arXiv:2306.03481

Transition role of entangled data in quantum machine learning

Xinbiao Wang, Yuxuan Du, Zhuozhuo Tu, Yong Luo, Xiao Yuan, & Dacheng Tao

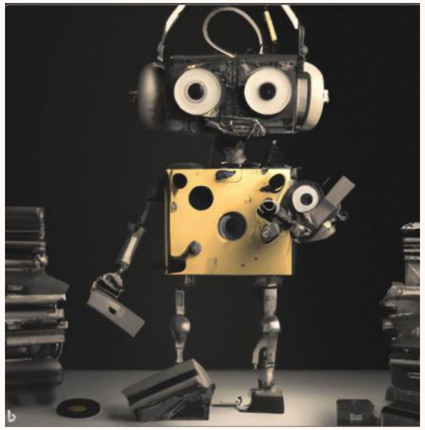
QTML2023, CERN, Geneva 20 Nov 2023

武漢大學 WUHAN UNIVERSITY

THE UNIVERSITY OF SYDNEY

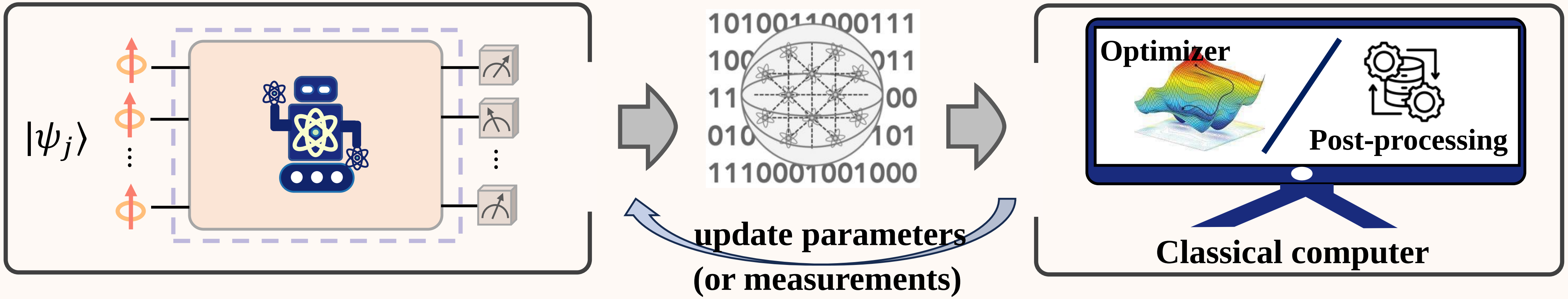
JD.COM

北京大學 PEKING UNIVERSITY



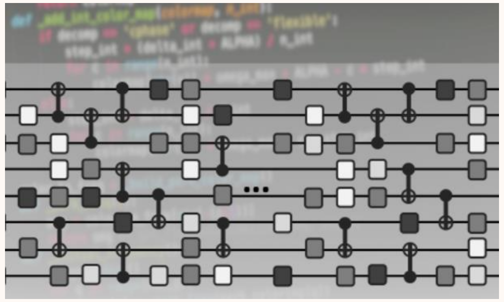
Quantum machine learning (QML)

Quantum-classical hybrid



Application

Classification (for Mnist or phase)



Unitary compiling

Quantum Fidelity

2-point Correlations

Entanglement Entropy

Local Observables

Quantum machine learning (QML)

A general formalism for quantum machine learning models

- the type of states
- the type of quantum circuits used by the learner
- the type of measurement done by the learner

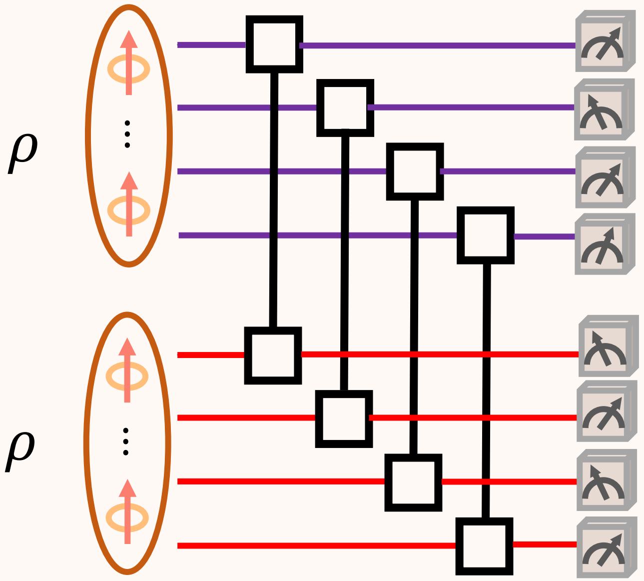
Modifying any one of these parameters can change the quantum learning model!

Evaluation metric for quantum learning models

- **Prediction error:** the ability to accurately make predictions on unseen data
- **Sample complexity:** the training data size used by the learning algorithm
- **Query complexity:** the number of total copies of the input states used by the learning algorithm

The power of entanglement in QML

Most quantum learning algorithms with quantum advantages share the common features: entanglement!



