Appendix for "Statistical Analysis of Quantum State Learning Process in Quantum Neural Networks"

A Preliminaries

A.1 Subspace Haar integration

The central technique used in our work is the subspace Haar integration, i.e., a series of formulas on calculating Haar integrals over a certain subspace of the given Hilbert space. In this section, we give a brief introduction to the common Haar integrals and then the basic formulas on subspace Haar integrals used in our work together with the proofs.

Haar integrals refer to the matrix integrals over the d-degree unitary group $\mathcal{U}(d)$ with the Haar measure $d\mu$, which is the unique uniform measure on $\mathcal{U}(d)$ such that

$$\int_{\mathcal{U}(d)} d\mu(V) f(V) = \int_{\mathcal{U}(d)} d\mu(V) f(VU) = \int_{\mathcal{U}(d)} d\mu(V) f(UV), \tag{S1}$$

for any integrand f and group element $U \in \mathcal{U}(d)$. If an ensemble $\mathbb{V}$ of unitaries V matches the Haar measure up to the t-degree moment, i.e.,

$$\mathbb{E}_{V\in\mathbb{V}}[p_{t,t}(V)] = \int_{\mathcal{U}(d)} d\mu(V) p_{t,t}(V), \tag{S2}$$

then $\mathbb{V}$ is called a unitary t-design [67]. $p_{t,t}(V)$ denotes an arbitrary polynomial of degree at most t in the entries of V and at most t in those of $V^\dagger$. $\mathbb{E}_{V\in\mathbb{V}}[\cdot]$ denotes the expectation over the ensemble $\mathbb{V}$. The Haar integrals over polynomials can be analytically solved and expressed into closed forms according to the following lemma.

Lemma S1 *Let $\varphi : \mathcal{U}(d) \to \mathrm{GL}(\mathbb{C}^{d'})$ be an arbitrary representation of unitary group $\mathcal{U}(d)$. Suppose that the direct sum decomposition of φ to irreducible representations is $\varphi = \bigoplus_{j,k} \phi_k^{(j)}$, where $\phi_k^{(j)}$ denotes the k^{th} copy of the irreducible representation $\phi^{(j)}$. A set of orthonormal basis in the representation space of $\phi_k^{(j)}$ is denoted as $\{|v_{j,k,l}\rangle\}$. For an arbitrary linear operator $A : \mathbb{C}^{d'} \to \mathbb{C}^{d'}$, the following equality holds* [68]

$$\int_{\mathcal{U}(d)} \varphi(U) A \varphi(U)^\dagger d\mu(U) = \sum_{j,k,k'} \frac{\operatorname{tr}(Q_{j,k,k'}^\dagger A)}{\operatorname{tr}(Q_{j,k,k'}^\dagger Q_{j,k,k'})} Q_{j,k,k'}, \tag{S3}$$

where $Q_{j,k,k'} = \sum_l |v_{j,k,l}\rangle\langle v_{j,k',l}|$ is the transfer operator from the representation subspace of $\phi_{k'}^{(j)}$ to that of $\phi_k^{(j)}$. The denominator on the right hand side of (S3) *can be simplified as $\operatorname{tr}(Q_{j,k,k'}^\dagger Q_{j,k,k'}) = \operatorname{tr}(P_{j,k'}) = d_j$, where $P_{j,k} = \sum_l |v_{j,k,l}\rangle\langle v_{j,k,l}| = Q_{j,k,k}$ is the projector to the representation subspace of $\phi_k^{(j)}$ and d_j is the dimension of the representation space of $\phi^{(j)}$.*

By choosing different representations of the unitary group $\mathcal{U}(d)$, some commonly used equalities can be derived, such as

$$\int_{\mathcal{U}(d)} V A V^\dagger d\mu(V) = \frac{\operatorname{tr}(A)}{d} I, \tag{S4}$$

$$\int_{\mathcal{U}(d)} V^\dagger A V B V^\dagger C V d\mu(V) = \frac{\operatorname{tr}(AC)\operatorname{tr} B}{d^2} I + \frac{d \operatorname{tr} A \operatorname{tr} C - \operatorname{tr}(AC)}{d(d^2-1)} \left(B - \frac{\operatorname{tr} B}{d} I\right), \tag{S5}$$

where I is the identity operator on the d-dimensional Hilbert space $\mathcal{H}$. A, B and C are arbitrary linear operators on $\mathcal{H}$. According to the linearity of the integrals, the following equalities can be further derived

$$\int_{\mathcal{U}(d)} \operatorname{tr}(VA)\operatorname{tr}(V^\dagger B) d\mu(V) = \frac{\operatorname{tr}(AB)}{d}, \tag{S6}$$

$$\int_{\mathcal{U}(d)} \operatorname{tr}(V^\dagger AVB)\operatorname{tr}(V^\dagger CVD)d\mu(V) = \frac{\operatorname{tr} A \operatorname{tr} B \operatorname{tr} C \operatorname{tr} D + \operatorname{tr}(AC)\operatorname{tr}(BD)}{d^2-1} - \frac{\operatorname{tr}(AC)\operatorname{tr} B \operatorname{tr} D + \operatorname{tr} A \operatorname{tr} C \operatorname{tr}(BD)}{d(d^2-1)}, \tag{S7}$$

where A, B, C and D are arbitrary linear operators on $\mathcal{H}$.

The subspace Haar integration can be regarded as a simple generalization of the formulas above. Suppose that $\mathcal{H}_{\text{sub}}$ is a subspace with dimension d_{sub} of the Hilbert space $\mathcal{H}$. $\mathbb{U}$ is an ensemble whose elements are unitaries in $\mathcal{H}$ with a block-diagonal structure $U = \bar{P} + PUP$. P is the projector from $\mathcal{H}$ to $\mathcal{H}_{\text{sub}}$ and PUP is a random unitary with the Haar measure on $\mathcal{H}_{\text{sub}}$. $\bar{P} = I - P$ is the projector from $\mathcal{H}$ to the orthogonal complement of $\mathcal{H}_{\text{sub}}$. Integrals with respect to such an ensemble $\mathbb{U}$ are dubbed as "*subspace Haar integrals*", which can be reduced back to the common Haar integrals by taking $\mathcal{H}_{\text{sub}} = \mathcal{H}$. The corresponding formulas of subspace Haar integrals are developed in the following lemmas, where $\mathbb{E}_{U\in\mathbb{U}}[\cdot] = \mathbb{E}_{\mathbb{U}}[\cdot]$ denotes the expectation with respect to the ensemble $\mathbb{U}$.

Lemma S2 *The expectation of a single element $U \in \mathbb{U}$ with respect to the ensemble $\mathbb{U}$ equals to the projector to the orthogonal complement, i.e.,*

$$\mathbb{E}_{\mathbb{U}}[U] = I - P. \tag{S8}$$

Proof The fact that Haar integrals of inhomogenous polynomials $p_{t,t'}$ with $t \neq t'$ over the whole space equals to zero leads to the vanishment of the block in $\mathcal{H}_{\text{sub}}$, i.e.,

$$\mathbb{E}_{\mathbb{U}}[U] = \mathbb{E}_{\mathbb{U}}\left[\bar{P} + PUP\right] = \bar{P}, \tag{S9}$$

which is just the projector to the orthogonal complement $\mathcal{H}_{\text{sub}}$. ■

Similarly, we know all the subspace Haar integrals involving only U or $U^\dagger$ will leave a projector after integration. For example, it holds that $\mathbb{E}_{\mathbb{U}}[UAU] = \bar{P}A\bar{P}$ for an arbitrary linear operator A.

Lemma S3 *For an arbitrary linear operator A on $\mathcal{H}$, the expectation of $U^\dagger AU$ with respect to the random variable $U \in \mathbb{U}$ is*

$$\mathbb{E}_{\mathbb{U}}\left[U^\dagger AU\right] = \frac{\operatorname{tr}(PA)}{d_{\text{sub}}}P + (I-P)A(I-P). \tag{S10}$$

Proof Eq. (S10) can be seen as a special case of Lemma S1 since U can be seen as the complete reducible representation of $\mathcal{U}(d_{\text{sub}})$ composed of $(d - d_{\text{sub}})$ trivial representations $\phi_k^{(1)}$ with $k = 1, ..., (d - d_{\text{sub}})$ and one natural representation $\phi^{(2)}$. This gives rise to

$$\begin{aligned} \sum_{k,k'} \frac{\operatorname{tr}(Q_{1,k,k'}^\dagger A)}{\operatorname{tr}(Q_{1,k,k'}^\dagger Q_{1,k,k'})} Q_{1,k,k'} &= \bar{P}A\bar{P}, \\ \frac{\operatorname{tr}(Q_2^\dagger A)}{\operatorname{tr}(Q_2^\dagger Q_2)} Q_2 &= \frac{\operatorname{tr}(PA)}{d_{\text{sub}}}P. \end{aligned} \tag{S11}$$

Alternatively, Eq. (S10) can just be seen as a result of the block matrix multiplication, i.e.,

$$\begin{aligned} \mathbb{E}_{\mathbb{U}}[U^\dagger AU] &= \mathbb{E}_{\mathbb{U}}[(\bar{P} + PU^\dagger P)A(\bar{P} + PUP)] \\ &= \mathbb{E}_{\mathbb{U}}[\bar{P}A\bar{P} + \bar{P}APUP + PU^\dagger PA\bar{P} + PU^\dagger PAPUP] \\ &= \bar{P}A\bar{P} + \frac{\operatorname{tr}(PAP)}{d_{\text{sub}}}P, \end{aligned} \tag{S12}$$

where $\operatorname{tr}(PAP) = \operatorname{tr}(P^2A) = \operatorname{tr}(PA)$. ■

Corollary S4 *Suppose $|\varphi\rangle$ is a Haar-random pure state in $\mathcal{H}_{\text{sub}}$. For arbitrary linear operators A on $\mathcal{H}$, the following equality holds*

$$\mathbb{E}_{\varphi}\left[\langle\varphi|A|\varphi\rangle\right] = \frac{\operatorname{tr}(PA)}{d_{\text{sub}}}, \tag{S13}$$

where $\mathbb{E}_{\varphi}[\cdot]$ is the expectation with respect to the random state $|\varphi\rangle$.

