APPENDIX A PRELIMINARIES IN QUANTUM INFORMATION

In this appendix, we write more details on quantum computation and quantum information for readers to better understand quantum computing.

A.1 QUANTUM COMPUTATION AND QUANTUM INFORMATION BASICS

We use $\|\|_p$ to denote the l_p-norm for vectors and the Schatten-p norm for matrices. The common-used linear algebra notations include complex conjugate transpose $A^\dagger$, the trace of matrix $\mathrm{Tr}[A]$. The μ-th component of the vector $\boldsymbol{\theta}$ is denoted as θ_μ. The derivative with respect to θ_μ is then represented as $\partial_\mu := \frac{\partial}{\partial\theta_\mu}$. The big-O notation $\mathcal{O}$ implies the asymptotic notation of upper bounds.

Quantum information is encoded and processed via the fundamental cells, namely, qubits, and described as quantum states. An n-qubit state can be mathematically represented by a $2^n \times 2^n$ positive semi-definite density matrix ρ, i.e., $\rho \succeq 0$ over the complex field and $\mathrm{Tr}[\rho] = 1$. A pure state, in this formulation, satisfy $\mathrm{Rank}\,(\rho) = 1$ and can be expressed in *Dirac bra-ket* notation as $\rho = |\psi\rangle\langle\psi|$ where $|\psi\rangle \in \mathbb{C}^{2^n}$ denotes a *Hilbert space* unit column vector with the corresponding *dual vector* $\langle\psi|^\dagger = |\psi\rangle$ and $\dagger$ denoting the complex conjugate transpose operation. A mixed state satisfies $\mathrm{Rank}\,(\rho) > 1$, and based on *Spectral theorem*, it has a decomposition form $\rho = \sum_j p_j |\psi_j\rangle\langle\psi_j|$ where $p_j > 0$ denotes the probability of observing $|\psi_j\rangle\langle\psi_j|$ in ρ and $\sum_j p_j = 1$.

Based on *Uhlmann's theorem* Nielsen & Chuang (2010) for every mixed state ρ acting as a linear operator on a Hilbert space A, there exists a purified state $|AR\rangle$ (i.e, pure state) in the composite system AR such that $\mathrm{Tr}_R[|AR\rangle\langle AR|] = \rho$, where $\mathrm{Tr}_R[\cdot]$ denotes the partial trace operation tracing out the ancillary system R. The purification $|AR\rangle$ has a *Schmidt decomposition* form $|AR\rangle = \sum_j \sqrt{p_j}|\psi_j\rangle \otimes |j_R\rangle$ for some orthonormal set $|j_R\rangle$ in R.

The partial trace operation in the above statement plays an important role in quantum computation and information. Given a composite quantum system described by a tensor product of Hilbert spaces, $\mathcal{H}_A \otimes \mathcal{H}_B$, or simply denoted as AB, where $\mathcal{H}_A$ and $\mathcal{H}_B$ represent the Hilbert spaces of subsystems A and B, respectively, the partial trace operation allows us to focus on subsystem A while tracing out the degrees of freedom associated with subsystem B. The partial trace of an operator ρ with respect to subsystem B is denoted as $\mathrm{Tr}_B[\rho]$ and is defined as follows:

$$\mathrm{Tr}_B[\rho] = \sum_i (I_A \otimes \langle i|_B) \cdot \rho \cdot (I_A \otimes |i\rangle_B)$$

Where I_A is the identity operator on $\mathcal{H}_A$; $|i\rangle_B$ forms an orthonormal basis for $\mathcal{H}_B$ and $\langle i|_B$ represents the conjugate transpose of $|i\rangle_B$.

The evolution of a quantum state ρ is realized by applying a series of quantum gates, which are mathematically described as unitary operators. The state ρ' that undergoes transformation via a quantum gate U can be obtained through direct matrix multiplication, expressed as $\rho' = U\rho U^\dagger$. Common single-qubit gates include the Pauli rotations $\{R_P(\theta) = e^{-i\frac{\theta}{2}P} | P \in \{X, Y, Z\}\}$, which are in the matrix exponential form of Pauli matrices

$$X := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, Y := \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, Z := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Common two-qubit gates include controlled-X gate CX (or CNOT) $= I \oplus X$ and controlled-Z gate CZ$= I \oplus Z$ where $\oplus$ denotes the direct sum operation. An n-qubit operator generally lives in the linear operator space $\mathcal{L}(\mathbb{C}^{2^n})$ over the complex field. Quantum measurements are then applied at the end of the quantum circuits, extracting classical information by projecting the quantum states onto its classical shadow.

A.2 FUNDAMENTAL OF QUANTUM NEURAL NETWORKS

In quantum machine learning, quantum neural networks (QNNs) are usually represented as parameterized unitaries consisting of a bunch of single-qubit rotation gates and several two-qubit gates, denoted as $\mathbf{U}(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are the trainable parameters. The model is trained using a classical optimizer according to a minimization process on some cost function $C(\boldsymbol{\theta})$ based on the quantum measurement results.

QNNs can be used to handle a variety of computational tasks, which is usually seen as a quantum version of classical neural networks. In the most general form, a QNN model can be expressed as $\mathbf{U}(\boldsymbol{\theta}) = \prod_{k=1}^{M} U_k(\boldsymbol{\theta}_k)$ for some sub-network layers $U_k(\boldsymbol{\theta}_k)$ where each layer can also be seen as a combination of parameterised circuits as $U_k(\boldsymbol{\theta}_k) = \prod_{j=1}^{d} U_j(\theta_j^{(k)})W_j$, where $U_j(\theta_j^{(k)}) = e^{-ig_j\theta_j^{(k)}}$ is a parameterised gate with a Hermitian generator g_j. W_j is usually non-parameterised, such as the networks of CNOT and CZ gates. The product $\prod_k$ here is, by default, in the increasing order from the right to the left in the above representations.

The idea of quantum neural networks has obtained massive attention since its birth Tóth et al. (1996). Various QNN architectures have been introduced to address a diverse range of computational challenges, spanning both classical and quantum problem domains Rebentrost et al. (2018); Zhao et al. (2019); Liu et al. (2013); Cong et al. (2019); Killoran et al. (2019), thereby pioneering an entirely novel realm of machine learning models. Recent literature focusing on the trainability theory of QNNs indicates a prospective direction for coping with barren plateaus by reducing the expressibility of QNN architectures Cerezo et al. (2021); Liu et al. (2022a). Beyond that, some strategies have been proposed under certain conditions, for example, adopting clever initialization strategy Grant et al. (2019); Kulshrestha & Safro (2022), using adaptive algorithms Grimsley et al. (2019); Zhang et al. (2021); Skolik et al. (2021); Grimsley et al. (2022), making parameterization generalization Volkoff & Coles (2021); Friedrich & Maziero (2022) and choosing different cost forms and circuit architectures Cerezo et al. (2021); Kieferova et al. (2021); Liu et al. (2022b).

APPENDIX B EFFECTIVENESS OF QSSM STATE LEARNING

In this section, we give proof of the effectiveness of QSSM based on Schmidt decomposition, *Uhlmann's theorem* and the properties of purification.

B.1 DEGREES OF FREEDOM IN PURIFICATION

One of the implications of Uhlmann's theorem is that it ensures the degrees of freedom for quantum state purification Nielsen & Chuang (2010). Purification is a commonly used mathematical procedure in quantum computing. For an arbitrary quantum state, its purification is not unique. However, we could bridge these purification states via unitary transformations, which we call freedom in purification.

Lemma S1 *Let $|\psi\rangle$ and $|\phi\rangle$ be two purifications of a state ρ acting on a composite system AE. Then there exists a unitary U_E locally acting on E s.t.,*

$$|\psi\rangle = (I_A \otimes U_E)|\phi\rangle.$$

The proof is simply inspired by the Schmidt decomposition. Let $|\psi\rangle$ and $|\phi\rangle$ be the purifications of ρ acting on AE. Write the Schmidt decomposition of these two states,

$$|\psi\rangle = \sum_j \sqrt{\lambda_j}|j_A\rangle|j_E\rangle \quad |\phi\rangle = \sum_k \sqrt{\eta_k}|k_A\rangle|k_E\rangle.$$

Notice $\mathrm{Tr}_E[\psi] = \rho = \mathrm{Tr}_E[\phi]$, which then induces,

$$\sum_j \lambda_j |j_A\rangle\langle j_A| = \sum_k \eta_k |k_A\rangle\langle k_A|.$$

By linear algebra, we could easily extend both $\{|j_A\rangle\}_j$ and $\{|k_E\rangle\}_k$ to the basis set of $\mathcal{H}_E$, via Gram-Schmidt method, and hence proves the existence of a unitary U_E s.t,

$$U_E|k_E\rangle = |j_E\rangle,$$

which is then substituted into the above equations to prove the lemma. Based on the freedom in purification, we could prove the lemma S2, and therefore prove the effectiveness of our QSSM.