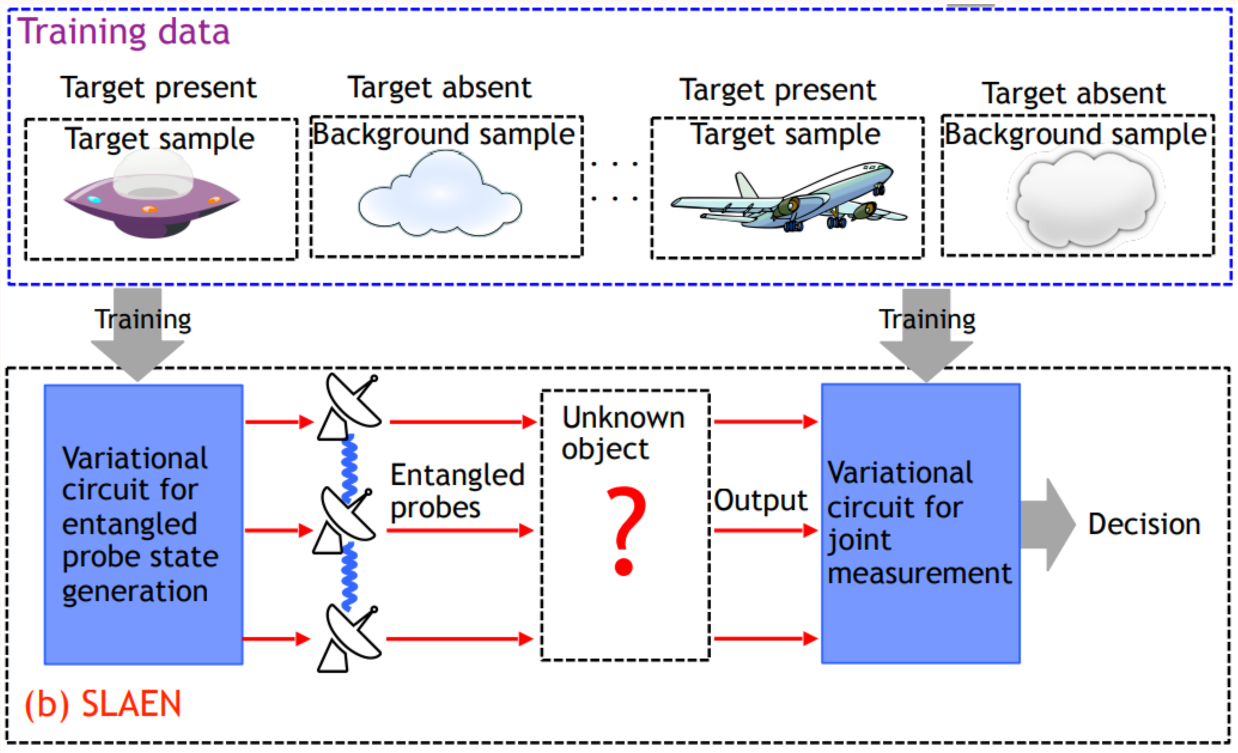
: Clliford gate

Predicting for any O:

$\mathrm{Tr}(\rho O)$

Using **entanglement** in quantum **measurements** [1,2]

Using **entanglement** in quantum **dynamics** [3]



[1] Huang, Hsin-Yuan, et al. "Quantum advantage in learning from experiments." Science 376.6598 (2022): 1182-1186.