Proof Suppose $|\varphi_0\rangle$ is an arbitrary fixed state in $\mathcal{H}_{\rm sub}$. The random state $|\varphi\rangle$ can be written in terms of $U \in \mathbb{U}$ as $|\varphi\rangle = U|\varphi_0\rangle$ such that

$$\mathbb{E}_\varphi\left[\langle\varphi|A|\varphi\rangle\right] = \mathbb{E}_{U\in\mathbb{U}}\left[\langle\varphi_0|U^\dagger AU|\varphi_0\rangle\right]. \tag{S14}$$

Eq. (S13) is naturally obtained from Lemma S3 by taking the expectation over $|\varphi_0\rangle$ which satisfies $P|\varphi_0\rangle = |\varphi_0\rangle$ and $\bar{P}|\varphi_0\rangle = 0$. ■

Lemma S5 *For arbitrary linear operators A, B, C on $\mathcal{H}$ and $U \in \mathbb{U}$, the following equality holds*

$$\begin{aligned}
&\mathbb{E}_{\mathbb{U}}\left[U^\dagger AUBU^\dagger CU\right] \\
&= \bar{P}A\bar{P}B\bar{P}C\bar{P} + \frac{\mathrm{tr}\,(PB)}{d_{\rm sub}}\bar{P}APC\bar{P} + \frac{\mathrm{tr}\,(PC)}{d_{\rm sub}}\bar{P}A\bar{P}BP + \frac{\mathrm{tr}\,(PA)}{d_{\rm sub}}PB\bar{P}C\bar{P} \\
&\quad + \frac{\mathrm{tr}\left(PA\bar{P}B\bar{P}C\right)}{d_{\rm sub}}P + \frac{\mathrm{tr}(PAPC)\,\mathrm{tr}(PB)}{d_{\rm sub}^2}P \\
&\quad + \frac{d_{\rm sub}\,\mathrm{tr}(PA)\,\mathrm{tr}(PC) - \mathrm{tr}(PAPC)}{d_{\rm sub}(d_{\rm sub}^2-1)}\left(PBP - \frac{\mathrm{tr}(PB)}{d_{\rm sub}}P\right).
\end{aligned} \tag{S15}$$

Proof Here we simply employ the block matrix multiplication to prove this equality. We denote the 2×2 blocks with indices $\begin{pmatrix} 11 & 12 \\ 21 & 22 \end{pmatrix}$ respectively where the index 2 corresponds to $\mathcal{H}_{\rm sub}$. Thus the random unitary U can be written as $U = \begin{pmatrix} I_{11} & 0 \\ 0 & U_{22} \end{pmatrix}$ where I_{11} is the identity matrix on the orthogonal complement of $\mathcal{H}_{\rm sub}$ and U_{22} is a Haar-random unitary on $\mathcal{H}_{\rm sub}$. The integrand becomes

$$U^\dagger AUBU^\dagger CU = \begin{pmatrix} A_{11} & A_{12}U_{22} \\ U_{22}^\dagger A_{21} & U_{22}^\dagger A_{22}U_{22} \end{pmatrix}\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}\begin{pmatrix} C_{11} & C_{12}U_{22} \\ U_{22}^\dagger C_{21} & U_{22}^\dagger C_{22}U_{22} \end{pmatrix}. \tag{S16}$$

The four matrix elements of the multiplication results are

$$\begin{aligned}
&11:\ A_{11}B_{11}C_{11} + A_{12}U_{22}B_{21}C_{11} + A_{11}B_{12}U_{22}^\dagger C_{11} + A_{12}U_{22}B_{22}U_{22}^\dagger C_{21}, \\
&12:\ A_{11}B_{11}C_{12}U_{22} + A_{12}U_{22}B_{21}C_{12}U_{22} + A_{11}B_{12}U_{22}^\dagger C_{22}U_{22} + A_{12}U_{22}B_{22}U_{22}^\dagger C_{22}U_{22}, \\
&21:\ U_{22}^\dagger A_{21}B_{11}C_{11} + U_{22}^\dagger A_{22}U_{22}B_{21}C_{11} + U_{22}^\dagger A_{21}B_{12}U_{22}^\dagger C_{21} + U_{22}^\dagger A_{22}U_{22}B_{22}U_{22}^\dagger C_{21}, \\
&22:\ U_{22}^\dagger A_{21}B_{11}C_{12}U_{22} + U_{22}^\dagger A_{22}U_{22}B_{21}C_{12}U_{22} + U_{22}^\dagger A_{21}B_{12}U_{22}^\dagger C_{22}U_{22} \\
&\qquad + U_{22}^\dagger A_{22}U_{22}B_{22}U_{22}^\dagger C_{22}U_{22}.
\end{aligned} \tag{S17}$$

Since inhomogeneous Haar integrals always vanish on $\mathcal{H}_{\rm sub}$, the elements above can be reduced to

$$\begin{aligned}
&11:\ A_{11}B_{11}C_{11} + A_{12}U_{22}B_{22}U_{22}^\dagger C_{21}, \\
&12:\ A_{11}B_{12}U_{22}^\dagger C_{22}U_{22}, \quad 21:\ U_{22}^\dagger A_{22}U_{22}B_{21}C_{11}, \\
&22:\ U_{22}^\dagger A_{21}B_{11}C_{12}U_{22} + U_{22}^\dagger A_{22}U_{22}B_{22}U_{22}^\dagger C_{22}U_{22}.
\end{aligned} \tag{S18}$$

Let $d_2 = d_{\rm sub} = \dim\mathcal{H}_{\rm sub}$ and I_{22} be the identity matrix in $\mathcal{H}_{\rm sub}$. Utilizing Eqs. (S4) and (S5), the expectation of each block becomes

$$\begin{aligned}
&11:\ A_{11}B_{11}C_{11} + \frac{\mathrm{tr}\,B_{22}}{d_2}A_{12}C_{21}, \\
&12:\ \frac{\mathrm{tr}\,C_{22}}{d_2}A_{11}B_{12}, \quad 21:\ \frac{\mathrm{tr}\,A_{22}}{d_2}B_{21}C_{11}, \\
&22:\ \frac{\mathrm{tr}\left(A_{21}B_{11}C_{12}\right)}{d_2}I_{22} + \frac{\mathrm{tr}(A_{22}C_{22})\,\mathrm{tr}(B_{22})}{d_2^2}I_{22} \\
&\qquad + \frac{d_2\,\mathrm{tr}(A_{22})\,\mathrm{tr}(C_{22}) - \mathrm{tr}(A_{22}C_{22})}{d_2(d_2^2-1)}\left(B_{22} - \frac{\mathrm{tr}(B_{22})}{d_2}I_{22}\right).
\end{aligned} \tag{S19}$$

Written in terms of subspace projectors P and $\bar{P}$, the results become exactly as Eq. (S15). ■

Corollary S6 *Suppose $|\varphi\rangle$ is a Haar-random pure state in $\mathcal{H}_{\rm sub}$. For arbitrary linear operators A on $\mathcal{H}$, the following equality holds*

$$\mathbb{E}_\varphi\left[\langle\varphi|A|\varphi\rangle^2\right]=\frac{\operatorname{tr}((PA)^2)+(\operatorname{tr}(PA))^2}{d_{\rm sub}(d_{\rm sub}+1)},\tag{S20}$$

where $\mathbb{E}_\varphi[\cdot]$ is the expectation with respect to the random state $|\varphi\rangle$.

Proof Suppose $|\varphi_0\rangle$ is an arbitrary fixed state in $\mathcal{H}_{\rm sub}$. The random state $|\varphi\rangle$ can be written in terms of $U\in\mathbb{U}$ as $|\varphi\rangle=U|\varphi_0\rangle$ such that

$$\mathbb{E}_\varphi\left[(\langle\varphi|A|\varphi\rangle)^2\right]=\mathbb{E}_{U\in\mathbb{U}}\left[\langle\varphi_0|U^\dagger AU|\varphi_0\rangle\langle\varphi_0|U^\dagger AU|\varphi_0\rangle\right].\tag{S21}$$

Eq. (S20) is naturally obtained from Lemma S5 by taking $C=A$ and $B=|\varphi_0\rangle\langle\varphi_0|$ which satisfies $\bar{P}B=B\bar{P}=0$, $PBP=B$ and $\operatorname{tr}B=1$. ■

Lemma S7 *For arbitrary linear operators A,B,C,D on $\mathcal{H}$ and $U\in\mathbb{U}$, the following equality holds*

$$\begin{aligned}
&\mathbb{E}_{\mathbb{U}}\left[\operatorname{tr}(U^\dagger AUB)\operatorname{tr}(U^\dagger CUD)\right]=\operatorname{tr}(\bar{P}A\bar{P}B)\operatorname{tr}(\bar{P}C\bar{P}D)\\
&+\frac{\operatorname{tr}(\bar{P}A\bar{P}B)\operatorname{tr}(PC)\operatorname{tr}(PD)}{d_{\rm sub}}+\frac{\operatorname{tr}(\bar{P}C\bar{P}D)\operatorname{tr}(PA)\operatorname{tr}(PB)}{d_{\rm sub}}\\
&+\frac{\operatorname{tr}(PB\bar{P}APC\bar{P}D)}{d_{\rm sub}}+\frac{\operatorname{tr}(PA\bar{P}BPD\bar{P}C)}{d_{\rm sub}}\\
&+\frac{\operatorname{tr}(PA)\operatorname{tr}(PB)\operatorname{tr}(PC)\operatorname{tr}(PD)+\operatorname{tr}(PAPC)\operatorname{tr}(PBPD)}{d_{\rm sub}^2-1}\\
&-\frac{\operatorname{tr}(PAPC)\operatorname{tr}(PB)\operatorname{tr}(PD)+\operatorname{tr}(PA)\operatorname{tr}(PC)\operatorname{tr}(PBPD)}{d_{\rm sub}(d_{\rm sub}^2-1)}.
\end{aligned}\tag{S22}$$

Proof Similarly with the proof of Lemma S5, the block matrix multiplication gives

$$\begin{aligned}
\operatorname{tr}(U^\dagger AUB)&=\operatorname{tr}\left[\begin{pmatrix}A_{11}&A_{12}U_{22}\\U_{22}^\dagger A_{21}&U_{22}^\dagger A_{22}U_{22}\end{pmatrix}\begin{pmatrix}B_{11}&B_{12}\\B_{21}&B_{22}\end{pmatrix}\right]\\
&=\operatorname{tr}(A_{11}B_{11})+\operatorname{tr}(A_{12}U_{22}B_{21})+\operatorname{tr}(U_{22}^\dagger A_{21}B_{12})+\operatorname{tr}(U_{22}^\dagger A_{22}U_{22}B_{22}).
\end{aligned}\tag{S23}$$

Hence we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{U}}\left[\operatorname{tr}(U^\dagger AUB)\operatorname{tr}(U^\dagger CUD)\right]=\mathbb{E}_{\mathbb{U}}\big[\operatorname{tr}(A_{11}B_{11})\operatorname{tr}(C_{11}D_{11})\\
&+\operatorname{tr}(A_{11}B_{11})\operatorname{tr}(U_{22}^\dagger C_{22}U_{22}D_{22})+\operatorname{tr}(C_{11}D_{11})\operatorname{tr}(U_{22}^\dagger A_{22}U_{22}B_{22})\\
&+\operatorname{tr}(A_{12}U_{22}B_{21})\operatorname{tr}(U_{22}^\dagger C_{21}D_{12})+\operatorname{tr}(U_{22}^\dagger A_{21}B_{12})\operatorname{tr}(C_{12}U_{22}D_{21})\\
&+\operatorname{tr}(U_{22}^\dagger A_{22}U_{22}B_{22})\operatorname{tr}(U_{22}^\dagger C_{22}U_{22}D_{22})\big].
\end{aligned}\tag{S24}$$

where all inhomogeneous terms have been ignored. Utilizing Eqs. (S4), (S6) and (S7), the expectation becomes

$$\begin{aligned}
&\mathbb{E}_{\mathbb{U}}\left[\operatorname{tr}(U^\dagger AUB)\operatorname{tr}(U^\dagger CUD)\right]=\operatorname{tr}(A_{11}B_{11})\operatorname{tr}(C_{11}D_{11})\\
&+\frac{\operatorname{tr}(A_{11}B_{11})\operatorname{tr}(C_{22})\operatorname{tr}(D_{22})}{d_2}+\frac{\operatorname{tr}(C_{11}D_{11})\operatorname{tr}(A_{22})\operatorname{tr}(B_{22})}{d_2}\\
&+\frac{\operatorname{tr}(B_{21}A_{12}C_{21}D_{12})}{d_2}+\frac{\operatorname{tr}(A_{21}B_{12}D_{21}C_{12})}{d_2}\\
&+\frac{1}{d_2^2-1}(\operatorname{tr}(A_{22})\operatorname{tr}(B_{22})\operatorname{tr}(C_{22})\operatorname{tr}(D_{22})+\operatorname{tr}(A_{22}C_{22})\operatorname{tr}(B_{22}D_{22}))\\
&-\frac{1}{d_2(d_2^2-1)}(\operatorname{tr}(A_{22}C_{22})\operatorname{tr}(B_{22})\operatorname{tr}(D_{22})+\operatorname{tr}(A_{22})\operatorname{tr}(C_{22})\operatorname{tr}(B_{22}D_{22}))
\end{aligned}\tag{S25}$$

