JDT 京东科技

The Capabilities and Limitations of Quantum Learning Models

Yuxuan Du

2021.04.10

Contents

JDT 京东科技

I. The power of QNN in the view of optimization and learning theory *[On the learnability of quantum neural networks, arXiv:2007.12369]*

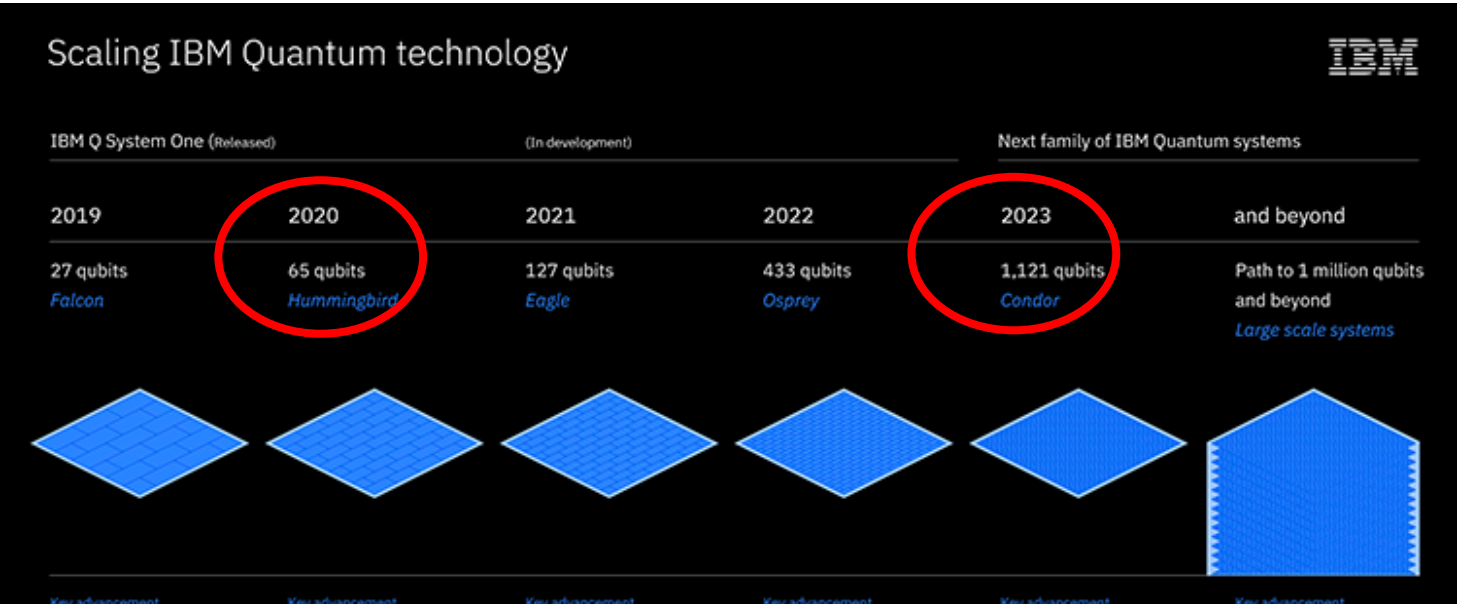
II. QAS: An efficient scheme to enhance the trainability of QNN and supress its error *[Quantum circuit architecture search: error mitigation and trainability enhancement for variational quantum solvers, arXiv:2010.10217]*

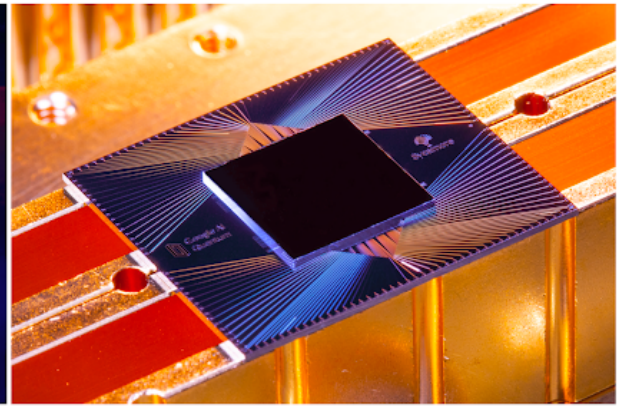
III. The power of quantum kernels in the NISQ era *[Towards understanding the power of quantum kernels in the NISQ era, arXiv:2103.16774]*

Noisy intermediate-scale quantum (NISQ) era

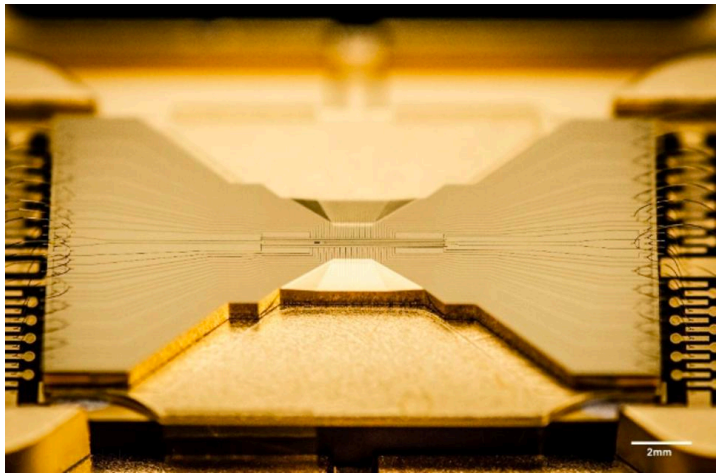
JDT 京东科技

Super-conducting

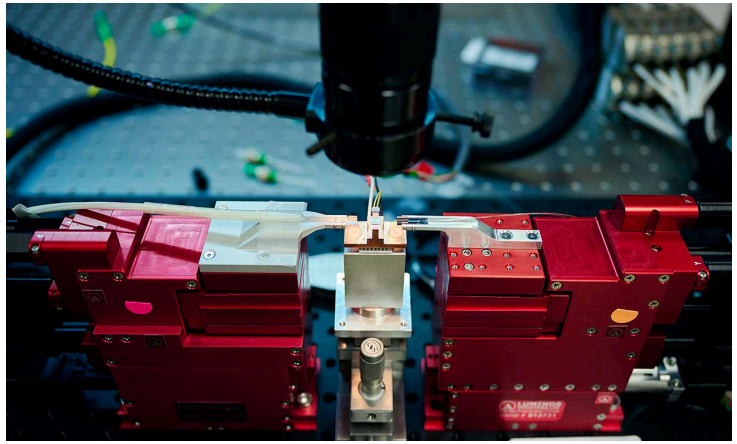




Trap-ion



Photonic chips



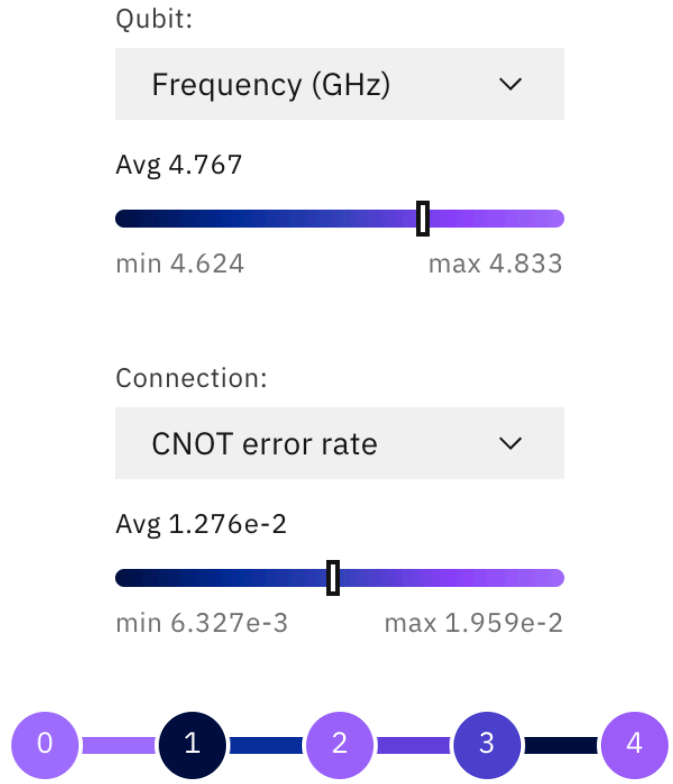
https://www.ibm.com/blogs/research/2020/09/ibm-quantum-roadmap/, https://www.forbes.com/sites/ibm/2020/01/16/the-quantum-computing-era-is-here-why-it-mattersand-how-it-may-change-our-world/, https://ai.googleblog.com/2019/10/quantum-supremacy-using-programmable.html, https://www.forbes.com/sites/moorinsights/2020/10/07/ionq-releases-a-new-32-qubit-trapped-ion-quantum-computer-with-massive-quantum-volume-claims, https://spectrum.ieee.org/tech-talk/computing/hardware/photonic-quantum

Noisy intermediate-scale quantum (NISQ) era

JDT 京东科技

The caveats of NISQ processors are:

- a limited number of qubits;
- shallow circuit depth;
- connectivity restriction;
- unavoidable system noise.

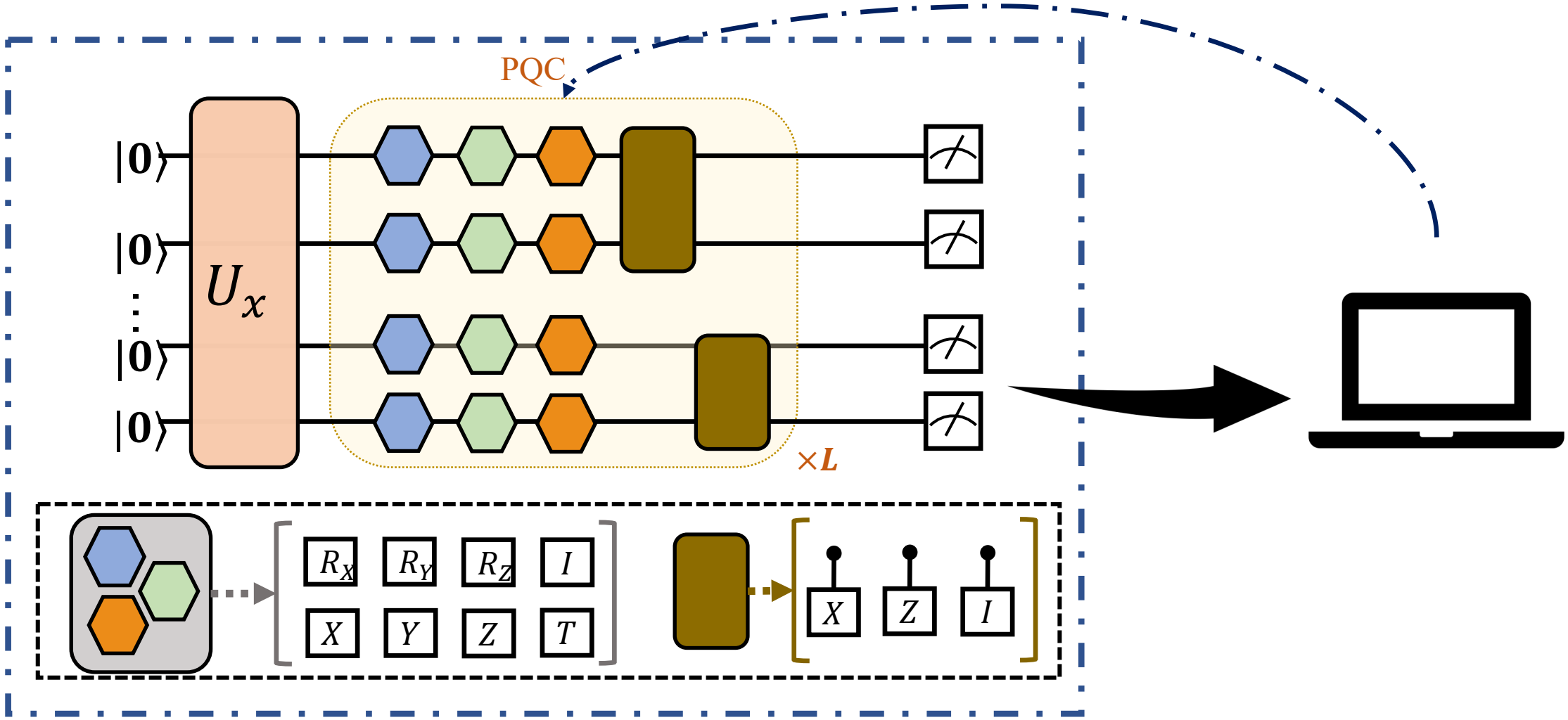


https://quantum-computing.ibm.com/

Noisy intermediate-scale quantum (NISQ) era

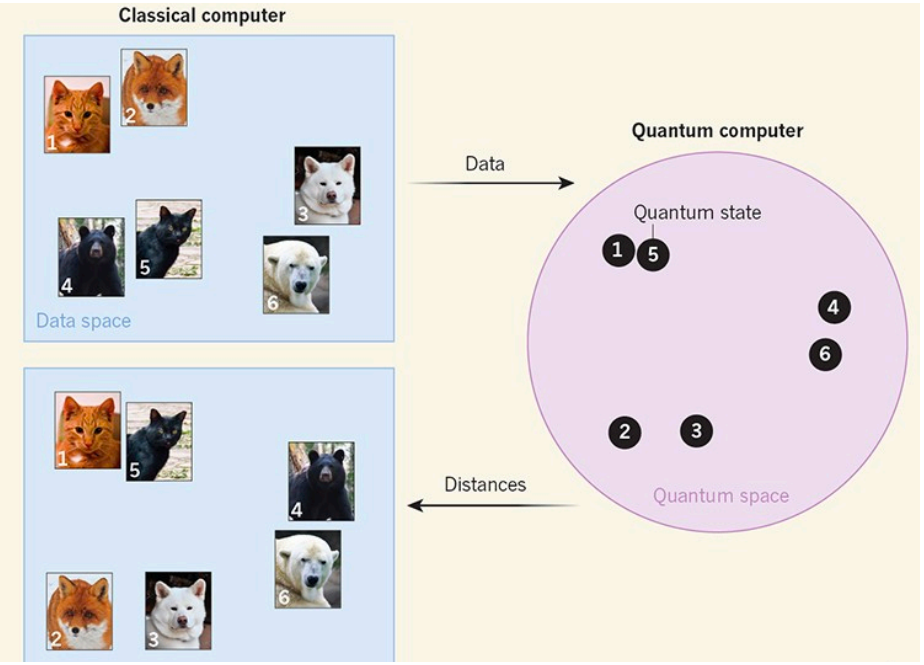
JDT 京东科技

Variational quantum algorithm



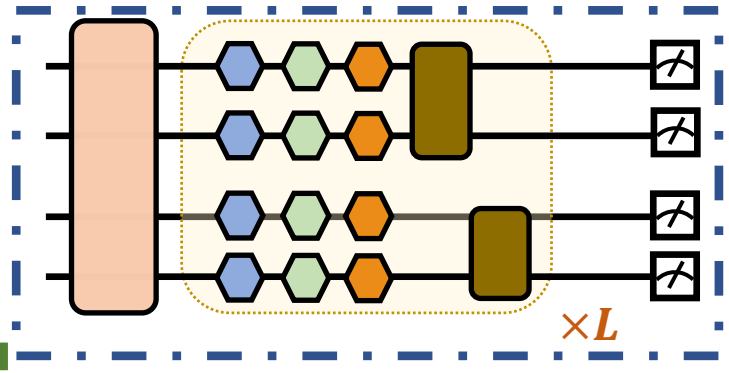
Noisy intermediate-scale quantum (NISQ) applications

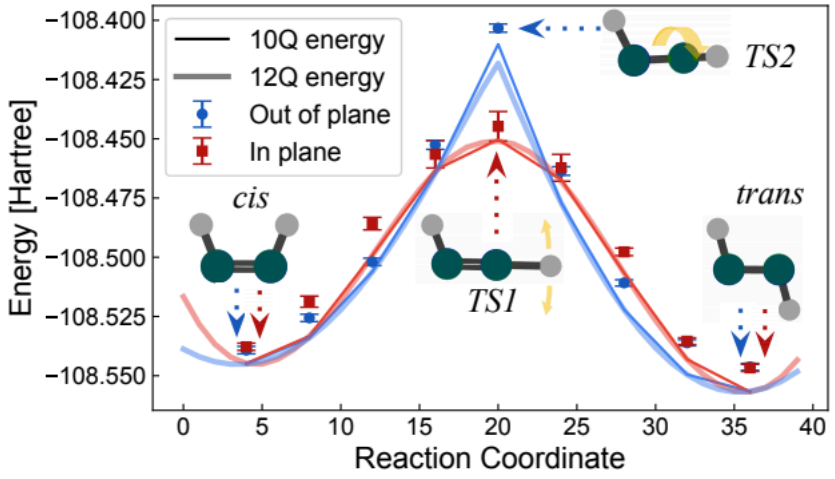
JDT 京东科技



Machine learning

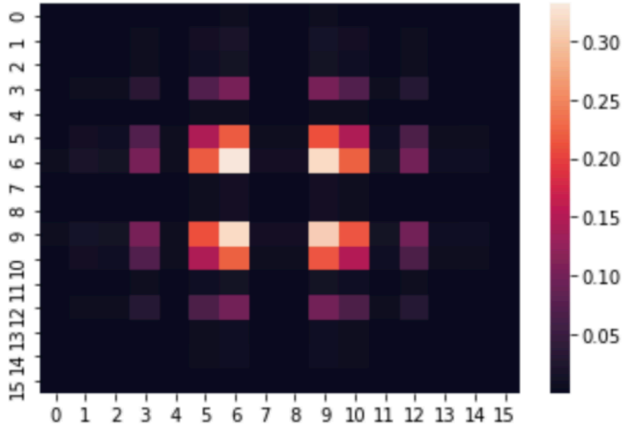
Variational quantum algorithm





Chemistry

Quantum information processing,
Quantum metrology, …



https://science.sciencemag.org/content/369, https://www.nature.com/articles/d41586-019-00771-0, https://arxiv.org/pdf/2003.02989.pdf

Problem setup

JDT 京东科技

ERM: The empirical risk minimization (ERM) principle is a learning paradigm that has been broadly employed to benchmark the performance of supervised learning algorithms.

When QNN is employed to tackle ERM, the optimization yields:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) := \frac{1}{n}\sum_{j=1}^{n} l\left(y_j, \hat{y}_j\right) + r(\boldsymbol{\theta}).$$

$\boldsymbol{z} = \{\boldsymbol{z}_j\}_{j=1}^{N} \in \mathcal{Z}$: Input dataset with N samples;

$\boldsymbol{z}_j = (\boldsymbol{x}_j, y_j)$: The j-th sample, the feature vector as $\boldsymbol{x}_j \in \mathbb{R}^{D_c}$, the label as $\boldsymbol{y}_j \in \mathbb{R}^1$,

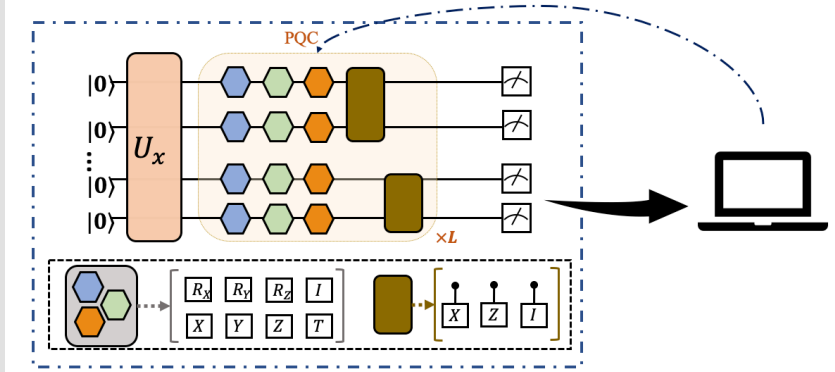
$\hat{y}_j = \mathrm{Tr}\left((\boldsymbol{\Pi}|\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|\right)$: The predicted label of the j-th sample;

$|\psi(\boldsymbol{\theta})\rangle$: quantum states prepared by $U(\boldsymbol{\theta})$ and $\boldsymbol{z}_j$;

$U(\boldsymbol{\theta})$: The trainable quantum circuit as.

$l(\cdot,\cdot)$: the loss function measures the difference between y_j and $\hat{y}_j$;

$r(\boldsymbol{\theta})$: the regularizer.



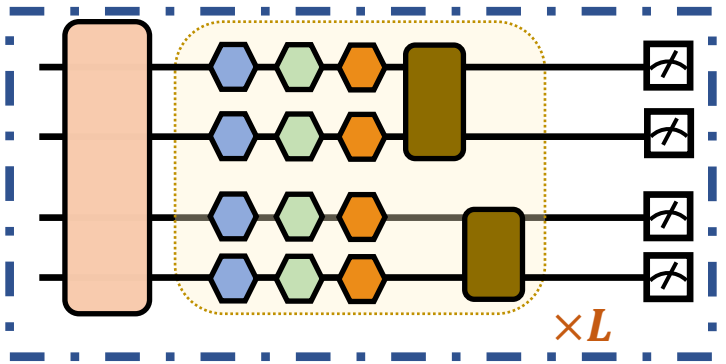
Challenges of QNN

JDT 京东科技

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) \coloneqq \frac{1}{n}\sum_{j=1}^{n} l\left(y_j, \hat{y}_j\right) + r(\boldsymbol{\theta}).$$

The difficulties to theoretically understand QNN are:

- the versatile structures;
- the non-convex optimization landscapes;
- the unavoidable gate noise and measurement errors.



Part I. Understanding the power of QNN

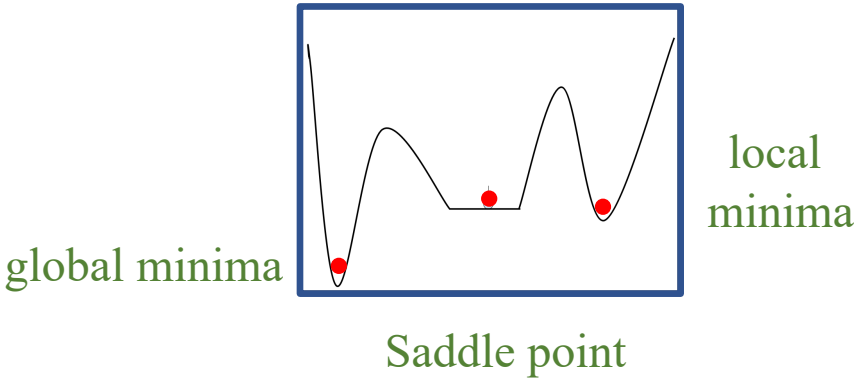
Challenges of QNN

JDT 京东科技

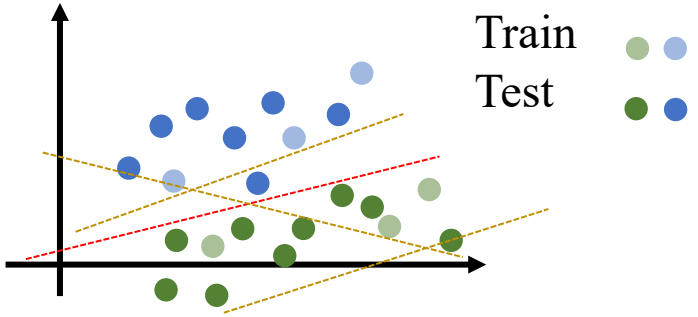
Key issue: what is the learnability of QNN, where

$$learnability = \text{`}trainability\text{'} + \text{`}generalization\text{'}.$$

- *Trainability* ←→ stationary point;



- *Generalization* ←→ find a good hypothesis in a poly sample complexity.



Allen-Zhu, Zeyuan, Yuanzhi Li, and Yingyu Liang. "Learning and generalization in overparameterized neural networks, going beyond two layers." Advances in neural information processing systems. 2019. https://machinelearningmastery.com/why-training-a-neural-network-is-hard/

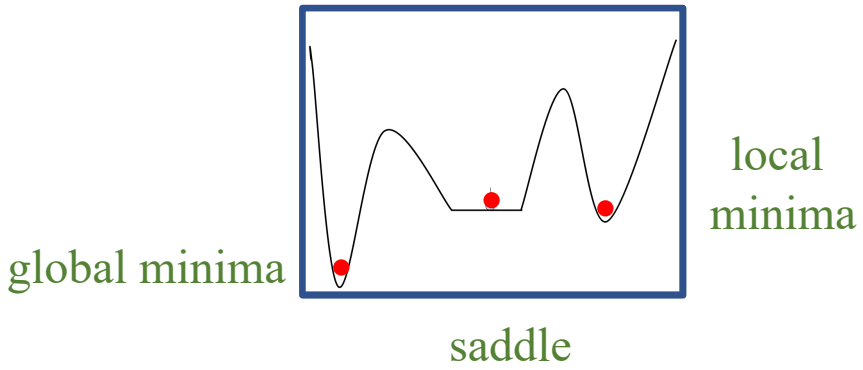
Why we need to understand the learnability of QNN

JDT 京东科技

Trainability:

- ❖ Evaluation of the performance of various QNN-based learning algorithms;

- ❖ Guide to design better quantum learning protocols to avoid barren plateaus.