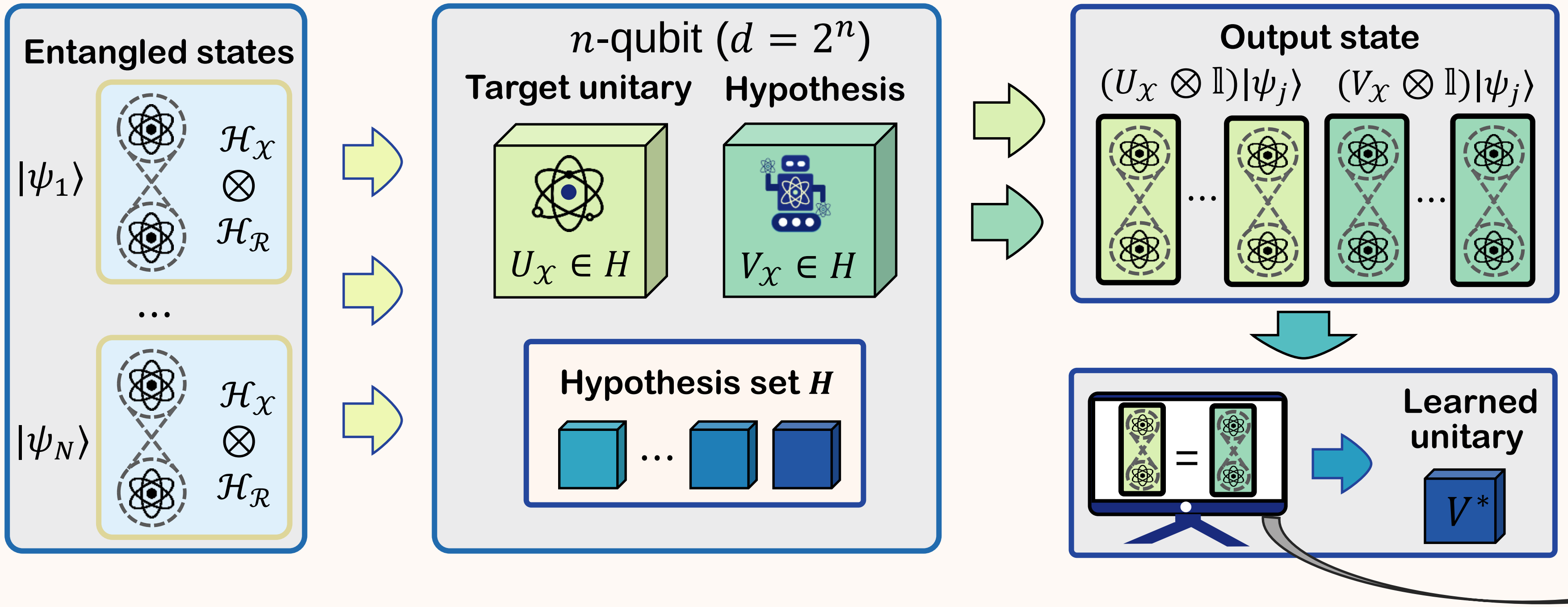
[2] Huang, Hsin-Yuan, et al. "Information-theoretic bounds on quantum advantage in machine learning." Physical Review Letters (2021)

[3] Zhuang, Quntao,et al. "Physical-layer supervised learning assisted by an entangled sensor network." Physical Review X (2019)

Entangled data in QML

How about the case when incorporating entanglement into training data?

- [4] shows the using entangled data can exponentially reduce the sample complexity for achieving zero prediction error.



Entangled states:

$$|\psi_j\rangle = \sum_{k=1}^{r} \sqrt{c_{j,k}}\, |\xi_{j,k}\rangle_\mathcal{X} |\zeta_{j,k}\rangle_\mathcal{R}$$

Training dataset:

$$\mathcal{S}_Q = \{(|\psi_j\rangle, |\phi_j\rangle) : |\psi_j\rangle, |\phi_j\rangle = (U \otimes \mathbb{I}_d)|\psi_j\rangle \in \mathcal{H}_{\mathcal{X},\mathcal{R}}\}_{j=1}^{N}$$

Perfect training (zero training error):

$$\langle \psi_j | V_\mathcal{X}^\dagger U_\mathcal{X} \otimes \mathbb{I} | \psi_j \rangle^2 = 1$$

Risk function: $\widetilde{R}_U(V_{\mathcal{S}_Q}) := \int_{\phi \sim \mathrm{Haar}} \mathrm{d}\phi \frac{1}{4} \left\| U|\phi\rangle\langle\phi|U^\dagger - V_{\mathcal{S}_Q}|\phi\rangle\langle\phi|V_{\mathcal{S}_Q}^\dagger \right\|_1^2$

Lower bound: $\mathbb{E}_U \mathbb{E}_{\mathcal{S}_Q} \widetilde{R}_U(V_{\mathcal{S}_Q}) \geq 1 - \frac{d + r^2 N^2}{d(d+1)}.$

[4] Sharma, Kunal, et al. "Reformulation of the no-free-lunch theorem for entangled datasets." Physical Review Letters 128.7 (2022)

Entangled data in QML

★ Caveats in previous work:

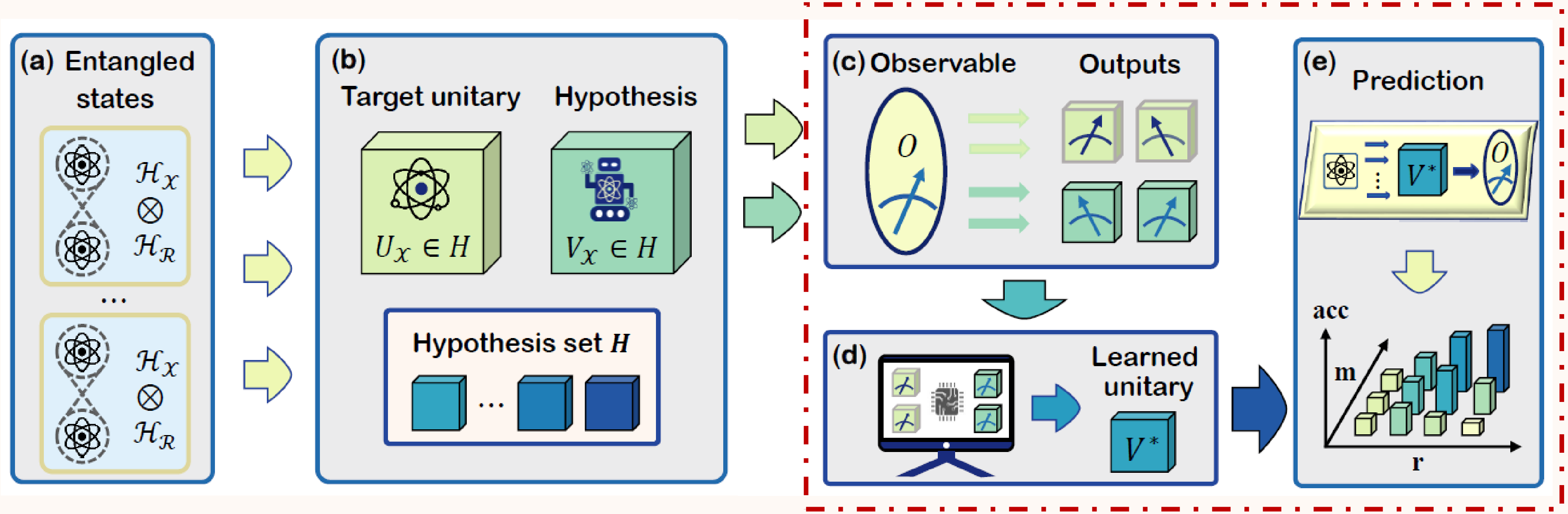
(1) Infinite number of measurements

(2) Coherent learning protocol

(3) Perfect training assumption （zero training error）

How about the power of entangled data in a more realistic setting?

Realistic problem setting



Learning task: $f_U(|\psi\rangle) = Tr(U|\psi\rangle\langle\psi|U^\dagger O)$ (We adopt projective measurement $O = |o\rangle\langle o|$)

Training Dataset: $\mathcal{S} = \left\{|\psi_j\rangle, o_j) \colon |\psi_j\rangle \in \mathcal{H}_{\mathcal{X}\mathcal{R}}, o_j = \frac{1}{m}\sum_{k=1}^{m} o_{jk} \right\}_{j=1}^{N}$

Risk function: $R_U(V_\mathcal{S}) = \mathbb{E}_{|\psi\rangle\sim Haar} Tr\left(O(V_\mathcal{S}|\psi\rangle\langle\psi|V_\mathcal{S}{}^\dagger - U|\psi\rangle\langle\psi|U^\dagger)\right)^2$

Transition role of entangled data

In the setting of finite number of measurements：Does entangled data contribute to quantum advantage？

We show that [5]： Assuming the training error is less than ε, the averaged risk function is lower bounded by

$$\mathbb{E}_U \mathbb{E}_{\mathcal{S}} R_U(V_{\mathcal{S}}) \geq \Omega\left(\frac{\tilde{\varepsilon}^2}{4^n}\left(1 - \frac{N \cdot \min\{m/(rc_1),\ rn\}}{2^n c_2}\right)\right) \qquad (\tilde{\varepsilon} = \Theta(2^n \varepsilon))$$

The implications from this lower bound in terms of r, N, m :

For Schmidt rank r: Entangled data has a **dual effect** in the prediction error：

Positive effect: For **a large number** of measurements $m \geq c_1 r^2 n$, entangled data leads to a **small prediction error.**

$r = 2^n$ can achieve an **exponential reduction** in terms of training data size N compared with $r = 1$.

This echoes with the result achieved in [4]

[4] Sharma, Kunal, et al. "Reformulation of the no-free-lunch theorem for entangled datasets." Physical Review Letters 128.7 (2022)
[5] Wang, Xinbiao, et al. "Transition role of entangled data in quantum machine learning." Arxiv:2306_03481 (2023)

Transition role of entangled data

In the setting of finite number of measurements：Does entangled data contribute to quantum advantage？

We show that [5]： Assuming the training error is less than ε, the averaged risk function is lower bounded by

$$\mathbb{E}_U \mathbb{E}_S R_U(V_S) \geq \Omega\left(\frac{\tilde{\varepsilon}^2}{4^n}\left(1 - \frac{N \cdot \min\{m/(rc_1),\ rn\}}{2^n c_2}\right)\right) \qquad (\tilde{\varepsilon} = \Theta(2^n \varepsilon))$$

The implications from this lower bound：

For Schmidt rank r: Entangled data has a **dual effect** in the prediction error：

Negative effect: For **a small number** of measurements $m < c_1 r^2 n$,

highly entangled data not only requires *a large amount of quantum resource* for preparing,

but also leads to **a large prediction error**.

[4] Sharma, Kunal, et al. "Reformulation of the no-free-lunch theorem for entangled datasets." Physical Review Letters 128.7 (2022)

[5] Wang, Xinbiao, et al. “Transition role of entangled data in quantum machine learning.” Arxiv:2306_03481 (2023)

Transition role of entangled data

We show that [5] : Assuming the training error is less than ε, the averaged prediction error is lower bounded by

$$\mathbb{E}_U \mathbb{E}_{\mathcal{S}} R_U(V_{\mathcal{S}}) \geq \Omega\left(\frac{\tilde{\varepsilon}^2}{4^n}\left(1 - \frac{N \cdot \min\{m/(rc_1),\ rn\}}{2^n c_2}\right)\right) \quad (c_1 = \Theta(2^n/\tilde{\varepsilon}^2))$$

The implications from this lower bound：

- **For training data size $\boldsymbol{N}$**: increasing N can **constantly decrease** the prediction error.

