Supplementary Information for:
Quantum automated learning with provable and explainable trainability

CONTENTS

I. THEORETICAL BACKGROUND

A. Supervised learning

We start by introducing the basic framework of supervised learning [1]. Let $\mathcal{X}$ be the set of input data and $\mathcal{Y} = \{1, 2, \cdots, k\}$ be the set of labels. We assume that every input data $\boldsymbol{x} \in \mathcal{X}$ has a deterministic label $y(\boldsymbol{x}) \in \mathcal{Y}$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$. The goal of supervised learning is to find an algorithm $\mathcal{A}(\cdot)$ (probably randomized in quantum machine learning) such that, input a sample $\boldsymbol{x} \sim \mathcal{D}$, output the label $y(\boldsymbol{x})$ with high probability. To achieve this goal, we parametrize the learning model by parameters $\boldsymbol{\theta}$ and optimize the average loss

$$R(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} L(\boldsymbol{x}, y(\boldsymbol{x}); \boldsymbol{\theta}). \tag{S1}$$

Here $L(\boldsymbol{x}, y; \boldsymbol{\theta})$ is some loss function, usually a metric of the difference between the output distribution of $\mathcal{A}(\boldsymbol{x}; \boldsymbol{\theta})$ and the correct label y. $R(\boldsymbol{\theta})$ is called the risk or the prediction error of the model $\mathcal{A}(\cdot\,; \boldsymbol{\theta})$. However, the distribution $\mathcal{D}$ is unknown, so we cannot directly calculate $R(\boldsymbol{\theta})$. Instead, we sample a training dataset $S = \{(\boldsymbol{x}_i, y_i = f(\boldsymbol{x}_i))\}_{i=1}^m$ from $\mathcal{D}$, and optimize the following empirical risk or training error:

$$\hat{R}_S(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{x}_i, y_i; \boldsymbol{\theta}). \tag{S2}$$

According to the simple decomposition $R(\boldsymbol{\theta}) = \hat{R}_S(\boldsymbol{\theta}) + (R(\boldsymbol{\theta}) - \hat{R}_S(\boldsymbol{\theta}))$, the success of supervised learning depends on two important factors: trainability and generalization. In short, trainability asks whether we can efficiently find $\boldsymbol{\theta}$ with low empirical risk, while generalization asks whether the generalization gap $\mathrm{gen}_S(\boldsymbol{\theta}) = R(\boldsymbol{\theta}) - \hat{R}_S(\boldsymbol{\theta})$ is upper bounded, i.e., whether the good performance on the training set S can be generalized to unseen data.

B. Variational quantum learning models

For conventional gradient-based quantum learning approaches [2], a learning algorithm $\mathcal{A}(\boldsymbol{x};\boldsymbol{\theta})$ executes a variational quantum circuit $U(\boldsymbol{\theta})$ to a data-encoded state $|\phi(\boldsymbol{x})\rangle$ before performing certain measurements to make the prediction. Assuming the measurement observable to be O_M, the output from the variational circuit is the expectation value $\left\langle\phi(\boldsymbol{x})|U(\boldsymbol{\theta})^\dagger O_M U(\boldsymbol{\theta})|\phi(\boldsymbol{x})\right\rangle$. The loss function is often defined as a function of this value, where commonly used forms include mean square error and cross-entropy. For a training task, the average loss value over a given set of training data is defined as the empirical risk, where schemes based on gradient descents are widely exploited to minimize it and find the optimal parameters $\boldsymbol{\theta}^*$. In the quantum machine learning realm, there are various methods proposed to calculate the gradients with respect to circuit parameters, including finite differences, the parameter-shift rules, and quantum natural gradients [3–5].

Quantum neural networks have demonstrated promising generalization capabilities in various learning settings [6, 7]. Intuitively, when the number of training data points exceeds the degrees of freedom in the parameter space, the generalization gap of the optimized parameters is typically bounded by a small constant. However, the practical trainability of quantum neural networks remains a significant challenge. A key bottleneck lies in the computational cost of estimating gradients with respect to the circuit parameters. For instance, computing the gradient of a single parameter accurately often requires executing the variational circuit thousands of times, even when employing comparably efficient parameter-shift rules. This process becomes increasingly time-consuming and impractical as the number of parameters grows.

Furthermore, the loss landscape of quantum neural networks can be highly non-convex and challenging to navigate. As shown in ref. [8], the loss function of quantum neural networks exhibits exponentially many local minima, which can trap optimization algorithms and hinder convergence. In parallel, the phenomenon of barren plateaus, first identified in ref. [9], poses a critical issue: the gradients of the loss function tend to vanish exponentially with the number of qubits, especially in deep quantum circuits. In such cases, the loss landscape becomes effectively flat, making it extremely difficult to identify a direction for optimization. The presence of barren plateaus is closely tied to the randomness and entanglement structure of the quantum circuit, as well as the choice of cost function and initial parameterization, which severely threads the scalability and practical utility of quantum neural networks for large-scale problems [9–14].

II. THE AUTOMATED LEARNING STRATEGY

In this section, we provide more technical details about the quantum automated learning strategy.

A. Choose an appropriate number of qubits

To carry out the QAL protocol, the first step is to decide an appropriate number of qubits n. Since an n-qubit state lives in a $O(2^n)$-dimensional Hilbert space and thus bears $O(2^n)$ degrees of freedom, one natural choice is $n = O(\log|\boldsymbol{x}|)$, where $|\boldsymbol{x}|$ is the dimension of data sample $\boldsymbol{x}$. However, we remark that the choice of n is much more flexible. For example, if we are classifying Hamiltonian data, it is more natural to set n to be the system size of the Hamiltonian. If the data are images of size $L \times L$, setting $n = L$ may align with the two-dimensional structure better. On the other hand, sometimes it is possible to set n to be even much smaller than $\log(|\boldsymbol{x}|)$, since realistic data samples are usually believed to lie in a low-dimension manifold.

B. Encode data into unitaries

Once we pin down the number of qubits n, the next step is to encode data into n-qubit unitaries. Here we present the detailed data encoding schemes for quantum automated learning, which incorporates three distinct categories of data: classical data, Hamiltonian data, and quantum state data. An overview of the encoding methods is provided in Fig. S1, which summarizes the key approaches before delving into the detailed descriptions of each scheme.

Classical data, including images, text, or audio, can be transformed into a vector of numerical values, denoted as $\boldsymbol{x}$. The data vector $\boldsymbol{x}$ is encoded into parameterized quantum circuit. Specifically, each element of $\boldsymbol{x}$ is mapped into rotation angles of single-qubit gates. A single-qubit gate is parametrized as $G(\alpha,\beta,\gamma) = R_y(\alpha)R_z(\beta)R_y(\gamma)$, where $R_y(\alpha)$ and $R_z(\beta)$ are the rotations around the Y and Z axes of the Bloch sphere by angle α and β, respectively. Therefore, for a n-qubit quantum circuit, a layer of single-qubit gates can encode up to $3n$ entries of the vector $\boldsymbol{x}$. If we denote the dimension of $\boldsymbol{x}$ as l, then it is necessary to employ $\lceil\frac{l}{3n}\rceil$ layers of single-qubit gates. More concretely, considering a $3n$-dimensional vector $\mathbf{y}$, we define the encoding of a single-qubit layer as: $G(\mathbf{y}) = \otimes_{i=1}^{n} G_i(y_{2n+i}, y_{n+i}, y_i)$, where G_i acts on the i-th qubit, as illustrated in Fig. S2**a**. Then the k-th layer single-qubit encoding of the data vector $\boldsymbol{x}$ is defined as $G\left(\boldsymbol{x}_{3n(k-1)+1:3nk}\right)$, where $\boldsymbol{x}_{i:j}$ denotes the abbreviation of $(x_i, x_{i+1}, \ldots, x_j)$. In cases where the number of elements in $\boldsymbol{x}$ does not exactly divide by $3n$, padding with zeros is used to ensure uniformity.

