint arXiv:1909.07344* (2019).

30 Huang, H. Y., Kueng, R., and Preskill, J. Predicting many properties of a quantum system from very few measurements. *Nature Physics* (2020), 1–40.

31 Islam, R., Ma, R., Preiss, P. M., Tai, M. E., Lukin, A., Rispoli, M., and Greiner, M. Measuring entanglement entropy in a quantum many-body system. *Nature 528*, 7580 (2015), 77–83.

32 Jaynes, E. T. Information theory and statistical mechanics. *Physical review 106*, 4 (1957), 620.

33 Jiang, H., Lee, Y. T., Song, Z., and Wong, S. C.-w. An Improved Cutting Plane Method for Convex Optimization, Convex-Concave Games and its Applications. *arXiv preprint arXiv:2004.04250* (apr 2020).

34 Jones, T., Endo, S., McArdle, S., Yuan, X., and Benjamin, S. C. Variational quantum algorithms for discovering hamiltonian spectra. *Physical Review A 99*, 6 (2019), 062304.

35 Kalai, A. T., and Vempala, S. Simulated Annealing for Convex Optimization. *Mathematics of Operations Research 31*, 2 (may 2006), 253–266.

36 Kandala, A., Mezzacapo, A., Temme, K., Takita, M., Brink, M., Chow, J. M., and Gambetta, J. M. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature 549*, 7671 (2017), 242–246.

37 Karmarkar, N. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing* (1984), pp. 302–311.

38 Kelley, Jr, J. E. The cutting-plane method for solving convex programs. *Journal of the society for Industrial and Applied Mathematics 8*, 4 (1960), 703–712.

39 Kieferová, M., and Wiebe, N. Tomography and generative training with quantum Boltzmann machines. *Physical Review A 96*, 6 (dec 2017), 062327.

40 Ladd, T. D., Jelezko, F., Laflamme, R., Nakamura, Y., Monroe, C., and O'Brien, J. L. Quantum computers. *Nature 464*, 7285 (mar 2010), 45–53.

41 LaRose, R., Tikku, A., O'Neel-Judy, É., Cincio, L., and Coles, P. J. Variational quantum state diagonalization. *npj Quantum Information 5*, 1 (2019), 1–10.

42 Lee, Y. T., Sidford, A., and Vempala, S. S. Efficient Convex Optimization with Membership Oracles. *arXiv preprint arXiv:1706.07357* (2017).

43 Lee, Y. T., Sidford, A., and Wong, S. C.-W. A Faster Cutting Plane Method and its Implications for Combinatorial and Convex Optimization. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science* (oct 2015), vol. 2015-Decem, IEEE, pp. 1049–1065.

44 Long, P. M., and Servedio, R. A. Restricted boltzmann machines are hard to approximately evaluate or simulate. *ICML* (2010).

45 McClean, J. R., Romero, J., Babbush, R., and Aspuru-Guzik, A. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics 18*, 2 (2016), 023023.

46 Mitarai, K., Negoro, M., Kitagawa, M., and Fujii, K. Quantum circuit learning. *Physical Review A 98*, 3 (2018), 032309.

47 Nakanishi, K. M., Mitarai, K., and Fujii, K. Subspace-search variational quantum eigensolver for excited states. *Physical Review Research 1*, 3 (oct 2019), 033062.

48 Nielsen, M. A., and Chuang, I. Quantum computation and quantum information, 2002.

49 Peruzzo, A., McClean, J., Shadbolt, P., Yung, M.-H., Zhou, X.-Q., Love, P. J., Aspuru-Guzik, A., and O'brien, J. L. A variational eigenvalue solver on a photonic quantum processor. *Nature communications 5* (2014), 4213.

50 Preskill, J. Quantum Computing in the NISQ era and beyond. *Quantum 2*, July (aug 2018), 79.

51 Rattew, A. G., Hu, S., Pistoia, M., Chen, R., and Wood, S. A domain-agnostic, noise-resistant, hardware-efficient evolutionary variational quantum eigensolver. *arXiv preprint arXiv:1910.09694* (2019).