Written in terms of subspace projectors P and $\bar{P}$, the results become exactly as Eq. (S22). ■

Finally, similar to the unitary t-design, we introduce the concept of "subspace t-design". If an ensemble $\mathbb{W}$ of unitaries V matches the ensemble $\mathbb{U}$ rotating the subspace $\mathcal{H}_{\rm sub}$ up to the t-degree moment, then $\mathbb{W}$ is called a subspace unitary t-design with respect to $\mathcal{H}_{\text{sub}}$. In the main text, the ensemble comes from the unknown target state. Alternatively, if a random QNN $\mathbf{U}(\boldsymbol{\theta})$ with some constraints such as keeping the loss function constant $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_0$, i.e.,

$$\mathbb{W} = \mathbf{U}(\Theta), \quad \Theta = \{\boldsymbol{\theta} \mid \mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_0\}, \tag{S26}$$

forms a approximate subspace 2-design, then similar results as in the main text can be established yet with a different interpretation: there is an exponentially large proportion of local minima on a constant-loss-section of the training landscape.

A.2 Perturbation on positive definite matrices

To identify whether a parameter point is a local minimum, we need to check whether the Hessian matrix is positive definite, where the following sufficient condition is used in the proof of our main theorem in the next section.

Lemma S8 *Suppose X is a positive definite matrix and Y is a Hermitian matrix. If the distance between Y and X is smaller than the minimal eigenvalue of X, i.e., $\|Y - X\|_\infty < \|X^{-1}\|_\infty^{-1}$, then Y is positive definite. Here $\|\cdot\|_\infty$ denotes the Schatten-∞ norm.*

Proof For an arbitrary vector $|v\rangle$, we have

$$\langle v|Y|v\rangle = \langle v|X|v\rangle + \langle v|Y - X|v\rangle \geq \|X^{-1}\|_\infty^{-1} - \|Y - X\|_\infty > 0. \tag{S27}$$

Note that $\|X^{-1}\|_\infty^{-1}$ just represents the minimal eigenvalue of the positive matrix X. Thus, Y is positive definite. ■

A.3 Tail inequalities

In order to bound the probability of avoiding local minima, we need to use some "tail inequalities" in probability theory, especially the generalized Chebyshev's inequality for matrices, which we summarize below for clarity.

Lemma S9 (Markov's inequality) *For a non-negative random variable X and $a > 0$, the probability that X is at least a is upper bounded by the expectation of X divided by a, i.e.,*

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}. \tag{S28}$$

Proof The expectation can be rewritten and bounded as

$$\begin{aligned} \mathbb{E}[X] &= \Pr[X < a] \cdot \mathbb{E}[X \mid X < a] + \Pr[X \geq a] \cdot \mathbb{E}[X \mid X \geq a] \\ &\geq \Pr[X \geq a] \cdot \mathbb{E}[X \mid X \geq a] \geq \Pr[X \geq a] \cdot a. \end{aligned} \tag{S29}$$

Thus we have $\Pr[X \geq a] \leq \mathbb{E}[X]/a$. ■

Lemma S10 (Chebyshev's inequality) *For a real random variable X and $\varepsilon > 0$, the probability that X deviates from the expectation $\mathbb{E}[X]$ by ε is upper bounded by the variance of X divided by ε^2, i.e.,*

$$\Pr[|X - \mathbb{E}[X]| \geq \varepsilon] \leq \frac{\mathrm{Var}[X]}{\varepsilon^2}. \tag{S30}$$

Proof Applying Markov's inequality in Lemma S9 to the random variable $(X - \mathbb{E}[X])^2$ gives

$$\Pr[|X - \mathbb{E}[X]| \geq \varepsilon] = \Pr[(X - \mathbb{E}[X])^2 \geq \varepsilon^2] \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{\varepsilon^2} = \frac{\mathrm{Var}[X]}{\varepsilon^2}. \tag{S31}$$

Alternatively, the proof can be carried out similarly as in Eq. (S29) with respect to $(X - \mathbb{E}[X])^2$. ■

Lemma S11 (Chebyshev's inequality for matrices) *For a random matrix X and $\varepsilon > 0$, the probability that X deviates from the expectation $\mathbb{E}[X]$ by ε in terms of the norm $\|\cdot\|_\alpha$ satisfies*

$$\Pr\left[\|X - \mathbb{E}[X]\|_\alpha \geq \varepsilon\right] \leq \frac{\sigma_\alpha^2}{\varepsilon^2} \tag{S32}$$

where $\sigma_\alpha^2 = \mathbb{E}[\|X - \mathbb{E}[X]\|_\alpha^2]$ denotes the variance of X in terms of the norm $\|\cdot\|_\alpha$.

Proof Applying Markov's inequality in Lemma S9 to the random variable $\|X - \mathbb{E}[X]\|_\alpha^2$ gives

$$\Pr[\|X - \mathbb{E}[X]\|_\alpha \geq \varepsilon] = \Pr[\|X - \mathbb{E}[X]\|_\alpha^2 \geq \varepsilon^2] \leq \frac{\mathbb{E}[\|X - \mathbb{E}[X]\|_\alpha^2]}{\varepsilon^2} = \frac{\sigma_\alpha^2}{\varepsilon^2}. \tag{S33}$$

Note that here the expectation $\mathbb{E}[X]$ is still a matrix while the "variance" σ_α^2 is a real number. ■

A.4 Quantum Fisher information matrix

Given a parameterized pure quantum state $|\psi(\boldsymbol{\theta})\rangle$, the quantum Fisher information (QFI) matrix $\mathcal{F}_{\mu\nu}$ [56] is defined as the Riemannian metric induced from the Bures fidelity distance $d_{\mathrm{f}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1 - |\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}')\rangle|^2$ (up to a factor 2 depending on convention), i.e.,

$$\mathcal{F}_{\mu\nu}(\boldsymbol{\theta}) = \left.\frac{\partial^2}{\partial\delta_\mu\partial\delta_\nu} d_{\mathrm{f}}(\boldsymbol{\theta}, \boldsymbol{\theta}+\boldsymbol{\delta})\right|_{\boldsymbol{\delta}=0} = -2\,\mathrm{Re}\left[\langle\partial_\mu\partial_\nu\psi|\psi\rangle + \langle\partial_\mu\psi|\psi\rangle\langle\psi|\partial_\nu\psi\rangle\right]. \tag{S34}$$

Note that $|\partial_\mu\psi\rangle$ actually refers to $\frac{\partial}{\partial\theta_\mu}|\psi(\boldsymbol{\theta})\rangle$. Using the normalization condition

$$\begin{aligned}
&\langle\psi|\psi\rangle = 1, \\
&\partial_\mu(\langle\psi|\psi\rangle) = \langle\partial_\mu\psi|\psi\rangle + \langle\psi|\partial_\mu\psi\rangle = 2\,\mathrm{Re}\left[\langle\partial_\mu\psi|\psi\rangle\right] = 0, \\
&\partial_\mu\partial_\nu(\langle\psi|\psi\rangle) = 2\,\mathrm{Re}\left[\langle\partial_\mu\partial_\nu\psi|\psi\rangle + \langle\partial_\mu\psi|\partial_\nu\psi\rangle\right] = 0,
\end{aligned} \tag{S35}$$

the QFI can be rewritten as

$$\mathcal{F}_{\mu\nu} = 2\,\mathrm{Re}\left[\langle\partial_\mu\psi|\partial_\nu\psi\rangle - \langle\partial_\mu\psi|\psi\rangle\langle\psi|\partial_\nu\psi\rangle\right]. \tag{S36}$$

The QFI characterizes the sensibility of a parameterized quantum state to a small change of parameters, and can be viewed as the real part of the quantum geometric tensor.

B Detailed proofs

In this section, we provide the detailed proofs of Lemma 1, Theorem 2 and Proposition 3 in the main text. Here we use d to denote the dimension of the Hilbert space. For a qubit system with N qubits, we have $d = 2^N$. As in the main text, we represent the value of a certain function at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ by appending the superscript "$*$" for simplicity of notation, e.g., $\nabla\mathcal{L}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ as $\nabla\mathcal{L}^*$ and $H_{\mathcal{L}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ as $H_{\mathcal{L}}^*$. In addition, for a parameterized quantum circuit $\mathbf{U}(\boldsymbol{\theta}) = \prod_{\mu=1}^{M} U_\mu(\theta_\mu) W_\mu$, we introduce the notation $V_{\alpha\to\beta} = \prod_{\mu=\alpha}^{\beta} U_\mu W_\mu$ if $\alpha \leq \beta$ and $V_{\alpha\to\beta} = I$ if $\alpha > \beta$. Note that the product $\prod_\mu$ is by default in the increasing order from the right to the left. The derivative with respect to the parameter θ_μ is simply denoted as $\partial_\mu = \frac{\partial}{\partial\theta_\mu}$. We remark that our results hold for all kinds of input states into QNNs in spite that we use $|0\rangle^{\otimes N}$ in the definition of $|\psi(\boldsymbol{\theta})\rangle$ for simplicity.