	Classical data	**Hamiltonian data**	**Quantum state data**
Original data	$\boldsymbol{x}$	$H_{\boldsymbol{x}}$	$\lvert\boldsymbol{x}\rangle$
Encoding unitary	Parameterized quantum circuit $U(\boldsymbol{x})$	Real-time evolution $\mathrm{e}^{-iH_{\boldsymbol{x}}t}$	$H_{\lvert\boldsymbol{x}\rangle} = (\langle\boldsymbol{x}\rvert \otimes \mathbf{I}_{\mathrm{n}})H(\langle\boldsymbol{x}\rvert \otimes \mathbf{I}_{\mathrm{n}})$ $\mathrm{e}^{-iH_{\lvert\boldsymbol{x}\rangle}}$

FIG. S1: **Overview of three different data encoding methods.** We encode classical data $\boldsymbol{x}$ into parameterized quantum circuit $U(\boldsymbol{x})$; encode Hamiltonian data $H_{\boldsymbol{x}}$ into its real-time evolution $e^{-iH_{\boldsymbol{x}}t}$; and encode quantum state $|\boldsymbol{x}\rangle$ first into an n-qubit Hamiltonian: $H_{|\boldsymbol{x}\rangle} = (\langle\boldsymbol{x}| \otimes \mathbf{I}_n)H(|\boldsymbol{x}\rangle \otimes \mathbf{I}_n)$, and then encode it to a unitary $e^{-iH_{|\boldsymbol{x}\rangle}}$.

Between two layers of single-qubit gates, we insert a layer of two-qubit gates to entangle the qubits, leading to the spread of information. This layer of two-qubit gates is composed of a CNOT-gate block A and a CZ-gate block B. Each block consists of two layers of two-qubit gates: in the first layer, the odd-numbered qubits act as the control qubits, while in the second layer, the even-numbered qubits serve as the control qubits. In both layers, each control qubit targets the subsequent qubit in the sequence. Mathematically, we define: $A = \left(\otimes_{i=1}^{\lfloor\frac{n-1}{2}\rfloor}\mathrm{CNOT}_{2i,2i+1}\right)\left(\otimes_{i=1}^{\lfloor\frac{n}{2}\rfloor}\mathrm{CNOT}_{2i-1,2i}\right)$ and $B = \left(\otimes_{i=1}^{\lfloor\frac{n-1}{2}\rfloor}\mathrm{CZ}_{2i,2i+1}\right)\left(\otimes_{i=1}^{\lfloor\frac{n}{2}\rfloor}\mathrm{CZ}_{2i-1,2i}\right)$, as shown in Fig. S2**a**. To ensure that all elements of the data vector can influence the measured qubits used for prediction, we add additional $\lfloor\frac{n}{2}\rfloor - 1$ layers of two-qubit gates.

The final unitary encoding $U(\boldsymbol{x})$ for classical data is then given by a sequence of single- and two-qubit gates:

$$(BA)^{\lfloor\frac{n}{2}\rfloor-1}G(\boldsymbol{x}_{3n(d-1)+1:3nd})\cdots BAG(\boldsymbol{x}_{3n+1:6n})BAG(\boldsymbol{x}_{1:3n}), \tag{S3}$$

where $d = \lceil\frac{l}{3n}\rceil$ and $\boldsymbol{x}$ is padded with zeros if $3nd > l$. as illustrated in Fig. S2**b**.

For Hamiltonian data, we encode the Hamiltonian $H_{\boldsymbol{x}}$ through its real-time evolution $e^{-iH_{\boldsymbol{x}}t}$. This encoding inherently captures the time evolution of quantum states governed by the Schrödinger equation. Both $e^{-iH_{\boldsymbol{x}}t}$ and its reverse evolution $e^{iH_{\boldsymbol{x}}t}$ can be implemented through quantum Hamiltonian simulation techniques [15–18]. One may exploit other encoding schemes for $H_{\boldsymbol{x}}$. In practice, we find that our real-time evolution encoding works well, as shown in our numerical simulations (Fig. **??**)

For quantum state classification, each datum is a quantum state $|\boldsymbol{x}\rangle$ of s qubits. In order to carry out the QAL protocol for classifying $|\boldsymbol{x}\rangle$, we need to encode $|\boldsymbol{x}\rangle$ into a unitary $U_{|\boldsymbol{x}\rangle}$. We fix an $(s+n)$-qubit Hamiltonian H. We first encode the quantum state $|\boldsymbol{x}\rangle$ into an n-qubit Hamiltonian: $H_{|\boldsymbol{x}\rangle} = (\langle\boldsymbol{x}| \otimes \mathbf{I}_n)H(|\boldsymbol{x}\rangle \otimes \mathbf{I}_n)$, and then encode it to a unitary $U_{|\boldsymbol{x}\rangle} = e^{-iH_{|\boldsymbol{x}\rangle}}$. This unitary can be efficiently implemented using copies of $|\boldsymbol{x}\rangle$ and real-time evolution e^{-iHt}, inspired by the Lloyd-Mohseni-Rebentrost protocol [19] that implements $e^{-i\rho t}$ from copies of ρ. More concretely, for any state ρ straightforward calculations yield

$$\begin{aligned}\mathrm{tr}_{\leq l}(e^{-iH\Delta t}(|\boldsymbol{x}\rangle\langle\boldsymbol{x}| \otimes \rho)e^{iH\Delta t}) &= \rho + \mathrm{tr}_{\leq l}(-iH\Delta t(|\boldsymbol{x}\rangle\langle\boldsymbol{x}| \otimes \rho) + (|\boldsymbol{x}\rangle\langle\boldsymbol{x}| \otimes \rho)iH\Delta t)) + O((\Delta t)^2)\\ &= \rho - i\Delta t[H_{|\boldsymbol{x}\rangle}, \rho] + O((\Delta t)^2)\\ &= e^{-iH_{|\boldsymbol{x}\rangle}\Delta t}\rho e^{iH_{|\boldsymbol{x}\rangle}\Delta t} + O((\Delta t)^2).\end{aligned} \tag{S4}$$

Therefore, if we apply $e^{-iH\Delta t}$ to $|\boldsymbol{x}\rangle\langle\boldsymbol{x}| \otimes \rho$, we effectively apply $e^{-iH_{|\boldsymbol{x}\rangle}\Delta t}$ to ρ, up to a second-order error $O((\Delta t)^2)$. Repeating this procedure $1/\Delta t$ times, we effectively apply $e^{-iH_{|\boldsymbol{x}\rangle}}$ to ρ up to error $O(\Delta t)$. By choosing Δt sufficiently small, we can approximate $e^{-iH_{|\boldsymbol{x}\rangle}}$ up to any precision. In our numerical simulations, we choose H as a random 4-local Hamiltonian. Each term of H is a 4-body Pauli interaction on 4 random positions with a random interaction strength between -1 and 1 (Fig. **??**).