52 Roberts, A. W. Convex functions. In *Handbook of Convex Geometry*. Elsevier, 1993, pp. 1081–1104.

53 Seth Lloyd. Universal Quantum Simulators. *Science 273*, 5278 (1996), 1073–1078.

54 Sharma, K., Khatri, S., Cerezo, M., and Coles, P. J. Noise resilience of variational quantum compiling. *New Journal of Physics 22*, 4 (apr 2020), 043006.

55 Shingu, Y., Seki, Y., Watabe, S., Endo, S., Matsuzaki, Y., Kawabata, S., Nikuni, T., and Hakoshima, H. Boltzmann machine learning with a variational quantum algorithm. *arXiv preprint arXiv:2007.00876* (jul 2020), 1–8.

56 Sim, S., Johnson, P. D., and Aspuru-Guzik, A. Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms. *Advanced Quantum Technologies 2*, 12 (dec 2019), 1900070.

57 Sweke, R., Wilde, F., Meyer, J., Schuld, M., Fährmann, P. K., Meynard-Piganeau, B., and Eisert, J. Stochastic gradient descent for hybrid quantum-classical optimization. *arXiv preprint arXiv:1910.01155* (2019), 1–23.

58 Valenti, A., van Nieuwenburg, E., Huber, S., and Greplova, E. Hamiltonian learning for quantum error correction. *Physical Review Research 1*, 3 (nov 2019), 033092.

59 Wang, J., Paesani, S., Santagati, R., Knauer, S., Gentile, A. A., Wiebe, N., Petruzzella, M., O'brien, J. L., Rarity, J. G., Laing, A., and Thompson, M. G. Experimental quantum Hamiltonian learning. *Nature Physics 13*, 6 (2017), 551–555.

60 Wang, X., Song, Z., and Wang, Y. Variational Quantum Singular Value Decomposition. *Quantum 5*, 1 (jun 2021), 483.

61 Wang, Y., Li, G., and Wang, X. Variational quantum Gibbs state preparation with a truncated Taylor series. *arXiv preprint arXiv:2005.08797* (2020), 1–23.

62 Wiebe, N., Granade, C., and Cory, D. G. Quantum bootstrapping via compressed quantum Hamiltonian learning. *New Journal of Physics 17*, 2 (feb 2015), 022005.

63 Wiebe, N., Granade, C., Ferrie, C., and Cory, D. Quantum Hamiltonian learning using imperfect quantum resources. *Physical Review A - Atomic, Molecular, and Optical Physics 89*, 4 (2014), 1–18.

64 Wiebe, N., Granade, C., Ferrie, C., and Cory, D. G. Hamiltonian Learning and Certification Using Quantum Resources. *Physical Review Letters 112*, 19 (may 2014), 190501.

65 Wu, J., and Hsieh, T. H. Variational thermal quantum simulation via thermofield double states. *Physical review letters 123*, 22 (2019), 220502.

66 Xu, X., Sun, J., Endo, S., Li, Y., Benjamin, S. C., and Yuan, X. Variational algorithms for linear algebra. *arXiv preprint arXiv:1909.03898 2*, 2 (2019), 1–10.

67 Yuan, X., Endo, S., Zhao, Q., Li, Y., and Benjamin, S. C. Theory of variational quantum simulation. *Quantum 3* (2019), 191.

68 Zanardi, P. Quantum entanglement in fermionic lattices. *Physical Review A 65*, 4 (mar 2002), 042101.

69 Zhang, S.-X., Hsieh, C.-Y., Zhang, S., and Yao, H. Differentiable quantum architecture search. *arXiv preprint arXiv:2010.08561* (2020).

Appendix A Proofs for Eqs. (9)-(10)

Consider a Hamiltonian $H \in \mathbb{C}^{N\times N}$ and a constant $\beta > 0$, then the system's free energy is given by $F(\rho) = \mathrm{Tr}(H\rho) - \beta^{-1}S(\rho)$. Recall the fact [48] that

$$S(\rho) \leqslant -\sum_{j=1}^{N} \rho_{jj} \log \rho_{jj}, \tag{S1}$$

where ρ_{jj} are the diagonal elements of quantum state ρ. Using this fact, for any state ρ, we can find a lower bound on free energy in the sense that

$$F(\rho) \geqslant \mathrm{Tr}(H\rho) + \beta^{-1}\sum_{j=1}^{N} \rho_{jj} \log \rho_{jj}. \tag{S2}$$