Lemma S2 *Given a target state ρ acting on system A and B, we suppose it can be purified on system ABE where E is an environment. For any pure state $|\psi\rangle$ acting on ABE, s.t.,*

$$\mathrm{Tr}_{BE}[|\psi\rangle\langle\psi|] = \mathrm{Tr}_B[\rho].$$

There always exists a local unitary U_{BE}, s.t.,

$$\mathrm{Tr}_E[(I_A \otimes U_{BE})|\psi\rangle\langle\psi|(I_A \otimes U_{BE}^\dagger)] = \rho.$$

From the definition, $|\psi\rangle\langle\psi|$ and ρ have the same reduced state acting on A. Suppose the state $|\phi\rangle$ is the purification of ρ on system ABE. Thus, it is also a purification of $\rho_A = \mathrm{Tr}_B[\rho]$. We have $|\phi\rangle$ and $|\psi\rangle$ acting on the composite system ABE. By lemma S1, there exists a U_{BE} s.t.,

$$|\phi\rangle\langle\phi| = (I_A \otimes U_{BE})|\psi\rangle\langle\psi|(I_A \otimes U_{BE}^\dagger).$$

Now since $|\phi\rangle$ is the purification of ρ we have,

$$\mathrm{Tr}_E[(I_A \otimes U_{BE})|\psi\rangle\langle\psi|(I_A \otimes U_{BE}^\dagger)] = \rho,$$

as required. Moreover, based on the Schmidt decomposition between AB and E, the dimensionality of system E clearly determines the maximum rank of the output states. For $\mathrm{Rank}[\rho] = r$. It is sufficient and necessary to construct such a unitary U_{BE} so that the last equation in lemma S2 can hold when $\dim[E] \geq \log_2 r$.

B.2 EFFECTIVENESS PROPOSITION OF QSSM

Before we move to the effectiveness proposition of QSSM state learning, we first define some symbols for a better layout of our demonstration of QSSM effectiveness. A k-th partition of ρ separates the state into bipartite subsystems $\mathcal{A}_k$ and $\bar{\mathcal{A}}_k$ covering the first k qubits and the remaining, respectively, where $1 \leq k \leq n$. For $k = n$, $\bar{\mathcal{A}}_k$ becomes trivial and $\mathcal{A}_k = \mathbb{C}^{2^n}$. We then could define the rank sequence of a given target state ρ in the following sense. A sketch of this has been figured out in Fig. S1

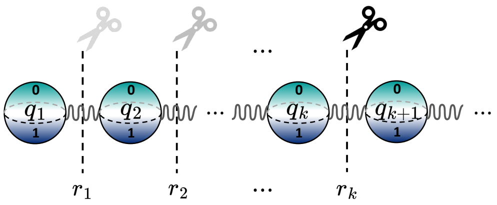


Fig S1: The sketch for illustrating the rank sequence of a given quantum state ρ.

Definition S1 *Given an n-qubit quantum state ρ represented by n ordered quantum registers labeled as $q_1, q_2, \cdots, q_n$, denoting ρ_k as the k-th reduced density matrix of the first k-register state, i.e., $\rho_k = \mathrm{Tr}_{q_{k+1}:q_n}[\rho]$ for $1 \leq k \leq n$ where the operation $\mathrm{Tr}_{q_i:q_j}[\cdot]$ representing a partial tracing over registers q_i to q_j, the (Schmidt) rank sequence of ρ is an ordered list $\mathcal{R}_\rho$,*

$$\mathcal{R}_\rho = \{r_1, r_2, \cdots, r_{n-1}, r_n\},$$

where r_k indicates $\mathrm{Rank}[\rho_k]$. In particular, if ρ is pure, then $r_n = 1$ since ρ can be represented as $|\phi\rangle\langle\phi|$ for some pure state vector $|\phi\rangle$.

Here for clarification, by setting up the k-partition of ρ, $\mathcal{A}_k$ contains the registers $q_1 : q_k$ and $\bar{\mathcal{A}}_k$ contains the registers $q_{k+1} : q_n$ which is the reason why we use this notation to represent the corresponding partial trace operations. We are now ready to prove the effectiveness proposition of the main results.

Proposition S3 *[Effectiveness] For a given n-qubit pure target state ρ represented by n ordered quantum registers $q_1, q_2, \cdots, q_n$, if the rank sequence of ρ is $\mathcal{R}_\rho = \{r_1, r_2, \cdots r_{n-1}, r_n\}$. Then there exists a quantum algorithm 1, based on QSSM, that could produce a state σ exactly satisfying $\sigma = \rho$, if and only if the k-th scattering layer $U_k(\boldsymbol{\theta}_k)$ of QSSM has a width w_k scales $\mathcal{O}(\lceil \log_2 r_k \rceil)$.*

To prove the above Proposition, we first suppose an n-qubit pure target $\rho = |\phi\rangle\langle\phi|$, and at the k-th step,

$$\sigma_k = \mathrm{Tr}_{\bar{\mathcal{A}}_k}[|\psi_k\rangle\langle\psi_k|] = \mathrm{Tr}_{\bar{\mathcal{A}}_k}[\rho] = \rho_k$$

We call this the *k-th perfect learning* condition of QSSM state learning. Then, by lemma S2, there exists a local unitary such that,

$$\mathrm{Tr}_{\bar{\mathcal{A}}_{k+1}}[(I_k \otimes U_{k+1})|\psi_k\rangle\langle\psi_k|(I_k \otimes U_{k+1}^\dagger)] = \mathrm{Tr}_{\bar{\mathcal{A}}_{k+1}}[\rho],$$

where the existence of U_{k+1} ensures the effectiveness of QSSM. We call it a *perfect learning* assumption of QSSM state learning if all the k-th perfect learning can be achieved.

Now, we are ready to deliver the proof of the effectiveness of QSSM. The proof assumes sufficient computational resources, ensuring perfect learning for each step's reduced target. We divide the entire learning task into three main stages based on the algorithm setup.

(1), in the beginning, a state $|0\rangle$ is initialized for the model. We denote the step as $k = 1$ for learning the reduced state acting on $\mathcal{A}_1$ of a single qubit. Notice that for any single-qubit state ρ_1 has an eigendecomposition,

$$\rho_1 = \lambda_1^{(1)}|0^{(1)}\rangle\langle 0^{(1)}| + \lambda_2^{(1)}|1^{(1)}\rangle\langle 1^{(1)}|,$$

where the states $|0^{(1)}\rangle$ and $|1^{(1)}\rangle$ are not necessary the computational basis elements. There exists a purification unitary $U_{\mathcal{A}_1\mathcal{A}_2}$,

$$U_{\mathcal{A}_1\mathcal{A}_2}|00\rangle = \sqrt{\lambda_1^{(1)}}|0^{(1)}\rangle|0^{(2)}\rangle + \sqrt{\lambda_2^{(1)}}|1^{(1)}\rangle|1^{(2)}\rangle.$$

Such a unitary should have the following components. The rest of the matrix can be extended using the Gram-Schmidt process. We could write out the computational basis representation of $U_{\mathcal{A}_1\mathcal{A}_2}$,

$$[U_{\mathcal{A}_1\mathcal{A}_2}]_{mn} = \begin{pmatrix} \sqrt{\lambda_1^{(1)}}\langle 00|0^{(1)}0^{(2)}\rangle + \sqrt{\lambda_2^{(1)}}\langle 00|1^{(1)}1^{(2)}\rangle & \cdots \\ \sqrt{\lambda_1^{(1)}}\langle 01|0^{(1)}0^{(2)}\rangle + \sqrt{\lambda_2^{(1)}}\langle 01|1^{(1)}1^{(2)}\rangle & \cdots \\ \sqrt{\lambda_1^{(1)}}\langle 10|0^{(1)}0^{(2)}\rangle + \sqrt{\lambda_2^{(1)}}\langle 10|1^{(1)}1^{(2)}\rangle & \cdots \\ \sqrt{\lambda_1^{(1)}}\langle 11|0^{(1)}0^{(2)}\rangle + \sqrt{\lambda_2^{(1)}}\langle 11|1^{(1)}1^{(2)}\rangle & \cdots \end{pmatrix}.$$

(2), now for $1 < k \leq \lceil n/2 \rceil$, by the assumption of ideal learning of state ρ_{k-1}, a purification, denoted as $|\psi_{k-1}\rangle$ of it would be imported from the $(k-1)$-th step. The reduced state ρ_k would generally require at least k extra ancillary qubits to be purified, which is why a width control $w_k = k + 1$ is settled in the worst case. Moreover, if the $\mathcal{R}_\rho$ is given as above, the rank values give better choices of layer widths as $w_k = \min\{k+1, \lceil \log_2 r_k \rceil\}$.