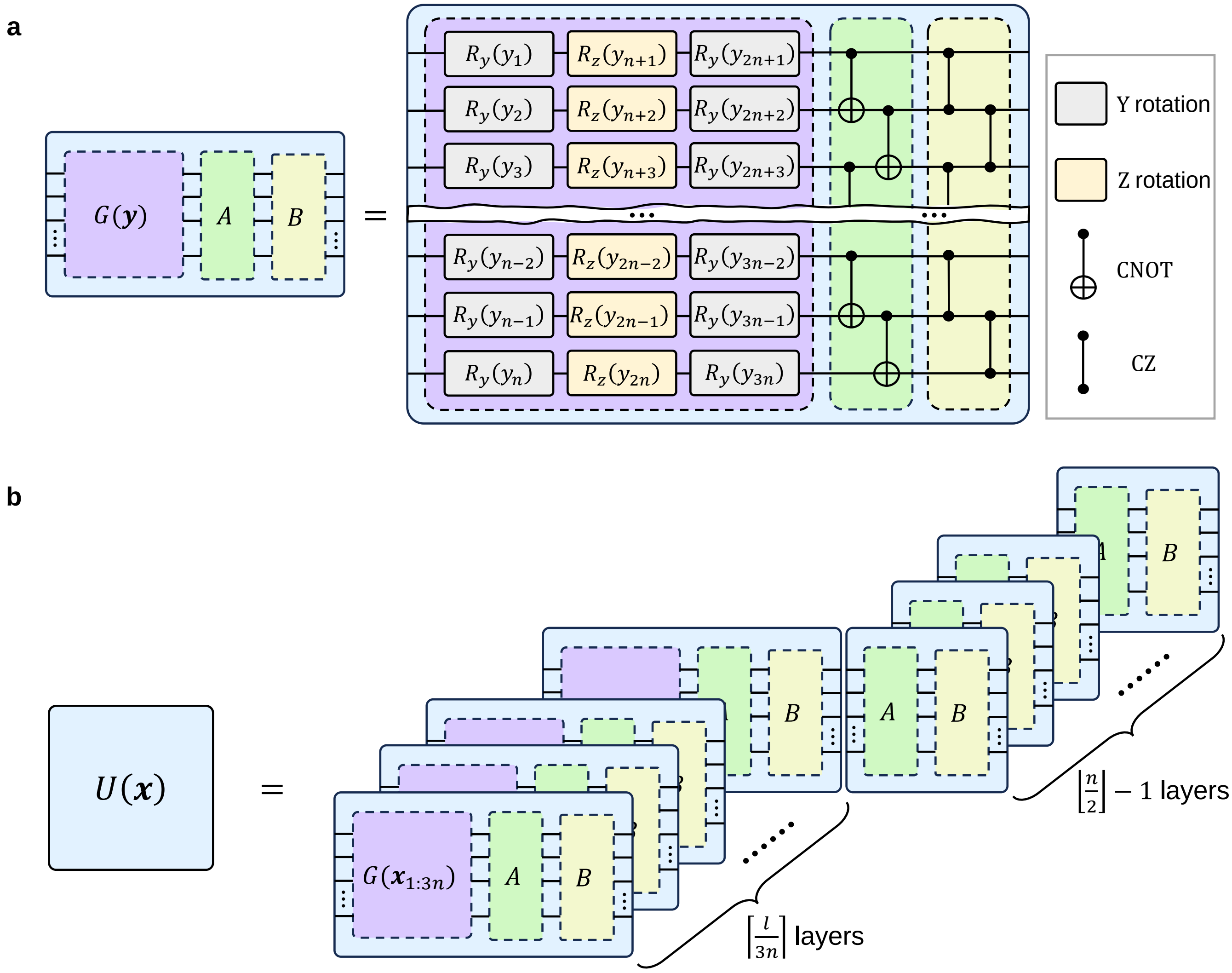

FIG. S2: **The classical data encoding scheme. a**, Illustrates an encoding layer for a $3n$-dimensional vector **y** ($G(\mathbf{y})$). The layer consists of three layers of single-qubit gates (R_y, R_z, R_y), denoted by the purple block; two layers of CNOT two-qubit gates, denoted by the green block; and two layers of CZ two-qubit gates, denoted by the yellow block. **b**, Illustrates the complete encoding for an l-dimensional vector $\boldsymbol{x}$. The encoding scheme involves $\lceil \frac{l}{3n} \rceil$ layers of single-qubit and two-qubit gates ($BAG\left(\boldsymbol{x}_{3n(k-1)+1:3nk}\right)$), followed by $\lfloor \frac{n}{2} \rfloor - 1$ layers of entangling gates (BA).

C. Training process and efficient compiling of U_y

In this subsection, we show that the required unitary U_y to implement the target-oriented perturbation can be compiled in an efficient way, with the count of CNOT gates scales logarithmically with the number of classes k. Recalling that $U_y = M_y \otimes Z + \sqrt{I - M_y^2} \otimes X$ acts on $\lceil \log k \rceil + 1$ qubits, and $M_y = |y\rangle\langle y| + (1-\eta)(\mathbf{I} - |y\rangle\langle y|)$, we arrive at the following decomposition:

$$\begin{aligned} U_y &= |y\rangle\langle y| \otimes Z + (I - |y\rangle\langle y|) \otimes ((1-\eta)Z + \sqrt{2\eta - \eta^2}X) \\ &= [I \otimes ((1-\eta)Z + \sqrt{2\eta - \eta^2}X)][|y\rangle\langle y| \otimes ((1-\eta)Z + \sqrt{2\eta - \eta^2}X)Z + (I - |y\rangle\langle y|) \otimes I]. \end{aligned} \tag{S5}$$

The first part of the above equation $I \otimes ((1-\eta)Z + \sqrt{2\eta - \eta^2}X)$ is a single-qubit gate and can be compiled into a constant number of gates. The second part is a $\lceil \log k \rceil$-controlled $SU(2)$ gate, as $|y\rangle\langle y|$ involves $\lceil \log k \rceil$ qubits and $(1-\eta)Z + \sqrt{2\eta - \eta^2}X)Z$ is a $SU(2)$ gate. Utilizing the techniques in Ref. [20], such a multi-controlled $SU(2)$ gate can be compiled into $O(\log k)$ CNOT gates. This leads to the conclusion that the whole U_y can be compiled in an efficient way with about $\log k$ two-qubit gates.