- **For number of measurements $\boldsymbol{m}$:** While m contributes to a small prediction error, it is not ***decisive*** to the ultimate performance of the prediction error, which is determined by N and r.

  - ⭐ At least $\mathrm{m} \geq r_2 c_1 n$ measurements are required to *fully utilize the power* of entangled data

Transition role of entangled data

We show that [5] : Assuming the training error is less than ε, the averaged prediction error is lower bounded by

$$\mathbb{E}_U \mathbb{E}_{\mathcal{S}} R_U(V_{\mathcal{S}}) \geq \Omega\left(\frac{\tilde{\varepsilon}^2}{4^n}\left(1-\frac{N\cdot\min\{m/(rc_1),\ rn\}}{2^n c_2}\right)\right) \qquad (c_1=\Theta(2^n/\tilde{\varepsilon}^2))$$

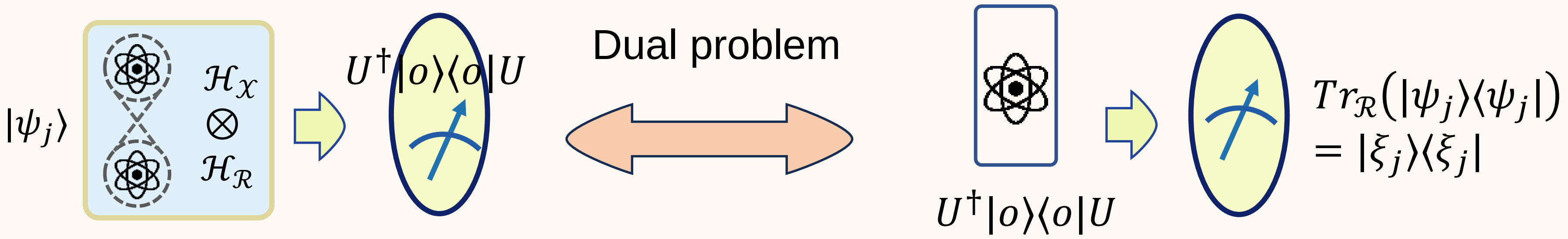
The implications from this lower bound :

For query complexity $\boldsymbol{mN}$: The lower bound of query complexity for achieving sufficiently small prediction error is $\Omega(4^n r/\tilde{\varepsilon}^2)$.

When $r=1$, this ***matches the optimal*** lower bound, for *quantum state tomography* with single-copy non-adaptive measurements [6].

learning $U^\dagger|o\rangle\langle o|U$

tomography pure state $U^\dagger|o\rangle\langle o|U)$



Lower bound for pure state:

$\Omega(4^n/\ell\tilde{\varepsilon}^2)$.

ℓ: number of outcomes

[6] Lowe, Angus, et al. "Lower bounds for learning quantum states with single-copy measurements." ArXiv:2207.14438 (2022).

Proof ideas: Discretizing the hypothesis space

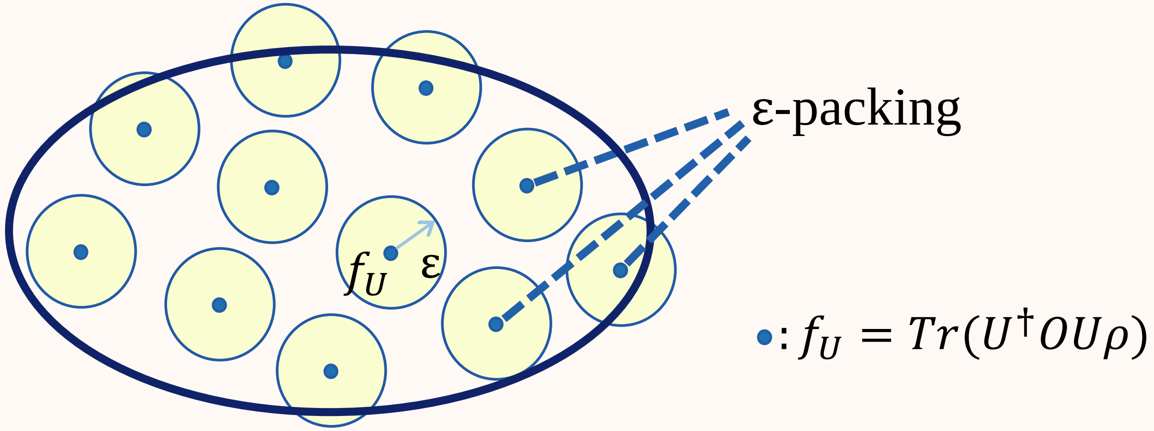
Aim: learning $f_U = Tr(U^\dagger O U \rho)$ from hypothesis set

$$\mathcal{F} = \{f_V(\rho) = Tr(V^\dagger O V \rho) | V \in \mathbb{SU}(d)\}$$

This task is hard when $\mathcal{F}$ contains a large amount of very different operators!