Lemma 1 *The expectation and variance of the gradient $\nabla\mathcal{L}$ and Hessian matrix $H_{\mathcal{L}}$ of the fidelity loss function $\mathcal{L}(\boldsymbol{\theta}) = 1 - |\langle\phi|\psi(\boldsymbol{\theta})\rangle|^2$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ with respect to the target state ensemble $\mathbb{T}$ satisfy*

$$\mathbb{E}_{\mathbb{T}}\left[\nabla\mathcal{L}^*\right] = 0, \quad \mathrm{Var}_{\mathbb{T}}[\partial_\mu\mathcal{L}^*] = f_1(p,d)\mathcal{F}_{\mu\mu}^*. \tag{S37}$$

$$\mathbb{E}_{\mathbb{T}}\left[H_{\mathcal{L}}^*\right] = \frac{dp^2-1}{d-1}\mathcal{F}^*, \quad \mathrm{Var}_{\mathbb{T}}[\partial_\mu\partial_\nu\mathcal{L}^*] \leq f_2(p,d)\|\Omega_\mu\|_\infty^2\|\Omega_\nu\|_\infty^2. \tag{S38}$$

where $\mathcal{F}$ denote the QFI matrix. f_1 and f_2 are functions of the overlap p and the Hilbert space dimension d, i.e.,

$$f_1(p,d) = \frac{p^2(1-p^2)}{d-1}, \quad f_2(p,d) = \frac{32(1-p^2)}{d-1}\left[p^2 + \frac{2(1-p^2)}{d}\right]. \tag{S39}$$

Proof Using the decomposition in Eq. (4), the loss function can be expressed by

$$\mathcal{L} = 1 - \langle\phi|\varrho|\phi\rangle = 1 - p^2\langle\psi^*|\varrho|\psi^*\rangle - (1-p^2)\langle\psi^\perp|\varrho|\psi^\perp\rangle - 2p\sqrt{1-p^2}\,\mathrm{Re}\left(\langle\psi^\perp|\varrho|\psi^*\rangle\right), \tag{S40}$$

where $\varrho(\boldsymbol{\theta}) = |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|$ denotes the density matrix of the output state from the QNN. According to Lemma S2 and Corollary S4, the expectation of the loss function with respect to the ensemble $\mathbb{T}$ can be calculated as

$$\begin{aligned}\mathbb{E}_{\mathbb{T}}\left[\mathcal{L}(\boldsymbol{\theta})\right] &= 1 - p^2\langle\psi^*|\varrho|\psi^*\rangle - (1-p^2)\frac{\operatorname{tr}[(I - |\psi^*\rangle\langle\psi^*|)\varrho]}{d-1} \\ &= 1 - p^2 + \frac{dp^2-1}{d-1}g(\boldsymbol{\theta}),\end{aligned} \tag{S41}$$

where $g(\boldsymbol{\theta}) = 1 - \langle\psi^*|\varrho(\boldsymbol{\theta})|\psi^*\rangle$ denotes the fidelity distance between the output states at $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. By definition, $g(\boldsymbol{\theta})$ takes the global minimum at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, i.e., at $\varrho = |\psi^*\rangle\langle\psi^*|$. Thus the commutation between the expectation and differentiation gives

$$\begin{aligned}\mathbb{E}_{\mathbb{T}}\left[\nabla\mathcal{L}^*\right] &= \nabla\left(\mathbb{E}_{\mathbb{T}}\left[C\right]\right)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \frac{dp^2-1}{d-1}\,\nabla g(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = 0, \\ \mathbb{E}_{\mathbb{T}}\left[H_{\mathcal{L}}^*\right] &= \frac{dp^2-1}{d-1}\,H_g(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \frac{dp^2-1}{d-1}\,H_g(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \frac{dp^2-1}{d-1}\mathcal{F}^*.\end{aligned} \tag{S42}$$

Note that $H_g(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ is actually the QFI matrix $\mathcal{F}$ of $|\psi(\boldsymbol{\theta})\rangle$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ (see Appendix A.4), which is always positive semidefinite. To estimate the variance, we need to calculate the expression of derivatives first due to the non-linearity of the variance, unlike the case of Eq. (S42) where the operations of taking the expectation and derivative is exchanged. The first order derivative can be expressed by

$$\partial_\mu\mathcal{L} = -\langle\phi|D_\mu|\phi\rangle = -p^2\langle\psi^*|D_\mu|\psi^*\rangle - q^2\langle\psi^\perp|D_\mu|\psi^\perp\rangle - 2pq\operatorname{Re}\left(\langle\psi^\perp|D_\mu|\psi^*\rangle\right). \tag{S43}$$

where $q = \sqrt{1-p^2}$ and $D_\mu = \partial_\mu\varrho$ is a traceless Hermitian operator since $\operatorname{tr} D_\mu = \partial_\mu(\operatorname{tr}\varrho) = 0$. At $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, the operator D_μ is reduced to D_μ^* which satisfies several useful properties

$$\begin{aligned}&D_\mu^* = [\partial_\mu(|\psi\rangle\langle\psi|)]^* = |\partial_\mu\psi^*\rangle\langle\psi^*| + |\psi^*\rangle\langle\partial_\mu\psi^*|, \\ &\langle\psi^*|D_\mu^*|\psi^*\rangle = \langle\psi^*|\partial_\mu\psi^*\rangle + \langle\partial_\mu\psi^*|\psi^*\rangle = \partial_\mu(\langle\psi|\psi\rangle)^* = 0, \\ &\langle\psi^\perp|D_\mu^*|\psi^\perp\rangle = \langle\psi^\perp|\partial_\mu\psi^*\rangle\langle\psi^*|\psi^\perp\rangle + \langle\psi^\perp|\psi^*\rangle\langle\partial_\mu\psi^*|\psi^\perp\rangle = 0, \\ &\langle\psi^\perp|D_\mu^*|\psi^*\rangle = \langle\psi^\perp|\partial_\mu\psi^*\rangle\langle\psi^*|\psi^*\rangle + \langle\psi^\perp|\psi^*\rangle\langle\partial_\mu\psi^*|\psi^*\rangle = \langle\psi^\perp|\partial_\mu\psi^*\rangle,\end{aligned} \tag{S44}$$

where we have used the facts of $\langle\psi|\psi\rangle = 1$ and $\langle\psi^*|\psi^\perp\rangle = 0$. Note that $|\partial_\mu\psi^*\rangle$ actually refers to $(\partial_\mu|\psi\rangle)^*$. Thus the variance of the first order derivative at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ becomes

$$\begin{aligned}\operatorname{Var}_{\mathbb{T}}[\partial_\mu\mathcal{L}^*] &= \mathbb{E}_{\mathbb{T}}\left[(\partial_\mu\mathcal{L}^* - \mathbb{E}_{\mathbb{T}}[\partial_\mu\mathcal{L}^*])^2\right] = \mathbb{E}_{\mathbb{T}}\left[(\partial_\mu\mathcal{L}^*)^2\right] \\ &= 4p^2q^2\,\mathbb{E}_{\mathbb{T}}\left[\left(\operatorname{Re}\langle\psi^\perp|\partial_\mu\psi^*\rangle\right)^2\right].\end{aligned} \tag{S45}$$

According to Lemma S2 and Corollary S4, it holds that

$$\begin{aligned}&\mathbb{E}_{\mathbb{T}}\left[\left(\operatorname{Re}\langle\psi^\perp|\partial_\mu\psi^*\rangle\right)^2\right] = \frac{1}{2}\mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|\partial_\mu\psi^*\rangle\langle\partial_\mu\psi^*|\psi^\perp\rangle\right] \\ &= \frac{\langle\partial_\mu\psi^*|\partial_\mu\psi^*\rangle - \langle\psi^*|\partial_\mu\psi^*\rangle\langle\partial_\mu\psi^*|\psi^*\rangle}{2(d-1)} = \frac{\mathcal{F}_{\mu\mu}^*}{4(d-1)},\end{aligned} \tag{S46}$$

where $\mathcal{F}_{\mu\mu}$ is the QFI diagonal element. Using the generators in the PQC, $\mathcal{F}_{\mu\mu}$ could be expressed as

$$\mathcal{F}_{\mu\mu} = 2\left(\langle\psi|\tilde{\Omega}_\mu^2|\psi\rangle - \langle\psi|\tilde{\Omega}_\mu|\psi\rangle^2\right), \tag{S47}$$

where $\tilde{\Omega}_\mu = V_{\mu\to M}\Omega_\mu V_{\mu\to M}^\dagger$. Finally, the variance of $\partial_\mu\mathcal{L}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ equals to

$$\operatorname{Var}_{\mathbb{T}}[\partial_\mu\mathcal{L}^*] = 4p^2q^2\cdot\frac{\mathcal{F}_{\mu\mu}^*}{4(d-1)} = \frac{p^2(1-p^2)}{d-1}\mathcal{F}_{\mu\mu}^*. \tag{S48}$$

The second order derivative can be expressed by

$$\begin{aligned}(H_{\mathcal{L}})_{\mu\nu} &= \frac{\partial^2\mathcal{L}}{\partial\theta_\mu\partial\theta_\nu} = \partial_\mu\partial_\nu\mathcal{L} = -\langle\phi|D_{\mu\nu}|\phi\rangle \\ &= -p^2\langle\psi^*|D_{\mu\nu}|\psi^*\rangle - q^2\langle\psi^\perp|D_{\mu\nu}|\psi^\perp\rangle - 2pq\operatorname{Re}\left(\langle\psi^\perp|D_{\mu\nu}|\psi^*\rangle\right),\end{aligned} \tag{S49}$$

where $D_{\mu\nu} = \partial_\mu \partial_\nu \varrho$ is a traceless Hermitian operator since $\operatorname{tr} D_{\mu\nu} = \partial_\mu \partial_\nu (\operatorname{tr} \varrho) = 0$. Please do not confuse $D_{\mu\nu}$ with D_μ above. At $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, the $D_{\mu\nu}$ is reduced to $D^*_{\mu\nu}$ which satisfies the following properties

$$
\begin{aligned}
D^*_{\mu\nu} &= \partial_\mu \partial_\nu (|\psi\rangle\langle\psi|)^* = 2\operatorname{Re}\left[|\partial_\mu \partial_\nu \psi^*\rangle\langle\psi^*| + |\partial_\mu \psi^*\rangle\langle\partial_\nu \psi^*|\right], && \text{(S50)}\\
\langle\psi^*|D^*_{\mu\nu}|\psi^*\rangle &= 2\operatorname{Re}\left[\langle\psi^*|\partial_\mu \partial_\nu \psi^*\rangle + \langle\psi^*|\partial_\mu \psi^*\rangle\langle\partial_\nu \psi^*|\psi^*\rangle\right] = -\mathcal{F}_{\mu\nu}, && \text{(S51)}\\
\langle\psi^\perp|D^*_{\mu\nu}|\psi^\perp\rangle &= 2\operatorname{Re}\left[\langle\psi^\perp|\partial_\nu \psi^*\rangle\langle\partial_\mu \psi^*|\psi^\perp\rangle\right]. && \text{(S52)}
\end{aligned}
$$

Here the notation $2\operatorname{Re}[\cdot]$ of square matrix A actually means the sum of the matrix and its Hermitian conjugate, i.e., $2\operatorname{Re}[A] = A + A^\dagger$. From Eq. (S50) we know that the rank of $D_{\mu\nu}$ is at most 4. Substituting the expectation in Eq. (S42), the variance of the second order derivative at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ becomes

$$
\begin{aligned}
\operatorname{Var}_{\mathbb{T}}[\partial_\mu \partial_\nu \mathcal{L}^*] &= \mathbb{E}_{\mathbb{T}}\left[\left(\partial_\mu \partial_\nu \mathcal{L}^* - \mathbb{E}_{\mathbb{T}}\left[\partial_\mu \partial_\nu \mathcal{L}^*\right]\right)^2\right]\\
&= \left(\frac{q^2}{d-1}\mathcal{F}^*_{\mu\nu}\right)^2 + q^4 \mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|D^*_{\mu\nu}|\psi^\perp\rangle^2\right]\\
&\quad - \frac{2q^4}{d-1}\mathcal{F}^*_{\mu\nu}\mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|D^*_{\mu\nu}|\psi^\perp\rangle\right] + 4p^2q^2\mathbb{E}_{\mathbb{T}}\left[(\operatorname{Re}\langle\psi^\perp|D^*_{\mu\nu}|\psi^*\rangle)^2\right].
\end{aligned}
\tag{S53}
$$

where the inhomogeneous cross terms vanish after taking the expectation according to Lemma S2 and have been omitted. Using Corollaries S4 and S6, the expectations in Eq. (S53) can be calculated as

$$
\begin{aligned}
\mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|D^*_{\mu\nu}|\psi^\perp\rangle^2\right] &= \frac{\operatorname{tr}((D^*_{\mu\nu})^2) - 2\left(\langle\psi^*|(D^*_{\mu\nu})^2|\psi^*\rangle - \langle\psi^*|D^*_{\mu\nu}|\psi^*\rangle^2\right)}{d(d-1)},\\
\mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|D^*_{\mu\nu}|\psi^\perp\rangle\right] &= -\frac{\langle\psi^*|D^*_{\mu\nu}|\psi^*\rangle}{d-1} = \frac{\mathcal{F}^*_{\mu\nu}}{d-1},\\
\mathbb{E}_{\mathbb{T}}\left[(\operatorname{Re}\langle\psi^\perp|D^*_{\mu\nu}|\psi^*\rangle)^2\right] &= \frac{1}{2}\mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|D^*_{\mu\nu}|\psi^*\rangle\langle\psi^*|D^*_{\mu\nu}|\psi^\perp\rangle\right]\\
&= \frac{\langle\psi^*|(D^*_{\mu\nu})^2|\psi^*\rangle - \langle\psi^*|D^*_{\mu\nu}|\psi^*\rangle^2}{2(d-1)}.
\end{aligned}
\tag{S54}
$$