Now suppose a purification $|\phi_k\rangle$ of ρ_k. Since $\dim(|\psi_{k-1}\rangle) \leq \dim(|\phi_k\rangle)$, we could always extend $|\psi_k\rangle$ to $|\tilde{\psi_k}\rangle = |\psi_k\rangle|0\rangle$ so that the result pure state lives in the same dimensional Hilbert as $|\phi_k\rangle$. We could observe $|\tilde{\psi_k}\rangle$ and $|\phi_k\rangle$ are both purification of ρ_{k-1}. Based on the lemma S2, there exists U_k acting on the qubits index from $k+1$ to $w_k + k$ s.t.,

$$\mathrm{Tr}_{\bar{\mathcal{A}}_{k+1}}[(I_{\mathcal{A}_{k-1}} \otimes U_k)\tilde{\psi}_k(I_{\mathcal{A}_{k-1}} \otimes U_k^\dagger)] = \mathrm{Tr}_{\bar{\mathcal{A}}_{k+1}}[\phi_k] = \rho_k.$$

(3), at last, for $\lceil n/2 \rceil < k \leq n$. $|\phi_k\rangle$ becomes the pure state acting on the entire system of n qubit registers. The imported purification $|\psi_{k-1}\rangle$ of ρ_{k-1} is also a pure state of n qubits. The result follows by applying the lemma S2 again but with $w_k = \min\{n-k+1, \lceil \log_2 r_k \rceil\}$.

Above all, we have proven the effectiveness of QSSM. One important point to note here is that the width of each scattering layer can be carefully settled concerning the rank of ρ_k for $1 \leq k < n$ in order to obtain the perfect learning. However, exactly constructing those purification unitaries using scattering layers $U_k(\boldsymbol{\theta}_k)$ is not possible. In reality, if each scattering layer of QSSM forms an approximate local unitary t-design for sufficient large positive integer t. Then, given enough time for training, the scattering layers would approximate these purification unitaries to arbitrarily high accuracy.

Further, the proposition identifies a group of quantum states that can be learned more efficiently using QSSM. One notable exemplar within this proposition is the n-qubit GHZ state.

Remark 1 An n-qubit GHZ state Greenberger et al. (1989) has constant rank $r_k = 2$ for $1 \leq k < n$. Hence, setting $w_k = 2\ \forall k$ is sufficient to obtain perfect learning of QSSM state learning on GHZ state.

The above phenomenon suggests a connection between the amount of entanglement within a target state and the sufficient widths w_k to achieve perfect learning. The higher the ranks, the harder the target state could be learnt via QSSM.

APPENDIX C TRAINABILITY AND GRADIENT ANALYSIS OF QSSM

In this section, we give the proof for the proposition 2 stated about the trainability of QSSM in this paper. We first recall some useful lemmas to make the proof easy to read and emphasize important intermediate results. The following lemmas were derived from the studies of unitary t-design. These were originally computed in Cerezo et al. (2021).

Definition S2 *A unitary t-design of dimension d Dankert et al. (2009) with respect to the Haar measure is defined as a finite set of unitaries $\{U_k\}_{k=1}^{M}$ on a d-dimensional Hilbert space such that,*

$$\frac{1}{M} \cdot \sum_{k=1}^{M} P_{(t,t)}(U_k) = \int_{\mathcal{U}(d)} d\mu_{\mathit{Haar}}(U) P_{(t,t)}(U),$$

where $P_{(t,t)}(U)$ denotes a homogeneous polynomial of degree at most t on the elements of U and $U^\dagger$.

Lemma S4 *Suppose $X \subset \mathcal{U}(d)$ is unitary t-design, and A, B, C, D are arbitrary linear operators. If $t \geq 1$, then we have*

$$\frac{1}{|X|} \sum_{U \in X} \operatorname{Tr}[U^\dagger A U B] = \int_{\mathcal{U}(d)} \operatorname{Tr}[U^\dagger A U B] d\eta(U) = \frac{\operatorname{Tr}[A] \operatorname{Tr}[B]}{d} \tag{Appendix C.1}$$

If $t \geq 2$, then we have

$$\frac{1}{|X|} \sum_{U \in X} \operatorname{Tr}[U^\dagger A U B U^\dagger C U D] = \int_{\mathcal{U}(d)} \operatorname{Tr}[U^\dagger A U B U^\dagger C U D] d\eta(U) \tag{Appendix C.2}$$

$$= \frac{\operatorname{Tr}[A] \operatorname{Tr}[C] \operatorname{Tr}[BD] + \operatorname{Tr}[AC] \operatorname{Tr}[B] \operatorname{Tr}[D]}{d^2 - 1} - \frac{\operatorname{Tr}[AC] \operatorname{Tr}[BD] + \operatorname{Tr}[A] \operatorname{Tr}[B] \operatorname{Tr}[C] \operatorname{Tr}[D]}{d(d^2 - 1)} \tag{Appendix C.3}$$

Lemma S5 *Suppose A, B, C, D are arbitrary linear operators. Then,*

$$\int_{\mathcal{U}(d)} \operatorname{Tr}[U A U^\dagger B] \operatorname{Tr}[U C U^\dagger D] d\eta(U) = \frac{1}{d^2 - 1} (\operatorname{Tr}[A] \operatorname{Tr}[B] \operatorname{Tr}[C] \operatorname{Tr}[D] + \operatorname{Tr}[AC] \operatorname{Tr}[BD]) - \frac{1}{d(d^2 - 1)} (\operatorname{Tr}[AC] \operatorname{Tr}[B] \operatorname{Tr}[D] + \operatorname{Tr}[A] \operatorname{Tr}[C] \operatorname{Tr}[BD])$$

Lemma S6 *Let $\mathcal{H} = \mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{B}}$ be a bipartite Hilbert space of dimension $d = d_A d_B$, and for arbitrary linear operators $M, N : \mathcal{H} \to \mathcal{H}$, we have*

$$\int_{\mathcal{U}(d_B)} d\eta(U) (I_A \otimes U) M (I_A \otimes U^\dagger) N = \frac{\operatorname{Tr}_B[M] \otimes I_B}{d_B} N,$$

and

$$\int_{\mathcal{U}(d_B)} d\eta(U) \operatorname{Tr}[(I_A \otimes U) M (I_A \otimes U^\dagger) N] = \frac{\operatorname{Tr}[\operatorname{Tr}_B[M] \operatorname{Tr}_B[N]]}{d_B}.$$

Lemma S7 *Let $\mathcal{H} = \mathcal{H_A} \otimes \mathcal{H_B}$ be a bipartite Hilbert space of dimension $d = d_A d_B$ ($d = 2^n, d_A = 2^{n'}$), and for arbitrary linear operators $M, N, U : \mathcal{H} \to \mathcal{H}$, we have*

$$\mathrm{Tr}[(I_A \otimes U)M(I_A \otimes U^\dagger)N] = \sum_{p,q} \mathrm{Tr}[UM_{qp}U^\dagger N_{pq}],$$

where the summation runs over all bitstrings of length n', and where

$$M_{qp} = \mathrm{Tr}_A[(|p\rangle\langle q| \otimes I)M]$$
$$N_{pq} = \mathrm{Tr}_A[(|q\rangle\langle p| \otimes I)N].$$

With these lemmas, we can now start our proof by directly calculating the variance of gradients. The whole proof includes three parts indicating the gradient magnitude of different stages in the algorithm.