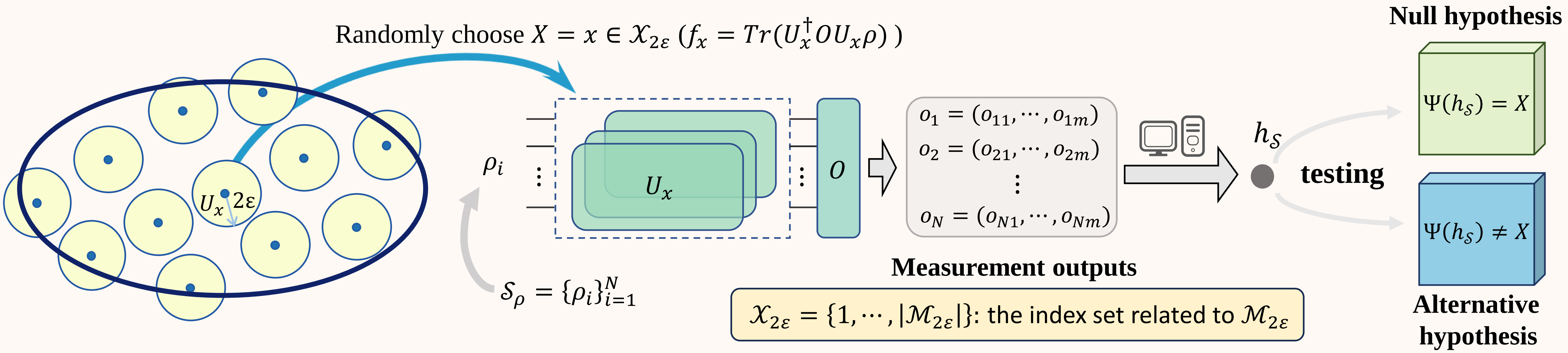
Solution: discretizing the hypothesis set by constructing the ε-packing

Definition (ε-packing): For a given set of functionals $\mathcal{F}$ and a distance metric ϱ on this set, the ε-packing $\mathcal{M}_\varepsilon(\mathcal{F}, \varrho)$ is a discrete subset of $\mathcal{F}$ whose elements are guaranteed to be distant from each other by a distance greater than or equal 2ε. Namely, for any element $f_1, f_2 \in \mathcal{M}_\varepsilon(\mathcal{F}, \varrho)$, the distance between f_1 and f_2 satisfies $\varrho(f_1, f_2) \geq 2\varepsilon$.



The points in the ε-packing are well distinguished!

Proof ideas: Information theoretical bound



Alice X (U_X) —encoding→ $\mathcal{S}$ —decoding→ $\hat{X} = \Psi(h_\mathcal{S})$ Bob

$$\mathbb{E}_U \mathbb{E}_\mathcal{S} R_U(V_\mathcal{S}) \geq \varepsilon^2 \left(1 - \frac{I(X;\hat{X}) + \log 2}{\log(|\mathcal{X}_{2\varepsilon}|)}\right)$$

upper bounding the mutual information $I(X;\hat{X})$

lower bounding the cardinality of 2ε-packing $\mathcal{X}_{2\varepsilon}$

(independent with r, m, N)

reduce to

Proof ideas: Bounding the mutual information $I(X;\hat{X})$

Lemma 3 (Upper bound of the mutual information $I(X;\hat{X})$). The average of mutual information over the training states $\{\rho_j\}_{j=1}^N$ yields

$$\mathbb{E}_{\rho_1,\cdots,\rho_N} I\left(X;\hat{X}\right) \leq N \cdot \min\left\{\frac{4m\tilde{\varepsilon}^2}{rd}, r\log(d)\right\}.$$

Intuitive understanding about the term $\min\left\{\frac{4m\tilde{\varepsilon}^2}{rd}, r\log(d)\right\}$ through Markov chain $X \to (U_X \otimes \mathbb{I})|\psi_j\rangle \to o_j \to \hat{X}$ $(N=1)$:

- $I\left(X;\hat{X}\right) \leq I\left(X;o_j\right) \leq \frac{4m\tilde{\varepsilon}^2}{rd}$:
  - Increasing the number of measurements enabling more information extraction
  - A large r decreases the information 'density' of the output states, and hence decreases the extracted information amount by single measurement.