D. Prediction and evaluataion

The predicted label of $\boldsymbol{x}$ is the outcome of measurements in the computational basis performed on $U(\boldsymbol{x})\,|\psi\rangle$. So the probability of correct prediction (i.e., the accuracy) is $\langle\psi|U(\boldsymbol{x})^\dagger\Pi_{f(\boldsymbol{x})}U(\boldsymbol{x})|\psi\rangle = 1-\langle\psi|H_{\boldsymbol{x}}|\psi\rangle$. This accuracy can be amplified by repetition. Indeed, once the data in each step has been sampled, the training process is a fixed quantum circuit (with post-selection). So we can run the circuit multiple times to obtain K copies of the final states $|\psi\rangle$. To predict the label of a new unseen data sample $\boldsymbol{x}$, we measure $U(\boldsymbol{x})\,|\psi\rangle$ in the computational basis for each copy and do a majority vote. For simplicity, we consider the binary classification problem and assume K is odd, then the probability of correct label (called K-accuracy) is

$$p_K(\boldsymbol{x}, |\psi\rangle) = \sum_{r=0}^{(K-1)/2} \binom{K}{r} \langle\psi|H_{\boldsymbol{x}}|\psi\rangle^r \left(1-\langle\psi|H_{\boldsymbol{x}}|\psi\rangle\right)^{K-r}.$$

When $K=1$, this reduces to the single-copy accuracy $1-\langle\psi|H_{\boldsymbol{x}}|\psi\rangle$. When $K=\infty$, the K-accuracy equals to the step function $1[\langle\psi|H_{\boldsymbol{x}}|\psi\rangle < 1/2]$. Throughout this paper, we call K the number of trials.

E. Gradient perspective

As mentioned in the main text, the dissipation process of quantum automated learning actually implements the gradient descent algorithm in an automated manner. In the training step (iii), the quantum system is evolved through $U(\boldsymbol{x})$, M_y and $U(\boldsymbol{x})^\dagger$. We define $M_y = \Pi_{y(\boldsymbol{x})} + (1-\eta)(\mathbf{I}-\Pi_{y(\boldsymbol{x})})$ and $H_{\boldsymbol{x}} = \mathbf{I} - U(\boldsymbol{x})^\dagger\Pi_{y(\boldsymbol{x})}U(\boldsymbol{x})$, where $\Pi_{y(\boldsymbol{x})}$ denotes the measurement projection corresponding to the state encoding the label y. Then we get the updated (unnormalized) state $U(\boldsymbol{x})^\dagger M_y U(\boldsymbol{x})|\psi\rangle = (\mathbf{I}-\eta H_{\boldsymbol{x}})|\psi\rangle$. As mentioned above, the non-unitary perturbation M_y in training step (iii) of the QAL protocol is implemented by block encoding into a unitary with an ancillary qubit combined with post-selection. As a result, this step effectively updates the state with unitary transformation:

$$|\psi\rangle \leftarrow \frac{(I-\eta H_{\boldsymbol{x}})\,|\psi\rangle}{\|(I-\eta H_{\boldsymbol{x}})\,|\psi\rangle\|}, \tag{S6}$$

where $\|(I-\eta H_{\boldsymbol{x}})\,|\psi\rangle\|$ is a normalization factor whose square gives the success probability of post-selection.

The probability of correct prediction of a datum $\boldsymbol{x}$ reads $\langle\psi|U(\boldsymbol{x})^\dagger\Pi_{y(\boldsymbol{x})}U(\boldsymbol{x})|\psi\rangle = 1-\langle\psi|H_{\boldsymbol{x}}|\psi\rangle$. As mentioned in the main text, we define the loss function as the average failure probability: $\hat{R}_S(\psi) = \mathbb{E}_{\boldsymbol{x}\sim S}\,\langle\psi|H_{\boldsymbol{x}}|\psi\rangle$, where $\mathbb{E}_{\boldsymbol{x}\sim S}$ denotes the expectation and $\boldsymbol{x}\sim S$ means $\boldsymbol{x}$ is uniformly sampled from the training set S. From the perspective of conventional machine learning, we may also regard $|\psi\rangle$ as a variational state parametrized by a complex vector $\boldsymbol{\psi}$. Given that the expectation value $\langle\psi|H_{\boldsymbol{x}}|\psi\rangle$ is real, we transform the complex vector $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{2^n})$ into a fully real representation: $(a_1, b_1, \ldots, a_{2^n}, b_{n})$, where a_i and b_i denote the real and imaginary components of ψ_i respectively. Owing to the Hermitian property of H_x, we derive the partial derivatives: $\frac{\partial\langle\psi|H_{\boldsymbol{x}}|\psi\rangle}{\partial a_i} = 2Re(\sum_j (H_x)_{i,j}\psi_j)$ and $\frac{\partial\langle\psi|H_{\boldsymbol{x}}|\psi\rangle}{\partial b_i} = 2Im(\sum_j (H_x)_{i,j}\psi_j)$. This allows us to define $\frac{\partial\langle\psi|H_{\boldsymbol{x}}|\psi\rangle}{\partial\psi_i} = 2\sum_j (H_x)_{i,j}\psi_j$. Consequently, the gradient of $\langle\psi|H_{\boldsymbol{x}}|\psi\rangle$ with respect to ψ can be succinctly expressed as $2H_{\boldsymbol{x}}\,|\psi\rangle$. Therefore, the update rule in Eq. (S6) essentially implements the stochastic projected gradient descent algorithm to minimize the loss function $\hat{R}_S(\psi)$ with a batch size one. Here we use the term "projected" to emphasize the normalization after each update. From the stochastic gradient descent perspective, one may conclude that an initial state $|\psi\rangle$ can exponentially converge to a local minimum through updating rule (S6) on expectation [21]. However, a rigorous proof of convergence to the global minimum is unattainable in general. In fact, this is an inherent drawback for conventional gradient-based quantum learning approaches. Whereas, owing to the quadratic form of the loss function and the clear physical interpretation, we can rigorously prove that $|\psi\rangle$ converges exponentially to the global minimum for the QAL protocol, as discussed in the main text and detailed in the following sections.

III. PHYSICAL INTERPRETATION AND ANALYTICAL RESULTS

Throughout this section, we use $\|A\|_1$, $\|A\|_\infty$ to denote the trace norm (the summation of singular values) and spectral norm (the largest singular value), respectively. For two Hermitian A, B, denote $A \preceq B$ if $B-A$ is positive semi-definite. By definition, for any $\boldsymbol{x}$, $0 \preceq H_{\boldsymbol{x}} \preceq I$, and thus $0 \preceq H_S \preceq I$. We will use the following fact.