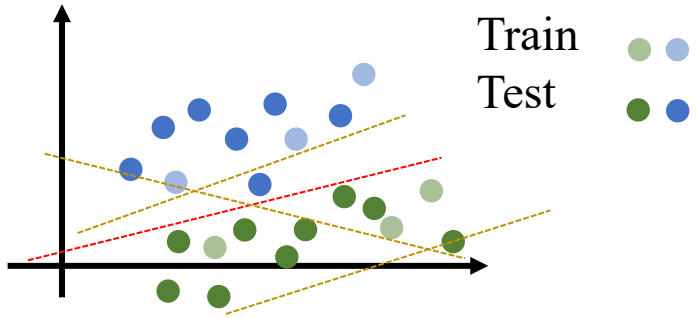
Why we need to understand the learnability of QNN

JDT 京东科技

Generalization:

- Answer whether there exists any class of concepts that can be efficiently learned by (noisy) QNN but are computationally hard for the classical learning models.

- Use QNN implemented on NISQ devices to accomplish certain tasks with theoretical advantages.



Problem setup of the trainability of QNN towards ERM

JDT 京东科技

QNN towards ERM:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{C}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) := \frac{1}{n}\sum_{j=1}^{n} l\left(y_j, \hat{y}_j\right) + r(\boldsymbol{\theta}).$$

Setting: the mean square error loss with $l\left(y_j, \hat{y}_j\right) = (y_j - \hat{y}_j)^2$;

the L2-norm regularizer $r(\boldsymbol{\theta}) = \frac{\lambda||\boldsymbol{\theta}||^2}{2}$,

depolarization noise.

Our analysis can be easily generalized to other loss functions, regularizer, noise model.

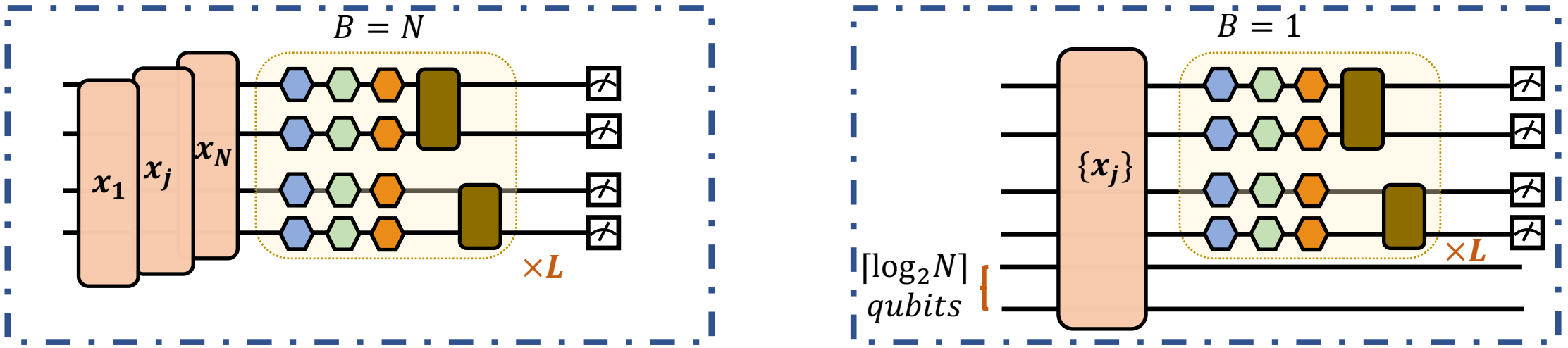
Problem setup of the trainability of QNN towards ERM

JDT 京东科技

The optimization rule is the batch gradient descent method. At the t-th iteration, the parameters are updated to

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \frac{\eta}{B}\sum_{i=1}^{B} \nabla \mathcal{L}(\boldsymbol{\theta}, B_i)\,,$$

where η is the learning rate and $\nabla \mathcal{L}(\boldsymbol{\theta}, B_i)$ is obtained by *the parameter shift rule*.



Problem setup of the trainability of QNN towards ERM

JDT 京东科技

The trainability (convergence performance) of QNN is measured by the following two standard utility metrics:

$$R_1\left(\boldsymbol{\theta}^{(T)}\right) := \mathbb{E}\left[\left\|\nabla \mathcal{L}(\boldsymbol{\theta}^{(T)})\right\|^2\right],$$

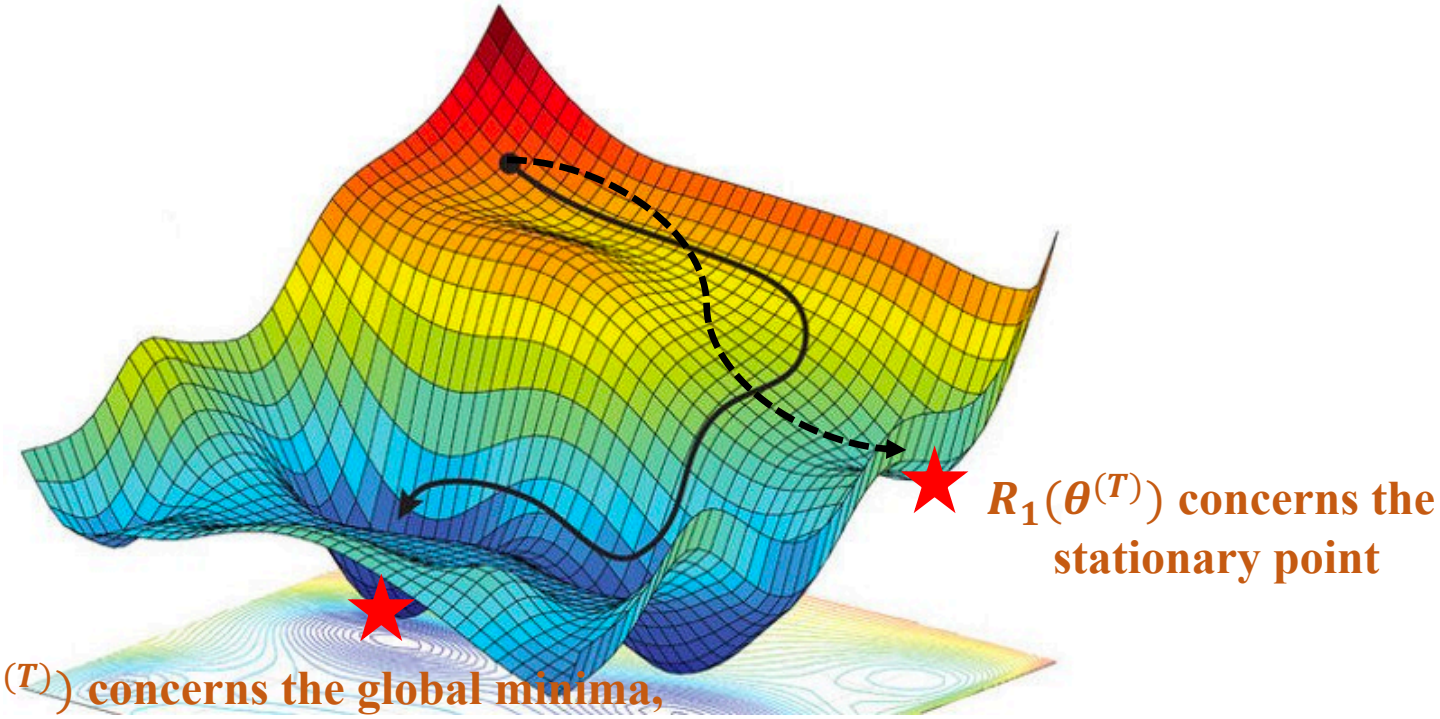
$$R_2\left(\boldsymbol{\theta}^{(T)}\right) := \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}^{(T)})] - \mathcal{L}(\boldsymbol{\theta}^*),$$

where the expectation is taken over the randomness of QNN resulted from the measurement error and gate noise.

Problem setup of the trainability of QNN towards ERM

JDT 京东科技

- $R_1(\boldsymbol{\theta}^{(T)})$ evaluates how far QNN is away from the stationary point, i.e., $\left|\left|\nabla\mathcal{L}\left(\boldsymbol{\theta}^{(T)}, \boldsymbol{z}\right)\right|\right| = 0$.
- $R_2(\boldsymbol{\theta}^{(T)})$ evaluates how far QNN is away from the optimal result $\mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{z})$.



https://medium.com/swlh/non-convex-optimization-in-deep-learning-26fa30a2b2b3

Main results of the trainability of QNN towards ERM

JDT 京东科技

Theorem 1. Let K be the number of measurements, L be the circuit depth, p be the depolarization noise of quantum gates, and B be the batch size. QNN outputs $\boldsymbol{\theta}^{(T)} \in \mathbb{R}^{\boldsymbol{d}}$ after training T iterations with the utility bound

$$R_1\left(\boldsymbol{\theta}^{(T)}\right) \leq \tilde{O}\left(poly\left(\frac{d}{TBK(1-p)^L}\right)\right).$$

When λ satisfies a technical assumption, QNN outputs $\boldsymbol{\theta}^{(T)} \in \mathbb{R}^{\boldsymbol{d}}$ after $T = \tilde{O}(d^3/(1-p)^L)$ training iterations with the utility bound

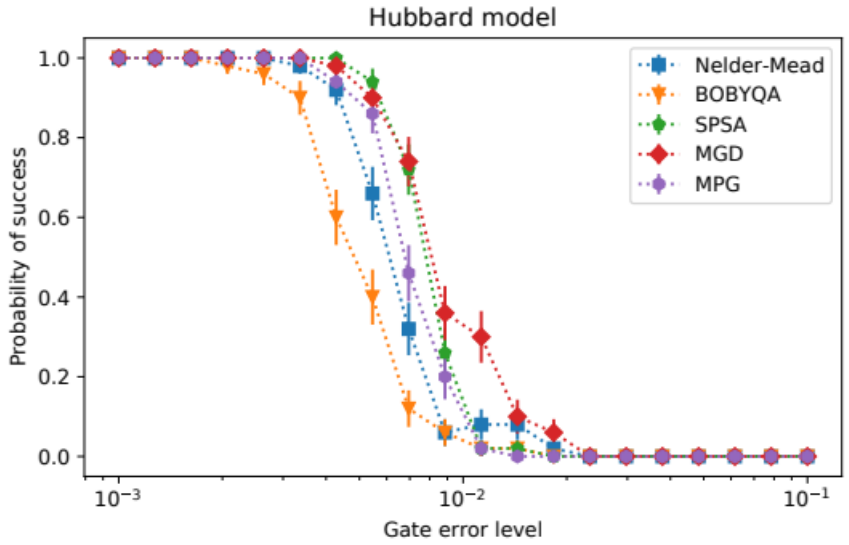
$$R_2\left(\boldsymbol{\theta}^{(T)}\right) \leq \tilde{O}\left(poly\left(\frac{d}{BK^2(1-p)^L}\right)\right).$$

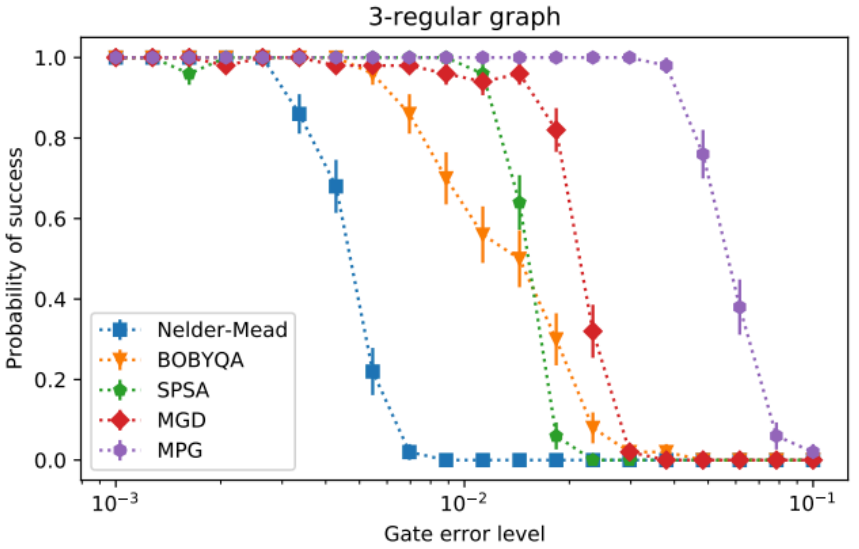
Main results of the trainability of QNN towards ERM

JDT 京东科技

Implications of Theorem 1:

- $K\uparrow$, B $\uparrow$, $p\downarrow$, $d\downarrow$, and L $\downarrow$ yield better utility bounds $R_1\left(\boldsymbol{\theta}^{(T)}\right)$ and $R_2\left(\boldsymbol{\theta}^{(T)}\right)$;





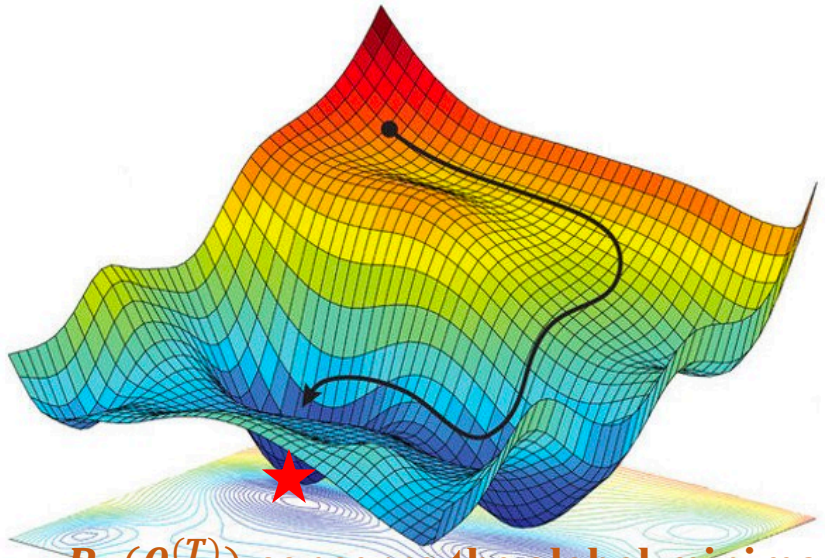
Sung, Kevin J., et al. "Using models to improve optimizers for variational quantum algorithms." *Quantum Science and Technology* 5.4 (2020): 044008.

Main results of the trainability of QNN towards ERM

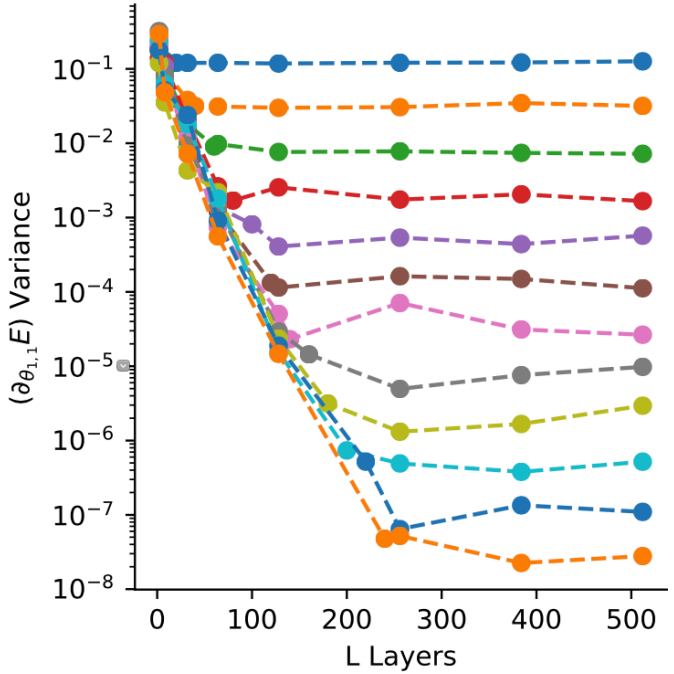
JDT 京东科技

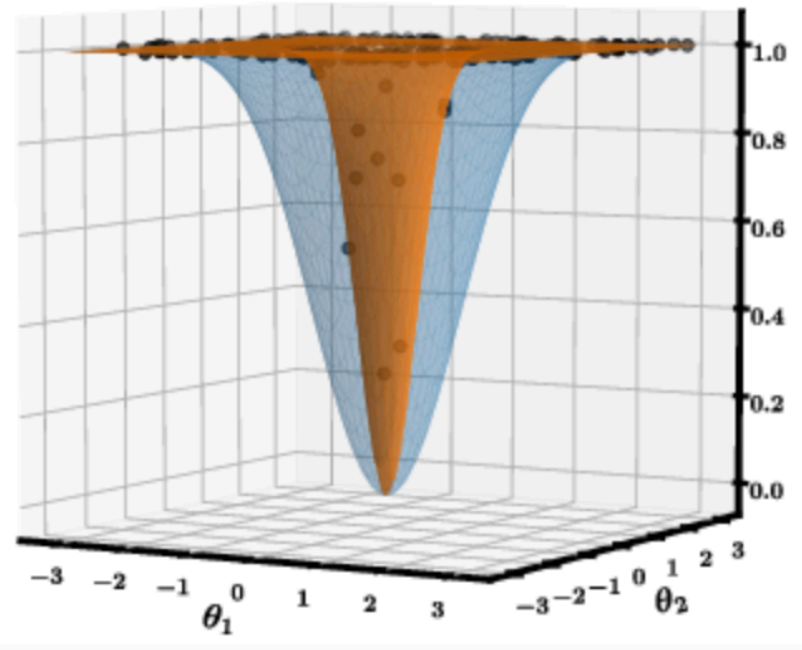
Implications of Theorem 1:

- $K\uparrow$, B $\uparrow$, $p\downarrow$, $d\downarrow$, and L $\downarrow$ yield better utility bounds $R_1\left(\boldsymbol{\theta}^{(T)}\right)$ and $R_2\left(\boldsymbol{\theta}^{(T)}\right)$;
- The convergence towards the global optima as shown in $R_2\left(\boldsymbol{\theta}^{(T)}\right)$ provides an insight about how to employ regularization techniques to avoid barren plateaus.



$R_2(\theta^{(T)})$ concerns the global minima, the target result of QNN





Sung, Kevin J., et al. "Using models to improve optimizers for variational quantum algorithms." *Quantum Science and Technology* 5.4 (2020): 044008.

Main results of the trainability of QNN towards ERM

JDT 京东科技

Implications of Theorem 1:

- $K\uparrow$, B $\uparrow$, $p\downarrow$, $d\downarrow$, and L $\downarrow$ yield better utility bounds $R_1\left(\boldsymbol{\theta}^{(T)}\right)$ and $R_2\left(\boldsymbol{\theta}^{(T)}\right)$;
- The convergence towards the global optima as shown in $R_2\left(\boldsymbol{\theta}^{(T)}\right)$ provides an insight about how to employ regularization techniques to avoid barren plateaus.
- The utility bound $R_1\left(\boldsymbol{\theta}^{(T)}\right)$ indicates that the optimization of QNN can diverge for large d and p, no matter how large T or K would become;

$$R_1 \leq \frac{2S(1+90\lambda d)}{T(1-\tilde{p})^2} + \frac{(2\tilde{p}-\tilde{p}^2)(2G+d)(1+10\lambda)^2}{(1-\tilde{p})^2} + \frac{6dK+8d}{(1-\tilde{p})^2 BK^2}\,.$$

Sketch of proof

JDT 京东科技

$$R_1 \leq \tilde{O}\left(poly\left(\frac{d}{T(1-p)^{L_Q}}, \frac{d}{BK(1-p)^{L_Q}}, \frac{d}{(1-p)^{L_Q}}\right)\right) \qquad R_1\left(\boldsymbol{\theta}^{(T)}\right) := \mathbb{E}\left[\left\|\nabla \mathcal{L}(\boldsymbol{\theta}^{(T)})\right\|^2\right]$$

The objective function $\mathcal{L}(\boldsymbol{\theta})$ is S-smooth, i.e.,

$$\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)}) \leq \langle \nabla \mathcal{L}(\boldsymbol{\theta}^{(t)}), \boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)} \rangle + \frac{S}{2}\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|^2 .$$

The optimization rule of **noisy** QNN at the t-th iteration follows

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \nabla \bar{\mathcal{L}}(\boldsymbol{\theta}^{(t)}) .$$

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)}) \\ \leq &-\frac{1}{S}\langle \nabla \mathcal{L}(\boldsymbol{\theta}^{(t+1)}), \nabla \bar{\mathcal{L}}(\boldsymbol{\theta}^{(t)}) \rangle \\ &+ \frac{1}{2S}\|\nabla \bar{\mathcal{L}}(\boldsymbol{\theta}^{(t)})\|^2 \end{aligned}$$

Sketch of proof

JDT 京东科技

$$R_1 \leq \tilde{O}\left(poly\left(\frac{d}{T(1-p)^{L_Q}}, \frac{d}{BK(1-p)^{L_Q}}, \frac{d}{(1-p)^{L_Q}}\right)\right) \qquad R_1\left(\boldsymbol{\theta}^{(T)}\right) := \mathbb{E}\left[\left\|\nabla \mathcal{L}(\boldsymbol{\theta}^{(T)})\right\|^2\right]$$

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)}) \\
\leq &-\frac{1}{S}\langle \nabla \mathcal{L}(\boldsymbol{\theta}^{(t+1)}), \nabla \bar{\mathcal{L}}(\boldsymbol{\theta}^{(t)})\rangle \\
&+\frac{1}{2S}\|\nabla \bar{\mathcal{L}}(\boldsymbol{\theta}^{(t)})\|^2
\end{aligned}$$

Theorem 3. *It follows that*

$$\nabla_j \bar{\mathcal{L}}_i(\boldsymbol{\theta}^{(t)}) = (1-\tilde{p})^2 \nabla_j \mathcal{L}_i(\boldsymbol{\theta}^{(t)}) + C_{j,1}^{(i,t)} + \varsigma_i^{(t,j)},$$

where $\tilde{p} = 1-(1-p)^{L_Q}$, L_Q *is the circuit depth, the constant* $C_{j,1}^{(i,t)}$ *only depends on* Y_i, $\boldsymbol{\theta}^{(t)}$, *and* $\tilde{p}$, *and* $\varsigma_i^{(t,j)}$ *follows the distribution* $\mathcal{P}_Q$ *that is formed by* Y_i, $\boldsymbol{\theta}^{(t)}$, *the number of measurements* K, *and* $\tilde{p}$ *with zero mean.*

$$\mathbb{E}_{\xi_i^{(t)}, \xi_i^{(t,j)}}[\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)})] \leq -\frac{1}{2S}(1-\tilde{p})^2 \|\nabla \mathcal{L}(\boldsymbol{\theta}^{(t)})\|^2 + \frac{2G+d}{2S}(2-\tilde{p})\tilde{p}(1+10\lambda)^2 + \frac{6dK+8d}{2SBK^2}.$$

Sketch of proof

JDT 京东科技

$$R_1 \leq \tilde{O}\left(poly\left(\frac{d}{T(1-p)^{L_Q}}, \frac{d}{BK(1-p)^{L_Q}}, \frac{d}{(1-p)^{L_Q}}\right)\right) \qquad R_1\left(\boldsymbol{\theta}^{(T)}\right) := \mathbb{E}\left[\left\|\nabla \mathcal{L}(\boldsymbol{\theta}^{(T)})\right\|^2\right]$$

$$\mathbb{E}_{\xi_i^{(t)}, \xi_i^{(t,j)}}[\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)})] \leq -\frac{1}{2S}(1-\tilde{p})^2 \|\nabla \mathcal{L}(\boldsymbol{\theta}^{(t)})\|^2 + \frac{2G+d}{2S}(2-\tilde{p})\tilde{p}(1+10\lambda)^2 + \frac{6dK+8d}{2SBK^2}.$$

By induction, with summing over t = 0, ..., T − 1 and taking expectation, we obtain

$$\mathbb{E}_t\left[\|\nabla \mathcal{L}(\boldsymbol{\theta}^{(t)})\|^2\right] \leq \frac{2S(1+90\lambda d)}{T(1-\tilde{p})^2} + \frac{(2\tilde{p}-\tilde{p}^2)(2G+d)(1+10\lambda)^2}{(1-\tilde{p})^2} + \frac{6dK+8d}{(1-\tilde{p})^2 BK^2}.$$

Sketch of proof

JDT 京东科技

$$R_2 \leq (1+90\lambda d)\exp\left(-\frac{\mu(1-\tilde{p})^2 T}{S}\right) + T\frac{(2\tilde{p}-\tilde{p}^2)(G+2d)(1+10\lambda)^2 BK^2 + 6dK + 8d}{2SBK^2}$$

$$R_2\left(\boldsymbol{\theta}^{(T)}\right) := \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}^{(T)})] - \mathcal{L}(\boldsymbol{\theta}^*),$$

PL condition: a function f satisfies PL condition if there exists $\mu > 0$ and for every possible $\boldsymbol{\theta} \in \boldsymbol{\mathcal{C}}$, $\left|\left|\nabla f(\boldsymbol{\theta})\right|\right|^2 \geq 2\mu(f(\boldsymbol{\theta}) - f^*)$, where $f^* = \min_{\boldsymbol{\theta}\in\boldsymbol{\mathcal{C}}} f(\boldsymbol{\theta})$.