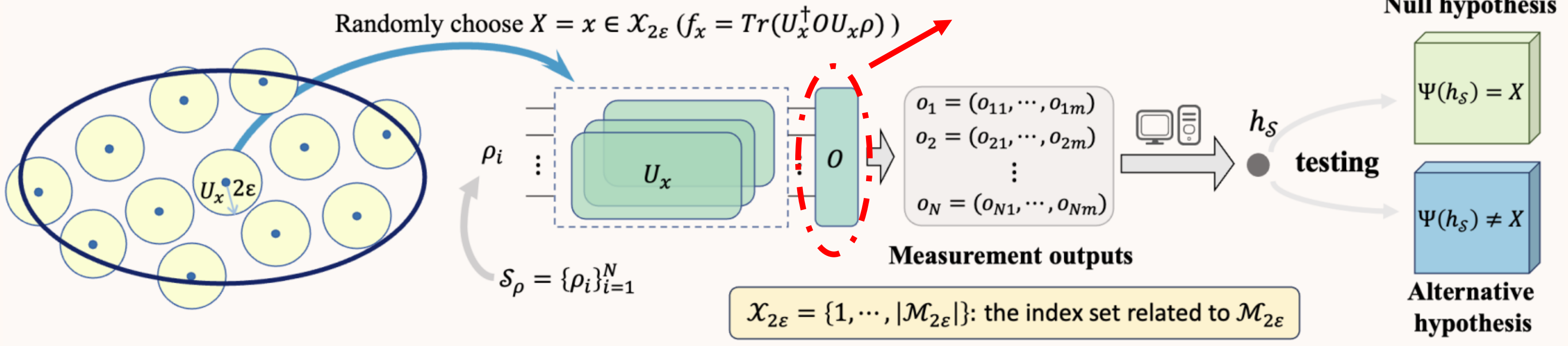
- $I\left(X;\hat{X}\right) \leq I\left(X;(U_X \otimes \mathbb{I})|\psi_j\rangle\right) \leq r\log(d)$: The information of the target unitary U contained in a single output state is limited. Meanwhile, a highly entangled output state contains more information about U than a lowly entangled output state.

The lower bound for POVM with ℓ-outcome

Theorem (Lower bound of $\mathbb{E}_U\mathbb{E}_\mathcal{S}R_U(V_\mathcal{S})$ ***).*** Let $\{f_{U_x}\}_{x\in\mathcal{X}_{2\varepsilon}}$ be a 2ε-packing of the function class $\mathcal{F}$ in the ϱ-metric. Denoting $\tilde{\boldsymbol{\varepsilon}} = \mathbf{4}\sqrt{\mathbf{2}\boldsymbol{d}(\boldsymbol{d}+\mathbf{1})}\boldsymbol{\varepsilon}$, the averaged risk function $\mathbb{E}_U\mathbb{E}_\mathcal{S}R_U(V_\mathcal{S})$ is lower bounded by

$$\mathbb{E}_U\mathbb{E}_\mathcal{S}R_U(V_\mathcal{S}) \geq \frac{\tilde{\varepsilon}^2}{8d(d+1)}\left(1-\frac{\min\{c_1 m\tilde{\varepsilon}^2/r(d+1)\ ,r\log(d)\}+\log(2)}{\log(|\mathcal{X}_{2\varepsilon}|)}\right)$$

POVM with ℓ-outcome



$$\mathbb{E}_U\mathbb{E}_\mathcal{S}R_U(V_\mathcal{S}) \geq \frac{\tilde{\varepsilon}^2}{8d(d+1)}\left(1-\frac{\min\{c_1 m\tilde{\varepsilon}^2/r, c_1 m\tilde{\varepsilon}^2\ell/r(d+1), r\log(d)\}+\log(2)}{\log(|\mathcal{X}_{2\varepsilon}|)}\right)$$

Increasing the number outcomes of POVM can ***exponentially reduce*** the number of measurements, but ***can not remove*** the effect of entangled data.

Numerical simulation: task description

Construction of target unitary set: $\mathcal{U} = \{U \in \mathbb{SU}(d) | U_{1j} = e^{i\gamma_j}, \gamma_j \in \mathbb{R}, j \in [d]\}$

Observable: $O = (|0\rangle\langle 0|)^{\otimes n}$

The substantial set of target operators: $\mathcal{U}_O = \{U \in \mathbb{SU}(d) | U^\dagger O U = |e_j\rangle\langle e_j| : j \in [d]\}$

Let U^* be the target unitary, learning $U^{*\dagger} O U^* = |e_{k^*}\rangle\langle e_{k^*}|$ ***is equivalent to*** identifying the unknown index $k^* \in [d]$.

Construction of entangled data: $|\psi_j\rangle = \sum_{k=1}^{r} \sqrt{c_{jk}} |\xi_{jk}\rangle_{\mathcal{X}} |\varsigma_{jk}\rangle_{\mathcal{R}}$ where $\sum_{k=1}^{r} c_{jk} = 1$

Observable O acts on the subsystem $\mathcal{X}$ ⇨ consider $\sigma_j = Tr_{\mathcal{R}}(|\psi_j\rangle\langle\psi_j|) = \sum_{k=1}^{r} c_{jk} |\xi_{jk}\rangle\langle\xi_{jk}|$

Mixed states set: $\widetilde{\mathcal{S}} = \left\{ \sigma = \sum_{k=1}^{r} c_k |\boldsymbol{e}_{\pi(k)}\rangle\langle\boldsymbol{e}_{\pi(k)}| : \pi \in S_d, |c\rangle = (\sqrt{c_1}, \cdots, \sqrt{c_r})^\top \in \mathbb{SU}(r), |\boldsymbol{e}_{\pi(k)}\rangle \in \mathcal{H}_{\mathcal{X}} \right\}$ ⇨ $U^* \sigma_j U^{*\dagger}$ ⇨ $o_j = \sum_{k=1}^{m} \frac{o_{jk}}{m}$

Collect the measurement outputs $\left\{\left(o_1^{(k)}, \cdots, o_N^{(k)}\right)\right\}_{k=1}^{d}$ over all possible index $k \in [d]$.