On the other hand, let U be a unitary such that $H = U\Lambda U^{\dagger}$, where $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_N)$ is a diagonal matrix. Let $\widetilde{\rho} = \mathrm{diag}(\rho_{11}, ..., \rho_{NN})$ be the diagonal matrix consisting of ρ's diagonal elements and let $\sigma = U^{\dagger}\widetilde{\rho}U$. It is easy to verify that $\mathrm{Tr}(H\rho) = \mathrm{Tr}(\Lambda\sigma)$. Furthermore, taking this relation into Eq. (S2)'s right hand side, we can find that

$$F(\rho) \geqslant \mathrm{Tr}(\Lambda\sigma) - \beta^{-1}S(\sigma). \tag{S3}$$

Notice that Eq. (S3)'s right-hand side is equal to $F(\widetilde{\rho})$, then we have

$$F(\rho) \geqslant F(\widetilde{\rho}). \tag{S4}$$

The inequality in Eq. (S4) shows that free energy's global optimum is commuting with the Hamiltonian H.

According to the above discussion, we can rewrite the optimization program of finding free energy's minimal value as follows

$$\min_{\rho} F(\rho) = \min_{\mathbf{p}} \sum_{j=1}^{N} \lambda_j p_j + \beta^{-1} \sum_{j=1}^{N} p_j \log p_j, \tag{S5}$$

where $\mathbf{p}$ represents an arbitrary probability distribution. Eq. (S5)'s right-hand side can be solved using the Lagrange multiplier method, and the optimum is given below:

$$\mathbf{p}^* = \frac{1}{Z}(e^{-\beta\lambda_1}, ..., e^{-\beta\lambda_N}), \tag{S6}$$

with $Z := \sum_{j=1}^{N} e^{-\beta\lambda_j}$.

Finally, the equalities in Eqs. (9)-(10) can be proved by taking $\mathbf{p}^*$ into Eq. (S5)'s right-hand side and computing the minimal value.

Appendix B Proof for Proposition 1

Lemma S1. For any parameterized Hamiltonian $H(\boldsymbol{\nu}) = \sum_{\ell=1}^{m} \nu_\ell E_\ell$ with $E_\ell \in \{X, Y, Z, I\}^{\otimes n}$, we have

$$\| H(\boldsymbol{\nu}) \| \leqslant \sqrt{m} \cdot \| \boldsymbol{\nu} \|_2 . \tag{S1}$$

where $\| \cdot \|$ denotes the spectral norm and $\| \cdot \|_2$ is the ℓ_2-norm.

Proof. Let U be the unitary that diagonalizes the Hamiltonian $H(\boldsymbol{\nu})$, and then we can use the following form to represent $H(\boldsymbol{\nu})$.

$$H(\boldsymbol{\nu}) = \sum_{j=1}^{N} \lambda_j \cdot U |\psi_j\rangle\langle\psi_j| U^\dagger, \tag{S2}$$

where $|\psi_1\rangle, ..., |\psi_N\rangle$ are the computational basis.

Typically, each eigenvalue is represented as follows:

$$\lambda_j = \langle\psi_j|U^\dagger H(\boldsymbol{\nu}) U|\psi_j\rangle \tag{S3}$$

$$= \sum_{\ell=1}^{m} \nu_\ell \langle\psi_j|U^\dagger E_\ell U|\psi_j\rangle \tag{S4}$$

Then, applying the Cauchy-Schwarz inequality leads to an upper bound on each eigenvalue:

$$(\lambda_j)^2 \leqslant \sum_{\ell=1}^{m} (\nu_\ell)^2 \cdot \sum_{\ell=1}^{m} (\langle\psi_j|U^\dagger E_\ell U|\psi_j\rangle)^2. \tag{S5}$$

Meanwhile, recalling that all E_ℓ are Pauli matrix tensor product, we can obtain an upper bound below:

$$(\lambda_j)^2 \leqslant m \sum_{\ell=1}^{m} (\nu_\ell)^2. \tag{S6}$$

Ranging j in $\{1, ..., N\}$ in Eq. (S6), the maximal eigenvalue is upper bounded by $\sqrt{m}\|\boldsymbol{\nu}\|_2$, validating the claim.