C.1 TRAINABILITY OF THE LAST LAYER

Proposition S8 *For a n-qubit target state ρ, assume we start from the $\hat{\sigma}$ such that $\mathrm{Tr}_n[\rho] = \mathrm{Tr}_n[\hat{\sigma}]$, where $\mathrm{Tr}_n[\rho]$ denotes partial trace over the last qubit of the state. And if the circuit is only acting on the last qubit and forms a 2-design, then $\mathbb{E}[\partial_\mu C_n] = 0$ and the variance $\mathrm{Var}[\partial_\mu C_n] \in [\frac{16}{27}, \frac{8}{9}]$.*

The proof is given by the following, suppose the output state is σ, then the cost function is

$$C_n(\boldsymbol{\theta}) = \mathrm{Tr}[(\rho - \sigma(\boldsymbol{\theta}))(\rho - \sigma(\boldsymbol{\theta}))^\dagger].$$

With a similar notation used in McClean's paper McClean et al. (2018), we can use U to denote the unitary representation of circuits. And we can write it as $U = U_+ e^{-i\theta_\mu H} U_-$, where H denotes the hermitian operator and in most cases it will be the Pauli matrices, and they are traceless. Since $\mathrm{Tr}_n[\rho] = \mathrm{Tr}_n[\hat{\sigma}]$, we have

$$\hat{\sigma} = (I_A \otimes V_B)\rho(I_A \otimes V_B^\dagger).$$

where V is a fixed unitary and system A denotes the first $n-1$ qubits and the system B denotes the last qubit. So $d_A = 2^{n-1}$ and $d_B = 2$. For simplicity, we will hide the subscript in the following proof.

We then arrive at

$$\sigma = (I \otimes UV)\rho(I \otimes V^\dagger U^\dagger).$$

Next, we compute the partial derivative of C_n w.r.t the k-th parameter. Notice that the trace is linear, the derivative operation could pass through the trace and hence we obtain,

$$\partial_\mu C_n = \partial_\mu(\mathrm{Tr}(\rho^2 + \sigma^2 - 2(\rho\sigma)) = -2\,\mathrm{Tr}(\rho\partial_\mu(\sigma)),$$

Now We start by calculating the mean of gradients, expanding the expression for σ, we could find,

$$\partial_\mu C_n = -2\,\mathrm{Tr}\left[\rho\left((I \otimes (\partial_\mu U)V)\rho(I \otimes V^\dagger U^\dagger) + (I \otimes UV)\rho(I \otimes V^\dagger(\partial_\mu U^\dagger))\right)\right],$$

by the chain rule of derivative. Since $U = U_+ e^{-i\theta_\mu H} U_-$, we could compute the derivatives as,

$$\begin{cases} \partial_\mu U = -iU_+ e^{-i\theta_\mu H} H U_- \\ \partial_\mu U^\dagger = iU_-^\dagger H e^{i\theta_\mu H} U_+^\dagger. \end{cases}$$

For convenient, we define $\tilde{U}_+ = U_+ e^{-i\theta_\mu H}$. Substituting the above into the expression of cost derivative to achieve,

$$\partial_\mu C_n = 2i\,\mathrm{Tr}\left[\rho\left((I \otimes \tilde{U}_+ H U_- V)\rho(I \otimes V^\dagger U^\dagger) - (I \otimes UV)\rho(I \otimes V^\dagger U_-^\dagger H \tilde{U}_+^\dagger)\right)\right].$$

Now we expand $U = \tilde{U}_+ U_-$, and assume the $\tilde{U}_- = U_- V$

$$\begin{aligned}
\partial_\mu C_n &= 2i\,\mathrm{Tr}\left[\rho\left((I \otimes \tilde{U}_+ H U_- V)\rho(I \otimes V^\dagger U_-^\dagger \tilde{U}_+^\dagger) - (I \otimes \tilde{U}_+ U_- V)\rho(I \otimes V^\dagger U_-^\dagger H \tilde{U}_+^\dagger)\right)\right] \\
&= 2i\,\mathrm{Tr}\left[\rho\left((I \otimes \tilde{U}_+ H \tilde{U}_-)\rho(I \otimes \tilde{U}_-^\dagger \tilde{U}_+^\dagger) - (I \otimes \tilde{U}_+ \tilde{U}_-)\rho(I \otimes \tilde{U}_-^\dagger H \tilde{U}_+^\dagger)\right)\right] \\
&= 2i\,\mathrm{Tr}\left[(I \otimes \tilde{U}_+^\dagger)\rho(I \otimes \tilde{U}_+)[I \otimes H, (I \otimes \tilde{U}_-)\rho(I \otimes \tilde{U}_-^\dagger)]\right].
\end{aligned}$$

where the $[A,B] = AB - BA$ denotes the commutator notation. Denote the commutator $[I \otimes H, (I \otimes \tilde{U}_-)\rho(I \otimes \tilde{U}_-^\dagger)]$ by T_-, thus we have

$$\partial_\mu C_n = 2i \operatorname{Tr}\left[(I \otimes \tilde{U}_+^\dagger)\rho(I \otimes \tilde{U}_+)T_-\right].$$

Then we integrate over $\tilde{U}_+$ by using the lemma S6,

$$\begin{aligned}\mathbb{E}[\partial_\mu C_n] &= 2i\frac{\operatorname{Tr}[\operatorname{Tr}_B[\rho]\operatorname{Tr}_B[T_-]]}{d_B}\\ &= i\operatorname{Tr}[\operatorname{Tr}_B[\rho]\operatorname{Tr}_B[T_-]].\end{aligned}$$

We can write the ρ as

$$\rho = \sum_{i,j} |i\rangle\langle j|_A \otimes X_{i,j}.$$

thus lead to

$$\begin{aligned}\operatorname{Tr}_B[T_-] &= \operatorname{Tr}_B[[I \otimes H, (I \otimes \tilde{U}_-)\rho(I \otimes \tilde{U}_-^\dagger)]]\\ &= \sum_{i,j}\operatorname{Tr}_B[[I \otimes H, (I \otimes \tilde{U}_-)(|i\rangle\langle j|_A \otimes X_{i,j})(I \otimes \tilde{U}_-^\dagger)]]\\ &= \sum_{i,j}\operatorname{Tr}_B[|i\rangle\langle j| \otimes H\tilde{U}_- X_{i,j}\tilde{U}_-^\dagger - |i\rangle\langle j| \otimes \tilde{U}_- X_{i,j}\tilde{U}_-^\dagger H]\\ &= \sum_{i,j}|i\rangle\langle j|(\operatorname{Tr}[H\tilde{U}_- X_{i,j}\tilde{U}_-^\dagger] - \operatorname{Tr}[\tilde{U}_- X_{i,j}\tilde{U}_-^\dagger H])\\ &= 0.\end{aligned} \tag{Appendix C.4}$$

Therefore, we have

$$\mathbb{E}[\partial_\mu C_n] = 0.$$

The mean of gradients is 0. Based on the fact that the mean of gradients is 0, we then only need to consider the $\mathbb{E}[(\partial_\mu C_n)^2]$ in order to determine the variance.

$$\operatorname{Var}[\partial_\mu C_n] = \mathbb{E}[(\partial_\mu C_n)^2] = -4\mathbb{E}_{\tilde{U}_+,\tilde{U}_-}\left[(\operatorname{Tr}[(I \otimes \tilde{U}_+^\dagger)\rho(I \otimes \tilde{U}_+)T_-])^2\right].$$