Lemma Assume $\lambda \in \left(0, \frac{1}{3\pi}\right) \cup \left(\frac{1}{\pi}, \infty\right)$ the loss $\mathcal{L}$ used in QNN satisfies PL condition with $\mu = \frac{(-1+\lambda\pi)^2}{1+\lambda d(3\pi)^2}$.

Sketch of proof

JDT 京东科技

$$R_2 \leq (1+90\lambda d)\exp\left(-\frac{\mu(1-\tilde{p})^2 T}{S}\right) + T\frac{(2\tilde{p}-\tilde{p}^2)(G+2d)(1+10\lambda)^2 BK^2 + 6dK + 8d}{2SBK^2}$$

$$R_2\left(\boldsymbol{\theta}^{(T)}\right) := \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}^{(T)})] - \mathcal{L}(\boldsymbol{\theta}^*),$$

Lemma Assume $\lambda \in \left(0, \frac{1}{3\pi}\right) \cup \left(\frac{1}{\pi}, \infty\right)$ the loss $\mathcal{L}$ used in QNN satisfies PL condition with $\mu = \frac{(-1+\lambda\pi)^2}{1+\lambda d(3\pi)^2}$.

$$+ \quad \left|\left|\nabla f(\boldsymbol{\theta})\right|\right|^2 \geq 2\mu(f(\boldsymbol{\theta}) - f^*)$$

$$\mathbb{E}_{\xi_i^{(t)},\xi_i^{(t,j)}}[\mathcal{L}(\boldsymbol{\theta}^{(t+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)})] \leq -\frac{1}{2S}(1-\tilde{p})^2\left\|\nabla\mathcal{L}(\boldsymbol{\theta}^{(t)})\right\|^2 + \frac{2G+d}{2S}(2-\tilde{p})\tilde{p}(1+10\lambda)^2 + \frac{6dK+8d}{2SBK^2}.$$

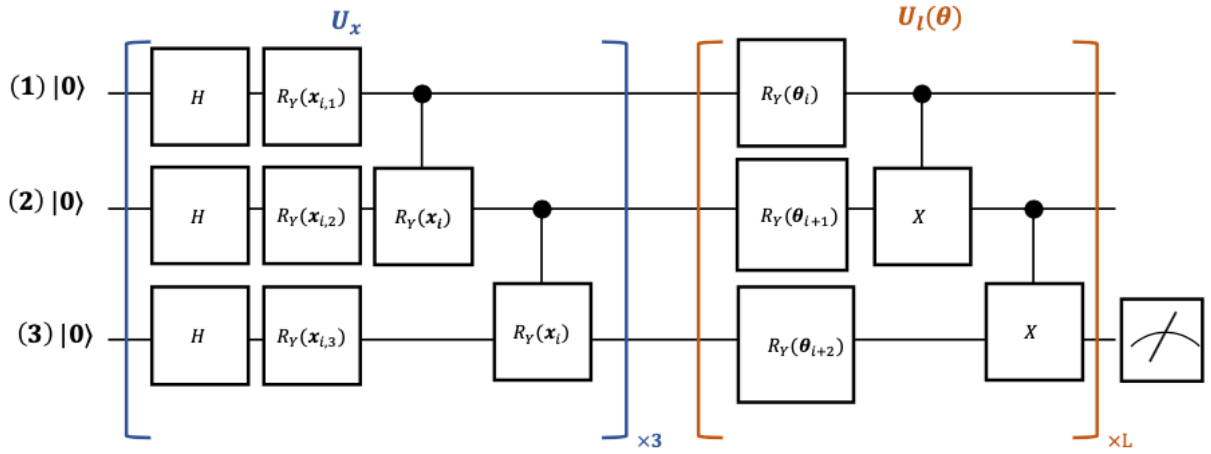
By induction, with summing over t, and taking expectation

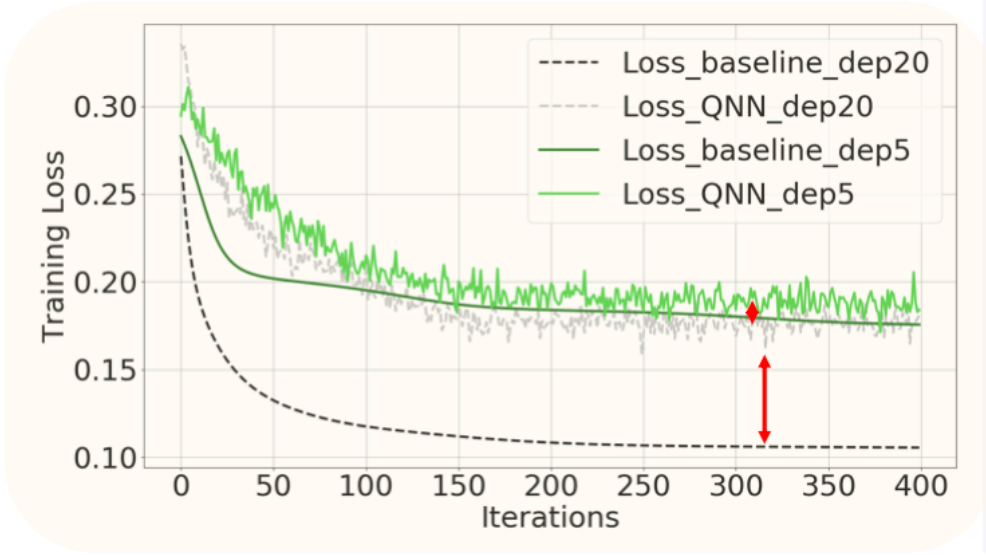
$$\mathbb{E}_{\boldsymbol{\varsigma}^{(t)}}[\mathcal{L}(\boldsymbol{\theta}^{(T)})] - \mathcal{L}^*$$

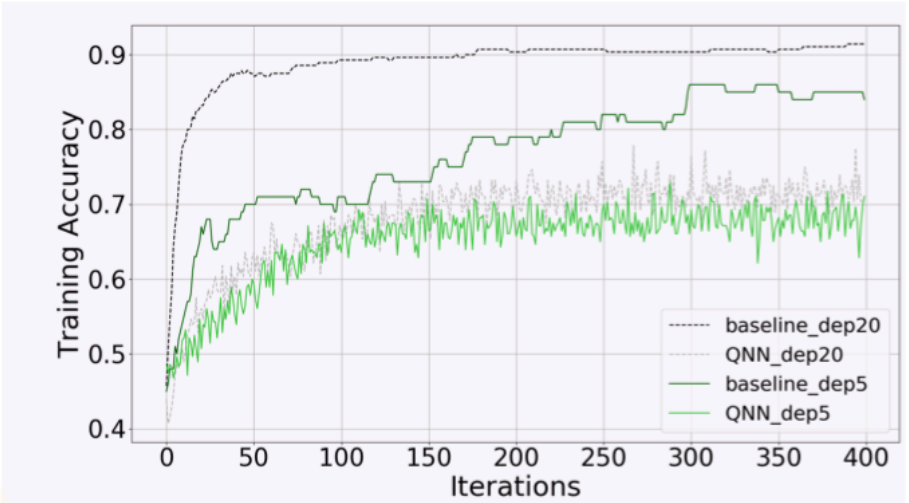
emma Assume $\lambda \in \left(0, \frac{1}{3\pi}\right) \cup \left(\frac{1}{\pi}, \infty\right)$ the loss $\mathcal{L}$ used in QNN satisfies PL condition with $= \frac{(-1+\lambda\pi)^2}{1+\lambda d(3\pi)^2}$.

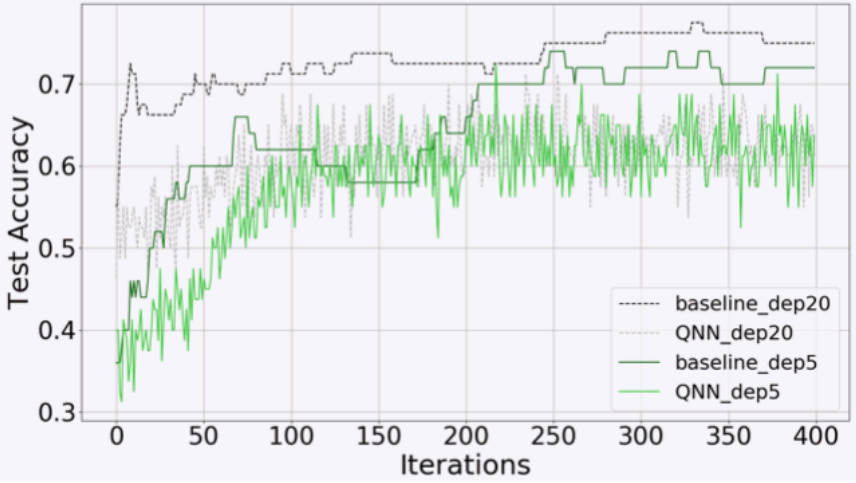
Simulation results

JDT 京东科技







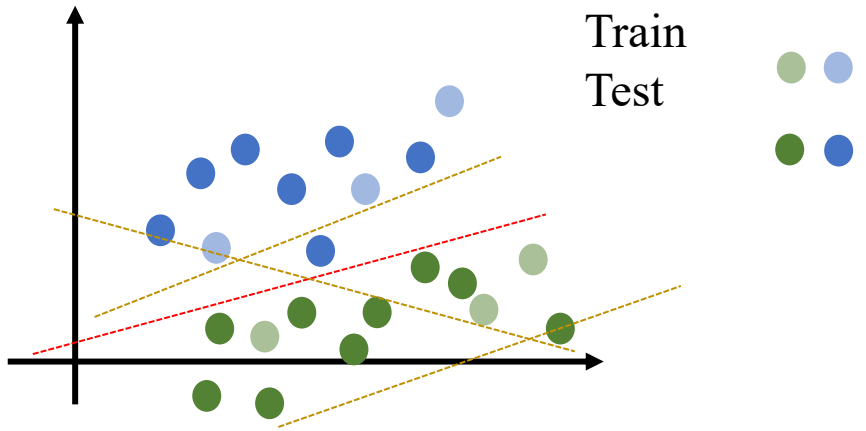


Problem setup of the generalization of QNN

JDT 京东科技

- **Generalization**: for a given learning problem, QNN requires a poly sample complexity to find a good hypothesis, while classical algorithms require an exp number of samples.

(Quantum) Leaning theory



Central tools to explore the generalization of QNN

JDT 京东科技

Quantum statistical query (QSQ) oracle: A QSQ oracle which takes a tolerance parameter τ and an observable $\mathbb{M} \in \mathbb{C}^{2^{N+1}\times 2^{N+1}}$ and returns a number α satisfies

$$|\alpha - \langle\psi_{c^*}|\mathbb{M}|\psi_{c^*}\rangle| \leq \tau.$$

$C \subseteq \{c: \{0,1\}^N \to \{0,1\}\}$: a concept class;

$D: \{0,1\}^N \to [0,1]$: an unknown distribution;

$|\psi_{c^*}\rangle = \sum_{\boldsymbol{x}\in\{\boldsymbol{0},\boldsymbol{1}\}^N} \sqrt{D(\boldsymbol{x})}|\boldsymbol{x}, \boldsymbol{c}^*(\boldsymbol{x})\rangle$: a quantum example.

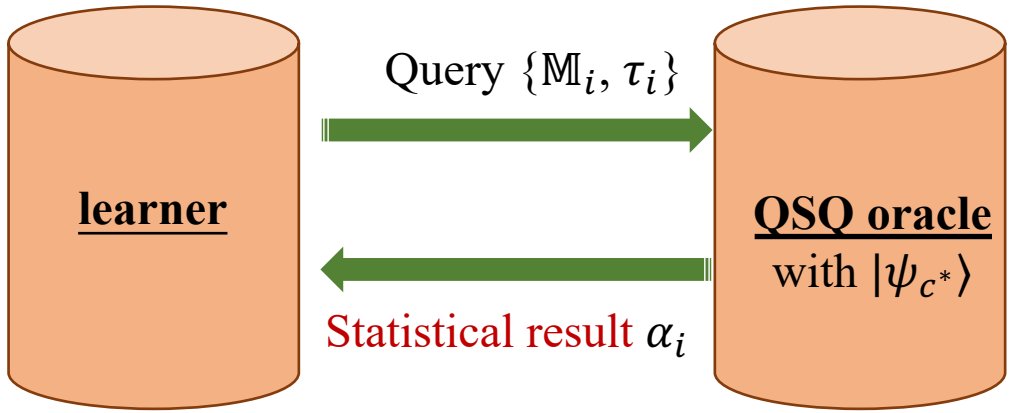
Arunachalam, Srinivasan, Alex B. Grilo, and Henry Yuen. "Quantum statistical query learning." arXiv preprint arXiv:2002.08240 (2020).

Central tools to explore the generalization of QNN

JDT 京东科技

Quantum statistical query (QSQ) learning algorithm: The QSQ learning algorithm adaptively feeds a sequence of $\{\mathbb{M}_i, \tau_i\}$ into a QSQ oracle, and exploits the responses of $\{\alpha_i\}$ to output a hypothesis $h: \{0,1\}^N \rightarrow \{0,1\}$.

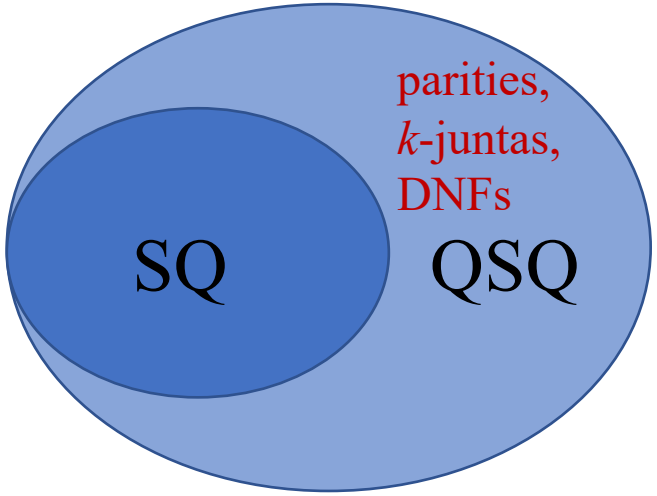
The goal of the learner is to achieve $\Pr_{\boldsymbol{x}\sim\boldsymbol{D}}\left(h(\boldsymbol{x}) = c^*(\boldsymbol{x})\right) \leq \epsilon$ for all possible D and c^*.

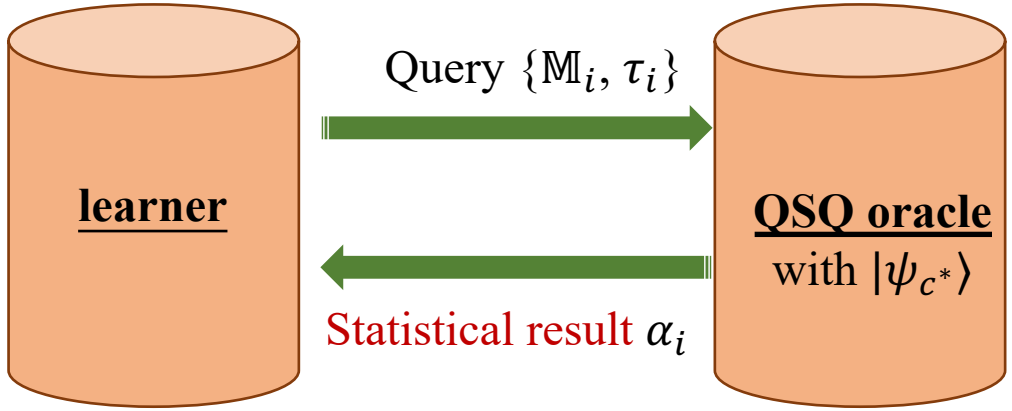


Central tools to explore the generalization of QNN

JDT 京东科技

Lemma 1. Let C be the concept class of parities, k-juntas, or poly(N)-sized DNFs (Disjunctive Normal Forms), then there exists a poly(N)-query QSQ algorithm with tolerance $\tau = \tilde{O}(\epsilon)$ that efficiently learns C under the uniform distribution.





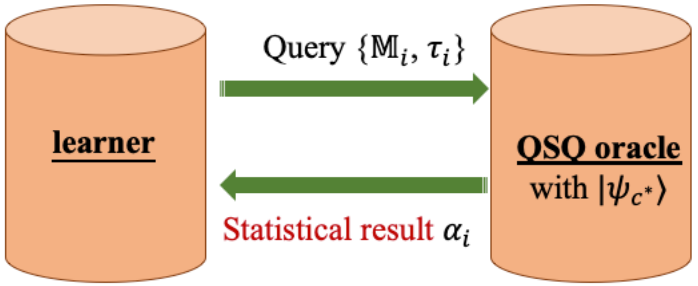
Arunachalam, Srinivasan, Alex B. Grilo, and Henry Yuen. "Quantum statistical query learning." arXiv preprint arXiv:2002.08240 (2020).

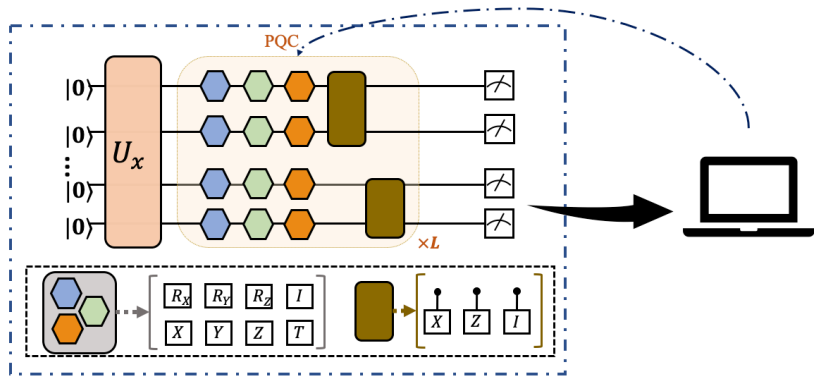
The generalization of QNN

JDT 京东科技

Motivation: the QSQ oracle yields a similar behavior of the variational quantum circuit used in QNN:

1) measurement statistics
2) noisy setting.





Main results of the generalization of QNN

JDT 京东科技

Theorem 2. A QSQ learning algorithm, where the distribution over the quantum example $|\psi_{c^*}\rangle$ is fixed to be uniform and the observable $\mathbb{M}$ can be implemented by at most poly(n) single and two-qubit gates, can be efficiently simulated by noisy QNN using **polynomial** samples.

Main results of the generalization of QNN

JDT 京东科技

Theorem 2. A QSQ learning algorithm, where the distribution over the quantum example $|\psi_{c^*}\rangle$ is fixed to be uniform and the observable $\mathbb{M}$ can be implemented by at most poly(n) single and two-qubit gates, can be efficiently simulated by noisy QNN using **polynomial** samples.

Implications:

- The noisy quantum circuit used in QNN has potential to tackle practical learning tasks, e.g., **support vector machines and linear and convex optimization**, with quantum advantages.

Sketch of Proof

JDT 京东科技

QSQ oracle: A QSQ oracle which takes a tolerance parameter τ and an observable $\mathbb{M} \in \mathbb{C}^{2^{N+1}\times 2^{N+1}}$ and returns a number α satisfies

$$|\alpha - \langle\psi_{c^*}|\mathbb{M}|\psi_{c^*}\rangle| \leq \tau.$$

ν

Let the encoding circuit U_x prepare the state $|\psi_{c^*}\rangle$ and the quantum measurement constructed from $\mathbb{M}$. Under the depolarization noise, the expectation value of quantum measurements of the noisy QNN yields

$$\tilde{\nu} = (1-\tilde{p})\nu + \frac{\tilde{p}\,\mathrm{Tr}(\mathbb{M})}{2^{N+1}}$$

Sketch of Proof

JDT 京东科技

Let the encoding circuit U_x prepare the state $|\psi_{c^*}\rangle$ and the quantum measurement constructed from $\mathbb{M}$. Under the depolarization noise, the expectation value of quantum measurements of the noisy QNN yields

$$\tilde{\nu} = (1-\tilde{p})\nu + \frac{\tilde{p}\,\mathrm{Tr}(\mathbb{M})}{2^{N+1}}$$

+

By the Chernoff-Hoeffding bound for real-valued variables, the relation between the sample mean $\frac{1}{K}\sum_{i=1} V_k$ and the result $\tilde{\nu}$ follows

$$\Pr\left(\left|\frac{1}{K}\sum_{i=1}^{K} V_k - \tilde{\nu}\right| \geq \frac{\delta}{2}\right) \leq 2\exp(-\delta^2 K/2)\ .$$

↓

$$\left|\frac{1}{K}\sum_{k=1}^{K} V_k - \nu\right| \leq \tilde{p}(\nu + \frac{1}{2^{N+1}}) + \frac{\delta}{2} \leq \frac{5}{4}\tilde{p} + \frac{\delta}{2} \leq \tau$$

→

With the successful probability $1-2b$, noisy QNN can efficiently simulate QSQ oracle using $K = 2\ln\left(\frac{2}{b}\right)/\delta^2$ examples.

Outlook

JDT 京东科技

- ✓ More measurements, lower noise, and shallower circuit depth contribute to a better performance of QNN.
- ✓ Theoretical guidance to devise more advanced QNN based learning models that are robust to inevitable gate noise and insensitive to the barren plateau phenomenon.
- ✓ QNN can efficiently learn parity, juntas, and DNF with quantum advantages even with gate noise.

❑ Exploit more advanced quantum models developed in quantum learning theory to explore the potential advantages that can be achieved by QNN.

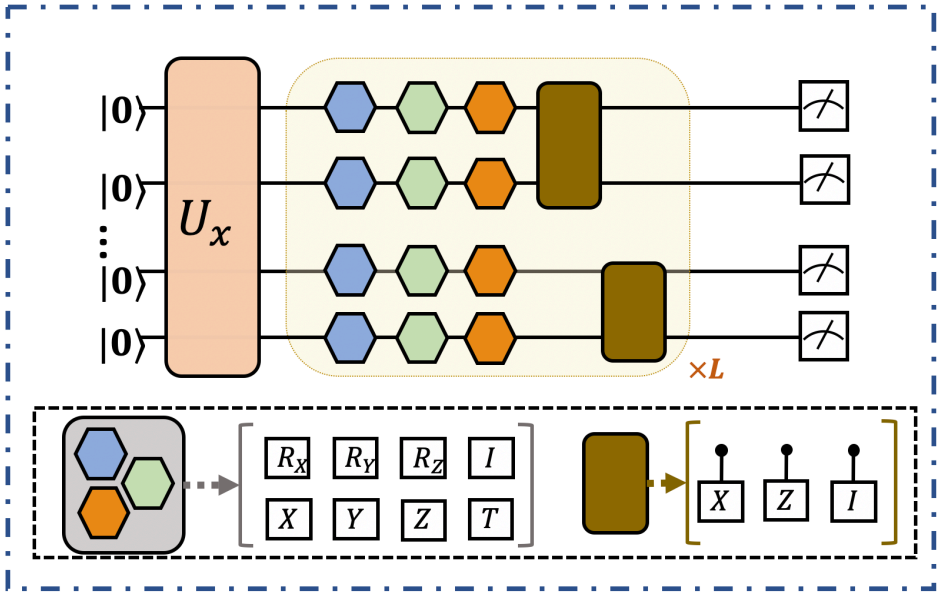
Part II. QAS: An efficient scheme to enhance the trainability of QNN and suppress its error

Noisy intermediate-scale quantum (NISQ) era

JDT 京东科技

The caveats of NISQ processors are:

- limited number of qubits;
- shallow circuit depth;
- connectivity restriction;
- unavoidable system noise.