Proposition 5. For any parameterized Hamiltonian $H(\boldsymbol{\nu}) = \sum_{\ell=1}^{m} \nu_\ell E_\ell$ with $E_\ell \in \{X, Y, Z, I\}^{\otimes n}$ and $\boldsymbol{\nu} \in \mathbb{R}^m$ and constant $\beta > 0$, suppose we are given access to a parameterized quantum circuit $U(\boldsymbol{\theta})$ that can learn $H(\boldsymbol{\nu})$'s eigenvectors, then the objective function $C(\mathbf{p})$ can be computed up to precision ϵ with probability larger than 2/3 by taking $T = O(m\|\boldsymbol{\nu}\|_2^2/\epsilon^2)$ samples. Furthermore, the probability can be improved to $1 - \eta$ costing an additional multiplicative factor of $O(\log(1/\eta))$.

Proof. Since the expression $\sum_{j=1}^{N} p_j \lambda_j$ is regarded as an expectation, then we can estimate it by the sample mean with high accuracy and probability. To be specific, let X denote a random variable that takes value λ_j with probability p_j. Then, this expression can be written as

$$\mathbf{E}[X] = \sum_{j=1}^{N} p_j \lambda_j. \tag{S7}$$

Furthermore, recall Chebyshev's inequality, then we have

$$\Pr\left(|\bar{X} - \mathbf{E}[X]| \leqslant \epsilon\right) \geqslant 1 - \frac{\mathbf{Var}[X]}{T\epsilon^2}. \tag{S8}$$

where $\bar{X} = \frac{1}{T}(X_1 + X_2 + ... + X_T)$ and $\mathbf{Var}[X]$ is the variance of X. Technically, we can set large T to increase the probability. Here, we only need to choose T such that

$$\frac{\mathbf{Var}[X]}{T\epsilon^2} = \frac{2}{3}. \tag{S9}$$

Note that the second moment $\mathbf{E}[X^2]$ bounds the variance $\mathbf{Var}[X]$. Meanwhile, the second moment of X is bounded by the squared spectral norm of H, shown below.

$$\mathbf{E}[X^2] = \sum_{j=1}^{N} p_j (\lambda_j)^2 \tag{S10}$$

$$\leqslant \sum_{j=1}^{N} p_j \|H(\boldsymbol{\nu})\|^2 \tag{S11}$$

$$= \|H(\boldsymbol{\nu})\|^2. \tag{S12}$$

The inequality is due to the fact that each eigenvalue is less than the spectral norm. Apply Lemma S1, then we will obtain an bound on T:

$$T = \frac{3\mathbf{Var}[X]}{2\epsilon^2} \leqslant \frac{3\mathbf{E}[X^2]}{2\epsilon^2} \leqslant \frac{3m\|\boldsymbol{\nu}\|_2^2}{2\epsilon^2}. \tag{S13}$$

Lastly, according to the Chernoff bound, we can boost the probability to $1-\eta$ for any $\eta > 0$ by repeatedly computing the sample mean $O(\log(1/\eta))$ times and taking the median of all sample means.

Appendix C Proof for Proposition 2

Lemma S2. Consider a parameterized Hamiltonian $H(\boldsymbol{\nu}) = \sum_{\ell=1}^{m} \nu_\ell E_\ell$ with $E_\ell \in \{X, Y, Z, I\}^{\otimes n}$. For any unitary U and state $|\psi\rangle$, estimating the value $\langle\psi|U^\dagger H(\boldsymbol{\nu})U|\psi\rangle$ up to precision ϵ with probability at least $1-\eta$ requires a sample complexity of

$$O\left(\frac{m\|\boldsymbol{\nu}\|_1^2 \log(m/\eta)}{\epsilon^2}\right). \tag{S1}$$

Proof. First, we rewrite the value $\langle\psi|U^\dagger H(\boldsymbol{\nu})U|\psi\rangle$ as follows:

$$\langle\psi|U^\dagger H(\boldsymbol{\nu})U|\psi\rangle = \sum_{\ell=1}^{m} \nu_\ell \langle\psi|U^\dagger E_\ell U|\psi\rangle. \tag{S2}$$