Lemma S1. *Let A, B, C be three Hermitian matrices such that $\|A\|_1 \le 1$, $0 \preceq B, C \preceq I$. Then*

$$\|BAB\|_1 \le 1, \|BAC + CAB\|_1 \le 2. \tag{S7}$$

Proof. We first prove the lemma when A is a normalized pure state $|a\rangle\langle a|$. Write $|b\rangle = B|a\rangle, |c\rangle = C|a\rangle$. Since $0 \preceq B, C \preceq I$, the norms of $|b\rangle, |c\rangle$ are at most 1. Then $\|BAB\|_1 = \||b\rangle\langle b|\|_1 \le 1$, $\|BAC + CAB\|_1 = \||b\rangle\langle c| + |c\rangle\langle b|\|_1 \le 2$. For general A, write the spectrum decomposition $A = \sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i|$. By triangle inequality, $\|BAB\|_1 \le \sum_i |\lambda_i| \|B|\lambda_i\rangle\langle\lambda_i| B\|_1 \le \sum_i |\lambda_i| = \|A\|_1 \le 1$ and $\|BAC + CAB\|_1 \le \sum_i |\lambda_i| \|B|\lambda_i\rangle\langle\lambda_i| C + C|\lambda_i\rangle\langle\lambda_i| B\|_1 \le \sum_i 2|\lambda_i| \le 2$. □

A. Formulation of the training process

In this subsection, we explain the training process from a physical perspective and derive an analytical characterization of the success probability of post-selection and the performance of the final model. Observe that the empirical risk is the energy of $|\psi\rangle$ under the Hamiltonian H_S, so finding the global minimum is equivalent to finding the ground state of H_S. We rewrite (S6) in the density matrix formalism:

$$\rho \xleftarrow{\boldsymbol{x}} (I - \eta H_{\boldsymbol{x}})\rho(I - \eta H_{\boldsymbol{x}}). \tag{S8}$$

Here we keep the post-state ρ unnormalized. Indeed, $\mathrm{Tr}(\rho)$ is the success probability of the post-selection. So the density matrix formalism helps us to keep track of the overall success probability. Another benefit of the density matrix formalism is that we can embed the randomness of the sample into the state. Since the datum $\boldsymbol{x}$ is uniformly sampled from S, the averaged post-state up to the second order term is

$$\begin{aligned} \rho &\leftarrow \mathbb{E}_{\boldsymbol{x}\sim S}(I - \eta H_{\boldsymbol{x}})\rho(I - \eta H_{\boldsymbol{x}}) \\ &\approx I - \eta(H_S\rho + \rho H_S) \\ &\approx e^{-\eta H_S}\rho e^{-\eta H_S}. \end{aligned} \tag{S9}$$

We make this approximation precise in the following lemma

Lemma S2.

$$\mathbb{E}_{\boldsymbol{x}\sim S}(I - \eta H_{\boldsymbol{x}})\rho(I - \eta H_{\boldsymbol{x}}) = e^{-\eta H_S}\rho e^{-\eta H_S} + \eta^2 O, \tag{S10}$$

where O is a Hermitian matrix with trace norm at most 4.

Proof. Let $R = (e^{-\eta H_S} - (I - \eta H_S))/\eta^2$. Since $0 \preceq H_S \preceq I$, all the eigenvalues of H_S are in $[0, 1]$. By the Taylor expansion of the exponential function, for any $x \in [0, 1]$, there exists $x^* \in [0, 1]$ such that $(e^{-\eta x} - (1 - \eta x))/\eta^2 = (x^*)^2/2 \in [0, 1/2]$. Therefore, R is a Hermitian matrix such that $0 \preceq R \preceq I/2$. By Lemma S1, we have

$$\begin{aligned} &\left\|\mathbb{E}_{\boldsymbol{x}\sim S}(I - \eta H_{\boldsymbol{x}})\rho(I - \eta H_{\boldsymbol{x}}) - e^{-\eta H_S}\rho e^{-\eta H_S}\right\|_1 \\ =&\left\|\rho - \eta(H_S\rho + \rho H_S) + \eta^2\mathbb{E}_{\boldsymbol{x}\sim S}H_{\boldsymbol{x}}\rho H_{\boldsymbol{x}} - (\eta^2 R + (I - \eta H_S))\rho(\eta^2 R + (I - \eta H_S))\right\|_1 \\ =&\eta^2\left\|\left(\mathbb{E}_{\boldsymbol{x}\sim S}H_{\boldsymbol{x}}\rho H_{\boldsymbol{x}} - H_S\rho H_S - \eta^2 R\rho R - (R\rho(I - \eta H_S) + (I - \eta H_S)\rho R)\right)\right\|_1 \\ \le&\eta^2(1 + 1 + \eta^2/4 + 1) < 4\eta^2. \end{aligned}$$

□

Up to the second order term, (S9) is the imaginary time evolution of ρ under H_S. Suppose the initial state is σ, then the averaged state after T epochs is $\rho = e^{-\eta T H_S}\sigma e^{-\eta T H_S}$. Let $\beta = \eta T$ be the summation of learning rates. We can approximate the success probability of possibility by $\mathrm{tr}\left(e^{-\beta H_S}\sigma e^{-\beta H_S}\right)$ and the loss by $\mathrm{tr}\left(H_S \frac{e^{-\beta H_S}\sigma e^{-\beta H_S}}{\mathrm{tr}\left(e^{-\beta H_S}\sigma e^{-\beta H_S}\right)}\right)$. We summarize and prove the results in the following theorem.

Theorem S1. *Suppose we train the QAL model with initial state σ for T steps, with learning rate η_t at step t. Define*

$$\beta = \sum_{i=1}^{T} \eta_t, \quad \gamma = \sum_{i=1}^{T} \eta_t^2, \quad \sigma(\beta) = e^{-\beta H_S}\sigma e^{-\beta H_S}. \tag{S11}$$

Averaging over choice of training samples, the success probability of post-selection is

$$\mathrm{tr}(\sigma(\beta)) + c_1\gamma, \tag{S12}$$

and the average loss conditioned on the success of post-selection is

$$\frac{\mathrm{tr}(H_S\sigma(\beta)) + c_2\gamma}{\mathrm{tr}(\sigma(\beta)) + c_1\gamma}. \tag{S13}$$

Here c_1, c_2 are two real numbers such that $|c_1|, |c_2| \le 4$.

Proof. Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T \sim S$ be the training samples in the T steps. We abbreviate $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t$ as $\boldsymbol{x}_{1:t}$. By (S8), the unnormalized state after step t is

$$\rho_t^{\boldsymbol{x}_{1:t}} = (I - \eta_t H_{\boldsymbol{x}_t}) \cdots (I - \eta_1 H_{\boldsymbol{x}_1}) \sigma (I - \eta_1 H_{\boldsymbol{x}_1}) \cdots (I - \eta_t H_{\boldsymbol{x}_t}). \tag{S14}$$