Problem setup

JDT 京东科技

Quantum neural networks (QNN) aims to find the optimal $\boldsymbol{\theta}^* \in \mathbb{R}^d$ that minimizes an objective function $\mathcal{L}$, i.e.,

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{C}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) \coloneqq \frac{1}{n}\sum_{j=1}^{n} l\left(y_j, \hat{y}_j\right) + r(\boldsymbol{\theta}).$$

$\boldsymbol{z} = \{\boldsymbol{z}_j\}_{j=1}^{N} \in \mathcal{Z}$: Input dataset with N samples;

$\boldsymbol{z}_j = (\boldsymbol{x}_j, y_j)$: The j-th sample, the feature vector as $\boldsymbol{x}_j \in \mathbb{R}^{D_c}$, the label as $\boldsymbol{y}_j \in \mathbb{R}^1$,

$U(\boldsymbol{\theta})$: The trainable quantum circuit as.

$\hat{y}_j = \mathrm{Tr}\left((\boldsymbol{\Pi}|\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|\right)$: The predicted label of the j-th sample;

$|\psi(\boldsymbol{\theta})\rangle$: quantum states prepared by $U(\boldsymbol{\theta})$ and $\boldsymbol{z}_j$;

$l(\cdot,\cdot)$: the loss function measures the difference between y_j and $\hat{y}_j$;

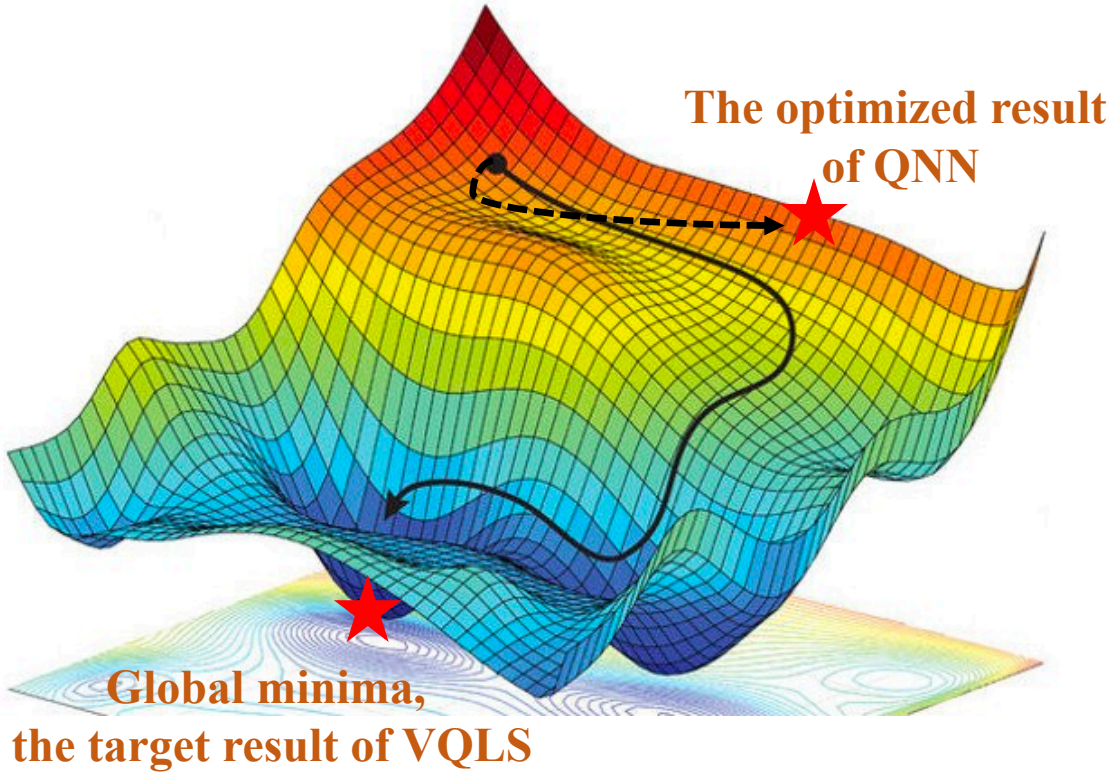
$r(\boldsymbol{\theta})$: the regularizer.

Challenges of VQLS

JDT 京东科技

Two key issues that destroy the optimization of QNN are:

- barren plateaus,
- accumulated system noise.



https://medium.com/swlh/non-convex-optimization-in-deep-learning-26fa30a2b2b3

Some solutions

JDT 京东科技

- To avoid barren plateaus:
  - ❖ modify cost functions (*Cerezo, M. et al. "Cost-Function-Dependent Barren Plateaus in Shallow Quantum Neural Networks."*),
  - ❖ adopt some strategies to initialize parameters (*Grant, Edward, et al. "An initialization strategy for addressing barren plateaus in parametrized quantum circuits."*).

- To suppress quantum system noise:
  - ❖ quantum error mitigation techniques (*Endo, Suguru, et al. "Hybrid quantum-classical algorithms and quantum error mitigation."*).

Motivation

JDT 京东科技

Q: Can we unify the elimination of barren plateaus and the mitigation of the quantum errors in **a single problem**?

Can we unify the elimination of barren plateaus and the mitigation of the quantum errors in **a single problem**?

Yes

Intuition

JDT 京东科技

The expressive power of $U(\boldsymbol{\theta})$ is improved;
The barren plateaus' phenomenon is serious;
The accumulated noise becomes larger;
The optimizated result $\boldsymbol{\theta}^{(T)}$ is far way from $\boldsymbol{\theta}^{*}$

Decreased circuit depth

Increased circuit depth

The expressive power of $U(\boldsymbol{\theta})$ is decreased;
The barren plateaus' phenomenon is negligible;
The accumulated noise is negligible;
The optimizated result $\boldsymbol{\theta}^{(T)}$ is far away to $\boldsymbol{\theta}^{*}$

Intuition

JDT 京东科技

The expressive power of $U(\boldsymbol{\theta})$ is improved;
The barren plateaus' phenomenon is serious;
The accumulated noise becomes larger;
The optimizated result $\boldsymbol{\theta}^{(T)}$ is far way from $\boldsymbol{\theta}^*$

Decreased circuit depth

The expressive power of $U(\boldsymbol{\theta})$ is decreased;
The barren plateaus' phenomenon is negligible;
The accumulated noise is negligible;
The optimizated result $\boldsymbol{\theta}^{(T)}$ is far away to $\boldsymbol{\theta}^*$

Target point

Increased circuit depth

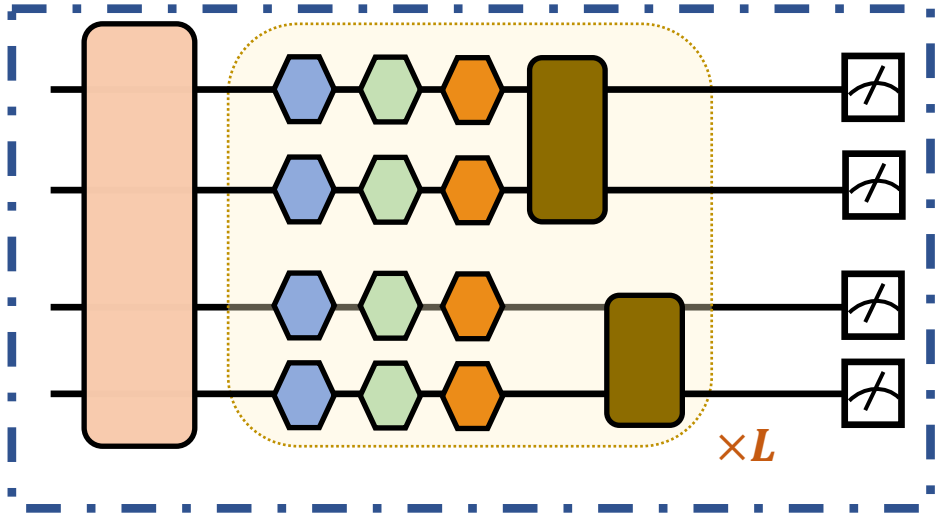
The expressive power of $U(\boldsymbol{\theta})$ is sufficient;
The accumulated noise is tolerable;
The optimizated result $\boldsymbol{\theta}^{(T)}$ is close to $\boldsymbol{\theta}^*$

Intuition

JDT 京东科技

The elimination of barren plateaus and the suppression of quantum noise amount to seeking an appropriate circuit architecture of $U(\boldsymbol{\theta})$ satisfying the following requirements:

- a good expressive power to cover the target result $\boldsymbol{\theta}^*$;
- using few number of quantum gates and shallow circuit depth;
- adapting to different connectiveity constraints of various quantum processors.



Problem setup

JDT 京东科技

Considering the circuit architecture information and the related noise, the objective of QNN is rewritten as

$$(\boldsymbol{\theta}^*, \boldsymbol{a}^*) = \arg\min_{\theta \in \boldsymbol{C}, \mathbf{a} \in \mathcal{S}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{a}, \mathcal{E}_{\boldsymbol{a}})$$

$\mathcal{S}$: the set that contains all possible circuit to build $U(\boldsymbol{\theta})$;

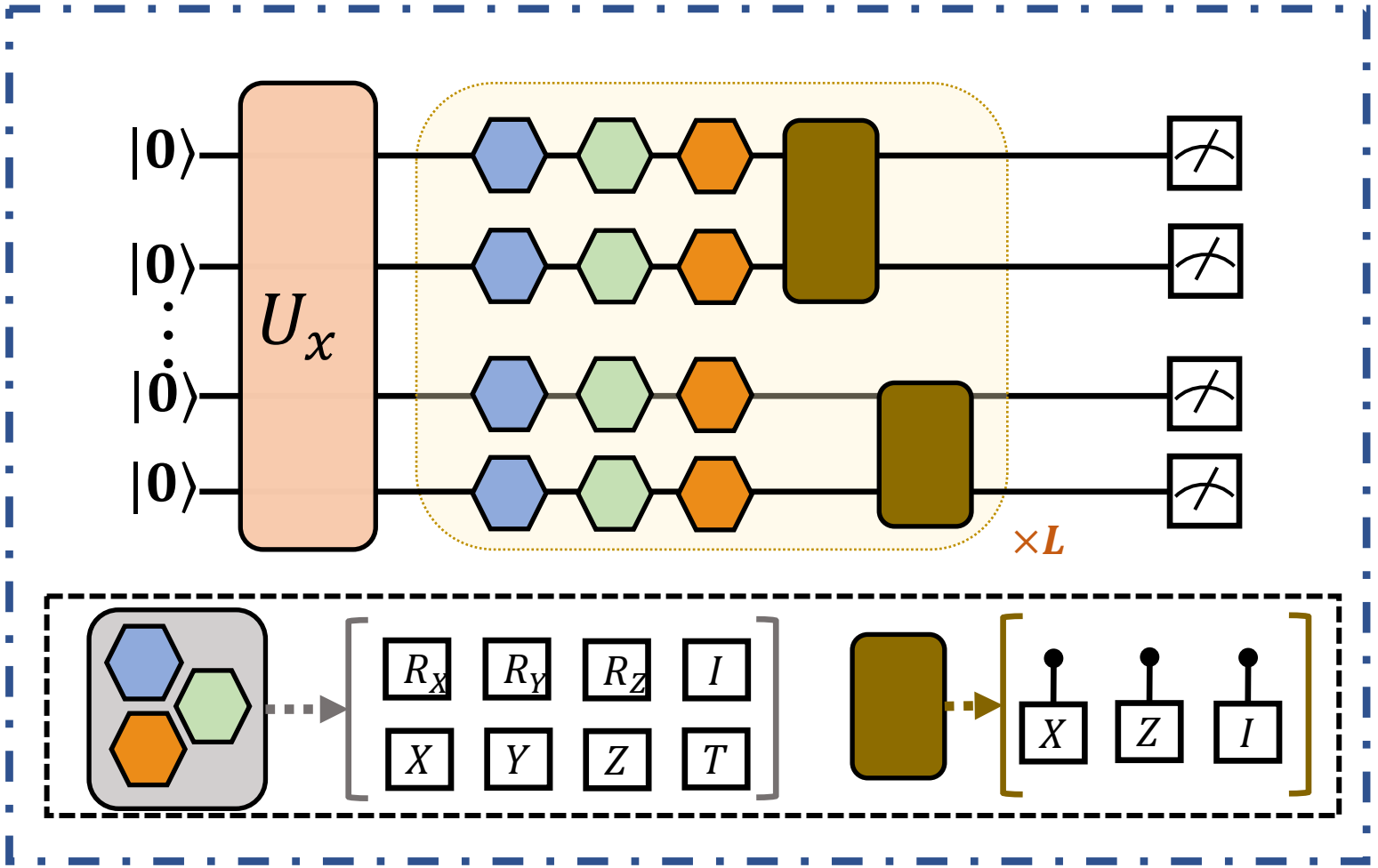
$\mathcal{E}_{\boldsymbol{a}}$: the quantum channel that simulates the quantum system noise induced by $\mathbf{a} \in \mathcal{S}$

Remarks

JDT 京东科技

The brute-force searching is infeasible, since the total number of circuit achitectures $|\mathcal{S}|$ **exponentially scales** with the qubits count N and the circuit depth L.

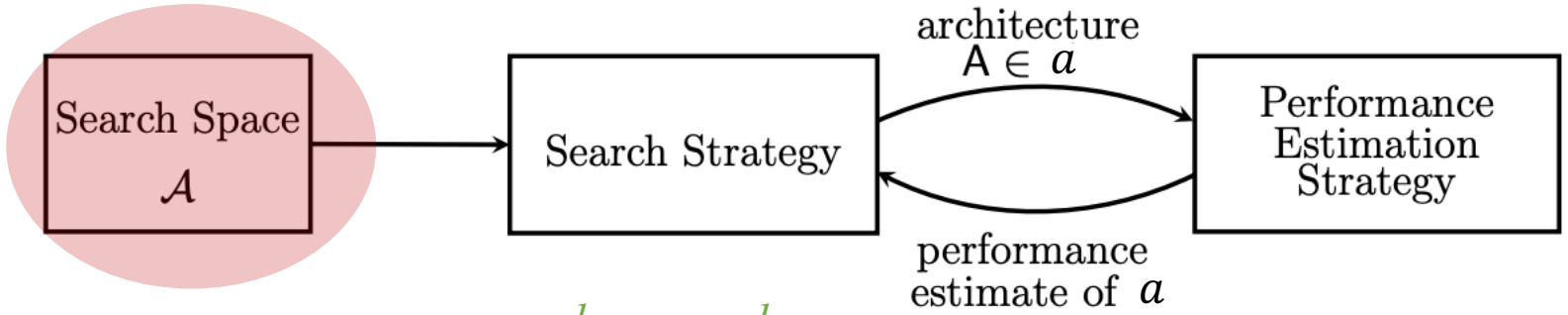
$$\mathrm{O}((8^{3N}\times 3^{N})^{L})$$



Classical analog

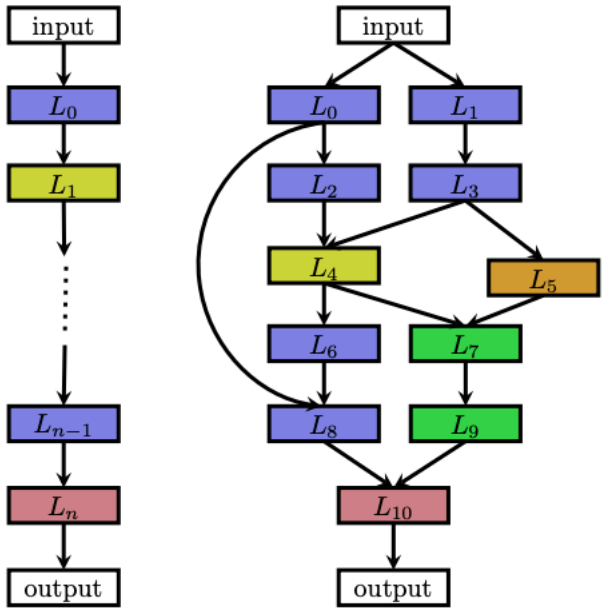
JDT 京东科技

AutoML: Neural network architecture search (NAS)



random search;
Bayesian;
evolutionary, RL, GD, ...

Different architecture spaces to search



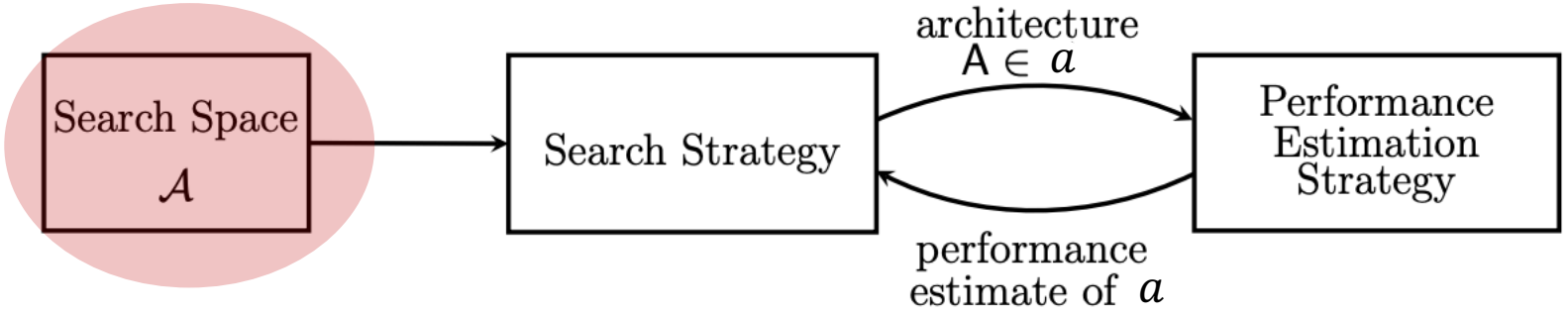
$$w_a = \underset{w}{\operatorname{argmin}}\, \mathcal{L}_{\text{train}}\left(\mathcal{N}(a, w)\right)$$

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \mathrm{ACC}_{\text{val}}\left(\mathcal{N}(a, w_a)\right)$$

Elsken, Thomas, Jan Hendrik Metzen, and Frank Hutter. "Neural Architecture Search: A Survey." Journal of Machine Learning Research 20 (2019): 1-21.

Classical analog

Exponentially large



$$w_a = \arg\min_{w} \mathcal{L}_{\text{train}}\left(\mathcal{N}(a, w)\right)$$

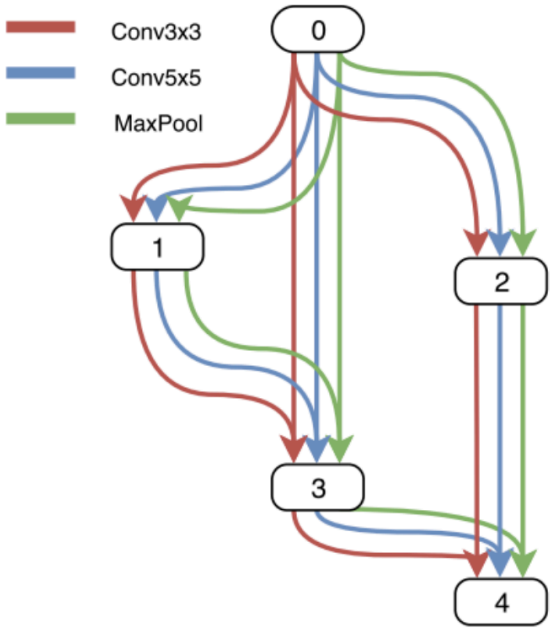
$$a^* = \arg\max_{a \in \mathcal{A}} \text{ACC}_{\text{val}}\left(\mathcal{N}(a, w_a)\right)$$

A strategy to avoid exponential overheard: Weight Sharing

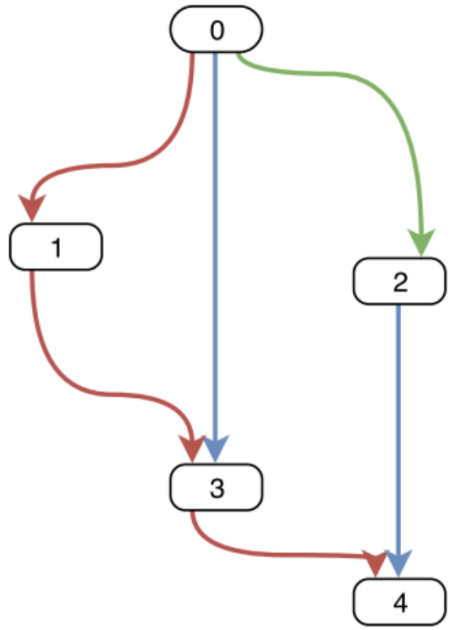
Optimization of NAS

$$W_{\mathcal{A}} = \arg\min_{W} \mathcal{L}_{\text{train}}\left(\mathcal{N}(\mathcal{A}, W)\right)$$

$$a^* = \arg\max_{a \in \mathcal{A}} \text{ACC}_{\text{val}}\left(\mathcal{N}(a, W_{\mathcal{A}}(a))\right)$$



Supergraph / supernet $\mathcal{A}$



subgraph / subnet a

Quantum circuit architecture search (QAS)

JDT 京东科技

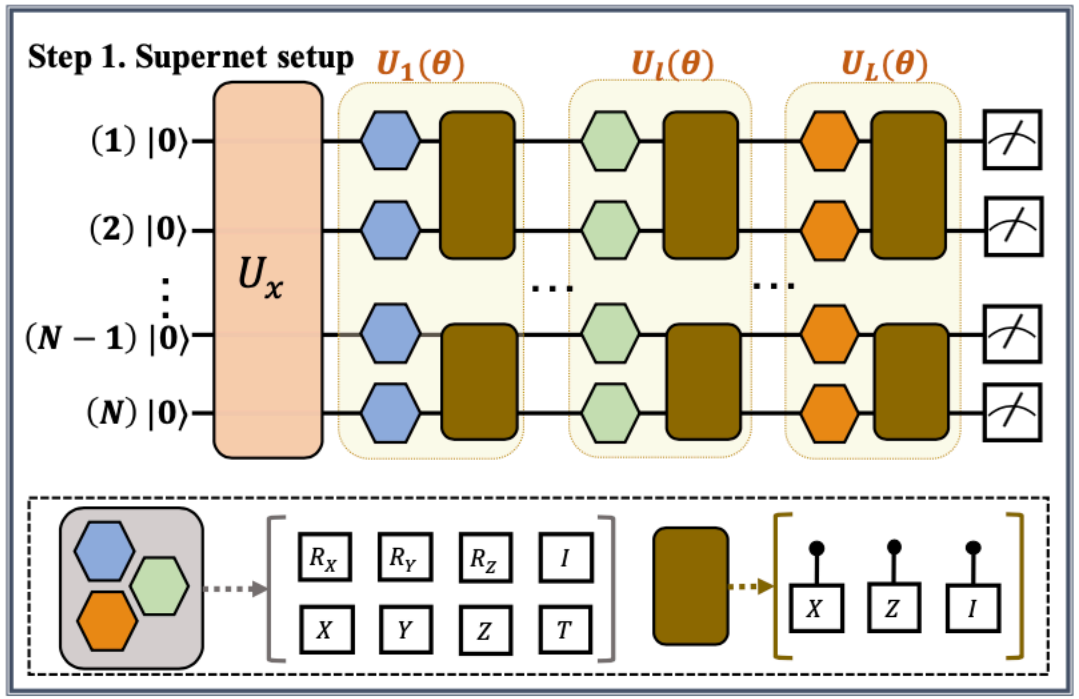
Here we propose the quantum circuit architecture search scheme (QAS) to tackle