Second, we count the required number of measurements to estimate the value $\langle\psi|U^\dagger E_\ell U|\psi\rangle$ up to precision $\epsilon/\|\boldsymbol{\nu}\|_1$ with probability at least $1-\eta/m$, where $\|\cdot\|_1$ denotes the ℓ_1-norm. Since the Pauli operator, E_ℓ, has eigenvalues ± 1, we can partition E_ℓ's eigenvectors into two sets, corresponding to positive and negative eigenvalues, respectively. For convenience, we call the measurement outcome corresponding to eigenvalue 1 as the positive measurement outcome and the rest as the negative measurement outcome. We define a random variable X in the sense that

$$X = \begin{cases} 1, & \Pr\text{[Positive measurement outcome]} \\ -1, & \Pr\text{[Negative measurement outcome]} \end{cases} \tag{S3}$$

It is easy to verify that $\mathbf{E}[X] = \langle\psi|U^\dagger E_\ell U|\psi\rangle$. Thus, an approach to compute value $\langle\psi|U^\dagger E_\ell U|\psi\rangle$ is computing an estimate for the expectation $\mathbf{E}[X]$. Meanwhile, consider that $\mathbf{E}[X^2] \leqslant 1$, then the required number of samples is $O(\|\boldsymbol{\nu}\|_1^2 \log(m/\eta)/\epsilon^2)$.

Lastly, for $\langle\psi|U^\dagger H(\boldsymbol{\nu})U|\psi\rangle$, the estimate's maximal error is $\|\boldsymbol{\nu}\|_1 \cdot \epsilon/\|\boldsymbol{\nu}\|_1 = \epsilon$. By union bound, the overall failure probability is less than $m \cdot \eta/m = \eta$. Thus, the claim is proved.

Proposition 6. Consider a parameterized Hamiltonian $H(\boldsymbol{\nu}) = \sum_{\ell=1}^{m} \nu_\ell E_\ell$ with Pauli operators $E_\ell \in \{X, Y, Z, I\}^{\otimes n}$ and constants $\nu_\ell \in [-1, 1]$. Given any constants $\epsilon > 0$, $\eta \in (0, 1)$, $\beta > 0$, the objective function $M(\boldsymbol{\theta})$ in SVQE can be estimated up to precision ϵ with probability at least $1-\eta$, costing TD samples with $T = O(m\|\boldsymbol{\nu}\|_2^2/\epsilon^2)$ and $D = O(\log(1/\eta))$. Besides, the total number of measurements is given below:

$$O\left(\frac{mTD\|\boldsymbol{\nu}\|_1^2(n + \log(m/\eta))}{\epsilon^2}\right). \tag{S4}$$

Proof. Let Y denote a random variable that takes value $\langle\psi_j|U^\dagger(\boldsymbol{\theta})H(\boldsymbol{\nu})U(\boldsymbol{\theta})|\psi_j\rangle$ with probability q_j, then the objective function $M(\boldsymbol{\theta})$ can be rewritten as

$$\mathbf{E}[Y] = M(\boldsymbol{\theta}). \tag{S5}$$

By Chebyshev's inequality, the expectation can be computed by taking enough samples of Y and averaging them. Note that the variance of Y determines the number of samples, and the absolute value Y is less than the spectral norm $\|H(\boldsymbol{\nu})\|$, i.e., $|Y| \leqslant \|H(\boldsymbol{\nu})\|$. Along with Lemma S1, it is easy to see that the required number of Y's samples for obtaining an estimate with error $\epsilon/2$ and probability larger than $2/3$ is $T = O(m\|\boldsymbol{\nu}\|_2^2/\epsilon^2)$. Furthermore, by Chernoff bounds, the probability can be improved to $1-\eta/2$ at an additional cost of multiplicative factor of $D = O(\log(1/\eta))$.