Thus the variance of the second order derivative at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ can be written as

$$
\begin{aligned}
\operatorname{Var}_{\mathbb{T}}[\partial_\mu \partial_\nu \mathcal{L}^*] =& q^4 \frac{\|D^*_{\mu\nu}\|_2^2 - 2\left(\langle\psi^*|(D^*_{\mu\nu})^2|\psi^*\rangle - \langle\psi^*|D^*_{\mu\nu}|\psi^*\rangle^2\right)}{d(d-1)}\\
&+ 2p^2q^2 \frac{\langle\psi^*|(D^*_{\mu\nu})^2|\psi^*\rangle - \langle\psi^*|D^*_{\mu\nu}|\psi^*\rangle^2}{d-1} - \left(\frac{q^2}{d-1}\mathcal{F}^*_{\mu\nu}\right)^2.
\end{aligned}
\tag{S55}
$$

Note that the factor

$$
\langle\psi^*|(D^*_{\mu\nu})^2|\psi^*\rangle - \langle\psi^*|D^*_{\mu\nu}|\psi^*\rangle^2 = \langle\psi^*|D^*_{\mu\nu}(I - |\psi^*\rangle\langle\psi^*|)D^*_{\mu\nu}|\psi^*\rangle, \tag{S56}
$$

is non-negative because the operator $(I - |\psi^*\rangle\langle\psi^*|)$ is positive semidefinite. Hence the variance can be upper bounded by

$$
\begin{aligned}
\operatorname{Var}_{\mathbb{T}}[\partial_\mu \partial_\nu \mathcal{L}^*] &\leq \frac{q^4}{d(d-1)}\|D^*_{\mu\nu}\|_2^2 + \frac{2p^2q^2}{d-1}\langle\psi^*|(D^*_{\mu\nu})^2|\psi^*\rangle\\
&\leq \frac{q^4}{d(d-1)}\|D^*_{\mu\nu}\|_2^2 + \frac{2p^2q^2}{d-1}\|(D^*_{\mu\nu})^2\|_\infty - \left(\frac{q^2}{d-1}\mathcal{F}^*_{\mu\nu}\right)^2\\
&\leq \frac{2q^2}{d-1}\left(p^2 + \frac{2q^2}{d}\right)\|D^*_{\mu\nu}\|_\infty^2,
\end{aligned}
\tag{S57}
$$

where we have used the properties

$$
\|D^*_{\mu\nu}\|_2 \leq \sqrt{\operatorname{rank}(D^*_{\mu\nu})}\|D^*_{\mu\nu}\|_\infty \leq 2\|D^*_{\mu\nu}\|_\infty, \quad \|(D^*_{\mu\nu})^2\|_\infty = \|D^*_{\mu\nu}\|_\infty^2. \tag{S58}
$$

Utilizing the quantum gates in the QNN, the operator $D_{\mu\nu}$ can be written as

$$
D_{\mu\nu} = V_{\nu+1\to M}[V_{\mu+1\to\nu}[V_{1\to\mu}|0\rangle\langle0|V^\dagger_{1\to\mu}, i\Omega_\mu]V^\dagger_{\mu+1\to\nu}, i\Omega_\nu]V^\dagger_{\nu+1\to M}, \tag{S59}
$$

where we assume $\mu \leq \nu$ without loss of generality. Thus $\|D_{\mu\nu}\|_\infty$ can be upper bounded by

$$\|D_{\mu\nu}\|_\infty \leq 4\|\Omega_\mu\Omega_\nu\|_\infty \leq 4\|\Omega_\mu\|_\infty\|\Omega_\nu\|_\infty. \tag{S60}$$

Finally, the variance of the second order derivative at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ can be bounded as

$$\mathrm{Var}_{\mathbb{T}}[\partial_\mu\partial_\nu\mathcal{L}^*] \leq f_2(p,d)\|\Omega_\mu\|_\infty^2\|\Omega_\nu\|_\infty^2. \tag{S61}$$

The factor $f_2(p,d)$ reads

$$f_2(p,d) = \frac{32(1-p^2)}{d-1}\left[p^2 + \frac{2(1-p^2)}{d}\right], \tag{S62}$$

which vanishes at least of order $1/d$. ■

Note that when $p \in \{0,1\}$, f_1, f_2 and hence the variances of the first and second order derivatives become exactly zero, indicating $\mathcal{L}^*$ takes the optimum in all cases. This is nothing but the fact that the range of the loss function is $[0,1]$, which reflects that the bound of $\mathrm{Var}_{\mathbb{T}}[\partial_\mu\partial_\nu\mathcal{L}^*]$ is tight in p.

We remark that the vanishing gradient here is both conceptually and technically distinct from barren plateaus [30]. Firstly, here we focus on a fixed parameter point $\boldsymbol{\theta}^*$ instead of a randomly chosen point on the training landscape. Other points apart from $\boldsymbol{\theta}^*$ is allowed to have a non-vanishing gradient expectation, which leads to prominent local minima instead of plateaus. Moreover, the ensemble $\mathbb{T}$ used here originates from the unknown target state instead of the random initialization. The latter typically demands a polynomially deep circuit to form a 2-design. Technically, a constant overlap p is assumed to construct the ensemble $\mathbb{T}$ instead of completely random over the entire Hilbert space. Thus our results apply to adaptive methods, while barren plateaus from the random initialization are not.

Theorem 2 *If the fidelity loss function satisfies $\mathcal{L}(\boldsymbol{\theta}^*) < 1 - 1/d$, the probability that $\boldsymbol{\theta}^*$ is not a local minimum of $\mathcal{L}$ up to a fixed precision $\epsilon = (\epsilon_1, \epsilon_2)$ with respect to the target state ensemble $\mathbb{T}$ is upper bounded by*

$$\mathrm{Pr}_{\mathbb{T}}\left[\neg\,\mathrm{LocalMin}(\boldsymbol{\theta}^*,\epsilon)\right] \leq \frac{2f_1(p,d)\|\boldsymbol{\omega}\|_2^2}{\epsilon_1^2} + \frac{f_2(p,d)\|\boldsymbol{\omega}\|_2^4}{\left(\frac{dp^2-1}{d-1}e^* + \epsilon_2\right)^2}, \tag{S63}$$

where e^* denotes the minimal eigenvalue of the QFI matrix at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. f_1 and f_2 are defined in Lemma 1 which vanish at least of order $1/d$.

Proof By definition in Eq. (5) in the main text, the probability $\mathrm{Pr}_{\mathbb{T}}\left[\neg\,\mathrm{LocalMin}(\boldsymbol{\theta}^*,\epsilon)\right]$ can be upper bounded by the sum of two terms: the probability that one of the gradient component is larger than ϵ_1, and the probability that the Hessian matrix is not positive definite up to the error ϵ_2, i.e.,

$$\begin{aligned}\mathrm{Pr}_{\mathbb{T}}\left[\neg\,\mathrm{LocalMin}(\boldsymbol{\theta}^*,\epsilon)\right] &= \mathrm{Pr}_{\mathbb{T}}\left[\bigcup_{\mu=1}^{M}\{|\partial_\mu\mathcal{L}^*| > \epsilon_1\} \cup \{H_{\mathcal{L}}^* \not\succ -\epsilon_2 I\}\right] \\ &\leq \mathrm{Pr}_{\mathbb{T}}\left[\bigcup_{\mu=1}^{M}\{|\partial_\mu\mathcal{L}^*| > \epsilon_1\}\right] + \mathrm{Pr}_{\mathbb{T}}\left[H_{\mathcal{L}}^* \not\succ -\epsilon_2 I\right].\end{aligned} \tag{S64}$$

The first term can be easily upper bounded by combining Lemma 1 and Chebyshev's inequality, i.e.,

$$\mathrm{Pr}_{\mathbb{T}}\left[\bigcup_{\mu=1}^{M}\{|\partial_\mu\mathcal{L}^*| > \epsilon_1\}\right] \leq \sum_{\mu=1}^{M}\mathrm{Pr}_{\mathbb{T}}\left[|\partial_\mu\mathcal{L}^*| > \epsilon_1\right] \leq \sum_{\mu=1}^{M}\frac{\mathrm{Var}_{\mathbb{T}}[\partial_\mu\mathcal{L}^*]}{\epsilon_1^2} = \frac{f_1(p,d)}{\epsilon_1^2}\,\mathrm{tr}\,\mathcal{F}^*, \tag{S65}$$

where the diagonal element of the QFI matrix is upper bounded as $\mathcal{F}_{\mu\mu} \leq 2\|\Omega_\mu\|_\infty^2$ by definition and thus $\mathrm{tr}\,\mathcal{F}^* \leq 2\|\boldsymbol{\omega}\|_2^2$. Here the generator norm vector $\boldsymbol{\omega}$ is defined as

$$\boldsymbol{\omega} = (\|\Omega_1\|_\infty, \|\Omega_2\|_\infty, \ldots, \|\Omega_M\|_\infty), \tag{S66}$$

so that the squared vector 2-norm of $\boldsymbol{\omega}$ equals to $\|\boldsymbol{\omega}\|_2^2 = \sum_{\mu=1}^{M}\|\Omega_\mu\|_\infty^2$. Thus we obtain the upper bound of the first term, i.e.,

$$\mathrm{Pr}_{\mathbb{T}}\left[\bigcup_{\mu=1}^{M}\{|\partial_\mu\mathcal{L}^*| > \epsilon_1\}\right] \leq \frac{2f_1(p,d)\|\boldsymbol{\omega}\|_2^2}{\epsilon_1^2}, \tag{S67}$$

It takes extra efforts to bound the second term. After assuming $p^2 > 1/d$ to ensure that $\mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^*]$ is positive semidefinite, a sufficient condition of the positive definiteness can be obtained by perturbing $\mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^*]$ using Lemma S8, i.e.,

$$\|H_{\mathcal{L}}^* - \mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^*]\|_\infty < \|\mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^* + \epsilon_2 I]^{-1}\|_\infty^{-1} \quad \Rightarrow \quad H_{\mathcal{L}}^* + \epsilon_2 I \succ 0, \tag{S68}$$