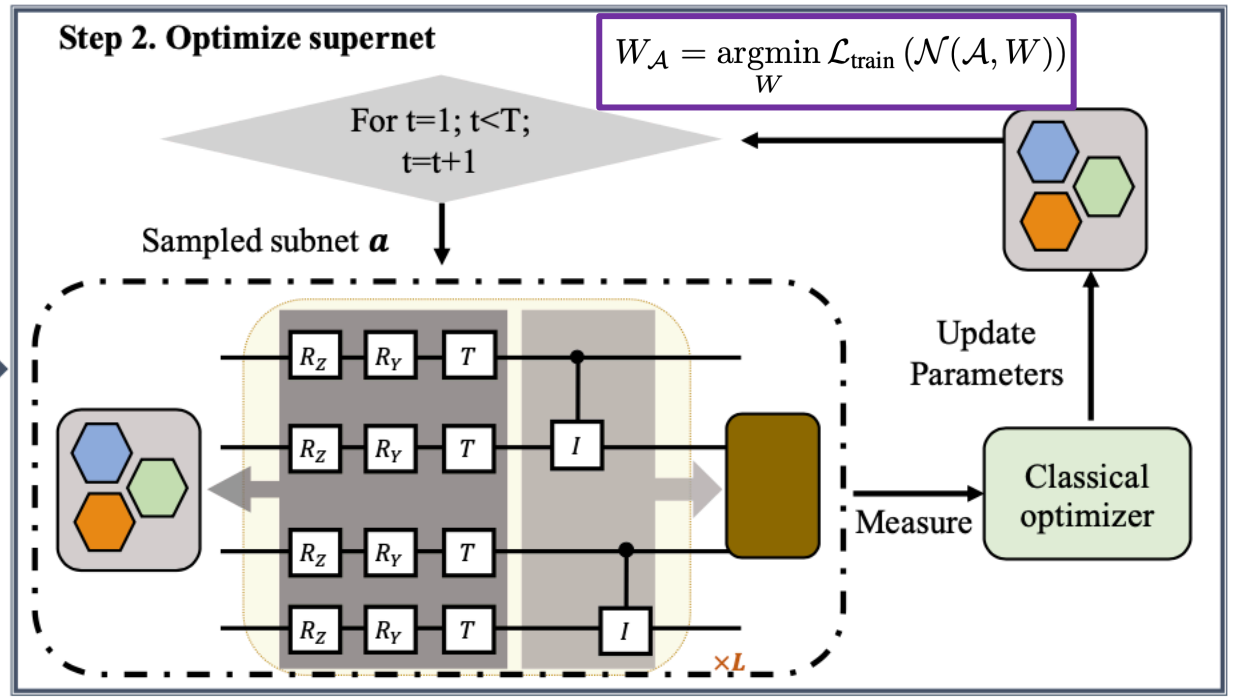
$$(\boldsymbol{\theta}^*, \boldsymbol{a}^*) = \arg\min_{\theta \in \boldsymbol{C}, \mathbf{a} \in \mathcal{S}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{a}, \mathcal{E}_{\boldsymbol{a}}).$$

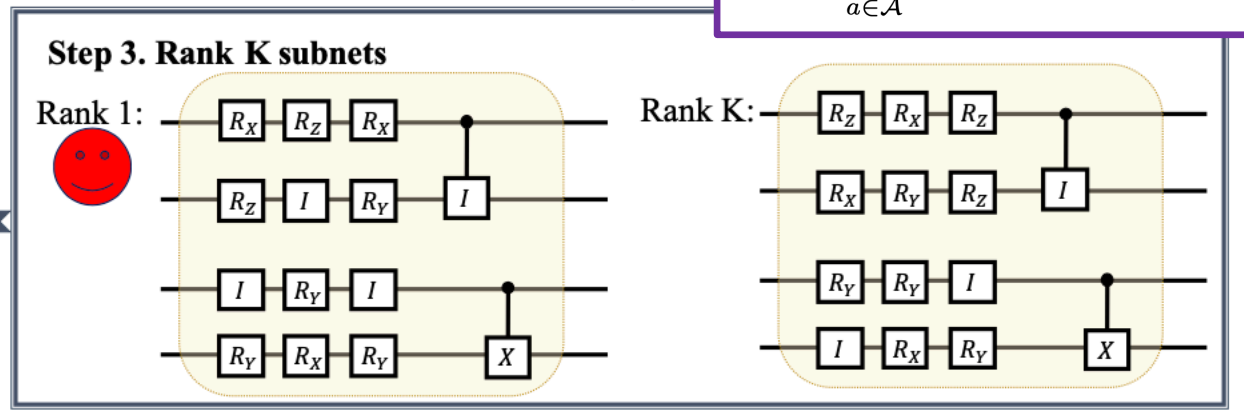
- The output of QAS avoids barren plateaus and suppresses errors;
- The runtime cost of QAS is almost same with conventional QNN-based algorithms.

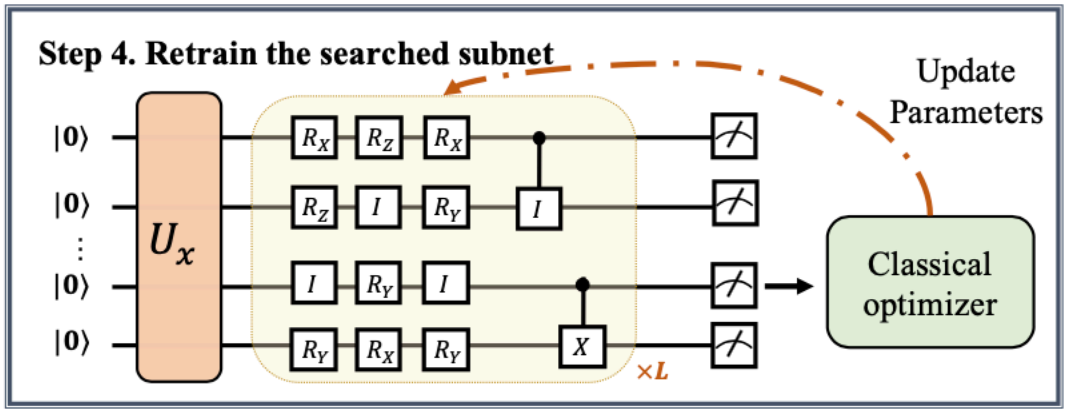
The paradigm of QAS

JDT 京东科技





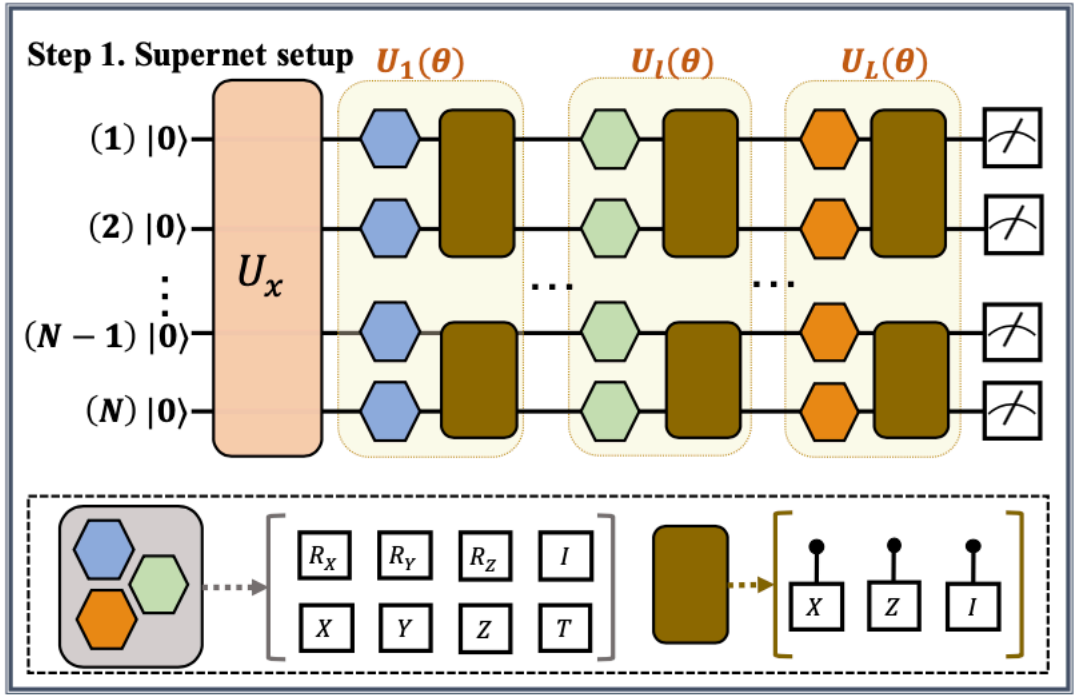




Step 1 of QAS: supernet setup + weight sharing strategy

JDT 京东科技

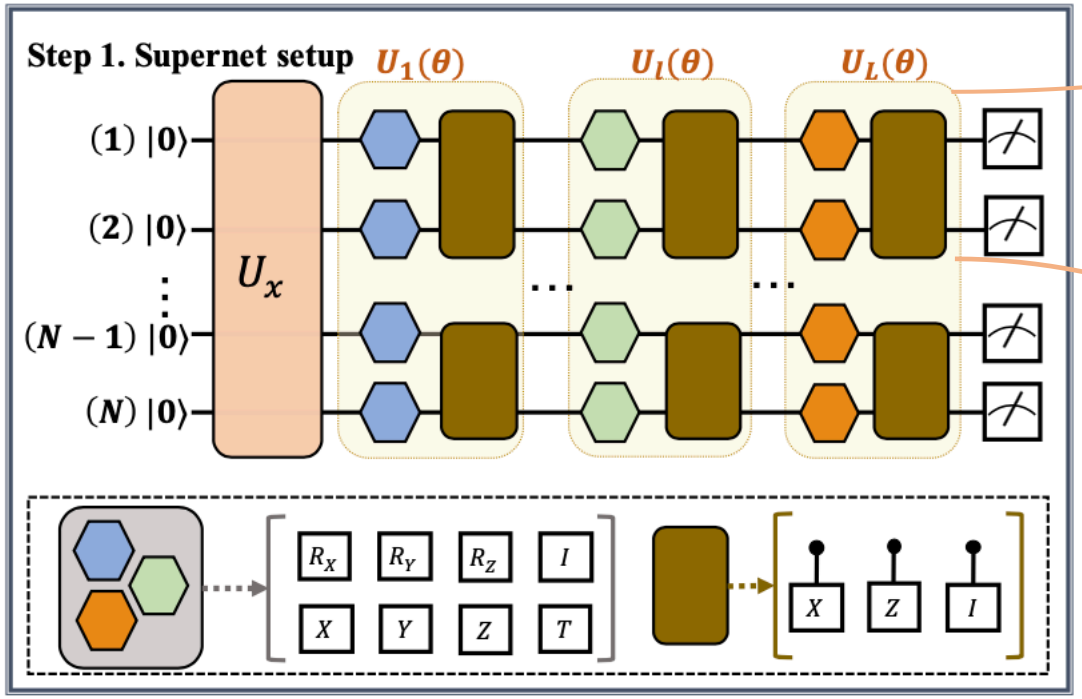
The aim of the supernet in QAS is **manipulating all possible variational quantum circuit architectures** of $U(\boldsymbol{\theta})$.

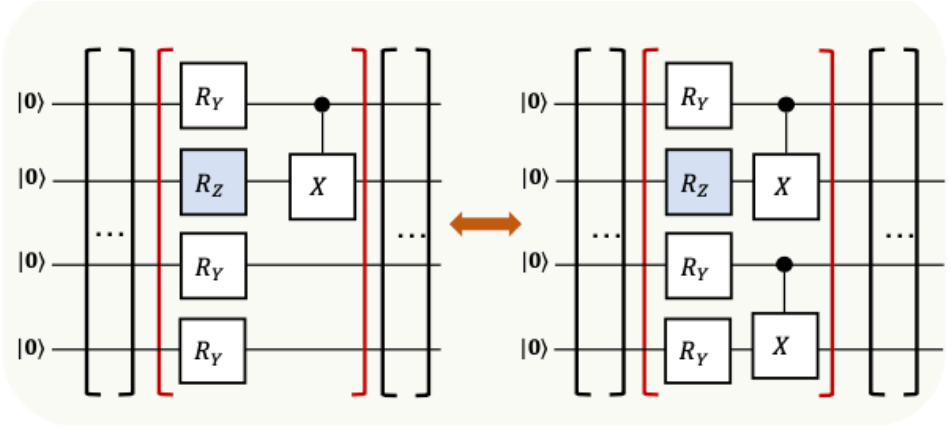


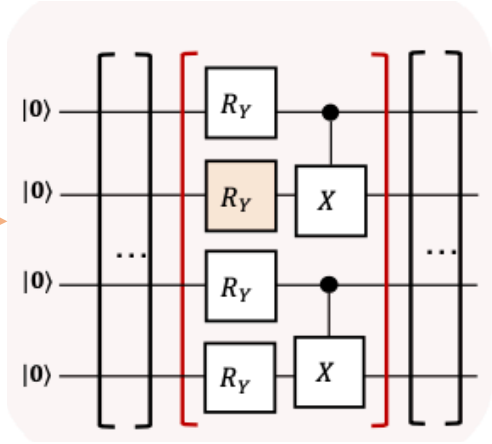
Step 1 of QAS: supernet setup + weight sharing strategy

JDT 京东科技

The weight sharing strategy: the weights in the supernet are shared across different architectures under the block level.







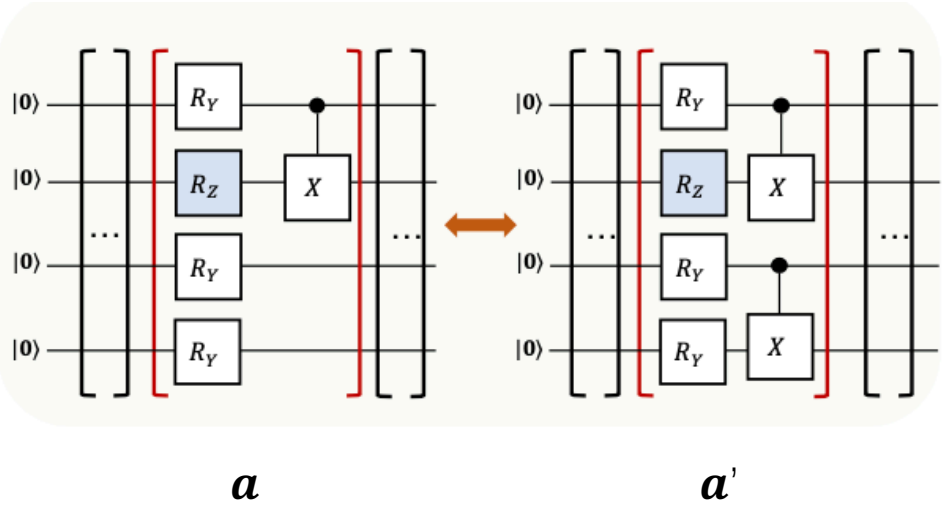
Step 2 of QAS: optimize supernet

JDT 京东科技

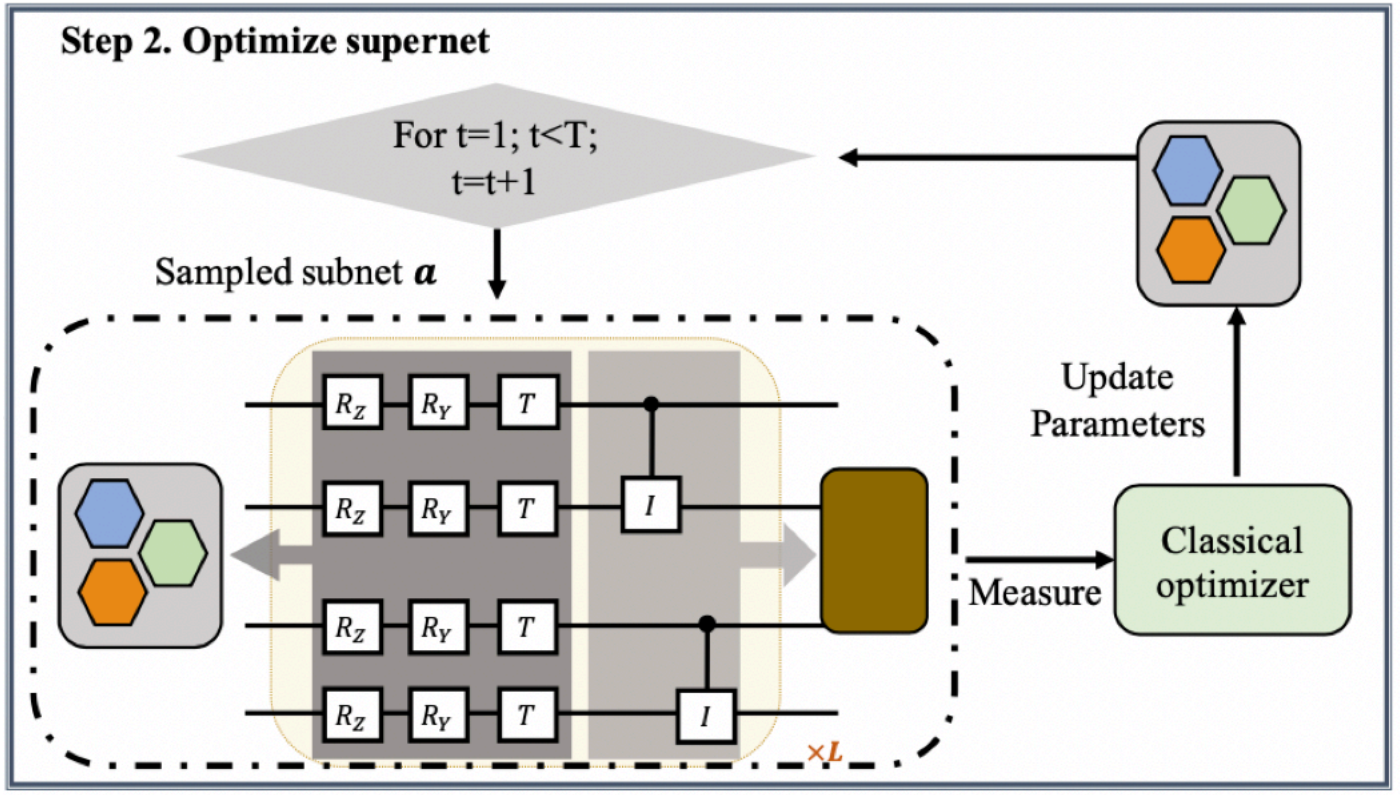
At the t-th iteration, QAS uniformly samples a subnet $\mathbf{a}^{(t)} \in \mathcal{S}$ to minimize $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{a}, \mathcal{E}_{\boldsymbol{a}})$, i.e., the parameters of the supernet relating to $\mathbf{a}^{(t)}$ is updated to

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \eta \frac{\partial \mathcal{L}(\boldsymbol{\theta}^{(t)}, \boldsymbol{z}, \boldsymbol{a}, \mathcal{E}_{\boldsymbol{a}})}{\partial \boldsymbol{\theta}^{(t)}},$$

where η is the learning rate.



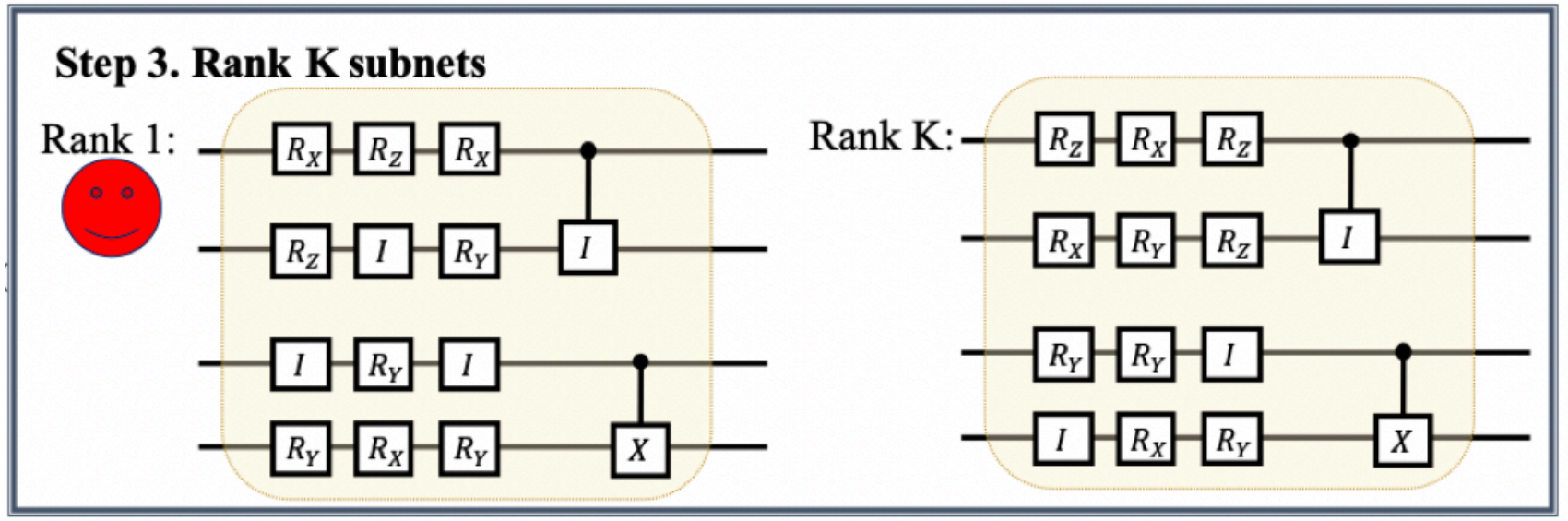
$\boldsymbol{a}$ $\boldsymbol{a}'$



Step 3 of QAS: Rank K subnet

JDT 京东科技

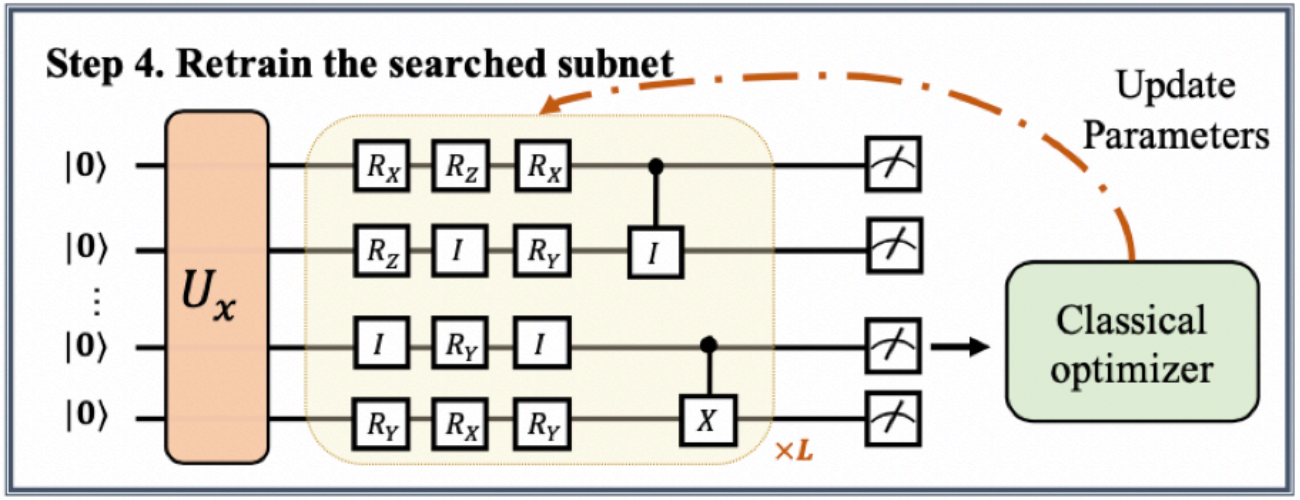
QAS uniformly samples K subnets, then ranks their performances, and assigns the subnet with the best performance as the output to approximate $\mathbf{a}^*$.



Step 4 of QAS: Rank K subnet

JDT 京东科技

QAS retrain the searched subnet to approximate the optimal parameters.

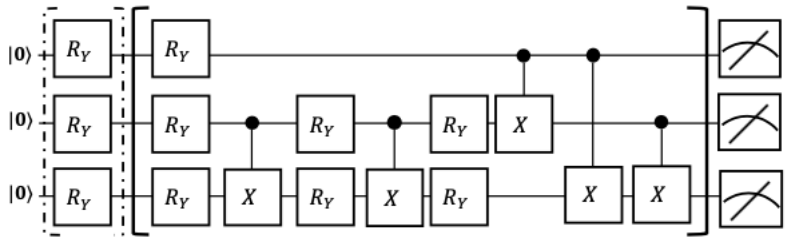


The fierce competition phenomenon of QAS

JDT 京东科技

Fierce competition in QAS: the performance of a specified subnet may be largely varied by independently training and QAS.