Given samples $\boldsymbol{x}_{1:T}$, $\operatorname{tr}(\rho_T^{\boldsymbol{x}_{1:T}})$ is the success probability of post-selection and $\operatorname{tr}\left(H_S \frac{\rho_T^{\boldsymbol{x}_{1:T}}}{\operatorname{tr}(\rho_T^{\boldsymbol{x}_{1:T}})}\right)$ is the loss conditioned on the success of post-selection. We now average over the choice of samples. Recursively apply Lemma S2 and Lemma S1, we have

$$\mathbb{E}_{\boldsymbol{x}_{1:T} \sim S^T} \rho_T^{\boldsymbol{x}_{1:T}} = \sigma(\beta) + \gamma O, \tag{S15}$$

for some Hermitian O with trace norm at most 4. The average success probability of post-selection is

$$\mathbb{E}_{\boldsymbol{x}_{1:T} \sim S^T} \operatorname{tr}(\rho_T^{\boldsymbol{x}_{1:T}}) = \operatorname{tr}(\sigma(\beta)) + c_1 \gamma, \tag{S16}$$

where $c_1 = \operatorname{tr}(O)$ satisfies $|c_1| \leq 4$. Now we calculate the averaged loss conditioned on the success of post-selection. For clarity, denote $q = \operatorname{tr}(\sigma(\beta)) + c_1\gamma$, $p = 1/|S|^T$ be the probability of sampling $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T$. Then conditioned on the success of post-selection, the conditional probability of sampling $\boldsymbol{x}_{1:T}$ is $p \operatorname{tr}(\rho_T^{\boldsymbol{x}_{1:T}})/q$. Therefore, the average loss conditioned on the success of post-selection is

$$\sum_{\boldsymbol{x}_{1:T} \sim S^T} \frac{p \operatorname{tr}(\rho_T^{\boldsymbol{x}_{1:T}})}{q} \operatorname{tr}\left(H_S \frac{\rho_T^{\boldsymbol{x}_{1:T}}}{\operatorname{tr}(\rho_T^{\boldsymbol{x}_{1:T}})}\right) = \frac{1}{q} \mathbb{E}_{\boldsymbol{x}_{1:T} \sim S^T} \operatorname{tr}(H_S \rho_T^{\boldsymbol{x}_{1:T}}) = \frac{\operatorname{tr}(H_S \sigma(\beta)) + c_2 \gamma}{\operatorname{tr}(\sigma(\beta)) + c_1 \gamma}, \tag{S17}$$

where $c_2 = \operatorname{tr}(H_S O)$ satisfies $|c_2| \leq 4$. □

According to the theorem, up to the second order term $c_1\gamma, c_2\gamma$, the training process behaves the same as the imaginary time evolution of σ under H_S. The effect of imaginary time evolution is clearer in the eigenbasis of H_S. Write the spectrum decomposition of H_S as $H_S = \sum_i E_i |E_i\rangle\langle E_i|$ and define $\sigma_i = \langle E_i|\sigma|E_i\rangle$ as the overlap of σ with the i-th eigenstate. Then

$$\sigma(\beta) = \sum_i \sigma_i e^{-2\beta E_i} |E_i\rangle\langle E_i|, \quad \frac{\operatorname{tr}(H_S \sigma(\beta))}{\operatorname{tr}(\sigma(\beta))} = \frac{\sum_i E_i \sigma_i e^{-2\beta E_i}}{\sum_i \sigma_i e^{-2\beta E_i}}. \tag{S18}$$

The weight of $|E_i\rangle\langle E_i|$, $\sigma_i e^{-2\beta E_i}$, decays exponentially with β. The decay is slower for lower energy eigenstates. Assume σ has a non-zero overlap with the ground space. As β goes up, eventually the weight of the ground space dominates, so $\rho_{\mathrm{n}}(\beta)$ converges to a ground state of H_S and the empirical risk converges to the global minimum. In the following, we will make this intuition rigorous in the presence of $c_1\gamma, c_2\gamma$.

B. Convergence to global minimum

Denote the ground energy of H_S (i.e., the global minimum of the loss) as g, the projector to the ground space as Π_g, and the gap between the ground energy and the first excited state as $\delta > 0$.

Theorem S2. *Suppose σ has a nonzero overlap with the ground space of H_S (that is, $\sigma_g = \operatorname{tr}(\Pi_g \sigma) > 0$). For any constant $c \in (0, 1)$, we can choose an appropriate η and T such that if we train the QAL model for T steps with learning rate η in each step, the averaged loss conditioned on the success of post-selection is at most $g + c$.*

Proof. According to Theorem S1, we only need to upper bound (S13) for $\beta = \eta T$ and $\gamma = \eta^2 T = \beta\eta$. Since

$$\begin{aligned}
\frac{\operatorname{tr}(H_S\sigma(\beta)) + c_2\beta\eta}{\operatorname{tr}(\sigma(\beta)) + c_1\beta\eta} &\leq \frac{\sigma_g e^{-2\beta g} g + (1-\sigma_g)e^{-2\beta(g+\delta)} + 4\beta\eta}{\sigma_g e^{-2\beta g} - 4\beta\eta} \\
&= g + \frac{(1-\sigma_g)e^{-2\beta(g+\delta)} + (4+4g)\beta\eta}{\sigma_g e^{-2\beta g} - 4\beta\eta} \\
&\leq g + \frac{e^{-2\beta\delta} + 8\beta\eta e^{2\beta g}}{\sigma_g - 4\beta\eta e^{2\beta g}}.
\end{aligned} \tag{S19}$$

Choose β such that $e^{-2\beta\delta} < \sigma_g c/4$, and then choose η such that $\beta\eta e^{2\beta g} < \sigma_g c/16 < \sigma_g/16$. Then the right hand side of (S19) is at most

$$g + \frac{\sigma_g c/4 + \sigma_g c/2}{\sigma_g - \sigma_g/4} = g + c. \tag{S20}$$

□

A randomly initialized state σ has a nonzero overlap with the ground space with probability 1. According to the theorem, the QAL model will converge to the global minimum of the loss function. However, this convergence is built on the success of post-selection, whose probability exponentially decays with the number of steps. Therefore, a more realistic question is whether we can build a reasonable trade-off between the success probability and the performance of the final model.

C. Convergence with constant probability

In this subsection, we will establish a practical trade-off between the accuracy of the final model and the success probability of post-selection when the initial state has a large overlap with the low-energy eigenspace of H_S.

Definition S1. *Let H be a Hamiltonian. The E low energy subspace of H is the subspace spanned by the eigenstates of H with energy at most E. Denote the projector to the E low energy subspace as Π_E^H. The overlap of a state σ with the E low energy subspace is defined as $\operatorname{tr}\big(\Pi_E^H\sigma\big)$.*

Throughout this section, we focus on the Hamiltonian H_S and omit the superscript H.

Theorem S3. *Let $c_1, c_3 \in (0,1), c_2 \in (0, 1/10)$ be three constants, g be the ground energy of H_S, and $\epsilon > 0$ such that $g/\epsilon \leq c_1$. Assume the overlap between the initial state σ and the $(g+\epsilon)$ low energy eigenspace of H_S, namely $\operatorname{tr}(\sigma\Pi_{g+\epsilon})$, is at least c_2. Then we can choose an appropriate η and T such that if we train the QAL model with the initial state σ for T steps with learning rate η in each step, the success probability of post-selection is at least c_4 and the averaged loss conditioned on the success of post-selection is at most $g + \epsilon + c_3$. Here c_4 is a constant that only depends on c_1, c_2, c_3.*