Note that $\|\mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^* + \epsilon_2 I]^{-1}\|_\infty^{-1} = \frac{dp^2-1}{d-1} e^* + \epsilon_2$, where e^* denotes the minimal eigenvalue of the QFI $\mathcal{F}^*$. A necessary condition for $H_{\mathcal{L}}^* + \epsilon_2 I \not\succ 0$ is hence obtained by the contrapositive, i.e.,

$$H_{\mathcal{L}}^* \not\succ -\epsilon_2 I \quad \Rightarrow \quad \|H_{\mathcal{L}}^* - \mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^*]\|_\infty \geq \frac{dp^2-1}{d-1} e^* + \epsilon_2. \tag{S69}$$

Thus the probability that $H_{\mathcal{L}}^*$ is not positive definite can be upper bounded by

$$\Pr_{\mathbb{T}}\left[H_{\mathcal{L}}^* \not\succ -\epsilon_2 I\right] \leq \Pr_{\mathbb{T}}\left[\|H_{\mathcal{L}}^* - \mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^*]\|_\infty \geq \frac{dp^2-1}{d-1} e^* + \epsilon_2\right]. \tag{S70}$$

The generalized Chebyshev's inequality in Lemma S11 regarding $H_{\mathcal{L}}^*$ and the Schatten-∞ norm gives

$$\Pr_{\mathbb{T}}\left[\|H_{\mathcal{L}}^* - \mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^*]\|_\infty \geq \varepsilon\right] \leq \frac{\sigma_\infty^2}{\varepsilon^2}, \tag{S71}$$

where the "norm variance" is defined as $\sigma_\infty^2 = \mathbb{E}_{\mathbb{T}}[\|H_{\mathcal{L}}^* - \mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^*]\|_\infty^2]$. By taking $\varepsilon = \frac{dp^2-1}{d-1} e^* + \epsilon_2$, we obtain

$$\Pr_{\mathbb{T}}\left[H_{\mathcal{L}}^* \not\succeq -\epsilon_2 I\right] \leq \frac{\sigma_\infty^2}{\left(\frac{dp^2-1}{d-1} e^* + \epsilon_2\right)^2}. \tag{S72}$$

Utilizing Lemma 1, σ_∞^2 can be further bounded by

$$\begin{aligned}
\sigma_\infty^2 \leq \sigma_2^2 = \mathbb{E}_{\mathbb{T}}[\|H_{\mathcal{L}}^* - \mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^*]\|_2^2] &= \sum_{\mu\nu} \mathbb{E}_{\mathbb{T}}\left[\left((H_{\mathcal{L}}^*)_{\mu\nu} - (\mathbb{E}_{\mathbb{T}}[H_{\mathcal{L}}^*])_{\mu\nu}\right)^2\right] \\
&= \sum_{\mu,\nu=1}^{M} \mathrm{Var}_{\mathbb{T}}\left[\partial_\mu \partial_\nu \mathcal{L}^*\right] \leq f_2(p,d) \sum_{\mu,\nu=1}^{M} \|\Omega_\mu\|_\infty^2 \|\Omega_\nu\|_\infty^2 \\
&= f_2(p,d)\left(\sum_{\mu=1}^{M} \|\Omega_\mu\|_\infty^2\right)^2 = f_2(p,d)\|\boldsymbol{\omega}\|_2^4.
\end{aligned} \tag{S73}$$

Combining Eqs. (S72) and (S73), we obtain the upper bound of the second term, i.e.,

$$\Pr_{\mathbb{T}}\left[H_{\mathcal{L}}^* \not\succ -\epsilon_2 I\right] \leq \frac{f_2(p,d)\|\boldsymbol{\omega}\|_2^4}{\left(\frac{dp^2-1}{d-1} e^* + \epsilon_2\right)^2}. \tag{S74}$$

Substituting the bounds for the first and second terms into Eq. (S64), one finally arrives at the desired upper bound for the probability that $\boldsymbol{\theta}^*$ is not a local minimum up to a fixed precision $\epsilon = (\epsilon_1, \epsilon_2)$. ∎

Proposition 3 *The expectation and variance of the fidelity loss function $\mathcal{L}$ with respect to the target state ensemble $\mathbb{T}$ can be exactly calculated as*

$$\begin{aligned}
\mathbb{E}_{\mathbb{T}}\left[\mathcal{L}(\boldsymbol{\theta})\right] &= 1 - p^2 + \frac{dp^2-1}{d-1} g(\boldsymbol{\theta}), \\
\mathrm{Var}_{\mathbb{T}}\left[\mathcal{L}(\boldsymbol{\theta})\right] &= \frac{1-p^2}{d-1} g(\boldsymbol{\theta})\left[4p^2 - \left(2p^2 - \frac{(d-2)(1-p^2)}{d(d-1)}\right) g(\boldsymbol{\theta})\right],
\end{aligned} \tag{S75}$$

where $g(\boldsymbol{\theta}) = 1 - |\langle \psi^* | \psi(\boldsymbol{\theta}) \rangle|^2$.

Proof The expression of the expectation $\mathbb{E}_{\mathbb{T}}[\mathcal{L}]$ has already been calculated in Eq. (S41). Considering Lemma S2, the variance of the loss function is

$$\begin{aligned}\mathrm{Var}_{\mathbb{T}}[\mathcal{L}] &= \mathbb{E}_{\mathbb{T}}\left[(\mathcal{L}-\mathbb{E}_{\mathbb{T}}[\mathcal{L}])^2\right]\\ &= \mathbb{E}_{\mathbb{T}}\left[\left(\frac{1-p^2}{d-1}(\langle\psi^*|\varrho|\psi^*\rangle-1)+q^2\langle\psi^\perp|\varrho|\psi^\perp\rangle+2pq\,\mathrm{Re}\left(\langle\psi^\perp|\varrho|\psi^*\rangle\right)\right)^2\right]\\ &= \frac{q^4}{(d-1)^2}(\langle\psi^*|\varrho|\psi^*\rangle-1)^2+\frac{2q^4}{d-1}(\langle\psi^*|\varrho|\psi^*\rangle-1)\,\mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|\varrho|\psi^\perp\rangle\right]\\ &\quad+q^4\,\mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|\varrho|\psi^\perp\rangle^2\right]+4p^2q^2\,\mathbb{E}_{\mathbb{T}}\left[\mathrm{Re}\left(\langle\psi^\perp|\varrho|\psi^*\rangle\right)^2\right],\end{aligned} \tag{S76}$$

where $q=\sqrt{1-p^2}$ and $\varrho(\boldsymbol{\theta})=|\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|$. According to Corollaries S4 and S6, the terms above can be calculated as

$$\begin{aligned}\mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|\varrho|\psi^\perp\rangle\right] &= \frac{1-\langle\psi^*|\varrho|\psi^*\rangle}{d-1},\\ \mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|\varrho|\psi^\perp\rangle^2\right] &= \frac{\left(\mathrm{tr}(\varrho^2)-2\langle\psi^*|\varrho^2|\psi^*\rangle+\langle\psi^*|\varrho|\psi^*\rangle^2\right)+(1-\langle\psi^*|\varrho|\psi^*\rangle)^2}{d(d-1)}\\ &= \frac{2(1-\langle\psi^*|\varrho|\psi^*\rangle)^2}{d(d-1)},\\ \mathbb{E}_{\mathbb{T}}\left[\mathrm{Re}\left(\langle\psi^\perp|\varrho|\psi^*\rangle\right)^2\right] &= \frac{1}{2}\mathbb{E}_{\mathbb{T}}\left[\langle\psi^\perp|\varrho|\psi^*\rangle\langle\psi^*|\varrho|\psi^\perp\rangle\right]=\frac{1-\langle\psi^*|\varrho|\psi^*\rangle^2}{2(d-1)}.\end{aligned} \tag{S77}$$

Thus the variance of the loss function becomes

$$\begin{aligned}\mathrm{Var}_{\mathbb{T}}[\mathcal{L}] &= -\frac{q^4(1-\langle\psi^*|\varrho|\psi^*\rangle)^2}{(d-1)^2}+\frac{2q^4(1-\langle\psi^*|\varrho|\psi^*\rangle)^2}{d(d-1)}+\frac{2p^2q^2(1-\langle\psi^*|\varrho|\psi^*\rangle^2)}{d-1}\\ &= \frac{q^2\left(1-\langle\psi^*|\varrho|\psi^*\rangle\right)}{d-1}\left[\frac{q^2(d-2)(1-\langle\psi^*|\varrho|\psi^*\rangle)}{d(d-1)}+2p^2(1+\langle\psi^*|\varrho|\psi^*\rangle)\right].\end{aligned} \tag{S78}$$

Substituting the relation $\langle\psi^*|\varrho|\psi^*\rangle=1-g(\boldsymbol{\theta})$, the desired expression is obtained. ■

If the quantum gate U_μ in the QNN satisfies the parameter-shift rule, the explicit form of the factor $g(\boldsymbol{\theta})$ could be known along the axis of θ_μ passing through $\boldsymbol{\theta}^*$, which is summarized in Corollary S12. We use $\theta_{\bar{\mu}}$ to represent the other components except for θ_μ, namely $\theta_{\bar{\mu}}=\{\theta_\nu\}_{\nu\neq\mu}$.

Corollary S12 *For QNNs satisfying the parameter-shift rule by $\Omega_\mu^2=I$, the expectation and variance of the fidelity loss function $\mathcal{L}$ restricted by only varying the parameter θ_μ from $\boldsymbol{\theta}^*$ with respect to the target state ensemble $\mathbb{T}$ can be exactly calculated as*

$$\begin{aligned}\mathbb{E}_{\mathbb{T}}\left[\mathcal{L}|_{\theta_{\bar{\mu}}=\theta^*_{\bar{\mu}}}\right] &= 1-p^2+\frac{dp^2-1}{d-1}g(\theta_\mu),\\ \mathrm{Var}_{\mathbb{T}}\left[\mathcal{L}|_{\theta_{\bar{\mu}}=\theta^*_{\bar{\mu}}}\right] &= \frac{1-p^2}{d-1}g(\theta_\mu)\left[4p^2-\left(2p^2-\frac{(d-2)(1-p^2)}{d(d-1)}\right)g(\theta_\mu)\right],\end{aligned} \tag{S79}$$

where $g(\theta_\mu)=\frac{1}{2}\mathcal{F}^*_{\mu\mu}\sin^2\left(\theta_\mu-\theta^*_\mu\right)$.