1st subnet $\boldsymbol{a}_1$



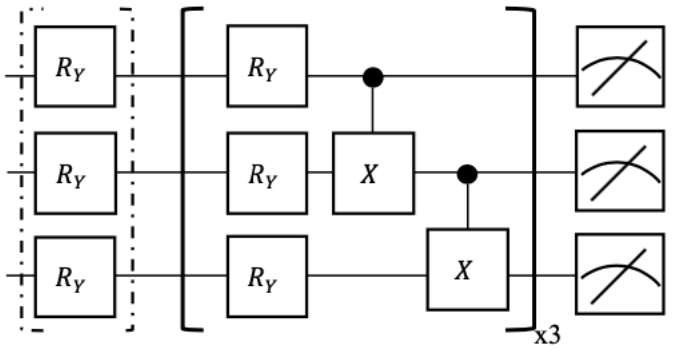
Independent training

99% acc | 94% acc

QAS with the weight sharing

91% acc | 93% acc

2cd subnet $\boldsymbol{a}_2$

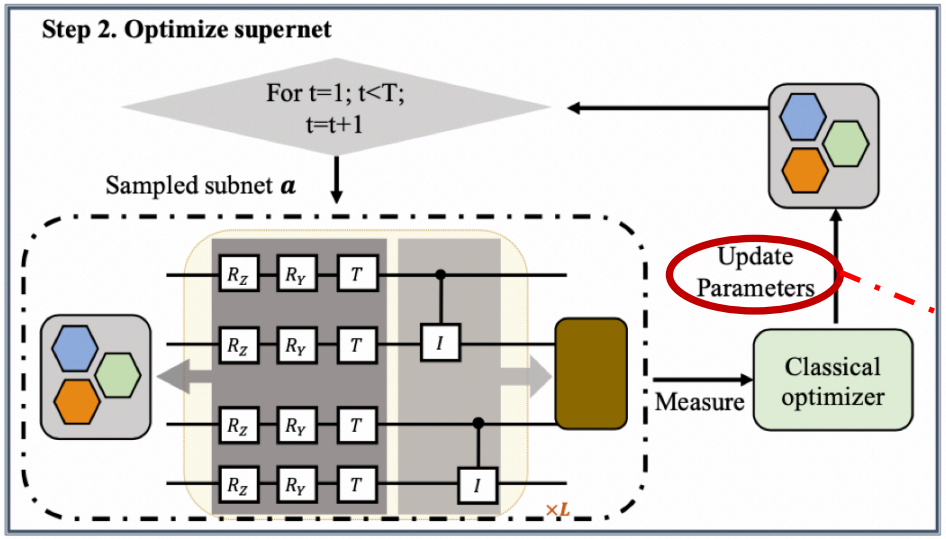


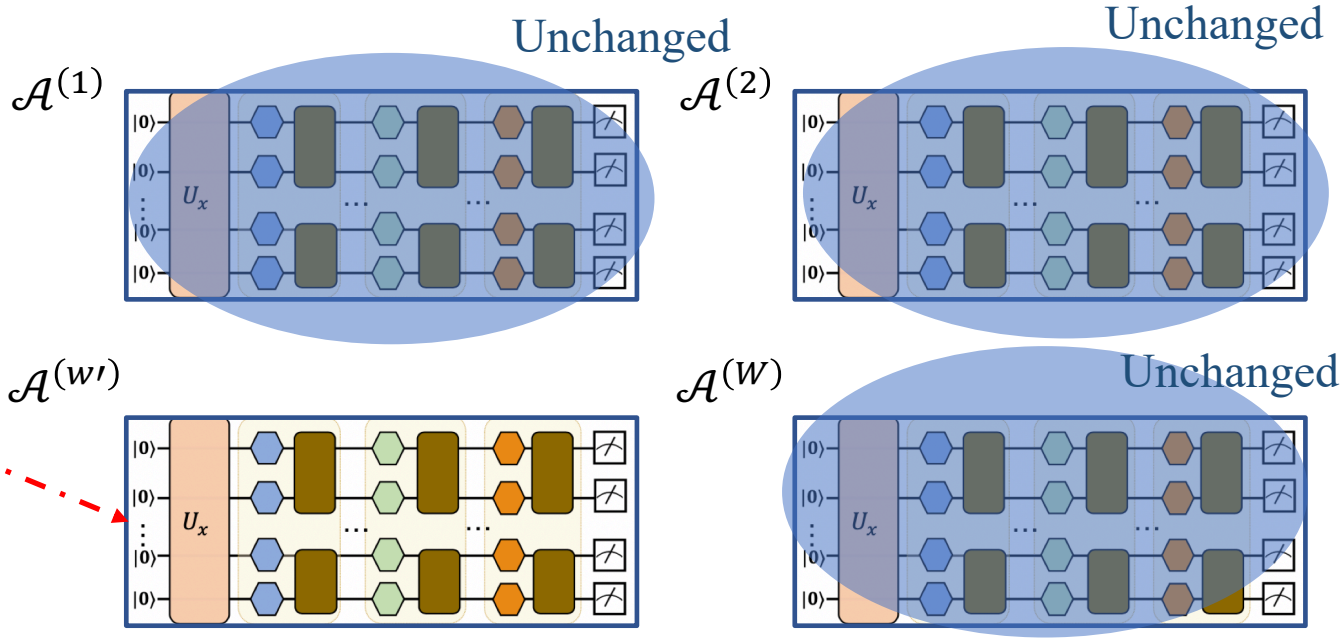
Solution to the fierce competition phenomenon of QAS

JDT 京东科技

Solution: Employ multiple supernets $\{\mathcal{A}^{(w)}\}_{w=1}^{W}$ and optimize them independently. In the stages II-III, the sampled subnet $\boldsymbol{a}^{(t)}$ is categorized into the w′-th supernet when

$$w' = \arg\min_{w\in[W]} \mathcal{L}\left(\boldsymbol{\theta}^{(t,w)}, \boldsymbol{z}, \boldsymbol{a}, \mathcal{E}_{\boldsymbol{a}}\right).$$





Relationship with the adversarial bandit problems

JDT 京东科技

Theorem 1. Let W and T be the number of supernets and iterations, respectively. Suppose that the subnet $\boldsymbol{a}^{(t)}$ is assigned to the $I_w^{(t)}$-th supernet $\mathcal{A}^{\left(I_w^{(t)}\right)}$ with $I_w^{(t)} \in [W]$ at the t-th iteration, where the corresponding objective function is $\mathcal{L}\left(\boldsymbol{\theta}^{(t,I_w^{(t)})}, \boldsymbol{z}, \boldsymbol{a}, \mathcal{E}_{\boldsymbol{a}}\right) \in [0,1]$. Define the regret as

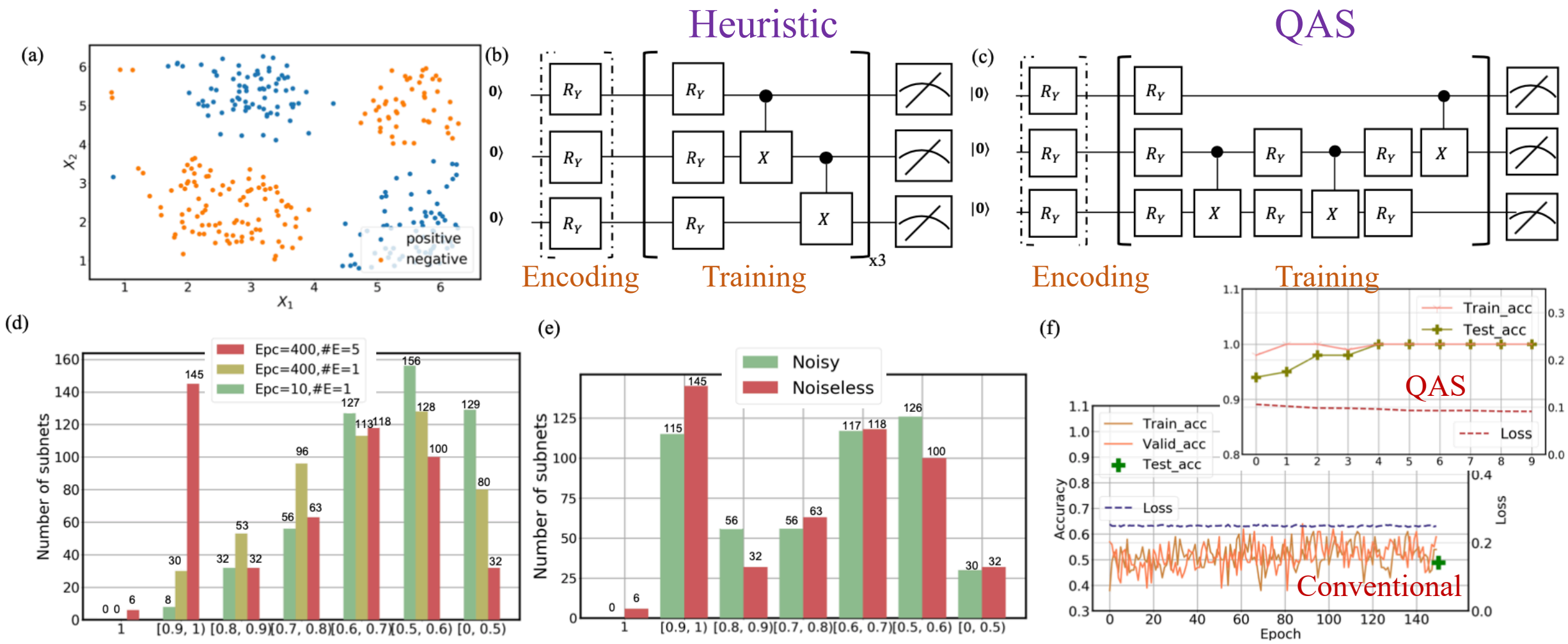
$$R_T = \sum_{t=1}^{T} \mathcal{L}\left(\boldsymbol{\theta}^{(t,I_w^{(t)})}, \boldsymbol{z}, \boldsymbol{a}, \mathcal{E}_{\boldsymbol{a}}\right) - \min_{\{w_t\}_{t=1}^{T}} \sum_{t=1}^{T} \mathcal{L}\left(\boldsymbol{\theta}^{\left(t,I_w^{(t)}\right)}, \boldsymbol{z}, \boldsymbol{a}, \mathcal{E}_{\boldsymbol{a}}\right).$$

The method used in QAS to determine $\{I_w^{(t)}\}$ promises the regret $R_T \leq 0$, while the regret for the best bandit algorithms is lower bounded by $R_T = \Omega(T)$.

Summary: The proposed strategy outperforms all bandit learning algorithms in the measure of the regret.

Experiment results for classification tasks

JDT 京东科技



Experiment results for the ground state energy estimation

JDT 京东科技

The molecular hydrogen Hamiltonian is formulated as

$$H_h = \mathrm{g} + \sum_{i=1}^{3} g_i Z_i + \sum_{i=1,k=1,i<k}^{3} g_{i,k} Z_i Z_k + g_a Y_0 X_1 X_2 Y_3 + g_b Y_0 Y_1 X_2 X_3 + g_c X_0 X_1 Y_2 Y_3 + g_d X_0 Y_1 Y_2 X_3,$$

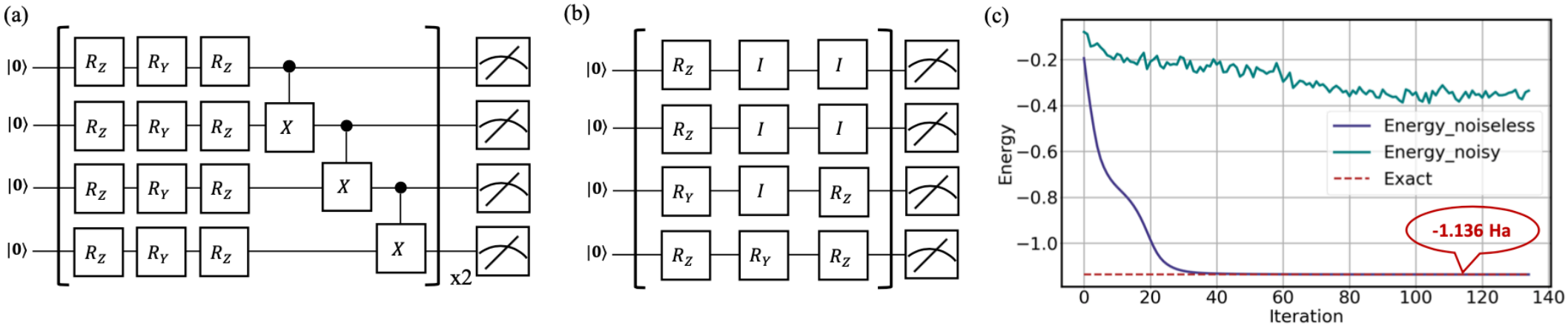
where $\{X_i, Y_i, Z_i\}$ denote the Pauli matrices acting on the i-th qubit and the real scalars g with or without subscripts are efficiently computable functions of the hydrogen-hydrogen bond length.

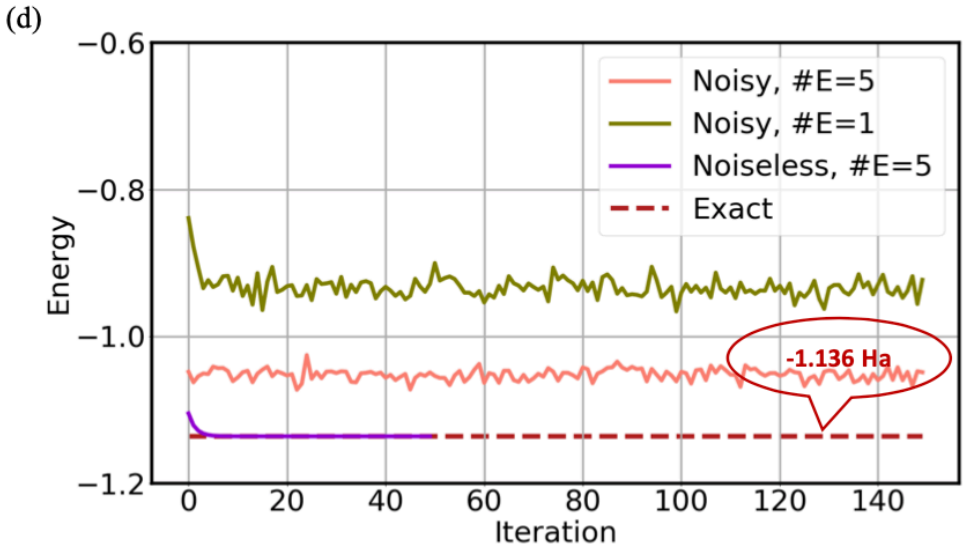
Experiment results for the ground state energy estimation

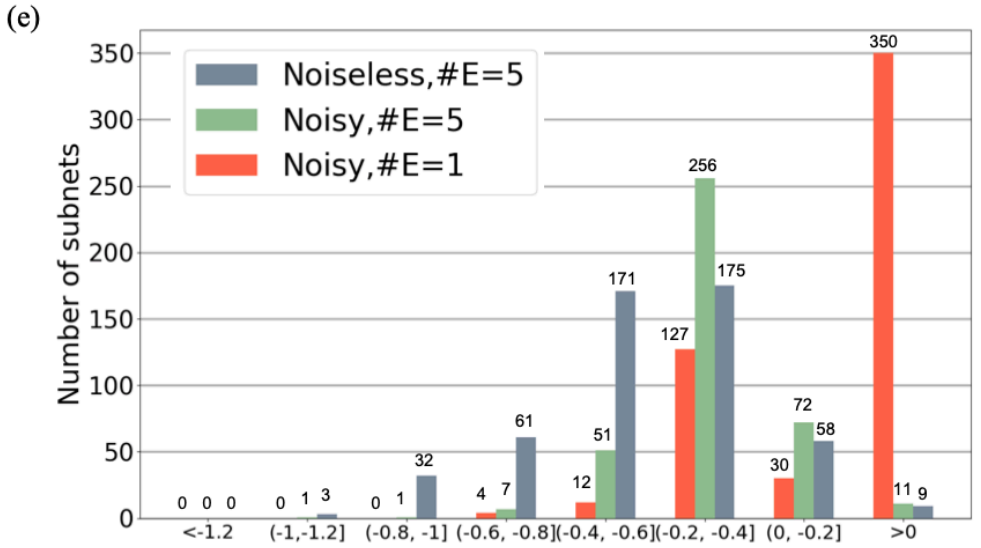
JDT 京东科技

Heuristic

QAS



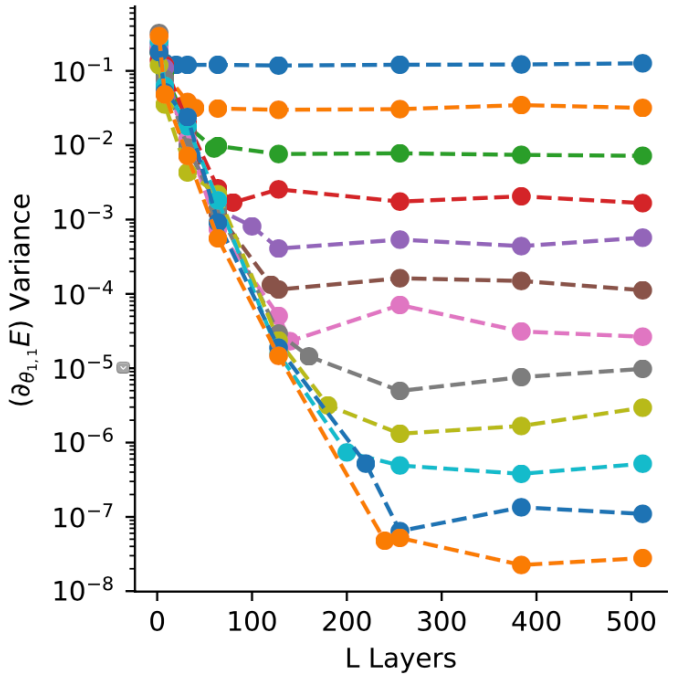


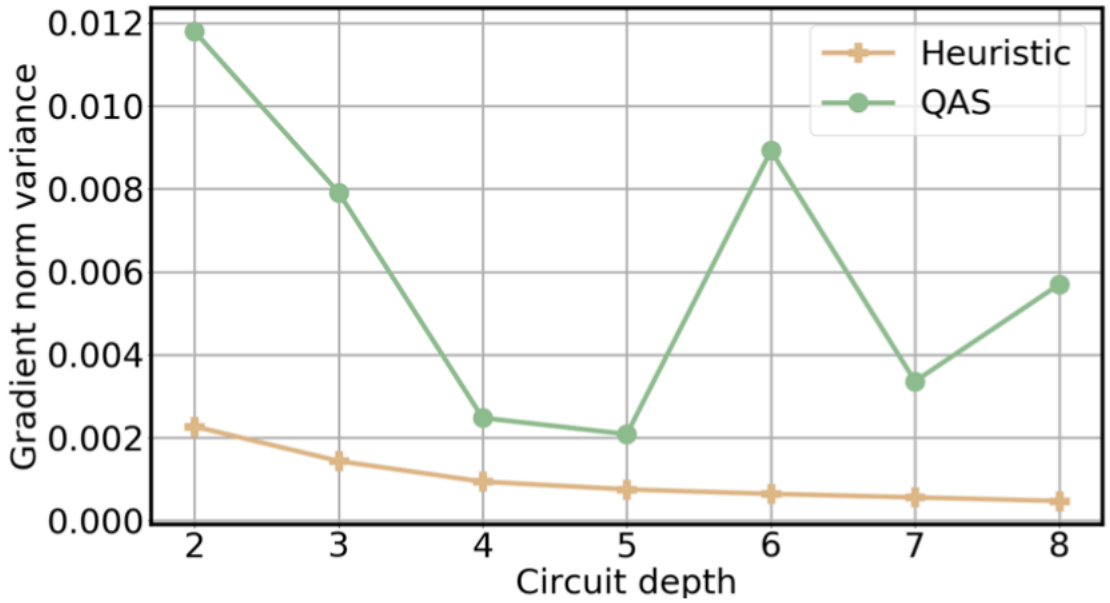


Experiment results for the ground state energy estimation

JDT 京东科技

QAS contributes to the alleviation of barren plateaus.

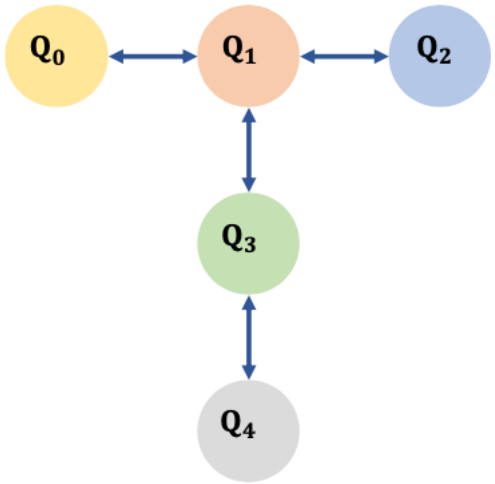




Experiment results for the ground state energy estimation

JDT 京东科技

We last carry out QAS and the conventional VQE on IBM's 5-qubit quantum machine, i.e., 'Ibmq_ourense', to accomplish the ground state energy estimation of H_h.

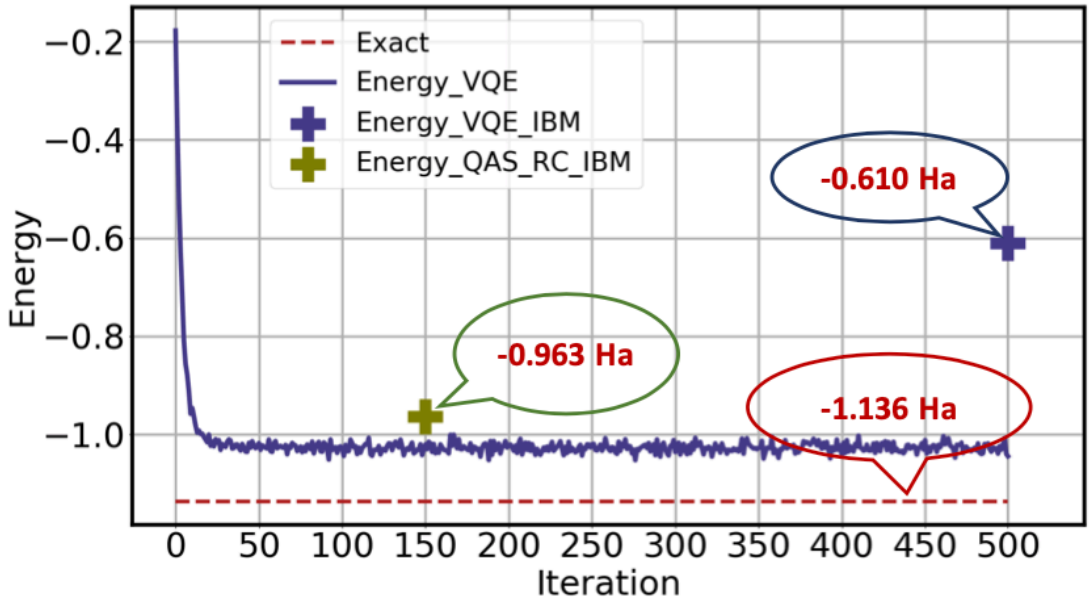


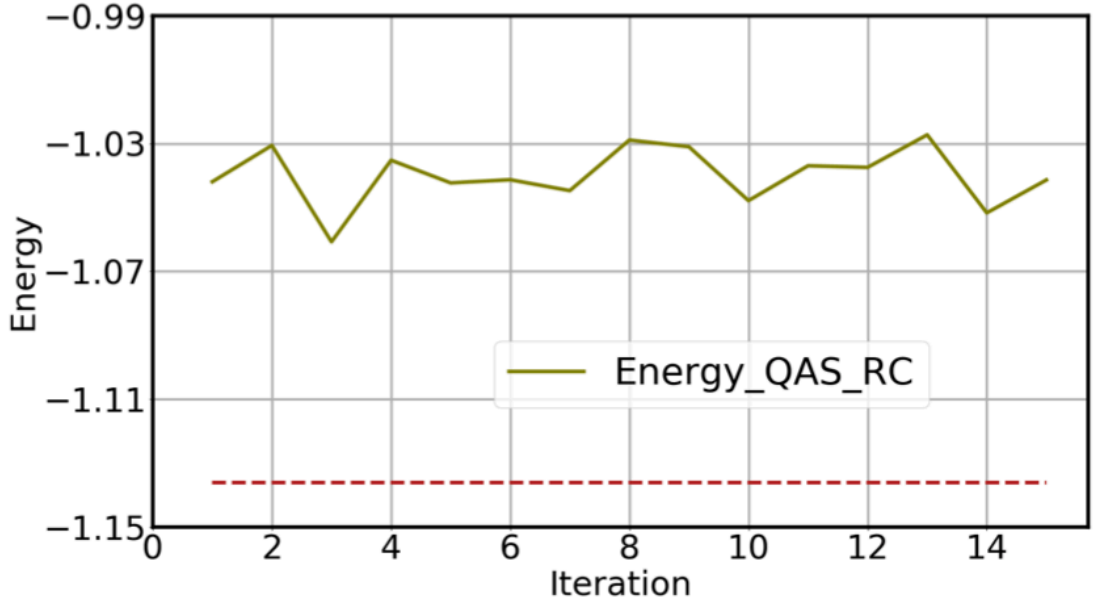
Qubit	T1(μs)	T2(μs)	Readout error	Single-qubit U2 error gate	CNOT error rate
Q0	75.75	50.81	1.65E-2	5.22E-4	cx0-1: 9.55E-3
Q1	78.47	27.56	2.38E-2	4.14E-4	cx1-0: 9.55E-3 cx1-2: 9.44E-3 cx1-3: 1.25E-2
Q2	101.51	107.00	1.57E-2	1.83E-4	cx2-1: 9.44E-3
Q3	79.54	78.38	3.95E-2	4.30E-4	cx3-1: 1.25E-2 cx3-4: 8.34E-3
Q4	74.27	30.00	4.74E-2	4.20E-4	cx4-3: 8.34E-3