Proof. By Theorem S1, we only need to lower bound the the success probability in (S12) and the conditional loss in (S13). We will follow the notation in Theorem S1, so that $\beta = \eta T$, $\gamma = \eta^2 T$, and $\sigma(\beta) = e^{-\beta H}\sigma e^{-\beta H}$. Here we write $H = H_S$ for simplicity. We will prove the theorem for $\beta = 3\ln(1+c_2)/(c_3\epsilon)$, $c_4 = e^{-6(1+c_1)\ln(1+c_2)/c_3}c_2/2$ and $\gamma = c_3c_4/40$. Accordingly, $\eta = \gamma/\beta$ is of order ϵ and $T = \beta^2/\gamma$ is of order $1/\epsilon^2$. By (S12), the success probability is at least

$$\begin{aligned}
\operatorname{tr}(\sigma(\beta)) - 4\gamma &\geq \operatorname{tr}(\Pi_{g+\epsilon}\sigma(\beta)) - 4\gamma \\
&= \operatorname{tr}\big(e^{-\beta H}\Pi_{g+\epsilon}e^{-\beta H}\sigma\big) - 4\gamma \\
&\geq e^{-2\beta(g+\epsilon)}\operatorname{tr}(\Pi_{g+\epsilon}\sigma) - 4\gamma \\
&\geq e^{-2\beta\epsilon(1+c_1)}c_2 - 4\gamma \\
&= 2c_4 - 4\gamma > c_4.
\end{aligned} \tag{S21}$$

By (S13), the averaged loss conditioned on the success of post-selection is at most

$$\begin{aligned}
\frac{\operatorname{tr}(H\sigma(\beta)) + 4\gamma}{\operatorname{tr}(\sigma(\beta)) - 4\gamma} &= g + \frac{\operatorname{tr}((H-gI)\sigma(\beta))}{\operatorname{tr}(\sigma(\beta)) - 4\gamma} + \frac{(4+4g)\gamma}{\operatorname{tr}(\sigma(\beta)) - 4\gamma} \\
&\leq g + \frac{\operatorname{tr}((H-gI)\sigma(\beta))}{\operatorname{tr}(\sigma(\beta))(1 - c_3/20)} + \frac{8\gamma}{c_4} \\
&\leq g + \frac{c_3}{5} + (1 + \frac{c_3}{5})\frac{\operatorname{tr}((H-gI)\sigma(\beta))}{\operatorname{tr}(\sigma(\beta))},
\end{aligned} \tag{S22}$$

where we use $4\gamma = c_3c_4/10 \leq c_3\operatorname{tr}(\sigma(\beta))/20$ in the second line and $(1 + c_3/5)(1 - c_3/20) \geq 1$ in the third line. So it suffices to upper bound $\operatorname{tr}((H-gI)\sigma(\beta))/\operatorname{tr}(\sigma(\beta))$. Write the spectrum decomposition of H as $H = \sum_i E_i |E_i\rangle\langle E_i|$ and let $x_i = 2\beta(E_i - g), \sigma_i = \langle E_i|\sigma|E_i\rangle$. We simplify the last term of (S22) to

$$\begin{aligned}
\frac{\operatorname{tr}((H-gI)\sigma(\beta))}{\operatorname{tr}(\sigma(\beta))} &= \frac{\sum_i (E_i - g)e^{-2\beta E_i}\sigma_i}{\sum_i e^{-2\beta E_i}\sigma_i} \\
&= \frac{\sum_i \sigma_i e^{-x_i}x_i}{\sum_i \sigma_i e^{-x_i}} \cdot \frac{1}{2\beta}.
\end{aligned} \tag{S23}$$

Split the Hilbert space into low energy and high energy eigenspaces, $L = \{i : E_i \leq g + \epsilon\}$ and $H = \{i : E_i > g + \epsilon\}$. Let $p_L = \operatorname{tr}(\sigma\Pi_{g+\epsilon}) = \sum_{i\in L}\sigma_i \geq c_2$ and $p_H = 1 - p_L$ be the overlaps of σ with the two subspaces. Since $f(y) = -y\ln(y)$ is concave, by Jensen's inequality,

$$\sum_{i\in L}\frac{\sigma_i}{p_L}f(e^{-x_i}) \leq f\big(\sum_{i\in L}\frac{\sigma_i}{p_L}e^{-x_i}\big). \tag{S24}$$

Let l be the number such that $e^{-l} = \sum_{i\in L}\sigma_i e^{-x_i}/p_L$. The inequality becomes $\sum_{i\in L}\sigma_i e^{-x_i}x_i \le p_L e^{-l}l$. Similarly, let $h = -\ln\big(\sum_{i\in H}\sigma_i e^{-x_i}/\sum_{i\in H}\sigma_i\big)$, then $\sum_{i\in H}\sigma_i e^{-x_i}x_i \le p_H e^{-h}h$. Therefore,

$$\frac{\sum_i \sigma_i e^{-x_i}x_i}{\sum_i \sigma_i e^{-x_i}} \le \frac{p_L e^{-l}l + p_H e^{-h}h}{p_L e^{-l} + p_H e^{-h}}. \tag{S25}$$

By definition, $x_i \in [0, 2\beta\epsilon]$ for $i \in L$ and $x_i > 2\beta\epsilon$ for $i \in H$. Since e^{-l} is a mixed of $e^{-x_i}(i \in L)$, we have $0 \le l \le 2\beta\epsilon$ and similarly $h \ge 2\beta\epsilon$. Denote $y = h - l$. Insert (S23) and (S25) to (S22).

$$\begin{aligned}\frac{\mathrm{tr}(H\sigma(\beta)) + 4\gamma}{\mathrm{tr}(\sigma(\beta)) - 4\gamma} &\le g + \frac{c_3}{5} + (1+\frac{c_3}{5})\frac{p_L e^{-l}l + p_H e^{-h}h}{p_L e^{-l} + p_H e^{-h}} \cdot \frac{1}{2\beta} \\ &= g + \frac{c_3}{5} + (1+\frac{c_3}{5})(l + \frac{p_H e^{-y}y}{p_L + p_H e^{-y}}) \cdot \frac{1}{2\beta} \\ &\le g + \frac{c_3}{5} + (1+\frac{c_3}{5})(l + \frac{e^{-y}y}{c_2 + e^{-y}}) \cdot \frac{1}{2\beta}.\end{aligned} \tag{S26}$$

By differentiating $g(y) = e^{-y}y/(c_2+e^{-y}) = y/(c_2e^y+1)$, we find that $g(y) \le g(y^*)$ for the $y^* > 0$ such that $c_2e^{y^*}(y^*-1) = 1$. For this y^* we have $g(y^*) = y^*-1$. Assume $y^* > \ln(1/c_2) > 2$, then $1 = c_2e^{y^*}(y^*-1) > c_2(1/c_2)(2-1) = 1$, a contradiction. So $g(y) \le g(y^*) = y^* - 1 \le \ln(1/c_2)$. By (S26), the averaged loss conditioned on the success of post-selection is at most

$$g + \frac{c_3}{5} + (1+\frac{c_3}{5})(2\beta\epsilon + \ln(1/c_2)) \cdot \frac{1}{2\beta} = g + \frac{c_3}{5} + (1+\frac{c_3}{5})(\epsilon + \frac{c_3\epsilon}{6}) < g + \epsilon + c_3. \tag{S27}$$

□

The theorem ensures the convergence of the QAL training process with a constant success probability, assuming a good initial state. Concretely, as long as the initial state has a large (at least c_2) overlap with the low energy eigenspace (with energy at most $g + \epsilon$), the QAL training process will converge to the low energy eigenspace (up to an arbitrarily small residue error c_3) with a constant success probability (c_4 that only depends on c_1, c_2, c_3).