Proof According to Proposition 3, we only need to calculate the factor $g(\boldsymbol{\theta})|_{\theta_{\bar{\mu}}=\theta^*_{\bar{\mu}}}$. We simply denote this factor as $g(\theta_\mu)$, the explicit expression of which could be calculated by just substituting the parameter-shift rule. Alternatively, the expression of $g(\theta_\mu)$ can be directly written down by considering the following facts. The parameter-shift rule ensures that $g(\theta_\mu)$ must take the form of linear combinations of 1, $\cos(2\theta_\mu)$ and $\sin(2\theta_\mu)$ since $U_\mu(\theta_\mu)=e^{-i\Omega_\mu\theta_\mu}=\cos\theta_\mu I-i\sin\theta_\mu\Omega_\mu$ and $g(\theta_\mu)$ takes the form of $U_\mu(\cdot)U_\mu^\dagger$. Furthermore, $g(\theta_\mu)$ takes its minimum at θ^*_μ so that it is an even function relative to $\theta_\mu=\theta^*_\mu$. Combined with the fact that $g(\theta_\mu)$ also takes zero at θ^*_μ, we know $g(\theta_\mu)\propto[1-\cos(2(\theta_\mu-\theta^*_\mu))]$. The coefficient can be determined by considering that the second order derivative of $g(\theta_\mu)$ equals to the QFI matrix element $\mathcal{F}^*_{\mu\mu}$ by definition, so that

$$g(\theta_\mu)=g(\boldsymbol{\theta})|_{\theta_{\bar{\mu}}=\theta^*_{\bar{\mu}}}=\frac{1}{4}\mathcal{F}^*_{\mu\mu}[1-\cos(2(\theta_\mu-\theta^*_\mu))]=\frac{1}{2}\mathcal{F}^*_{\mu\mu}\sin^2(\theta_\mu-\theta^*_\mu). \tag{S80}$$

The expressions of the expectation and variance of the loss function can be obtained by directly substituting Eq. (S80) into Proposition 3. ■

C Generalization to the local loss function

In the main text, we focus on the fidelity loss function, also known as the "global" loss function [31], where the ensemble construction and calculation are preformed in a clear and meaningful manner. However, there is another type of loss function called "local" loss function [31], such as the energy expectation in the variational quantum eigensolver (VQE) which aims to prepare the ground state of a physical system. The local loss function takes the form of

$$\mathcal{L}(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|H|\psi(\boldsymbol{\theta})\rangle, \tag{S81}$$

where H is the Hamiltonian of the physical system as a summation of Pauli strings. Eq. (S81) can formally reduce to the fidelity loss function by taking $H = I - |\phi\rangle\langle\phi|$. In this section, we generalize the results of the fidelity loss function to the local loss function and show that the conclusion keeps the same, though the ensemble construction and calculation are more complicated.

The ensemble we used in the main text decomposes the unknown target state into the learnt component $|\psi^*\rangle$ and the unknown component $|\psi^\perp\rangle$, and regards $|\psi^\perp\rangle$ as a Haar random state in the orthogonal complement of $|\psi^*\rangle$. This way of thinking seems to be more subtle in the case of the local loss function since the Hamiltonian is usually already known in the form of Pauli strings and hence it is unnatural to assume an unknown Hamiltonian. However, a known Hamiltonian does not imply a known target state, i.e., the ground state of the physical system. One needs to diagonalize the Hamiltonian to find the ground state, which requires an exponential cost in classical computers. That is to say, what one really does not know is the unitary used in the diagonalization, i.e., the relation between the learnt state $|\psi^*\rangle$ and the eigen-basis of the Hamiltonian. We represent this kind of uncertainty by a unitary V from the ensemble $\mathbb{V}$, where $\mathbb{V}$ comes from the ensemble $\mathbb{U}$ mentioned in Appendix A.1 by specifying $\bar{P} = |\psi^*\rangle\langle\psi^*|$. Such an ensemble $\mathbb{V}$ induces an ensemble of loss functions via

$$\mathcal{L}(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|V^\dagger H V|\psi(\boldsymbol{\theta})\rangle, \tag{S82}$$

similar with the loss function ensemble induced by the unknown target state in the main text. $\mathbb{V}$ can be interpreted as all of the possible diagonalizing unitaries that keeps the loss value $\mathcal{L}(\boldsymbol{\theta}^*)$ constant, denoted as $\mathcal{L}^*$. In the following, similar with those for the global loss function, we calculate the expectation and variance of the derivatives of the local loss function in Lemma S13 and bound the probability of avoiding local minima in Theorem S14. Hence, the results and relative discussions in the main text could generalize to the case of local loss functions.

Lemma S13 *The expectation and variance of the gradient $\nabla\mathcal{L}$ and Hessian matrix $H_{\mathcal{L}}$ of the local loss function $\mathcal{L}(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|H|\psi(\boldsymbol{\theta})\rangle$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ with respect to the ensemble $\mathbb{V}$ satisfy*

$$\begin{aligned} &\mathbb{E}_{\mathbb{V}}\left[\nabla\mathcal{L}^*\right] = 0, \quad \mathrm{Var}_{\mathbb{V}}[\partial_\mu\mathcal{L}^*] = f_1(H,d)\mathcal{F}^*_{\mu\mu}, \\ &\mathbb{E}_{\mathbb{V}}\left[H^*_{\mathcal{L}}\right] = \frac{\operatorname{tr} H - d\mathcal{L}^*}{d-1}\mathcal{F}^*, \quad \mathrm{Var}_{\mathbb{V}}\left[\partial_\mu\partial_\nu\mathcal{L}^*\right] \le f_2(H,d)\|\Omega_\mu\|_\infty^2\|\Omega_\nu\|_\infty^2, \end{aligned} \tag{S83}$$

where $\mathcal{F}$ denotes the QFI matrix. f_1 and f_2 are functions of the Hamiltonian H and the Hilbert space dimension d, i.e.,

$$f_1(H,d) = \frac{\langle H^2\rangle_* - \langle H\rangle_*^2}{d-1}, \quad f_2(H,d) = 32\left(\frac{\langle H^2\rangle_* - \langle H\rangle_*^2}{d-1} + \frac{2\|H\|_2^2}{d(d-2)}\right), \tag{S84}$$

where we introduce the notation $\langle\cdot\rangle_* = \langle\psi^*|\cdot|\psi^*\rangle$.

Proof Using Lemma S3, the expectation of the local loss function can be directly calculated as

$$\mathbb{E}_{\mathbb{V}}\left[\mathcal{L}(\boldsymbol{\theta})\right] = \mathcal{L}^* + \frac{\operatorname{tr} H - d\mathcal{L}^*}{d-1}g(\boldsymbol{\theta}). \tag{S85}$$

where $g(\boldsymbol{\theta}) = 1 - \langle\psi^*|\varrho(\boldsymbol{\theta})|\psi^*\rangle$ denotes the fidelity distance between the output states at $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. By definition, $g(\boldsymbol{\theta})$ takes the global minimum at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Thus the commutation between the expectation and differentiation gives

$$\begin{aligned}\mathbb{E}_{\mathbb{V}}\left[\nabla\mathcal{L}^*\right] &= \nabla\left(\mathbb{E}_{\mathbb{V}}\left[\mathcal{L}\right]\right)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \frac{\operatorname{tr}H - d\mathcal{L}^*}{d-1}\,\nabla g(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = 0,\\ \mathbb{E}_{\mathbb{V}}\left[H^*_{\mathcal{L}}\right] &= \frac{\operatorname{tr}H - d\mathcal{L}^*}{d-1}\,H_g(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \frac{\operatorname{tr}H - d\mathcal{L}^*}{d-1}\mathcal{F}^*.\end{aligned}\tag{S86}$$

By definition, $H_g(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ is actually the QFI matrix $\mathcal{F}^*$ of $|\psi(\boldsymbol{\theta})\rangle$ at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ (see Appendix A.4), which is always positive semidefinite. To estimate the variance, we need to calculate the expression of derivatives first due to the non-linearity of the variance. The first order derivative of the local loss function can be expressed by

$$\partial_\mu\mathcal{L} = \operatorname{tr}[V^\dagger HVD_\mu] = 2\operatorname{Re}\langle\psi|V^\dagger HV|\partial_\mu\psi\rangle,\tag{S87}$$

where $D_\mu = \partial_\mu\varrho$ is a traceless Hermitian operator since $\operatorname{tr}D_\mu = \partial_\mu(\operatorname{tr}\varrho) = 0$. By definition, we know that $|\psi\rangle^*$ is not changed by V, i.e., $V|\psi^*\rangle = |\psi^*\rangle$, which leads to the reduction $\partial_\mu\mathcal{L}^* = 2\operatorname{Re}(\langle\psi^*|HV|\partial_\mu\psi^*\rangle)$. Hence, the variance of the first order derivative at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ is

$$\begin{aligned}\operatorname{Var}_{\mathbb{V}}[\partial_\mu\mathcal{L}^*] &= \mathbb{E}_{\mathbb{V}}\left[(\partial_\mu\mathcal{L}^* - \mathbb{E}_{\mathbb{V}}[\partial_\mu\mathcal{L}^*])^2\right] = \mathbb{E}_{\mathbb{V}}\left[(\partial_\mu\mathcal{L}^*)^2\right]\\ &= \mathbb{E}_{\mathbb{V}}\left[(2\operatorname{Re}\langle\psi^*|HV|\partial_\mu\psi^*\rangle)^2\right]\\ &= \mathbb{E}_{\mathbb{V}}\left[\langle\psi^*|HV|\partial_\mu\psi^*\rangle^2\right] + \mathbb{E}_{\mathbb{V}}\left[\langle\partial_\mu\psi^*|V^\dagger H|\psi^*\rangle^2\right]\\ &\quad + 2\mathbb{E}_{\mathbb{V}}\left[\langle\psi^*|HV|\partial_\mu\psi^*\rangle\langle\partial_\mu\psi^*|V^\dagger H|\psi^*\rangle\right].\end{aligned}\tag{S88}$$

Utilizing Lemmas S2 and S3, we obtain

$$\begin{aligned}&\mathbb{E}_{\mathbb{V}}\left[\langle\psi^*|HV|\partial_\mu\psi^*\rangle^2\right] = \langle H\rangle_*^2\langle\psi^*|\partial_\mu\psi^*\rangle^2,\\ &\mathbb{E}_{\mathbb{V}}\left[\langle\partial_\mu\psi^*|V^\dagger H|\psi^*\rangle^2\right] = \langle H\rangle_*^2\langle\partial_\mu\psi^*|\psi^*\rangle^2,\\ &\mathbb{E}_{\mathbb{V}}\left[\langle\psi^*|HV|\partial_\mu\psi^*\rangle\langle\partial_\mu\psi^*|V^\dagger H|\psi^*\rangle\right]\\ &\quad = \frac{\langle H^2\rangle_* - \langle H\rangle_*^2}{2(d-1)}\mathcal{F}^*_{\mu\mu} + \langle H\rangle_*^2\langle\psi^*|\partial_\mu\psi^*\rangle\langle\partial_\mu\psi^*|\psi^*\rangle,\end{aligned}\tag{S89}$$

where we introduce the notation $\langle\cdot\rangle_* = \langle\psi^*|\cdot|\psi^*\rangle$ and hence $\mathcal{L}^* = \langle H\rangle_*$. The $1/2$ factor in the third line arises from the definition of the QFI matrix. Note that there are three terms above canceling each other due to the fact

$$\begin{aligned}&2\operatorname{Re}\left[\langle\partial_\mu\psi|\psi\rangle\right] = \langle\partial_\mu\psi|\psi\rangle + \langle\psi|\partial_\mu\psi\rangle = \partial_\mu(\langle\psi|\psi\rangle) = 0,\\ &\langle\psi|\partial_\mu\psi\rangle^2 + \langle\partial_\mu\psi|\psi\rangle^2 + 2\langle\psi|\partial_\mu\psi\rangle\langle\partial_\mu\psi|\psi\rangle = (2\operatorname{Re}\left[\langle\partial_\mu\psi|\psi\rangle\right])^2 = 0.\end{aligned}\tag{S90}$$

Therefore, the variance of the first order derivative at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ equals to

$$\operatorname{Var}_{\mathbb{V}}\left[\partial_\mu\mathcal{L}^*\right] = \mathbb{E}_{\mathbb{V}}\left[(2\operatorname{Re}\langle\psi^*|HV|\partial_\mu\psi^*\rangle)^2\right] = \frac{\langle H^2\rangle_* - \langle H\rangle_*^2}{d-1}\mathcal{F}^*_{\mu\mu}.\tag{S91}$$