Experiment results for the ground state energy estimation

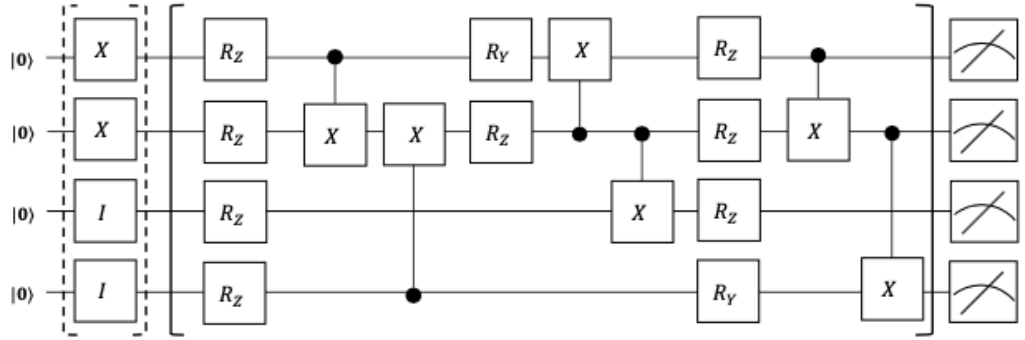
JDT 京东科技

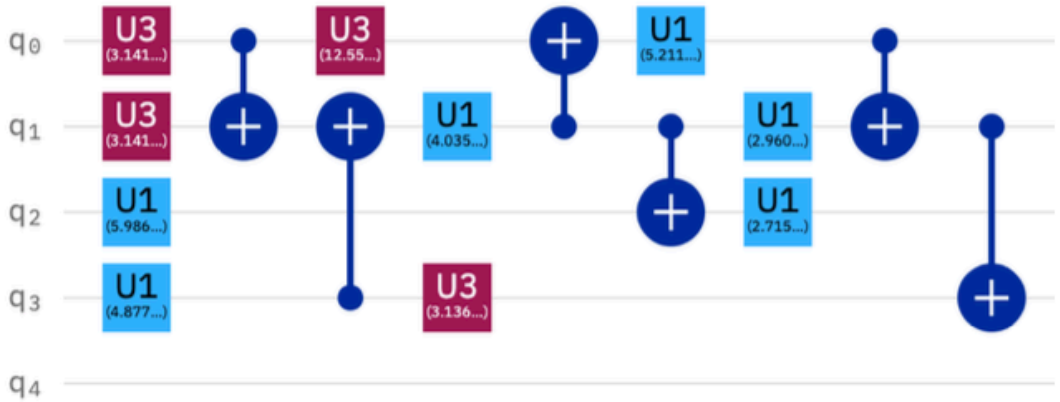
Experiment results of the ground state energy estimation of VQE and QAS.





The searched circuit architecture by QAS

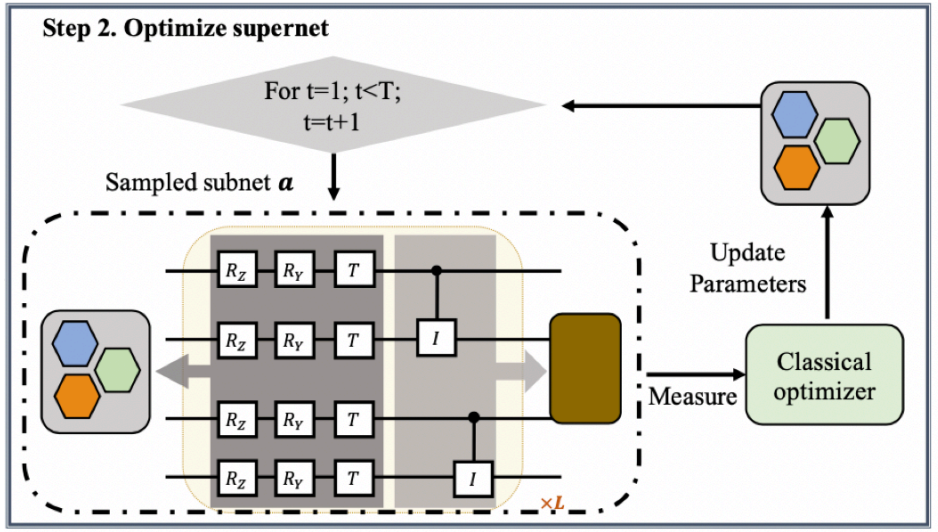




Improvements

JDT 京东科技

- Improving the ranking stage of QAS, i.e., replacing the uniformly sampling by evolutionary algorithms or reinforcement learning algorithms.



- Employing the adversarial bandit learning algorithms to train multiple supernets. Benefits: reduce the runtime cost. Disadvantages: inducing a relatively large regret bound.

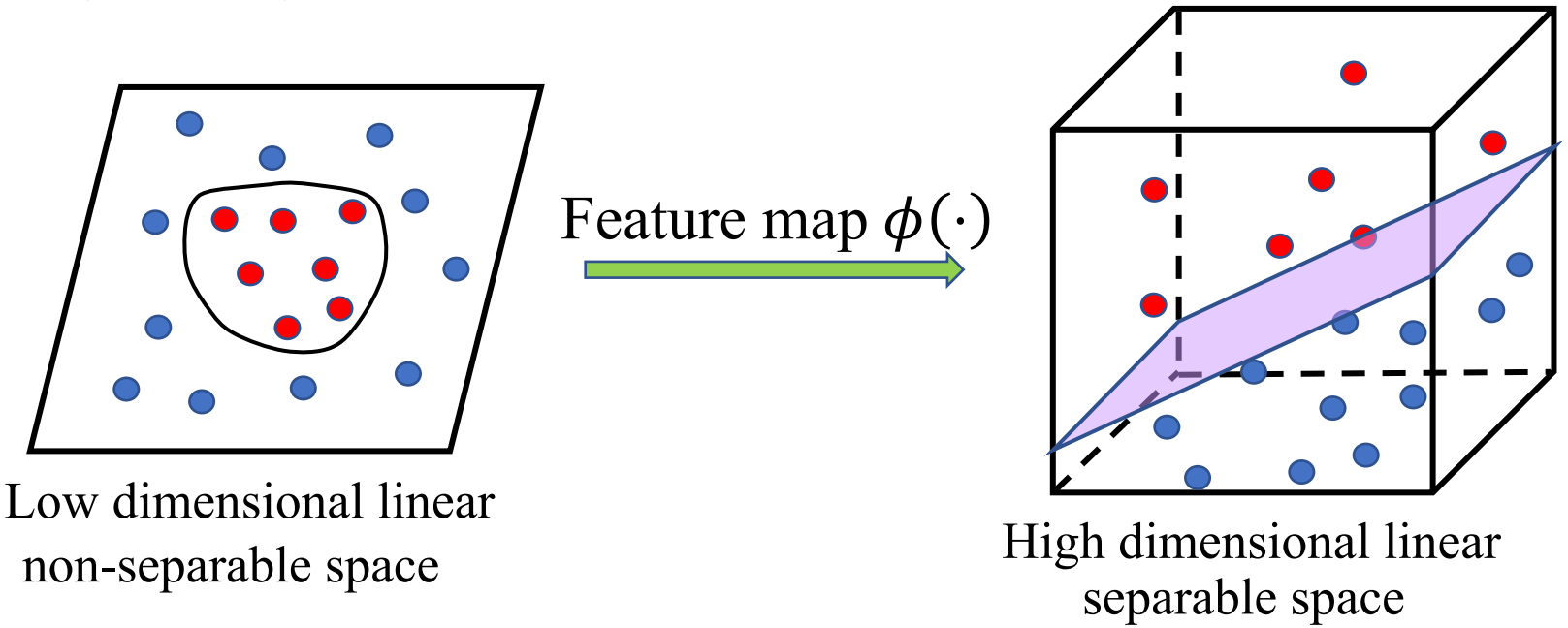
III. The power of quantum kernels in the NISQ era

Kernel methods

JDT 京东科技

Core idea:

- Mapping the given input $x^{(i)} \in \mathbb{R}^d$ into a high-dimensional feature space, i.e., $\phi(\cdot)$: $\mathbb{R}^d \rightarrow \mathbb{R}^q$ $(q \gg d)$.
- Constructing a kernel function $\kappa(x^{(i)}, x^{(j)}) = \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$ and the corresponding kernel matrix $\mathrm{K}_{ij} = \kappa(x^{(i)}, x^{(j)})$



Typical kernels: the radial basis function kernel, Gaussian kernel, and polynomial kernel.

Kernel methods

JDT 京东科技

Core idea:

For efficient optimization of w, we consider minimization of the following loss function

$$\min_{w} \lambda\langle w, w\rangle + \sum_{i=1}^{N} \left(\langle w, \phi(x)\rangle - \mathrm{Tr}(U^{\dagger}OU\rho(x_i))\right)^2,$$

$h(x_i)$ y_i

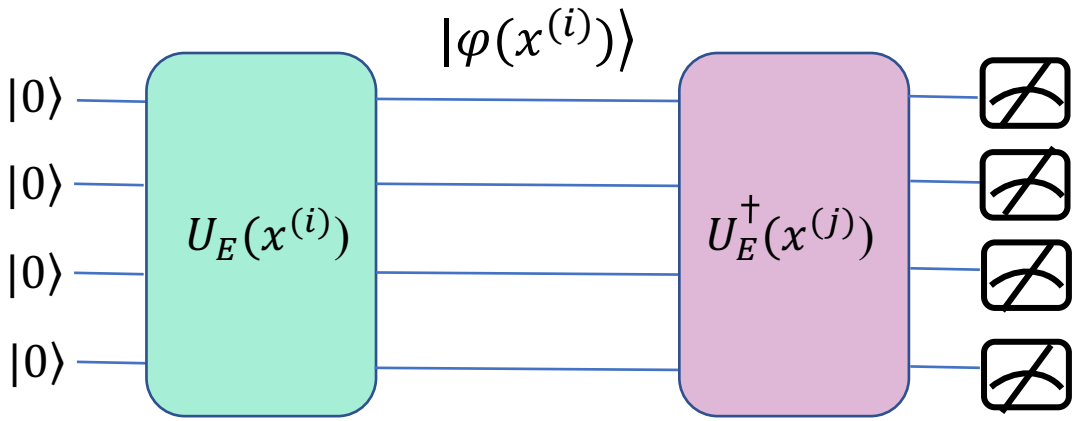
The optimal ω can be written down explicitly as

$$w = \sum_{i=1}^{N}\sum_{j=1}^{N} \phi(x_i)((K + \lambda I)^{-1})_{ij}\,\mathrm{Tr}(U^{\dagger}OU\rho(x_j)).$$

Quantum kernel methods

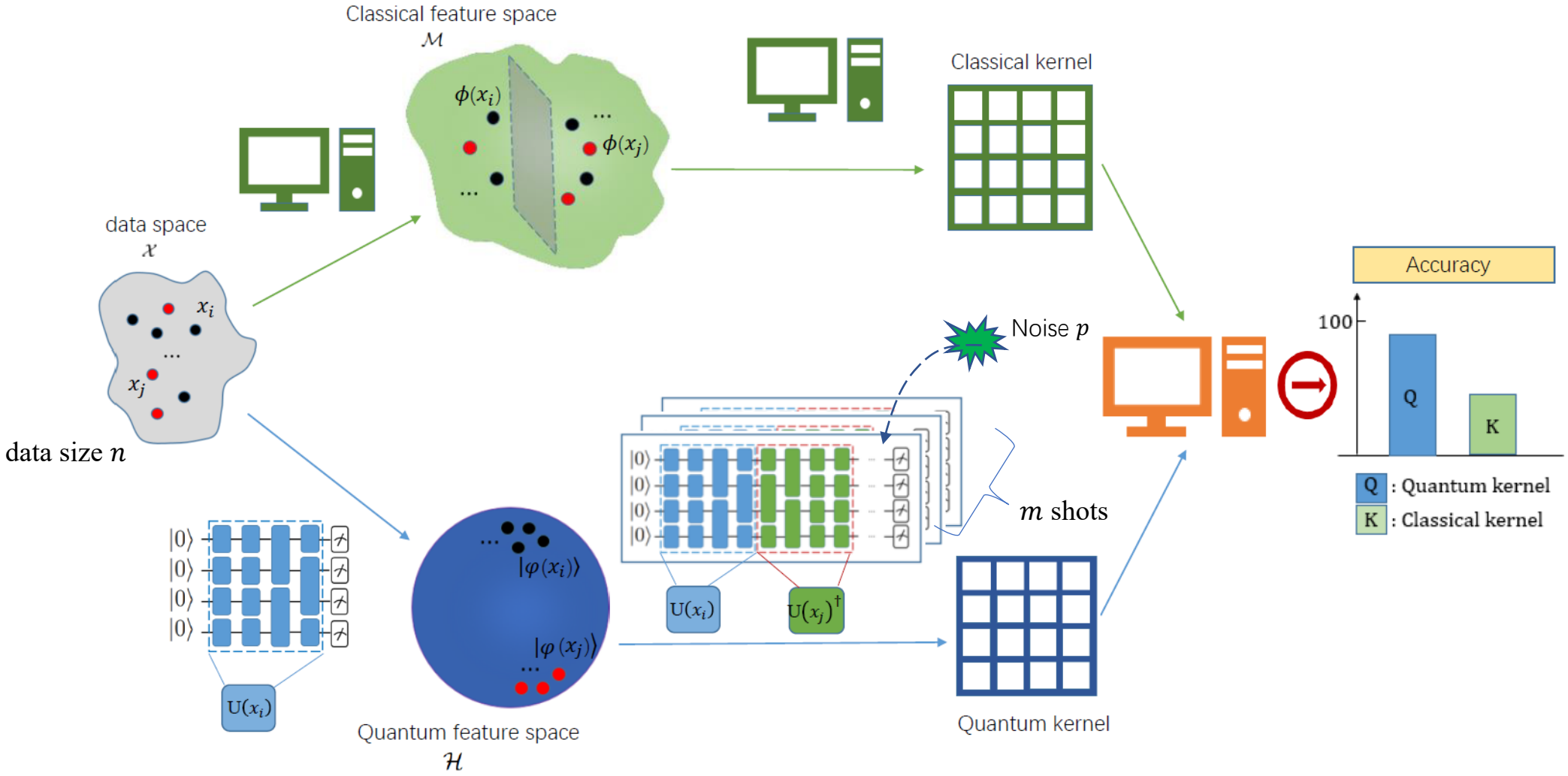
JDT 京东科技

- Replacing classical feature map $\phi(\cdot)$ with a variational quantum circuit $U_E(\cdot)$.
- Mapping the give classical data $x^{(i)} \in \mathbb{R}^d$ to exponentially large feature space, i.e., $|\phi(x^{(i)})\rangle = U_E(x^{(i)})\,|0^d\rangle \in \mathbb{R}^{2^d}$. Denote N as the number of qubits, then $N = d$.
- The result of kernel function $\kappa(\cdot,\cdot)$ coincides with applying measurements on the prepared. quantum states, then the (i,j)-th element in kernel matrix is $\mathbf{W}_{ij} = |\langle\varphi(\boldsymbol{x}^{(j)})|\varphi(\boldsymbol{x}^{(i)})\rangle|^2$



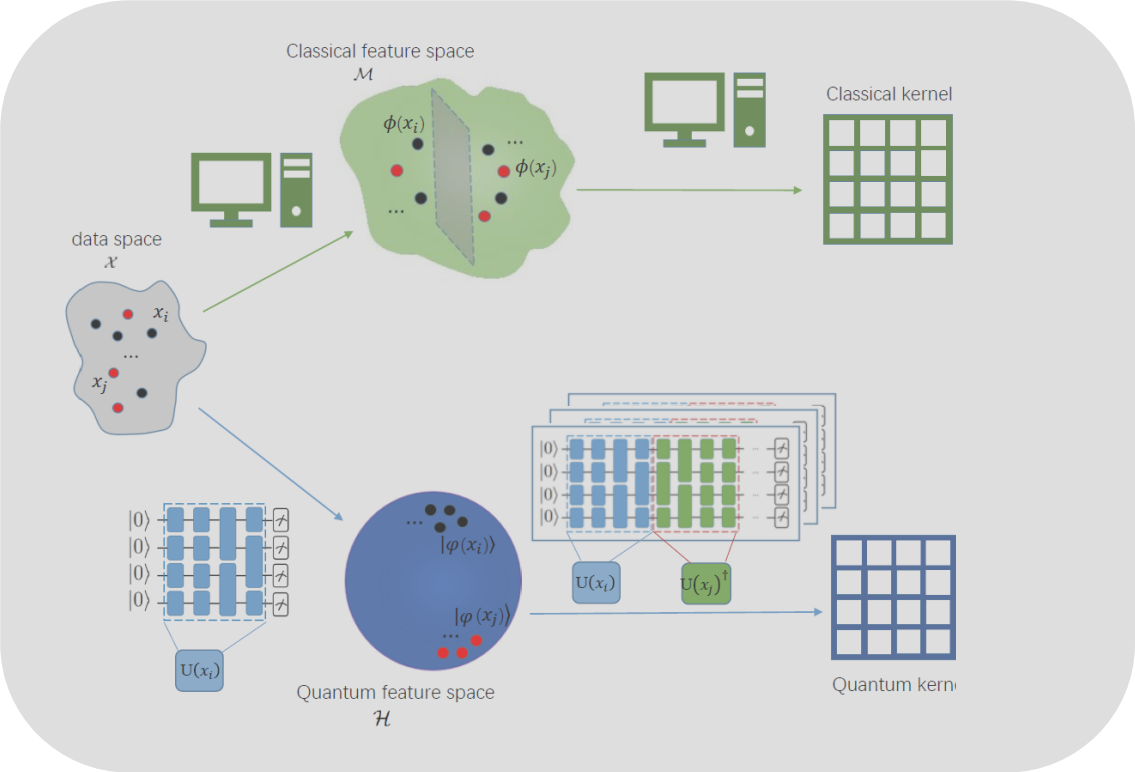
Overview

JDT 京东科技



The power of quantum kernels

JDT 京东科技



How quantum kernel methods outperform classical kernel methods in predicting the unseen data ?

The power of data in quantum kernels

JDT 京东科技

Problem setup.

Each data pair $(x^{(i)}, y^{(i)})$ is constructed by $y^{(i)} = f\left(x^{(i)}\right) = \mathrm{Tr}\left(OU(\boldsymbol{\theta}^*)\rho\left(x^{(i)}\right)U(\boldsymbol{\theta}^*)^\dagger\right)$, where $U(\boldsymbol{\theta}^*)$ a specified unitary evolution applied on the encoded quantum example $\rho\left(x^{(i)}\right)$, and O is a certain measurement operator.

Given N examples, the goal of a learning model is to infer a map $h(\cdot)$ such that $y^{(i)} = h\left(x^{(i)}\right)$.

The power of data in quantum kernels

JDT 京东科技

Results.

Huang et al. [arXiv:2011.01938] present a generalization error bound under **the ideal scenario,** i.e., given n examples, the generalization error of classical kernels follows

$$\mathbb{E}_{x\sim\mathcal{D}}|h(x)-f(x)| \leq c\sqrt{\frac{s_K(n)}{n}}$$

where $s_K(n)$ refers to the model complexity with

$$s_K(n) = \sum_{i=1}^{n}\sum_{j=1}^{n}(K^{-1})_{ij}\,\mathrm{Tr}(O^U\rho(x_i))\,\mathrm{Tr}(O^U\rho(x_j)).$$

The power of data in quantum Kernels

JDT 京东科技

$$\mathbb{E}_{x\sim\mathcal{D}}|h(x)-f(x)| \leq c\sqrt{\frac{s_K(n)}{n}} \qquad s_K(n) = \sum_{i=1}^{n}\sum_{j=1}^{n}(K^{-1})_{ij}\,\mathrm{Tr}(O^U\rho(x_i))\,\mathrm{Tr}(O^U\rho(x_j)).$$

Two insights

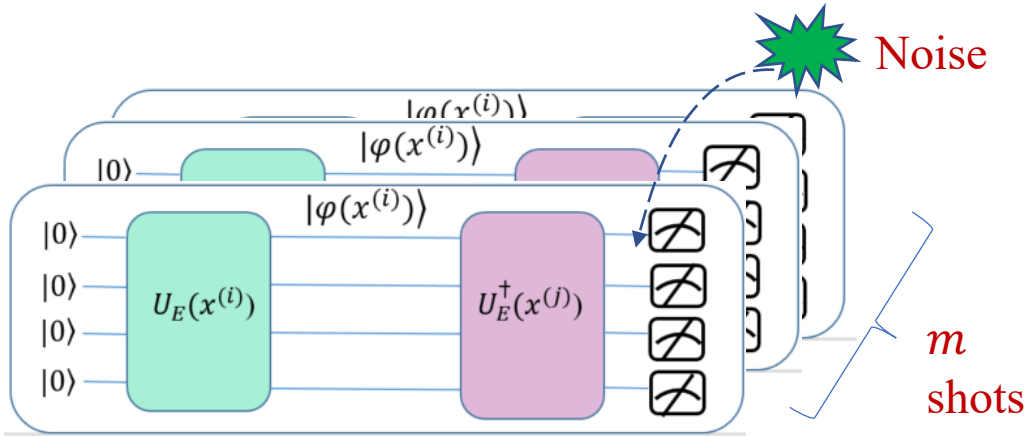
- Given the same sample size, a small $\frac{Y^\top \mathrm{W}^{-1} Y}{Y^\top \mathrm{K}^{-1} Y}$ leads to a large quantum advantage over the classical kernel methods.

- Under the same kernel method, a large n leads to a better prediction performance.

How about power of quantum kernels in the NISQ era?

JDT 京东科技

In the NISQ era, quantum kernels are affected by the system noise and finite shots.



- Consider the depolarization channel $\mathcal{N}_p(\rho(\boldsymbol{x}^{(i)})) = (1-p)\rho(\boldsymbol{x}^{(i)}) + \frac{p\mathbb{I}_{2^N}}{2^N}$, where p is the depolarization rate. Then the quantum kernel can be written as

$$\widetilde{\mathrm{W}} = (1-p)\,\mathrm{W} + \frac{p\mathbb{I}_{2^N}}{2^N}$$

- Consider the finite measurements, the estimated element of quantum kernels yields

$$\widehat{\mathrm{W}}_{ij} = \frac{1}{m}\sum_{k=1}^{m} V_k, \qquad V_k \sim \mathrm{Ber}(\widetilde{\mathrm{W}}_{ij})$$

The power of quantum kernels under the NISQ setting

JDT 京东科技

Under the NISQ setting, we give a generalization error bound for the estimate kernel $\widehat{\mathrm{W}}$.

Theorem 3. *Let the size of training dataset be n and the number of measurements is m. Define $Y = [y_1, \cdots, y_n]^\top$ as the label vector and $c_{\mathrm{W}} = \| \mathrm{W}^{-1} \|_2$. Suppose the system noise is modeled by $\mathcal{N}_p$. With probability at least $1-\delta$, the noisy quantum kernel $\widehat{W}$ can be used to infer a hypothesis $h(\boldsymbol{x})$ with generalization error*

$$\mathbb{E}_{\boldsymbol{x},\widehat{\mathrm{W}}} |h(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \tilde{O}\left(\sqrt{\frac{c_1}{n}} + \sqrt{\frac{1}{c_2}\frac{n}{\sqrt{m}}}\right) \tag{D10}$$

where $c_1 = Y^\top \mathrm{W}^{-1} Y$ *and* $c_2 = \max(c_{\mathrm{W}}^{-2}((\frac{1}{2}\log(\frac{4n^2}{\delta}))^{\frac{1}{2}} + m^{\frac{1}{2}} p\left(1 + \frac{1}{2^{N+1}}\right))^{-1} - \frac{n}{\sqrt{m}} c_{\mathrm{W}}^{-1}, 0)$.