On the other hand, each sample Y's value has to be determined by performing the measurement. Since $|\psi_j\rangle$ is a computational basis, hence Y can take at most 2^n different values. To ensure the probability for estimating $\mathbf{E}[Y]$ larger than $1-\eta$, the probability of each estimate $\langle\psi_j|U^\dagger(\boldsymbol{\theta})H(\boldsymbol{\nu})U(\boldsymbol{\theta})|\psi_j\rangle$ only needs to be at least $1-\eta/2^{n+1}$. By union bound, the overall failure probability is at most $\eta/2 + \eta \cdot \frac{TD}{2^{n+1}} < \eta$ (For large Hamiltonians, the number of samples TD can be significantly less than dimension 2^n). Besides, according to Lemma S2, $\langle\psi_j|U^\dagger(\boldsymbol{\theta})H(\boldsymbol{\nu})U(\boldsymbol{\theta})|\psi_j\rangle$'s estimate within accuracy $\epsilon/2$ and probability $1-\eta/2^{n+1}$ requires a sample complexity of $O(m\|\boldsymbol{\nu}\|_1^2(n + \log(m/\eta))/\epsilon^2)$. Thus, the overall number of measurements is the product of the number of samples $TD = O(m\|\boldsymbol{\nu}\|_2^2 \log(1/\eta)/\epsilon^2)$ and each sample's sample complexity $O(m\|\boldsymbol{\nu}\|_1^2(n + \log(m/\eta))/\epsilon^2)$. In other words, the objective function $M(\boldsymbol{\theta})$'s estimate within error ϵ and probability $1-\eta$ requires a sample complexity of

$$O\left(TD \cdot \frac{m\|\boldsymbol{\nu}\|_1^2(n + \log(m/\eta))}{\epsilon^2}\right) = O\left(\frac{m^2\|\boldsymbol{\nu}\|_1^2\|\boldsymbol{\nu}\|_2^2 \log(1/\eta)(n + \log(m/\eta))}{\epsilon^4}\right).$$

Appendix D Proof for Proposition 3

Lemma S3. Let $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_1, ..., \widehat{\lambda}_N)$ denote the estimated eigenvalues from SVQE and define a function $G(\mathbf{p})$ as follows:

$$G(\mathbf{p}) = \sum_{j=1}^{N} p_j \widehat{\lambda}_j + \beta^{-1} \sum_{j=1}^{N} p_j \log p_j. \tag{S1}$$

Let $\widehat{\mathbf{p}}^*$ be the global optimal point of $G(\mathbf{p})$, that is, for any probability distribution $\mathbf{p}$, we have $G(\widehat{\mathbf{p}}^*) \leqslant G(\mathbf{p})$. Meanwhile, suppose $\mathbf{p}^*$ is the global optimal point of $C(\mathbf{p})$. Then, we have

$$|G(\widehat{\mathbf{p}}^*) - C(\mathbf{p}^*)| \leqslant \max\left\{\mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|], \mathbf{E}_{\mathbf{p}^*}[|\widehat{\lambda} - \lambda|]\right\}, \tag{S2}$$

where

$$\mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|] = \sum_{j=1}^{N} \widehat{p}_j^* |\widehat{\lambda}_j - \lambda_j|, \tag{S3}$$

$$\mathbf{E}_{\mathbf{p}^*}[|\widehat{\lambda} - \lambda|] = \sum_{j=1}^{N} p_j^* |\widehat{\lambda}_j - \lambda_j|. \tag{S4}$$

Proof. Since functions $C(\mathbf{p})$ and $G(\mathbf{p})$ reach their global minimums at points $\mathbf{p}^*$ and $\widehat{\mathbf{p}}^*$ respectively, then we have

$$C(\widehat{\mathbf{p}}^*) \geqslant C(\mathbf{p}^*), \tag{S5}$$

$$G(\widehat{\mathbf{p}}^*) \leqslant G(\mathbf{p}^*). \tag{S6}$$

Besides, we also have another relation:

$$|C(\mathbf{p}) - G(\mathbf{p})| = \sum_{j=1}^{N} p_j |(\widehat{\lambda}_j - \lambda_j)|, \tag{S7}$$

where $\|\cdot\|_\infty$ denotes the maximum norm.

Combining the above inequalities, we have the following result:

$$C(\mathbf{p}^*) \leqslant C(\widehat{\mathbf{p}}^*) \leqslant G(\widehat{\mathbf{p}}^*) + \mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|] \leqslant G(\mathbf{p}^*) + \mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|] \leqslant C(\mathbf{p}^*) + \mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|] + \mathbf{E}_{\mathbf{p}^*}[|\widehat{\lambda} - \lambda|]. \tag{S8}$$

Then the inequality in Eq. (S2) is proved.