The second-order derivative can be expressed by

$$\partial_\mu\partial_\nu\mathcal{L} = (H_{\mathcal{L}})_{\mu\nu} = \operatorname{tr}\left[V^\dagger HVD_{\mu\nu}\right],\tag{S92}$$

where $D_{\mu\nu} = \partial_\mu\partial_\nu\varrho$ is a traceless Hermitian operator since $\operatorname{tr}D_{\mu\nu} = \partial_\mu\partial_\nu(\operatorname{tr}\varrho) = 0$. By direct expansion, the variance of the second order derivative at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ can be expressed as

$$\operatorname{Var}_{\mathbb{V}}[\partial_\mu\partial_\nu\mathcal{L}^*] = \mathbb{E}_{\mathbb{V}}\left[(\partial_\mu\partial_\nu\mathcal{L}^*)^2\right] - (\mathbb{E}_{\mathbb{V}}\left[\partial_\mu\partial_\nu\mathcal{L}^*\right])^2,\tag{S93}$$

where the second term is already obtained in Eq. (S86). Lemma S7 directly implies

$$\begin{aligned}
&\mathbb{E}_{\mathbb{V}}\left[(\partial_\mu\partial_\nu\mathcal{L})^2\right] = \mathbb{E}_{\mathbb{V}}\left[\operatorname{tr}(V^\dagger HVD_{\mu\nu})\operatorname{tr}(V^\dagger HVD_{\mu\nu})\right] \\
&= \langle H\rangle_*^2\langle D_{\mu\nu}\rangle_*^2 + \frac{2\langle H\rangle_*\langle D_{\mu\nu}\rangle_*}{d-1}(\operatorname{tr}H-\langle H\rangle_*)(\operatorname{tr}D_{\mu\nu}-\langle D_{\mu\nu}\rangle_*) \\
&+ \frac{2}{d-1}(\langle H^2\rangle_* - \langle H\rangle_*^2)(\langle D_{\mu\nu}^2\rangle_* - \langle D_{\mu\nu}\rangle_*^2) \\
&+ \frac{1}{d(d-2)}(\operatorname{tr}H-\langle H\rangle_*)^2(\operatorname{tr}D_{\mu\nu}-\langle D_{\mu\nu}\rangle_*)^2 \\
&+ \frac{1}{d(d-2)}(\operatorname{tr}(H^2)-2\langle H^2\rangle_*+\langle H\rangle_*^2)(\operatorname{tr}(D_{\mu\nu}^2)-2\langle D_{\mu\nu}^2\rangle_*+\langle D_{\mu\nu}\rangle_*^2) \\
&- \frac{1}{d(d-1)(d-2)}\left[(\operatorname{tr}(H^2)-2\langle H^2\rangle_*+\langle H\rangle_*^2)(\operatorname{tr}D_{\mu\nu}-\langle D_{\mu\nu}\rangle_*)^2\right] \\
&- \frac{1}{d(d-1)(d-2)}\left[(\operatorname{tr}H-\langle H\rangle_*)^2(\operatorname{tr}(D_{\mu\nu}^2)-2\langle D_{\mu\nu}^2\rangle_*+\langle D_{\mu\nu}\rangle_*^2)\right].
\end{aligned} \tag{S94}$$

According to Eq. (S86), $\langle D_{\mu\nu}^*\rangle_* = -\mathcal{F}_{\mu\nu}^*$ in Eq. (S51) and $\mathcal{L}^* = \langle H\rangle_*$, we have

$$(\mathbb{E}_{\mathbb{V}}[\partial_\mu\partial_\nu\mathcal{L}^*])^2 = \left(\frac{\operatorname{tr}H-d\mathcal{L}^*}{d-1}\mathcal{F}_{\mu\nu}^*\right)^2 = \left(\frac{\operatorname{tr}H-\langle H\rangle_*}{d-1}-\langle H\rangle_*\right)^2\langle D_{\mu\nu}^*\rangle_*^2. \tag{S95}$$

Combining Eqs. (S94) and (S95) together with the condition $\operatorname{tr}D_{\mu\nu}=0$, we obtain

$$\begin{aligned}
&\mathrm{Var}_{\mathbb{V}}[\partial_\mu\partial_\nu\mathcal{L}^*] = \frac{2}{d-1}(\langle H^2\rangle_*-\langle H\rangle_*^2)(\langle D_{\mu\nu}^{2*}\rangle_*-\langle D_{\mu\nu}^*\rangle_*^2) \\
&+ \frac{1}{d(d-2)}(\operatorname{tr}(H^2)-2\langle H^2\rangle_*+\langle H\rangle_*^2)(\operatorname{tr}(D_{\mu\nu}^{2*})-2\langle D_{\mu\nu}^{2*}\rangle_*+\langle D_{\mu\nu}^*\rangle_*^2) \\
&- \frac{1}{d(d-1)(d-2)}\left[(\operatorname{tr}(H^2)-2\langle H^2\rangle_*+\langle H\rangle_*^2)\langle D_{\mu\nu}^*\rangle_*^2\right] \\
&- \frac{1}{d(d-1)(d-2)}\left[(\operatorname{tr}H-\langle H\rangle_*)^2(\operatorname{tr}(D_{\mu\nu}^{2*})-2\langle D_{\mu\nu}^{2*}\rangle_*+\frac{d-2}{d-1}\langle D_{\mu\nu}\rangle_*^2))\right].
\end{aligned} \tag{S96}$$

Note that we always have $d\geq 2$ in qubit systems. If $\operatorname{rank}H\geq 2$, then it holds that

$$\operatorname{tr}(H^2)-2\langle H^2\rangle_*+\langle H\rangle_*^2 \geq \|H\|_2^2-2\|H\|_\infty^2 \geq (\operatorname{rank}H)\|H\|_\infty^2-2\|H\|_\infty^2\geq 0. \tag{S97}$$

Otherwise if $\operatorname{rank}H=1$ (the case of $\operatorname{rank}H=0$ is trivial), then we assume $H=\lambda|\phi\rangle\langle\phi|$ and it holds that

$$\operatorname{tr}(H^2)-2\langle H^2\rangle_*+\langle H\rangle_*^2 = \lambda^2-2\lambda^2|\langle\psi^*|\phi\rangle|^2+\lambda^2|\langle\psi^*|\phi\rangle|^4 = \lambda^2\left(1-|\langle\psi^*|\phi\rangle|^2\right)^2\geq 0. \tag{S98}$$

Hence we conclude that it always holds that

$$\operatorname{tr}(H^2)-2\langle H^2\rangle_*+\langle H\rangle_*^2\geq 0. \tag{S99}$$

Similarly, because $\operatorname{tr}D_{\mu\nu}=0$, we know $\operatorname{rank}D_{\mu\nu}\geq 2$ and thus

$$\operatorname{tr}(D_{\mu\nu}^2)-2\langle D_{\mu\nu}^2\rangle_*\geq(\operatorname{rank}D_{\mu\nu})\|D_{\mu\nu}\|_\infty^2-2\|D_{\mu\nu}\|_\infty^2\geq 0. \tag{S100}$$

Therefore, we can upper bound the variance by just discarding the last two terms in Eq. (S96)

$$\begin{aligned}
&\mathrm{Var}_{\mathbb{V}}[\partial_\mu\partial_\nu\mathcal{L}^*] \leq \frac{2}{d-1}(\langle H^2\rangle_*-\langle H\rangle_*^2)(\langle D_{\mu\nu}^{2*}\rangle_*-\langle D_{\mu\nu}^*\rangle_*^2) \\
&+ \frac{1}{d(d-2)}(\operatorname{tr}(H^2)-2\langle H^2\rangle_*+\langle H\rangle_*^2)(\operatorname{tr}(D_{\mu\nu}^{2*})-2\langle D_{\mu\nu}^{2*}\rangle_*+\langle D_{\mu\nu}^*\rangle_*^2).
\end{aligned} \tag{S101}$$

On the other hand, we have

$$\operatorname{tr}(H^2)-2\langle H^2\rangle_*+\langle H\rangle_*^2 = \operatorname{tr}(H^2)-\langle H^2\rangle_*-(\langle H^2\rangle_*-\langle H\rangle_*^2)\leq\operatorname{tr}(H^2), \tag{S102}$$

since $\langle H^2\rangle_*-\langle H\rangle_*^2 = \langle H(I-|\psi^*\rangle\langle\psi^*|)H\rangle_*\geq 0$. A similar inequality also holds for $D_{\mu\nu}$. Thus the variance can be further bounded by

$$\mathrm{Var}_{\mathbb{V}}[\partial_\mu\partial_\nu\mathcal{L}^*]\leq\frac{2}{d-1}(\langle H^2\rangle_*-\langle H\rangle_*^2)\|D_{\mu\nu}^*\|_\infty^2+\frac{4\|H\|_2^2\|D_{\mu\nu}^*\|_\infty^2}{d(d-2)}, \tag{S103}$$

where we have used the properties in Eq. (S58). Using the inequality in Eq. (S60) associated with the gate generators, the variance of the second order derivative at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ can be ultimately upper bounded by

$$\mathrm{Var}_{\mathbb{V}}[\partial_\mu \partial_\nu \mathcal{L}^*] \leq f_2(H, d) \|\Omega_\mu\|_\infty^2 \|\Omega_\nu\|_\infty^2, \tag{S104}$$

The factor $f_2(H, d)$ reads

$$f_2(H, d) = 32 \left(\frac{\langle H^2 \rangle_* - \langle H \rangle_*^2}{d-1} + \frac{2\|H\|_2^2}{d(d-2)} \right), \tag{S105}$$

which vanishes at least of order $\mathcal{O}(\mathrm{poly}(N) 2^{-N})$ with the qubit count $N = \log_2 d$ if $\|H\|_\infty \in \mathcal{O}(\mathrm{poly}(N))$. ■

Theorem S14 *If $\mathcal{L}^* < \frac{\operatorname{tr} H}{d}$, the probability that $\boldsymbol{\theta}^*$ is not a local minimum of the local cost function $\mathcal{L}$ up to a fixed precision $\epsilon = (\epsilon_1, \epsilon_2)$ with respect to the ensemble $\mathbb{V}$ is upper bounded by*

$$\mathrm{Pr}_{\mathbb{V}} \left[\neg\, \mathrm{LocalMin}(\boldsymbol{\theta}^*, \epsilon) \right] \leq \frac{2 f_1(H, d) \|\boldsymbol{\omega}\|_2^2}{\epsilon_1^2} + \frac{f_2(H, d) \|\boldsymbol{\omega}\|_2^4}{\left(\frac{\operatorname{tr} H - d\mathcal{L}^*}{d-1} e^* + \epsilon_2 \right)^2}, \tag{S106}$$

where e^* denotes the minimal eigenvalue of the QFI matrix at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. f_1 and f_2 are defined in Eq. (S84) which vanish at least of order $\mathcal{O}(\mathrm{poly}(N) 2^{-N})$ with the qubit count $N = \log_2 d$ if $\|H\|_\infty \in \mathcal{O}(\mathrm{poly}(N))$.

Proof Utilizing Lemma S13, the proof is exactly the same as that of Theorem 2 up to the different hessian expectation $\mathbb{E}_{\mathbb{V}}[H_{\mathcal{L}}^*]$ and coefficient functions f_1 and f_2. ■