Using lemma S7, we have

$$\begin{aligned}\mathbb{E}_{\tilde{U}_+,\tilde{U}_-}\left[(\operatorname{Tr}[(I \otimes \tilde{U}_+^\dagger)\rho(I \otimes \tilde{U}_+)T_-])^2\right] &= \mathbb{E}_{\tilde{U}_+,\tilde{U}_-}\left[(\sum_{p,q}\operatorname{Tr}[\tilde{U}_+\rho_{qp}\tilde{U}_+^\dagger T_{-pq}])(\sum_{m,n}\operatorname{Tr}[\tilde{U}_+\rho_{nm}\tilde{U}_+^\dagger T_{-mn}])\right]\\ &= \mathbb{E}_{\tilde{U}_+,\tilde{U}_-}\left[\sum_{p,q,m,n}\operatorname{Tr}[\tilde{U}_+\rho_{qp}\tilde{U}_+^\dagger T_{-pq}]\operatorname{Tr}[\tilde{U}_+\rho_{nm}\tilde{U}_+^\dagger T_{-mn}]\right]\\ &= \sum_{p,q,m,n}\mathbb{E}_{\tilde{U}_+,\tilde{U}_-}\left[\operatorname{Tr}[\tilde{U}_+\rho_{qp}\tilde{U}_+^\dagger T_{-pq}]\operatorname{Tr}[\tilde{U}_+\rho_{nm}\tilde{U}_+^\dagger T_{-mn}]\right].\end{aligned}$$

Then, according to lemma S5

$$\begin{aligned}&\sum_{p,q,m,n}\mathbb{E}_{\tilde{U}_+,\tilde{U}_-}\left[\operatorname{Tr}[\tilde{U}_+\rho_{qp}\tilde{U}_+^\dagger T_{-pq}]\operatorname{Tr}[\tilde{U}_+\rho_{nm}\tilde{U}_+^\dagger T_{-mn}]\right]\\ =&\sum_{p,q,m,n}\mathbb{E}_{\tilde{U}_-}(\frac{1}{d_B^2-1}(\operatorname{Tr}[\rho_{qp}]\operatorname{Tr}[T_{-pq}]\operatorname{Tr}[\rho_{nm}]\operatorname{Tr}[T_{-mn}] + \operatorname{Tr}[\rho_{qp}\rho_{nm}]\operatorname{Tr}[T_{-pq}T_{-mn}])\\ &-\frac{1}{d_B(d_B^2-1)}(\operatorname{Tr}[\rho_{qp}\rho_{nm}]\operatorname{Tr}[T_{-pq}]\operatorname{Tr}[T_{-mn}] + \operatorname{Tr}[\rho_{qp}]\operatorname{Tr}[\rho_{nm}]\operatorname{Tr}[T_{-pq}T_{-mn}])).\end{aligned} \tag{Appendix C.5}$$

Since

$$\begin{aligned}\operatorname{Tr}[\rho_{qp}] &= \operatorname{Tr}[\operatorname{Tr}_A[(|p\rangle\langle q| \otimes I)\rho]]\\ &= \operatorname{Tr}[(|p\rangle\langle q| \otimes I)\rho]\\ &= \operatorname{Tr}[|p\rangle\langle q|\operatorname{Tr}_B[\rho]]\\ &= \langle q|\operatorname{Tr}_B[\rho]|p\rangle,\end{aligned} \tag{Appendix C.6}$$

and

$$\begin{aligned}\operatorname{Tr}[T_{-pq}] &= \operatorname{Tr}[\operatorname{Tr}_A[(|q\rangle\langle p| \otimes I)T_-]] \\ &= \operatorname{Tr}[(|q\rangle\langle p| \otimes I)T_-] \\ &= \operatorname{Tr}[|q\rangle\langle p| \operatorname{Tr}_B[T_-]] \\ &= 0. \end{aligned} \tag{Appendix C.7}$$

where the Eq. Appendix C.7 holds because of Eq. Appendix C.4.

Thus the Eq. Appendix C.5 can be simplified as

$$\begin{aligned}&\sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_+,\tilde{U}_-} \left[\operatorname{Tr}[\tilde{U}_+\rho_{qp}\tilde{U}_+^\dagger T_{-pq}] \operatorname{Tr}[\tilde{U}_+\rho_{nm}\tilde{U}_+^\dagger T_{-mn}]\right] \\ =& \sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_-} \big(\frac{1}{d_B^2-1} \operatorname{Tr}[\rho_{qp}\rho_{nm}] \operatorname{Tr}[T_{-pq}T_{-mn}] - \frac{1}{d_B(d_B^2-1)} \operatorname{Tr}[\rho_{qp}] \operatorname{Tr}[\rho_{nm}] \operatorname{Tr}[T_{-pq}T_{-mn}]\big) \\ =& \sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_-} \left(\frac{1}{d_B(d_B^2-1)} \operatorname{Tr}[T_{-pq}T_{-mn}](d_B \operatorname{Tr}[\rho_{qp}\rho_{nm}] - \operatorname{Tr}[\rho_{qp}] \operatorname{Tr}[\rho_{nm}])\right) \\ =& \sum_{p,q,m,n} \frac{1}{d_B(d_B^2-1)}(d_B \operatorname{Tr}[\rho_{qp}\rho_{nm}] - \operatorname{Tr}[\rho_{qp}] \operatorname{Tr}[\rho_{nm}])\mathbb{E}_{\tilde{U}_-} \left(\operatorname{Tr}[T_{-pq}T_{-mn}]\right). \end{aligned}$$

We now need to evaluate the other integral w.r.t $\tilde{U}_-$. A simplification can be first done by noticing,

$$\begin{aligned}T_{-pq} &= \operatorname{Tr}_A[(|q\rangle\langle p| \otimes I)T_-] \\ &= \operatorname{Tr}_A[I \otimes H, (I \otimes \tilde{U}_-)(|q\rangle\langle p| \otimes I)\rho(I \otimes \tilde{U}_-^\dagger)] \\ &= [H, \tilde{U}_- \operatorname{Tr}_A[|q\rangle\langle p| \otimes I)\rho]\tilde{U}_-^\dagger] \\ &= [H, \tilde{U}_-\rho_{pq}\tilde{U}_-^\dagger], \end{aligned}$$

since $|p\rangle\langle q| \otimes I$ commutes with other operators. Therefore,

$$\begin{aligned}\operatorname{Tr}[T_{-pq}T_{-mn}] &= \operatorname{Tr}[[H, \tilde{U}_-\rho_{pq}\tilde{U}_-^\dagger][H, \tilde{U}_-\rho_{mn}\tilde{U}_-^\dagger]] \\ &= 2\operatorname{Tr}[H\tilde{U}_-\rho_{pq}\tilde{U}_-^\dagger H\tilde{U}_-\rho_{mn}\tilde{U}_-^\dagger] - \operatorname{Tr}[\tilde{U}_-\rho_{pq}\rho_{mn}\tilde{U}_-^\dagger H^2] - \operatorname{Tr}[\tilde{U}_-\rho_{mn}\rho_{pq}\tilde{U}_-^\dagger H^2]. \end{aligned}$$

So according to lemma S4,

$$\begin{aligned}&\mathbb{E}_{\tilde{U}_-} \left(\operatorname{Tr}[T_{-pq}T_{-mn}]\right) \\ =&\frac{2}{d_B^2-1}(\operatorname{Tr}[\rho_{pq}] \operatorname{Tr}[\rho_{mn}] \operatorname{Tr}[H^2] + \operatorname{Tr}[\rho_{pq}\rho_{mn}] \operatorname{Tr}^2[H]) \\ &- \frac{2}{d_B(d_B^2-1)}(\operatorname{Tr}[\rho_{pq}\rho_{mn}] \operatorname{Tr}[H^2] + \operatorname{Tr}[\rho_{pq}] \operatorname{Tr}[\rho_{mn}] \operatorname{Tr}^2[H]) - \frac{2}{d_B} \operatorname{Tr}[\rho_{pq}\rho_{mn}] \operatorname{Tr}[H^2] \\ =&\frac{-2}{d_B(d_B^2-1)}(d_B \operatorname{Tr}[\rho_{pq}\rho_{mn}] - \operatorname{Tr}[\rho_{pq}] \operatorname{Tr}[\rho_{mn}])(d_B \operatorname{Tr}[H^2] - \operatorname{Tr}^2[H]) \\ =&\frac{-2}{(d_B^2-1)} \operatorname{Tr}[H^2](d_B \operatorname{Tr}[\rho_{pq}\rho_{mn}] - \operatorname{Tr}[\rho_{pq}] \operatorname{Tr}[\rho_{mn}]). \end{aligned}$$