Proposition 7 (Correctness)**.** Consider a parameterized Hamiltonian $H(\boldsymbol{\nu})$ and its Gibbs state $\rho_\beta(\boldsymbol{\nu})$. Suppose the $U(\boldsymbol{\theta})$ from SVQE and $\widehat{\mathbf{p}}^*$ from log-partition function estimation procedure are optimal. Define a density operator ρ_β^* as follows:

$$\rho_\beta^* = \sum_{j=1}^{N} \widehat{p}_j^* \cdot U(\boldsymbol{\theta}) \, |\psi_j\rangle\langle\psi_j| \, U^\dagger(\boldsymbol{\theta}). \tag{S9}$$

where $\{|\psi\rangle_j\}$ denote the computational basis. Denote the estimated eigenvalues by $\widehat{\boldsymbol{\lambda}}$, where $\widehat{\lambda}_j = \langle\psi_j|U^\dagger(\boldsymbol{\theta})H(\boldsymbol{\nu})U(\boldsymbol{\theta})|\psi_j\rangle$. Then, ρ_β^* is an approximate of $\rho_\beta(\boldsymbol{\nu})$ in the sense that

$$D(\rho_\beta^*, \rho_\beta(\boldsymbol{\nu})) \leqslant \sqrt{2\beta \max\left\{\mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|], \mathbf{E}_{\mathbf{p}^*}[|\widehat{\lambda} - \lambda|]\right\}}. \tag{S10}$$

where $D(\cdot, \cdot)$ denotes the trace distance, $\boldsymbol{\lambda}$ represent $H(\boldsymbol{\nu})$'s true eigenvalues.

Proof. Recalling the expressions of $C(\mathbf{p}^*)$ and $G(\widehat{\mathbf{p}}^*)$ in Eqs. (12) & (S1), it is easy to verify the following inequalities:

$$F(\rho_\beta(\boldsymbol{\nu})) = C(\mathbf{p}^*), \tag{S11}$$

$$F(\rho_\beta^*) = G(\widehat{\mathbf{p}}^*). \tag{S12}$$

where F denotes the free energy, i.e., $F(\rho) = \mathrm{Tr}(H\rho) - \beta^{-1}S(\rho)$.

Using the result in Lemma S3, we will obtain the following inequality.

$$|F(\rho_\beta^*) - F(\rho_\beta(\boldsymbol{\nu}))| = |G(\widehat{\mathbf{p}}^*) - C(\mathbf{p}^*)| \leqslant \max\left\{\mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|], \mathbf{E}_{\mathbf{p}^*}[|\widehat{\lambda} - \lambda|]\right\}. \tag{S13}$$

In the meanwhile, a property of the free energy says that

$$F(\rho_\beta^*) = F(\rho_\beta(\boldsymbol{\nu})) + \beta^{-1} S(\rho_\beta^* \| \rho_\beta(\boldsymbol{\nu})). \tag{S14}$$

where $S(\rho_\beta^* \| \rho_\beta(\boldsymbol{\nu}))$ is the relative entropy. Rewriting the above equation as follows:

$$F(\rho_\beta^*) - F(\rho_\beta(\boldsymbol{\nu})) = \beta^{-1} S(\rho_\beta^* \| \rho_\beta(\boldsymbol{\nu})). \tag{S15}$$

Combining the relations in Eqs. (S13) and (S15), we obtain the following inequality:

$$S(\rho_\beta^* \| \rho_\beta(\boldsymbol{\nu})) \leqslant \beta \max\left\{\mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|], \mathbf{E}_{\mathbf{p}^*}[|\widehat{\lambda} - \lambda|]\right\}. \tag{S16}$$

Lastly, according to Pinsker's inequality, the above inequality immediately leads to a bound on the trace distance between ρ_β and ρ_β^* in the sense that

$$D(\rho_\beta^*, \rho_\beta(\boldsymbol{\nu})) \leqslant \sqrt{2S(\rho_\beta^* \| \rho_\beta(\boldsymbol{\nu}))} \leqslant \sqrt{2\beta \max\left\{\mathbf{E}_{\widehat{\mathbf{p}}^*}[|\widehat{\lambda} - \lambda|], \mathbf{E}_{\mathbf{p}^*}[|\widehat{\lambda} - \lambda|]\right\}}. \tag{S17}$$

The the claimed is proved.