$\hat{k}$ is determined by minimizing:

$$\hat{k} = \arg\min_{k \in [d]} \sum_{j=1}^{N} \left(\boldsymbol{o}_j^{(k)} - \boldsymbol{o}_j\right)^2$$

Numerical simulation: task description

- Conditions for correctly Identifying k^* (Training mixed states $\sigma_j = \sum_{k=1}^{r} c_{jk}|\xi_{jk}\rangle\langle\xi_{jk}|$)
  - The states set $\{|\xi_{jk}\rangle\langle\xi_{jk}|\}_{j,k=1}^{N,r}$ contains the target operator $U^{*\dagger}OU^* = |e_k^*\rangle\langle e_k^*|$
  - The measurement outputs $\{o_j\}_{j=1}^{N}$ closely approximate the corresponding Schmidt coefficient c_k^* of the operator $U^{*\dagger}OU^* = |e_k^*\rangle\langle e_k^*| \in \{|\xi_{jk}\rangle\langle\xi_{jk}|\}_{j,k=1}^{N,r}$
- Two extreme cases of $r = 1$ and $r = d$ when $N = 1$:
  - $r = 1$, $N = 1$ $(c_{11} = 1)$:

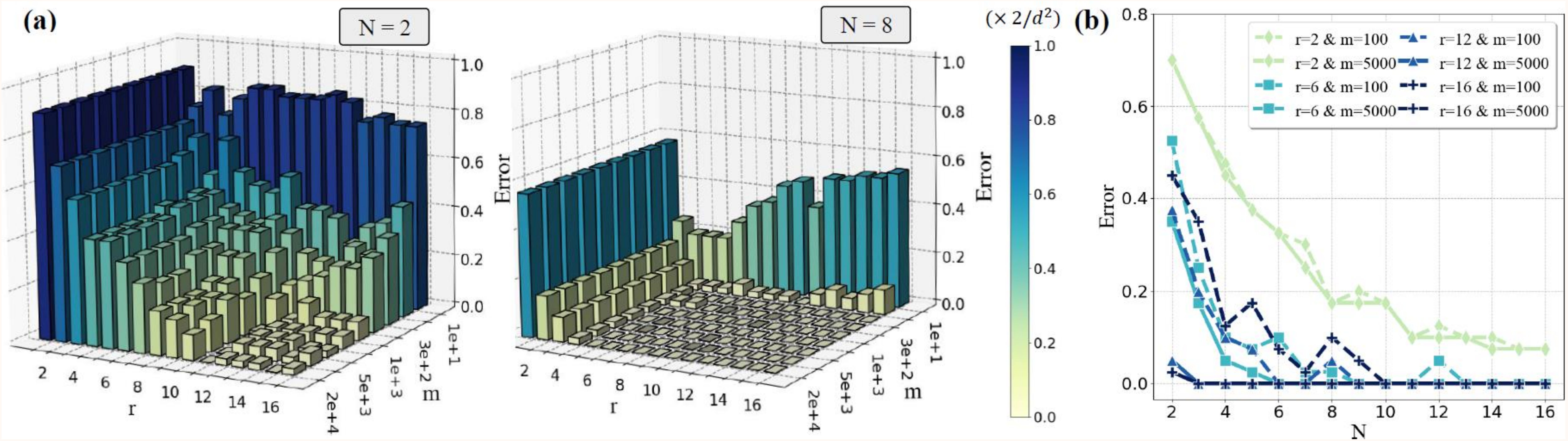
    $|\xi_{11}\rangle\langle\xi_{11}| \neq |e_k^*\rangle\langle e_k^*|$: the output o_1 is always 0

    $|\xi_{11}\rangle\langle\xi_{11}| = |e_k^*\rangle\langle e_k^*|$: few number of measurements can identify the target index k^*.
  - $r = 16$, $N = 1$ $(\mathbb{E}c_{jk} = 1/d)$:

    $|e_k^*\rangle\langle e_k^*| \in \{|\xi_{jk}\rangle\langle\xi_{jk}|\}_{j,k=1}^{N,r}$: the output o_1 is always $nonzero$, but a large number of measurements is required to identify the target index k^*.

Numerical results

Learning a 4-qubit unitary U



Simulation results with independent training states.

Questions & Answers!

Xinbiao Wang

Email: wangxb08@whu.edu.cn