Then, we go back to Eq. Appendix C.8,

$$\begin{aligned}&\sum_{p,q,m,n} \mathbb{E}_{\tilde{U}_+,\tilde{U}_-} \left[\operatorname{Tr}[\tilde{U}_+\rho_{qp}\tilde{U}_+^\dagger T_{-pq}] \operatorname{Tr}[\tilde{U}_+\rho_{nm}\tilde{U}_+^\dagger T_{-mn}]\right] \\ =& \sum_{p,q,m,n} \frac{-2}{d_B(d_B^2-1)^2} \operatorname{Tr}[H^2](d_B \operatorname{Tr}[\rho_{qp}\rho_{nm}] - \operatorname{Tr}[\rho_{qp}] \operatorname{Tr}[\rho_{nm}])(d_B \operatorname{Tr}[\rho_{pq}\rho_{mn}] - \operatorname{Tr}[\rho_{pq}] \operatorname{Tr}[\rho_{mn}]). \end{aligned}$$

First, we look at the $\mathrm{Tr}[\rho_{qp}\rho_{nm}]$

$$\begin{aligned}\mathrm{Tr}[\rho_{qp}\rho_{nm}] &= \mathrm{Tr}[\mathrm{Tr}_A[(|p\rangle\langle q|\otimes I)\rho]\,\mathrm{Tr}_A[(|m\rangle\langle n|\otimes I)\rho]] \\ &= \mathrm{Tr}[\sum_i (\langle i|\otimes I\,((|p\rangle\langle q|\otimes I)\rho)\,|i\rangle\otimes I)\sum_j(\langle j|\otimes I(|p\rangle\langle q|\otimes I)\rho|j\rangle\otimes I)] \\ &= \mathrm{Tr}[(\langle q|\otimes I)\rho(|p\rangle\langle n|\otimes I)\rho(|m\rangle\otimes I)] \\ &= \mathrm{Tr}[\langle q|\,\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]|m\rangle] \\ &= \langle q|\,\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]|m\rangle.\end{aligned}$$

Then,

$$\begin{aligned}&\sum_{p,q,m,n}\mathrm{Tr}[\rho_{qp}\rho_{nm}]\,\mathrm{Tr}[\rho_{pq}\rho_{mn}] \\ =&\sum_{p,q,m,n}\langle q|\,\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]|m\rangle\langle m|\,\mathrm{Tr}_B[\rho(|n\rangle\langle p|\otimes I)\rho]|q\rangle \\ =&\sum_{p,n}\mathrm{Tr}\left[\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]\,\mathrm{Tr}_B[\rho(|n\rangle\langle p|\otimes I)\rho]\right].\end{aligned}$$

Suppose the Schmidt decomposition of $|\phi\rangle$ is

$$|\phi\rangle=\sum_k\lambda_k|u_k\rangle_A|v_k\rangle_B. \tag{Appendix C.8}$$

where $\{|u_k\rangle\}$ are orthogonal basis on the system A and $\{|v_k\rangle\}$ are orthogonal basis on the system B. Therefore, we can write the ρ as

$$\rho=\sum_{i,j}\lambda_i\lambda_j|u_i\rangle\langle u_j|\otimes|v_i\rangle\langle v_j|. \tag{Appendix C.9}$$

We can expand the ρ in $\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]$

$$\begin{aligned}&\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho] \\ =&\mathrm{Tr}_B[(\sum_{i,j}\lambda_i\lambda_j|u_i\rangle\langle u_j|\otimes|v_i\rangle\langle v_j|)(|p\rangle\langle n|\otimes I)(\sum_{k,l}\lambda_k\lambda_l|u_k\rangle\langle u_l|\otimes|v_k\rangle\langle v_l|)] \\ =&\sum_{i,j,k,l}\lambda_i\lambda_j\lambda_k\lambda_l\,\mathrm{Tr}_B[|u_i\rangle\langle u_j||p\rangle\langle n||u_k\rangle\langle u_l|\otimes|v_i\rangle\langle v_j||v_k\rangle\langle v_l|] \\ =&\sum_{i,j}\lambda_i^2\lambda_j^2|u_i\rangle\langle u_j||p\rangle\langle n||u_j\rangle\langle u_i|.\end{aligned}$$

Thus, we arrive at

$$\begin{aligned}&\sum_{p,n}\mathrm{Tr}\left[\mathrm{Tr}_B[\rho(|p\rangle\langle n|\otimes I)\rho]\,\mathrm{Tr}_B[\rho(|n\rangle\langle p|\otimes I)\rho]\right] \\ =&\sum_{p,n}\mathrm{Tr}[(\sum_{i,j}\lambda_i^2\lambda_j^2|u_i\rangle\langle u_j||p\rangle\langle n||u_j\rangle\langle u_i|)(\sum_{k,l}\lambda_k^2\lambda_l^2|u_k\rangle\langle u_l||p\rangle\langle n||u_k\rangle\langle u_l|)] \\ =&\sum_{p,n}\mathrm{Tr}[\sum_{i,j,k,l}]\lambda_i^2\lambda_j^2\lambda_k^2\lambda_l^2|u_i\rangle\langle u_j||p\rangle\langle n||u_j\rangle\langle u_i||u_k\rangle\langle u_l||n\rangle\langle p||u_l\rangle\langle u_k| \\ =&\sum_{p,n}\sum_{i,j,l}\lambda_i^4\lambda_j^2\lambda_l^2\,\mathrm{Tr}[\langle u_j||p\rangle\langle n||u_j\rangle\langle u_l||n\rangle\langle p||u_l\rangle] \\ =&\sum_{i,j,l}\lambda_i^4\lambda_j^2\lambda_l^2\,\mathrm{Tr}[\mathrm{Tr}[|u_l\rangle\langle u_j|]\,\mathrm{Tr}[|u_j\rangle\langle u_l|]] \\ =&\sum_{i,j}\lambda_i^4\lambda_j^4 \\ =&(\sum_i\lambda_i^4)^2.\end{aligned}$$

Then we look at the $\mathrm{Tr}[\rho_{qp}]\,\mathrm{Tr}[\rho_{pq}]\,\mathrm{Tr}[\rho_{mn}]\,\mathrm{Tr}[\rho_{nm}]$,

$$\begin{aligned}
&\sum_{p,q,m,n} \mathrm{Tr}[\rho_{qp}]\,\mathrm{Tr}[\rho_{pq}]\,\mathrm{Tr}[\rho_{mn}]\,\mathrm{Tr}[\rho_{nm}] \\
=&\sum_{p,q,m,n} \langle q|\,\mathrm{Tr}_B[\rho]|p\rangle\langle p|\,\mathrm{Tr}_B[\rho]|q\rangle\langle m|\,\mathrm{Tr}_B[\rho]|n\rangle\langle n|\,\mathrm{Tr}_B[\rho]|m\rangle \\
=&\,\mathrm{Tr}[\mathrm{Tr}_B[\rho]\,\mathrm{Tr}_B[\rho]]\,\mathrm{Tr}[\mathrm{Tr}_B[\rho]\,\mathrm{Tr}_B[\rho]] \\
=&(\mathrm{Tr}[\mathrm{Tr}_B[\rho]\,\mathrm{Tr}_B[\rho]])^2 \\
=&(\sum_i \lambda_i^4)^2.
\end{aligned}$$

Now, we look at the $\mathrm{Tr}[\rho_{qp}\rho_{nm}]\,\mathrm{Tr}[\rho_{pq}]\,\mathrm{Tr}[\rho_{mn}]$

$$\mathrm{Tr}[\rho_{qp}\rho_{nm}] = \langle n|\,\mathrm{Tr}_B[\rho(|m\rangle\langle q|\otimes I)\rho]|p\rangle \tag{Appendix C.10}$$

$$= \sum_{i,j} \lambda_i^2\lambda_j^2 \langle n||u_i\rangle\langle u_j||m\rangle\langle q||u_j\rangle\langle u_i||p\rangle. \tag{Appendix C.11}$$