Appendix E Proof for Proposition 4

Proposition 8 (Sample complexity)**.** Given $\epsilon > 0$ and $\eta \in (0, 1)$, Algorithm 3 can compute an estimate for the gradient $\nabla L(\boldsymbol{\nu})$ up to precision ϵ with probability larger than $1 - \eta$. Particularly, the overall number of samples is $KD = O(\beta^2 \log(m/\eta)/\epsilon^2)$ with $K = O(\beta^2/\epsilon^2)$ and $D = O(\log(2m/\eta))$. Besides, the total number of measurements is $O(KD \cdot m\beta^2(n + \log(m/\eta))/\epsilon^2)$.

Proof. Let Z_ℓ denote the random variable that takes value $\langle\psi_j|U^\dagger(\boldsymbol{\theta})E_\ell U(\boldsymbol{\theta})|\psi_j\rangle$ with probability $\widehat{p}_j^*$, for all $\ell = 1, ..., m$. Then we have

$$\mathbf{E}[Z_\ell] = \sum_{j=1}^{N} \widehat{p}_j^* \cdot \langle\psi_j|U^\dagger(\boldsymbol{\theta})E_\ell U(\boldsymbol{\theta})|\psi_j\rangle. \tag{S1}$$

Thus partial derivative can be computed in the following way

$$\frac{\partial L(\boldsymbol{\nu})}{\partial \nu_\ell} \approx -\beta\mathbf{E}[Z_\ell] + \beta e_\ell. \tag{S2}$$

It implies that the estimate's error can be set as ϵ/β to ensure the gradient's maximal error less than ϵ.

Next, we determine the number of samples such that the overall failure probability for estimating the gradient is less than δ. Since the gradient has m partial derivatives, corresping to $\mathbf{E}[Z_\ell]$, thus it suffices to estimate each with probability larger than $1 - \delta/m$. Meanwhile, each mean $\mathbf{E}[Z_\ell]$ can be computed by sampling. Notice that all $|Z_\ell| \leqslant 1$, by Chebyshev's inequality, then it suffices to take $K = O(\beta^2/\epsilon^2)$ samples to compute an estimate for each $\mathbf{E}[Z_\ell]$ with precision $\epsilon/2\beta$ and probability larger than $2/3$. Furthermore, by Chernoff bounds, the probability can be improved to $1 - \eta/2m$ at an additional cost of multiplicative factor of $D = O(\log(2m/\eta))$. It is worth pointing out that, for each variable Z_ℓ, the samples are taken according to the same probability distribution $\widehat{\mathbf{p}}^*$, thus it is natural to use the sampled states $|\psi_{t_j^s}\rangle$ (cf. Algorithm 3) to compute all means $\mathbf{E}[Z_\ell]$. Then the total number of samples is $KD = O(\beta^2 \log(m/\eta)/\epsilon^2)$.

On the other hand, each value $\langle\psi_j|U^\dagger(\boldsymbol{\theta})E_\ell U(\boldsymbol{\theta})|\psi_j\rangle$ in Eq. (S1) has to be computed by performing the measurement. Note that there are 2^n values $\langle\psi_j|U^\dagger(\boldsymbol{\theta})E_\ell U(\boldsymbol{\theta})|\psi_j\rangle$ in all. To ensure the mean estimate's failure probability less than $\eta/2m$, it suffices to suppress each value's failure probability to $\eta/2^{n+1}m$. Following the same discussion in Lemma S2, the estimate for value $\langle\psi_j|U^\dagger(\boldsymbol{\theta})E_\ell U(\boldsymbol{\theta})|\psi_j\rangle$ can be computed up to precision $\epsilon/2\beta$ using $O(\beta^2 \log(2^{n+1}m/\eta)/\epsilon^2)$ measurements.

Regarding the failure probability, by union bound, the overall failure probability is at most $m \cdot (\eta/2m + KD \cdot \eta/2^{n+1}m)$, where KD is the number of samples $KD = O(\beta^2 \log(m/\eta)/\epsilon^2)$. Especially, for larger Hamiltonians, the number of measurements is usually less than the dimension 2^n. Thus, the overall failre probability is less than η.

Lastly, the total number of measurements is given below:

$$m \cdot KD \cdot O(\beta^2 \log(2^{n+1}m/\eta)/\epsilon^2) = O(m\beta^4 \log(m/\eta) \log(2^{n+1}m/\eta)/\epsilon^4). \tag{S3}$$