D. Heavy-tailed Hamiltonian

Theorem S3 highlights the importance of the initial state σ. However, without prior knowledge of H_S, we cannot do better than a random guess, or equivalently, starting from the maximally-mixed state $\sigma = I/2^n$. Therefore, we actually hope that H_S has a constant proportion of low energy eigenstates that do not scale up with n, the dimension of $\boldsymbol{x}$, and the size m of the training dataset. We formalize this intuition in the following definition.

Definition S2 (Heavy-tailed Hamiltonian)**.** *We say a Hamiltonian H is (E, c)-heavy-tailed if the proportion of eigenstates with energy at most E is at least c.*

Theorem S4. *Let $c_1, c_3 \in (0, 1), c_2 \in (0, 1/10)$ be three constants. Suppose H_S is $(g + \epsilon, c_2)$-heavy-tailed, where g is the ground energy of H_S and $\epsilon > 0$ such that $g/\epsilon \le c_1$. Then we can choose an appropriate η and T such that if we train the QAL model with a maximally-mixed initial state in computational basis for T steps with learning rate η in each step, the success probability of post-selection is at least c_4 and the averaged loss conditioned on the success of post-selection is at most $g+\epsilon+c_3$. Here c_4 is a constant that only depends on c_1, c_2, c_3.*

Proof. By definition of heavy-tailed Hamiltonian, the overlap between the initial state $\sigma = I/2^n$ and the $(g + \epsilon)$ low eigenspace of H_S is at least c_2. The theorem follows directly from Theorem S3. □

Therefore, the QAL training process is guaranteed to converge to the low energy eigenspace of a heavy-tailed H_S. We now argue that when H_S comes from a reasonable dataset, it is likely to be heavy-tailed due to the similarity of data. Consider the extremely simple example of classifying dogs and cats, where all dogs look similar and all cats look similar. The Hamiltonian H_S is approximately a mixture of two projectors of dimensions 2^{n-1}, H_{dogs} and H_{cats}. Regard H_{dogs} and H_{cats} as random projectors, then H_S has a constant proportion of near-zero eigenvalues. This assumption is supported by the numerical simulation.

Remark that while Theorem S4 applies to the maximally-mixed initial state, in reality we will use a random initial state in the computational basis. Once we sample an initial state better than the maximally-mixed state, we can stick to it and apply Theorem S3.

E. Generalization

The previous results establish the explainable trainability of QAL. While in training classical neural networks, each epoch is a full pass of the training dataset, in QAL, each step only involves a single datum. This indicates that the QAL model could be optimized using a few data points. In this subsection we rigorously demonstrate the generalization ability of QAL, showing that once the model achieves a good performance on a small dataset, the good performance will generalize to unseen data.

Recall that the training loss and the true loss of $|\psi\rangle$ are $\hat{R}_S(\psi) = \mathbb{E}_{x\sim S}\,\langle\psi|H_{\boldsymbol{x}}|\psi\rangle$ and $R(\psi) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\,\langle\psi|H_{\boldsymbol{x}}|\psi\rangle$, respectively.

Theorem S5. *With probability at least $1-\delta$ over the choice of S, the generalization gap is upper bounded by*

$$\max_{|\psi\rangle}\left(R(\psi) - \hat{R}_S(\psi)\right) \leq \sqrt{\frac{4\ln(2^{n+1}/\delta)}{m}}. \tag{S28}$$

Proof. By definition, the left-hand side is upper bounded by the spectral norm of $\mathbb{E}_{x\sim_u S} H_x - \mathbb{E}_{x\sim\mathcal{D}} H_x$, which can be bounded by matrix Bernstein inequality (see, e.g., [22, Theorem 6.1.1]):

$$\Pr_S[\|\mathbb{E}_{x\sim_u S} H_x - \mathbb{E}_{x\sim\mathcal{D}} H_x\|_2 \geq t] \leq 2^{n+1}\exp\bigl(-mt^2/4\bigr).$$

The theorem follows by setting $t = \sqrt{4\ln(2^{n+1}/\delta)/m}$. □

According to the theorem, as long as the size m of the training dataset is larger than $\Omega(n)$ (i.e., a logarithm of the degree of freedom), a parameter state $|\boldsymbol{\theta}\rangle$ with low training loss has a low true loss with high probability. This is better than quantum neural networks where the training dataset size has to be larger than the degree of freedom [6, 7]. From the proof of the theorem, it is clear that the good generalization stems from the simple quadratic form of the loss function.

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).

[2] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[3] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[4] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A **99**, 032331 (2019).

[5] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum Natural Gradient, Quantum **4**, 269 (2020).

[6] H. Cai, Q. Ye, and D.-L. Deng, Sample complexity of learning parametric quantum circuits, Quantum Sci. Technol. **7**, 025014 (2022).

[7] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, Generalization in quantum machine learning from few training data, Nat. Commun. **13**, 4919 (2022).

[8] X. You and X. Wu, Exponentially Many Local Minima in Quantum Neural Networks, in *Proceedings of the 38th International Conference on Machine Learning* (PMLR, 2021) pp. 12144–12155.

[9] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

[10] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost Function Dependent Barren Plateaus in Shallow Parametrized Quantum Circuits, Nat. Commun. **12**, 1791 (2021).

[11] C. Ortiz Marrero, M. Kieferová, and N. Wiebe, Entanglement-Induced Barren Plateaus, PRX Quantum **2**, 040316 (2021).

[12] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum **3**, 010313 (2022).

[13] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, Nat. Commun. **12**, 6961 (2021).

[14] M. Larocca, S. Thanasilp, S. Wang, K. Sharma, J. Biamonte, P. J. Coles, L. Cincio, J. R. McClean, Z. Holmes, and M. Cerezo, A Review of Barren Plateaus in Variational Quantum Computing, arXiv:2405.00781 (2024).

[15] A. M. Childs and N. Wiebe, Hamiltonian Simulation Using Linear Combinations of Unitary Operations, QIC **12**, 901 (2012).

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

[17] G. H. Low and I. L. Chuang, Optimal hamiltonian simulation by quantum signal processing, Phys. Rev. Lett. **118**, 010501 (2017).

[18] L. Clinton, J. Bausch, and T. Cubitt, Hamiltonian simulation algorithms for near-term quantum hardware, Nat. Commun. **12**, 4989 (2021).

[19] S. Lloyd, M. Mohseni, and P. Rebentrost, Quantum principal component analysis, Nat. Phys. **10**, 631 (2014).

[20] R. Vale, T. M. D. Azevedo, I. C. S. Araújo, I. F. Araujo, and A. J. da Silva, Decomposition of Multi-controlled Special Unitary Single-Qubit Gates, arXiv:2302.06377 (2023).

[21] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning* (MIT press, 2018).

[22] J. A. Tropp, An Introduction to Matrix Concentration Inequalities, arXiv:1501.01571 (2015).