and

$$\begin{aligned}
\mathrm{Tr}[\rho_{pq}]\,\mathrm{Tr}[\rho_{mn}] &= \langle p|\,\mathrm{Tr}_B[\rho]|q\rangle\langle m|\,\mathrm{Tr}_B[\rho]|n\rangle \\
&= \sum_{i,j} \lambda_i^2\lambda_j^2 \langle p||u_i\rangle\langle u_i||q\rangle\langle m||u_j\rangle\langle u_j||n\rangle.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{p,q,m,n} \mathrm{Tr}[\rho_{qp}\rho_{nm}]\,\mathrm{Tr}[\rho_{pq}]\,\mathrm{Tr}[\rho_{mn}] \\
=&\sum_{p,q,m,n} (\sum_{k,l} \lambda_k^2\lambda_l^2 \langle n||u_k\rangle\langle u_l||m\rangle\langle q||u_k\rangle\langle u_l||p\rangle)(\sum_{i,j} \lambda_i^2\lambda_j^2 \langle p||u_i\rangle\langle u_i||q\rangle\langle m||u_j\rangle\langle u_j||n\rangle) \\
=&\sum_{p,q,m,n} \sum_{i,j,k,l} \lambda_i^2\lambda_j^2\lambda_k^2\lambda_l^2 (\langle n||u_k\rangle\langle u_l||m\rangle\langle q||u_k\rangle\langle u_l||p\rangle\langle p||u_i\rangle\langle u_i||q\rangle\langle m||u_j\rangle\langle u_j||n\rangle) \\
=&\sum_{q,m} \sum_{i,j,k,l} \lambda_i^2\lambda_j^2\lambda_k^2\lambda_l^2\,\mathrm{Tr}[|u_k\rangle\langle u_l||m\rangle\langle q||u_k\rangle\langle u_l||u_i\rangle\langle u_i||q\rangle\langle m||u_j\rangle\langle u_j|] \\
=&\sum_{i,j,k,l} \lambda_i^2\lambda_j^2\lambda_k^2\lambda_l^2\,\mathrm{Tr}[|u_k\rangle\langle u_l||u_i\rangle\langle u_i|]\,\mathrm{Tr}[|u_j\rangle\langle u_j||u_k\rangle\langle u_l|] \\
=&\sum_i \lambda_i^8.
\end{aligned}$$

Therefore, we have,

$$\begin{aligned}
&\sum_{p,q,m,n} (d_B\,\mathrm{Tr}[\rho_{pq}\rho_{mn}] - \mathrm{Tr}[\rho_{pq}]\,\mathrm{Tr}[\rho_{mn}]) \\
=&(d_B^2+1)(\sum_i \lambda_i^4)^2 - 2d_B(\sum_i \lambda_i^8).
\end{aligned}$$

So,

$$\mathrm{Var}[\partial_\mu C_n] = \frac{8}{d_B(d_B^2-1)^2}\,\mathrm{Tr}[H^2]((d_B^2+1)(\sum_i \lambda_i^4)^2 - 2d_B(\sum_i \lambda_i^8)) \tag{Appendix C.12}$$

Since the d_B is 2, we can simplify the equation above as

$$\mathrm{Var}[\partial_\mu C_n] = \frac{4}{9}\,\mathrm{Tr}[H^2](\lambda_1^8+\lambda_2^8+10\lambda_1^4\lambda_2^4) \tag{Appendix C.13}$$

$$= \frac{8}{9}(c_1^4+c_2^4+10c_1^2c_2^2). \tag{Appendix C.14}$$

where the $c_1=\lambda_1^2$, $c_2=\lambda_2^2$ such that $c_1+c_2=1$, and $\mathrm{Tr}[H^2]=d_B=2$.

Therefore, we can simply get the range of the variance.

$$\frac{16}{27} \le \mathrm{Var}[\partial_\mu C_n] \le \frac{8}{9} \tag{Appendix C.15}$$

C.2 TRAINABILITY OF THE MIDDLE STEP

Lemma S9 *For the target pure state ρ_{ABC} on system ABC, suppose we start from a initial state $\hat{\sigma}$ such that $\mathrm{Tr}_{BC}[\rho] = \mathrm{Tr}_{BC}[\hat{\sigma}]$ and the output state is σ. If the cost function is*

$$C = \mathrm{Tr}[(\mathrm{Tr}_C[\rho] - \mathrm{Tr}_C[\sigma])(\mathrm{Tr}_C[\rho] - \mathrm{Tr}_C[\sigma])] \qquad \text{(Appendix C.16)}$$

and the circuit is acting on system BC while forming a local 4-design, then $\mathbb{E}[\partial_\mu C] = 0$ and the variance of cost gradient scales as $\mathrm{Var}[\partial_\mu C] \in \mathcal{O}(\frac{1}{d_B^3 d_C})$, where d_B, d_C denote the dimension of system B and C respectively.

Since $\mathrm{Tr}_{BC}[\rho] = \mathrm{Tr}_{BC}[\hat{\sigma}]$, there exist a fixed unitary V such that

$$\hat{\sigma} = (I_A \otimes V_{BC})\rho(I_A \otimes V_{BC}^\dagger). \qquad \text{(Appendix C.17)}$$

Then

$$\sigma = (I \otimes UV)\rho(I \otimes V^\dagger U^\dagger). \qquad \text{(Appendix C.18)}$$

Then, the cost gradient becomes,

$$\begin{aligned}\partial_\mu C &= 2\,\mathrm{Tr}[\mathrm{Tr}_C[\sigma]\partial_\mu \mathrm{Tr}_C[\sigma] - 2\,\mathrm{Tr}[\mathrm{Tr}_C[\rho]\partial_\mu \mathrm{Tr}_C[\sigma]] \\ &= 2i\,\mathrm{Tr}[\mathrm{Tr}_C[(I \otimes U_+U_-V)\rho(I \otimes V^\dagger U_-^\dagger U_+^\dagger) - \rho]\,\mathrm{Tr}_C[(I \otimes U_+U_-V)\rho(I \otimes V^\dagger U_-^\dagger H U_+^\dagger) \\ &- (I \otimes U_+HU_-V)\rho(I \otimes V^\dagger U_-^\dagger U_+^\dagger)]]\end{aligned}$$

We exploit the RTNI package Fukuda et al. (2019a) to calculate the mean of the cost gradient. It turns out that the mean of the cost gradient is zero.

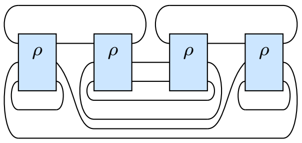
$$\mathbb{E}[\partial_\mu C] = 0$$

Then we consider the variance

$$\mathrm{Var}[\partial_\mu \mathrm{C}] = -\mathbb{E}[(\partial_\mu \mathrm{C})^2]$$

With the RTNI package Fukuda et al. (2019b), it turns out that the exact expression of the variance is dominant by

$$\mathrm{Var}[\partial_\mu C] \xrightarrow{d\to\infty} \frac{\mathrm{Tr}[H^2]}{d_B^2(d_B^2 d_C^2 - 1)} \cdot \Bigg(\qquad \Bigg)$$



We know that $\mathrm{Tr}[H^2] = d_B d_C$, thus we have

$$\mathrm{Var}[\partial_\mu C] \in \mathcal{O}(\frac{g(\rho)}{d_B^3 d_C}),$$

where $g(\rho)$ denotes the dominant factor from the tensor product illustrated above. Finally, we can conclude the following Proposition,

Proposition S10 *For the k-th learning step ($k \leq n$) in QSSM, the mean of cost gradient is 0, and the variance of cost gradient scales as $\mathrm{Var}[\partial_\mu C_k] \in \mathcal{O}(2^{-n_k})$, where n_k is the circuit width of k-th learning step.*

Suppose the target state is ρ and the input state for k-the learning step is $\hat{\sigma}$. We assume system A denotes the first $k-1$ qubits, system B denotes the k-th qubit and system C denotes the $(k+1)$-th qubit to the $(k+n_k-1)$-th qubit. With the definition of n_k claimed in the text, there exists a purification $\hat{\rho}_{ABC}$ of ρ_A on system ABC. According to lemma S9, we can easily know that

$$\mathrm{Var}[\partial_\mu C_k] \in \mathcal{O}(\frac{1}{2^{n_k+2}}) = \mathcal{O}(2^{-n_k})$$

Proposition S11 *[Trainability] Given the state learning algorithm stated in Proposition 1, for an* n*-qubit pure target state* ρ *represented by* n *ordered quantum registers* $q_1, q_2, \cdots, q_n$ *with a rank sequence* $\mathcal{R}_\rho = \{r_1, r_2, \cdots r_{n-1}, r_n\}$*, if one of the* $U_{\pm}^{(k)}$ *in the* k*-th scattering layer* U_k *forms at least local unitary* 4*-design, the expectation and the variance of* C_k *with respect to* θ_μ *can be upper bounded by,*

$$\mathbb{E}[\partial_\mu C_k] = 0; \quad \mathrm{Var}[\partial_\mu C_k] \in \mathcal{O}\left(\frac{g(\rho_k)}{r_k}\right),$$

where the expectation is computed regarding the Haar measure and the factor $g(\rho_k)$ *scales polynomially in* $\mathrm{Tr}[\rho_k^2]$ *known as the purity of* ρ_k*.*

Since we know that $2^{n_k - 1} \leq r_k \leq 2^{n_k}$, thus according to Proposition S10, we can get the proof. Notice that the factor $g(\rho_k)$ scales polynomially in $\mathrm{Tr}[\rho_k^2]$ due to the Cauchy-Schwartz inequality of density matrices. We then finish the proof of the Proposition.

APPENDIX D ANALYTIC EVALUATION OF COST FUNCTION AND GRADIENT

In this appendix, we provide a detailed analysis of the analytic gradient of our cost function C_k equation 3. We take the 2-norm squared cost function as our objective. At the k-th learning step, analyzing the exact form of $\partial_\mu C_k$ is necessary for further designing the training strategy of QSSM. Recalling the expression of C_k, we could derive the derivative form with respect to the parameter $\theta_\mu = \theta_k^\mu$. From here, we have concentrated on the k-th step and for convenience, we will omit the subscript k of the parameter in the following sections. The partial derivative of C_k with respect to θ_μ is then expressed as,

$$\partial_\mu C_k = 2\,\mathrm{Tr}(2\sigma_k \partial_\mu(\sigma_k)) - 2\,\mathrm{Tr}(\rho_k \partial_\mu(\sigma_k)), \qquad \text{(Appendix D.1)}$$

where $\sigma_k = \sigma_k(\boldsymbol{\theta})$ which is constructed via paramterized circuit $U_k(\boldsymbol{\theta})$, and ρ_k is the k-th step reduced target. In a practical sense, our U_k is composed of the quantum gates satisfying the parameter-shift rule and $U_k = U_l e^{-i\frac{\theta_\mu}{2}\Omega_\mu} U_r = \tilde{U}_l U_r$, where $\Omega_\mu^2 = I$. The k-th scattering layer has been shown in Fig. S2. Then the following lemma holds,

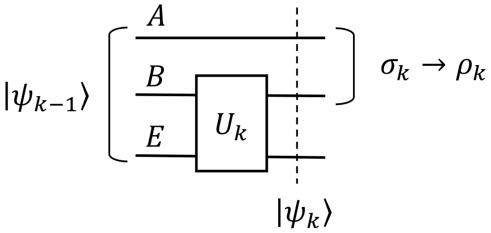


Fig S2: The k-th learning step layer. Based on adaptive learning processes, the previously learnt state $|\psi_{k-1}\rangle$ on system ABE must be pure where E is the additional system acted by the k-th step layer U_k. Under perfect learning situation, we have $\sigma_{k-1} = \mathrm{Tr}_{BE}(\psi_{k-1}) = \rho_{k-1}$.

Lemma S12 *The* k*-th step cost function* C_k *has the partial derivative form (w.r.t.* θ_μ *and evaluated at* $\boldsymbol{\theta} = \boldsymbol{\theta}^*$*),*

$$\partial_\mu C_k^* = \left\langle \Delta_k^* \otimes \frac{I_E}{d_E} \right\rangle_{\theta_\mu + \frac{\pi}{2}} - \left\langle \Delta_k^* \otimes \frac{I_E}{d_E} \right\rangle_{\theta_\mu - \frac{\pi}{2}}$$

where $\Delta_k = \sigma_k - \rho_k$ *with* $*$ *indicating the state difference evaluated at* $\boldsymbol{\theta}^*$*. The other symbols all match the settings in Fig. S2.*

By observing $\sigma_k = \mathrm{Tr}_E((I_A \otimes U_k) P_{\psi_{k-1}} (I_A \otimes U_k^\dagger))$, where $P_{\psi_{k-1}} = |\psi_{k-1}\rangle\langle\psi_{k-1}|$, we could compute the expression of $\partial_\mu \sigma_k$ based on the linearity of derivative operation,

$$\partial_\mu \sigma_k = \mathrm{Tr}_E((I_A \otimes \partial_\mu(U_k)) P_{\psi_{k-1}} (I_A \otimes U_k^\dagger)) + \mathrm{Tr}_E((I_A \otimes U_k) P_{\psi_{k-1}} (I_A \otimes \partial_\mu(U_k^\dagger))).$$

Recalling the expression of $\partial_\mu(U_k)$ and $\partial_\mu(U_k^\dagger)$, we have,

$$\begin{aligned}\partial_\mu\sigma_k &= -\frac{i}{2}\operatorname{Tr}_E((I_A\otimes\tilde{U}_l)[(I_A\otimes\Omega_\mu),(I_A\otimes U_r)P_{\psi_{k-1}}(I_A\otimes U_r^\dagger)](I_A\otimes\tilde{U}_l^\dagger)) \\ &= -\frac{i}{2}\operatorname{Tr}_E(\tilde{U}_l[\Omega_\mu,U_rP_{\psi_{k-1}}U_r^\dagger]\tilde{U}_l^\dagger)\end{aligned},$$

where we have abbreviated the '$I_A\otimes$' correspondence for simplicity, which the subsystem A would never join the optimizations during the k-th step. Since $U_\mu(\theta_\mu)=e^{-i\frac{\theta_\mu}{2}\Omega_\mu}$ satisfies the parameter-shift rule. we could use the gate identity,

$$i[\Omega_\mu,M]=U_\mu\left(-\frac{\pi}{2}\right)MU_\mu^\dagger\left(-\frac{\pi}{2}\right)-U_\mu\left(\frac{\pi}{2}\right)MU_\mu^\dagger\left(\frac{\pi}{2}\right)$$

for any linear operator M, and then derive the exact value of $\partial_\mu\sigma_k^*$ at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ as,

$$\partial_\mu(\sigma_k^*)=\frac{1}{2}\operatorname{Tr}_E\left(U_k(\theta_\mu^*+\frac{\pi}{2})P_{\psi_{k-1}}U_k^\dagger(\theta_\mu^*+\frac{\pi}{2})-U_k(\theta_\mu^*-\frac{\pi}{2})P_{\psi_{k-1}}U_k^\dagger(\theta_\mu^*-\frac{\pi}{2})\right).$$

Here $\partial_\mu(\sigma_k^*)=\partial_\mu(\sigma_k)|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$, and circuit $U_k(\theta_\mu^*+\alpha)$ intakes $\boldsymbol{\theta}^*$ and modifies the parameter θ_μ^* to $\theta_\mu^*+\frac{\pi}{2}$. Now, recalling the fact that,

$$\operatorname{Tr}(\operatorname{Tr}_B(\rho_{AB})\sigma_A)=\operatorname{Tr}\left(\rho_{AB}(\sigma_A\otimes\frac{I_B}{d_B})\right),$$

we have,

$$\begin{aligned}\operatorname{Tr}(\rho_k\partial_\mu(\sigma_k^*)) &= \left\langle\rho_k\otimes\frac{I_E}{d_E}\right\rangle_{\theta_\mu^*+\frac{\pi}{2}}-\left\langle\rho_k\otimes\frac{I_E}{d_E}\right\rangle_{\theta_\mu^*-\frac{\pi}{2}} \\ \operatorname{Tr}(\sigma_k^*\partial_\mu(\sigma_k^*)), &= \left\langle\sigma_k^*\otimes\frac{I_E}{d_E}\right\rangle_{\theta_\mu^*+\frac{\pi}{2}}-\left\langle\sigma_k^*\otimes\frac{I_E}{d_E}\right\rangle_{\theta_\mu^*-\frac{\pi}{2}}\end{aligned},$$

where $\langle M\rangle_\theta=\langle\psi_k(\theta)|M|\psi_k(\theta)\rangle$ and $|\psi_k(\theta)\rangle$ is derived by applying $U_k(\theta)$ on $|\psi_{k-1}\rangle$. Combining the above calculations to obtain the desired result in lemma S12 taking $\Delta^*=\sigma_k(\boldsymbol{\theta}^*)-\rho_k$. Finally, by taking the actual dimensional factors, we could derive the analytic form of the partial derivative as shown in Sec. 4.2.