An additional term $\sqrt{n/(c_2\sqrt{m})}$ compared to the bound of Huang et al. has following implications:

- $n/\sqrt{m}$ suggests that an increased number of data n will result in a higher generalization error.
- The performance of quantum kernels in the NISQ era are heavily depends on the number of measurements.
- The term c_2 indicates the negative role of the system noise. Moreover, the generalization error bound would be infinite once $p > 1/(nc_{\mathrm{W}}(1 + 1/2^{N+1}))$.

The power of quantum kernels under the NISQ setting

JDT 京东科技

Under the NISQ setting, we give a generalization error bound for the estimate kernel $\widehat{\mathrm{W}}$.

$$\mathbb{E}_{\boldsymbol{x},\widehat{\mathrm{W}}} |h(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \tilde{O}\left(\sqrt{\frac{c_1}{n}} + \sqrt{\frac{1}{c_2}\frac{n}{\sqrt{m}}}\right) \tag{D10}$$

where $c_1 = Y^\top \mathrm{W}^{-1} Y$ *and* $c_2 = \max(c_{\mathrm{W}}^{-2}((\frac{1}{2}\log(\frac{4n^2}{\delta}))^{\frac{1}{2}} + m^{\frac{1}{2}} p\left(1 + \frac{1}{2^{N+1}}\right))^{-1} - \frac{n}{\sqrt{m}} c_{\mathrm{W}}^{-1}, 0)$.

An additional term $\sqrt{n/(c_2\sqrt{m})}$ compared to the bound of Huang et al. has following implications:

- $n/\sqrt{m}$ suggests that an increased number of data n will result in a higher generalization error;
- The performance of noisy quantum kernels heavily depends on shots m;
- The term c_2 indicates the negative role of the system noise. Moreover, the generalization error bound would be infinite once $p > 1/(nc_{\mathrm{W}}(1 + 1/2^{N+1}))$.

Proof Sketches of Theorem 1

JDT 京东科技

Following the result of Huang et al., with probability $1-\delta/2$, the generalization error of $\widehat{\mathrm{W}}$ yields

$$\mathbb{E}_{\boldsymbol{x},\widehat{\mathrm{W}}}\left|f(\boldsymbol{x})-h(\boldsymbol{x})\right| \leq \widetilde{O}\left(\sqrt{\frac{Y^{\top}\widehat{\mathrm{W}}^{-1}Y}{n}}\right)$$

Note that $\sqrt{Y^{\top}\widehat{\mathrm{W}}^{-1}Y}=\sqrt{Y^{\top}\left(\widehat{\mathrm{W}}^{-1}-\mathrm{W}^{-1}\right)Y+Y^{\top}\mathrm{W}^{-1}Y} \leq \sqrt{\left\|\widehat{\mathrm{W}}^{-1}-\mathrm{W}^{-1}\right\|_2 Y^{\top}Y}+\sqrt{Y^{\top}\mathrm{W}^{-1}Y}$ (3)

and the connection of $\left\|\widehat{\mathrm{W}}^{-1}-\mathrm{W}^{-1}\right\|_2$ and $\left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_2$, that is

$$\left\|\mathrm{W}^{-1}-\widehat{\mathrm{W}}^{-1}\right\|_2 \leq \frac{c_{\mathrm{W}}^2\left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_2}{1-c_{\mathrm{W}}\left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_2} \quad .\left(c_{\mathrm{W}}=\left\|\mathrm{W}^{-1}\right\|_2\right) \tag{4}$$

To achieve $\|\mathrm{W}^{-1}-\widehat{\mathrm{W}}^{-1}\|_2 \leq \delta$, it is sufficient to show $\left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_2 \leq \frac{\delta}{c_{\mathrm{W}}(\delta+c_{\mathrm{W}})}$.

Meanwhile, utilizing the Chernoff-Hoeffding bound and the equation $\widetilde{\mathrm{W}}=(1-p)\,\mathrm{W}+\frac{p\mathbb{I}_{2^N}}{2^N}$ yields

$$\Pr\left(\left|\mathrm{W}_{ij}-\widehat{\mathrm{W}}_{ij}\right| \geq p\left(1+\frac{1}{2^{N+1}}\right)+\frac{\delta}{2}\right) \leq 2\exp\left(\frac{-\delta^2 m}{2}\right). \tag{5}$$

Proof Sketches of Theorem 1

JDT 京东科技

Then a similar concentration inequality in term of the norm $\left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_2$ can be written as

$$\Pr\left(\left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_2 \geq \delta'\right) \leq \Pr\left(\left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_{\mathrm{F}} \geq \delta'\right) \qquad \left(\left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_2 \leq \left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_{\mathrm{F}}\right)$$

$$= \Pr\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\mathrm{W}_{ij}-\widehat{\mathrm{W}}_{ij}\right|^2 \geq \delta'^2\right)$$

$$\leq \Pr\left(\bigcup_{i=1}^{n}\bigcup_{j=1}^{n}\left|\mathrm{W}_{ij}-\widehat{\mathrm{W}}_{ij}\right|^2 \geq \frac{\delta'^2}{n^2}\right) \qquad \text{(union bound)}$$

$$\leq 2n^2\exp\left(-2\left(\frac{\delta'}{n}-p\left(1+\frac{1}{2^{N+1}}\right)\right)^2 m\right) \qquad \text{(using Eqn. (4) and requiring } \delta' > np\left(1+\frac{1}{2^{N+1}}\right)\text{)}$$

Combing this inequality with Eqn. (4), we have

$$\Pr\left(\left\|\mathrm{W}^{-1}-\widehat{\mathrm{W}}^{-1}\right\|_2 \geq \delta\right) \leq \Pr\left(\left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_2 \geq \frac{\delta}{c_{\mathrm{W}}(\delta+c_{\mathrm{W}})}\right) \leq 2n^2\exp\left(-2\left(\frac{\delta}{nc_{\mathrm{W}}(\delta+c_{\mathrm{W}})}-p\left(1+\frac{1}{2^{N+1}}\right)\right)^2 m\right). \qquad (6)$$

Setting the right term in Eqn. (6) as $\delta'/2$ we can get that with probability at least $1-\delta'/2$, (needs to meet non-negative)

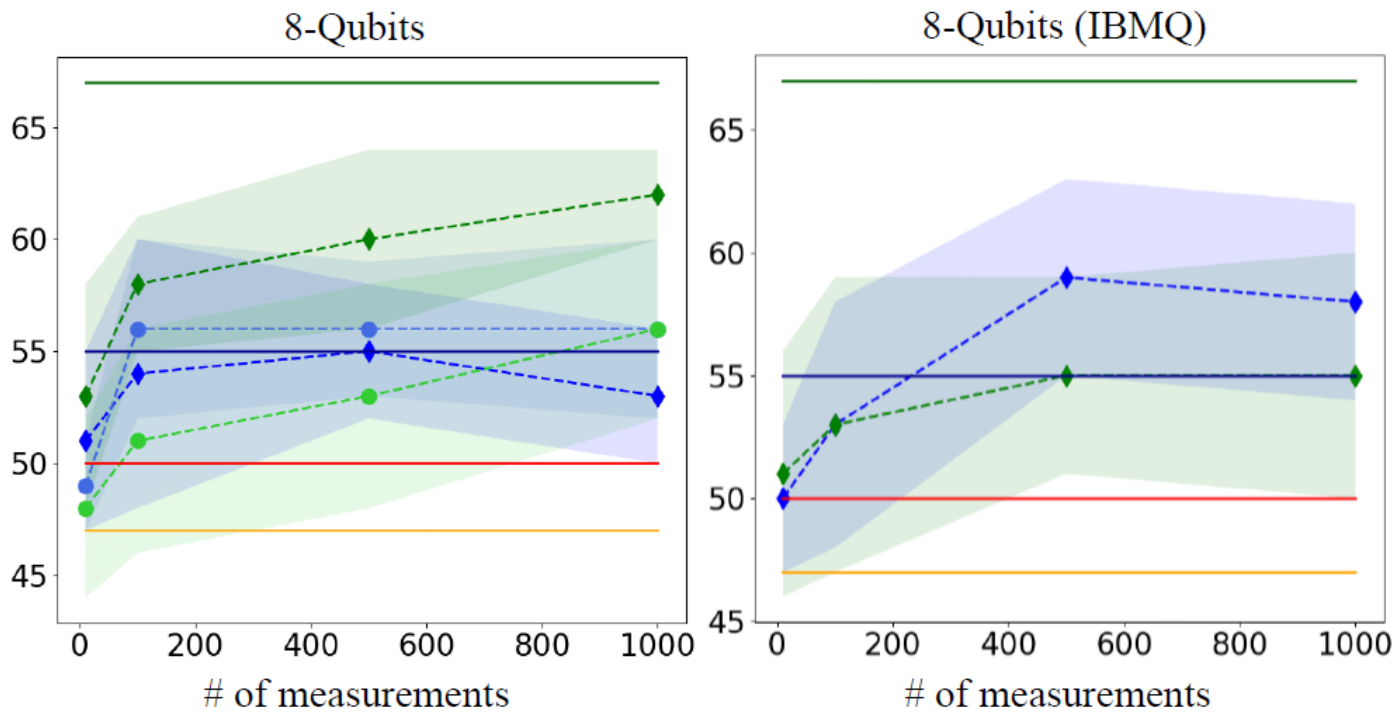
$$\left\|\mathrm{W}^{-1}-\widehat{\mathrm{W}}^{-1}\right\|_2 \leq \max\left(c_{\mathrm{W}}^{-2}\left(\left(\frac{1}{2}\log\left(\frac{4n^2}{\delta}\right)\right)^{\frac{1}{2}}+m^{\frac{1}{2}}p\left(1+\frac{1}{2^{N+1}}\right)\right)^{-1}-\frac{n}{\sqrt{m}}c_{\mathrm{W}}^{-1},0\right).$$

Note that this bound can meet all needed conditions. And this complete the proof.

Numerical evidence

JDT 京东科技

The numerical simulations conducted on the Fashion-Mnist dataset accord with our theoretical results.



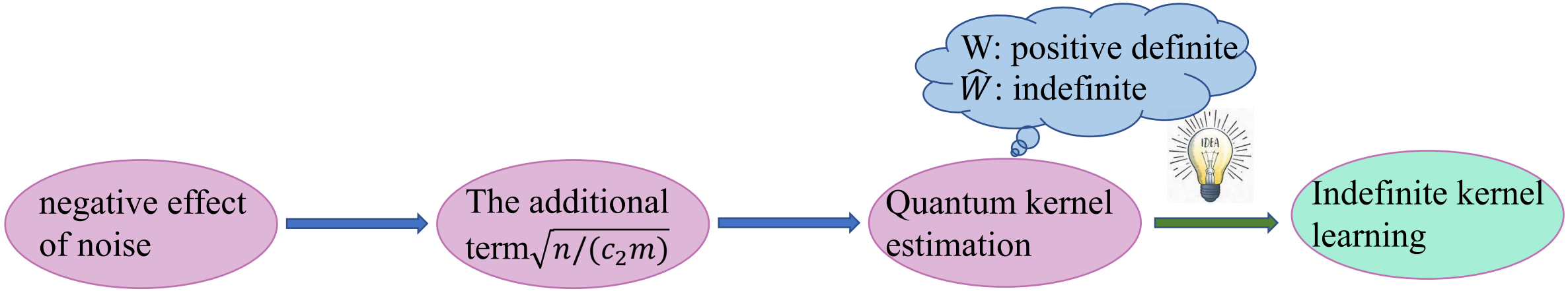
Enhance Performance of Noise Quantum Kernels

JDT 京东科技

how to preserve the superiority of quantum kernels carried on NISQ machines?

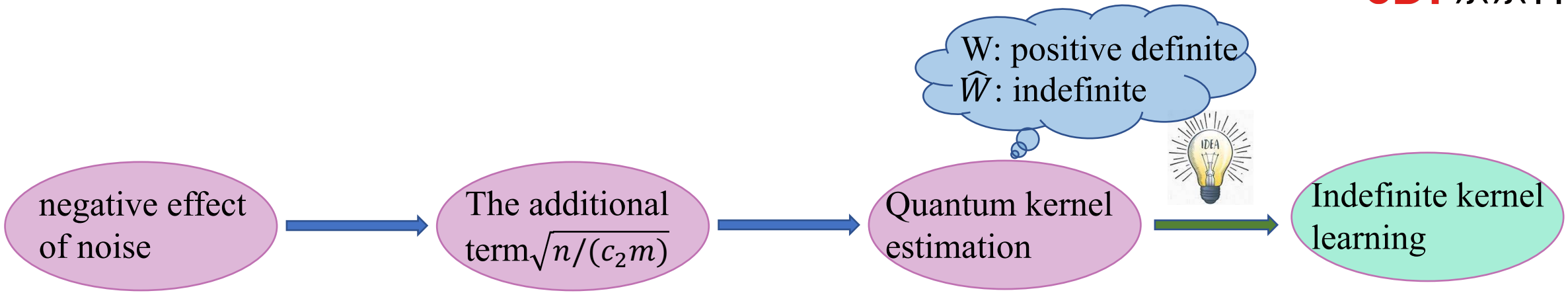
- Slimming the size of training data;
- Suppressing the effects of noise (HOW?)

Tracing back the source of such negative effect can get that



Enhance Performance of Noise Quantum Kernels

JDT 京东科技



Denote the eigenvalues of $\widehat{\mathrm{W}}$ as $\widehat{\lambda}_i$ ($\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_r \geq 0 \geq \cdots \geq \widehat{\lambda}_n$) and the eigenvectors as $\boldsymbol{u}_i$.

Three advanced spectral transformation techniques:

Clip: clipping all negative eigenvalues to zero, i.e., $\widehat{\mathrm{W}}_{\mathrm{c}} = \sum_{i=1}^{r} \widehat{\lambda}_i \boldsymbol{u}_i \boldsymbol{u}_i^{\top}$.

Flip: flipping the sigh of negative eigenvalues, i.e., $\widehat{\mathrm{W}}_{\mathrm{f}} = \sum_{i=1}^{r} \widehat{\lambda}_i \boldsymbol{u}_i \boldsymbol{u}_i^{\top} - \sum_{i=1}^{r} \widehat{\lambda}_i \boldsymbol{u}_i \boldsymbol{u}_i^{\top}$.

Shift: shifting all eigenvalues by a positive constant, i.e., $\widehat{\mathrm{W}}_s = \sum_{i=1}^{n} (\widehat{\lambda}_i - \widehat{\lambda}_n) \boldsymbol{u}_i \boldsymbol{u}_i^{\top}$.

Note that our goal is to shirk the distance between noiseless and noisy quantum kernels.

Enhance Performance of Noise Quantum Kernels

JDT 京东科技

Lemma 1. *Let* W *and* $\widehat{\mathrm{W}}$ *be the ideal and noisy quantum kernel, respectively. Applying the spectral transformation techniques to* $\widehat{\mathrm{W}}$*, the obtained kernel* $\widehat{\mathrm{W}}_\diamond \in \left\{\widehat{\mathrm{W}}_\mathrm{c}, \widehat{\mathrm{W}}_\mathrm{f}, \widehat{\mathrm{W}}_\mathrm{s}\right\}$ *yields*

$$\left\|\mathrm{W}-\widehat{\mathrm{W}}_\diamond\right\|_\mathrm{F} \leq \left\|\mathrm{W}-\widehat{\mathrm{W}}\right\|_\mathrm{F}, \tag{7}$$

where $\|\cdot\|_\mathrm{F}$ *refers to the Frobenius norm.*

Proof sketches: We give a brief proof for the clipping case and the other cases are similar.

Denote $\mathrm{W}=\sum_{i=1}^{n}\lambda_i \boldsymbol{v}_i\boldsymbol{v}_i^\top$, where λ_i refer to the eigenvalues a$\boldsymbol{v}_i$ are the corresponding eigenvectors.

Supported by the definition of the Frobenius norm, an equivalent of achieving Eqn. (6) is

$$\mathrm{Tr}\left(\left(\mathrm{W}-\widehat{\mathrm{W}}_\mathrm{c}\right)^2\right) \leq \mathrm{Tr}\left(\left(\mathrm{W}-\widehat{\mathrm{W}}\right)^2\right) \Longleftrightarrow \mathrm{Tr}\left(\widehat{\mathrm{W}}_\mathrm{c}^2\right)-2\,\mathrm{Tr}\left(\mathrm{W}\widehat{\mathrm{W}}_\mathrm{c}\right) \leq \mathrm{Tr}\left(\widehat{\mathrm{W}}^2\right)-2\,\mathrm{Tr}\left(\mathrm{W}\widehat{\mathrm{W}}\right)$$

Considering $\mathrm{Tr}(\widehat{\mathrm{W}}^2)$, $\mathrm{Tr}(\widehat{\mathrm{W}}^2)=\sum_{i=1}^{n}\hat{\lambda}_i^2 \geq \sum_{i=1}^{r}\hat{\lambda}_i^2 = \mathrm{Tr}(\widehat{\mathrm{W}}_\mathrm{c}^2)$. (8)

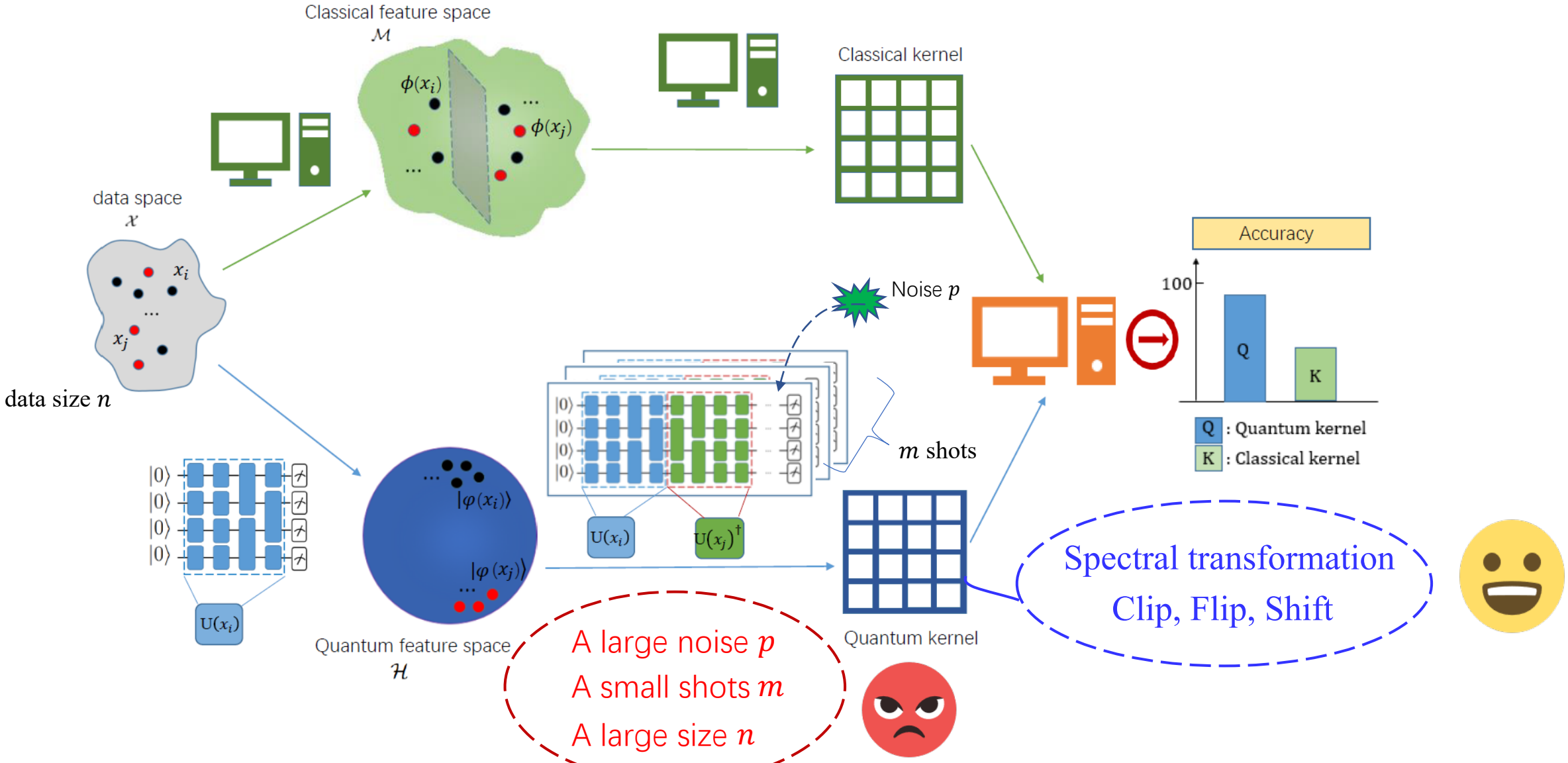
Considering $\mathrm{Tr}(\mathrm{W}\,\widehat{\mathrm{W}})$,

$$\mathrm{Tr}(\mathrm{W}\,\widehat{\mathrm{W}}) = \mathrm{Tr}\left(\left(\sum_{i=1}^{n}\lambda_i\boldsymbol{v}_i\boldsymbol{v}_i^\top\right)\left(\sum_{j=1}^{r}\hat{\lambda}_j\boldsymbol{u}_j\boldsymbol{u}_j^\top+\sum_{j=r+1}^{n}\hat{\lambda}_j\boldsymbol{u}_j\boldsymbol{u}_j^\top\right)\right) = \mathrm{Tr}\left(\mathrm{W}\,\widehat{\mathrm{W}}_\mathrm{c}\right)+\sum_{i=1}^{n}\sum_{j=r+1}^{n}\lambda_i\hat{\lambda}_j\left(\boldsymbol{u}_j^\top\boldsymbol{v}_i\right)^2 \leq \mathrm{Tr}\left(\mathrm{W}\,\widehat{\mathrm{W}}_\mathrm{c}\right). \tag{9}$$

In conjunction with Eqn. (8) and (9), we can achieve Eqn (7).

The power of noisy quantum kernels

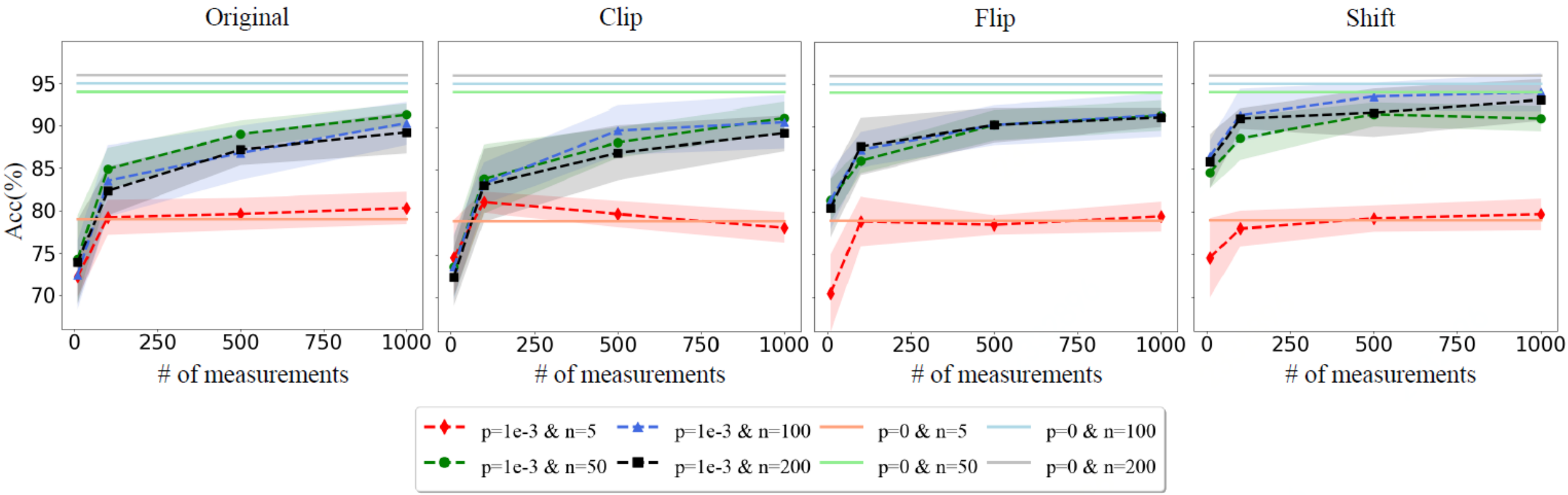
JDT 京东科技



Enhance Performance of Noise Quantum Kernels

JDT 京东科技

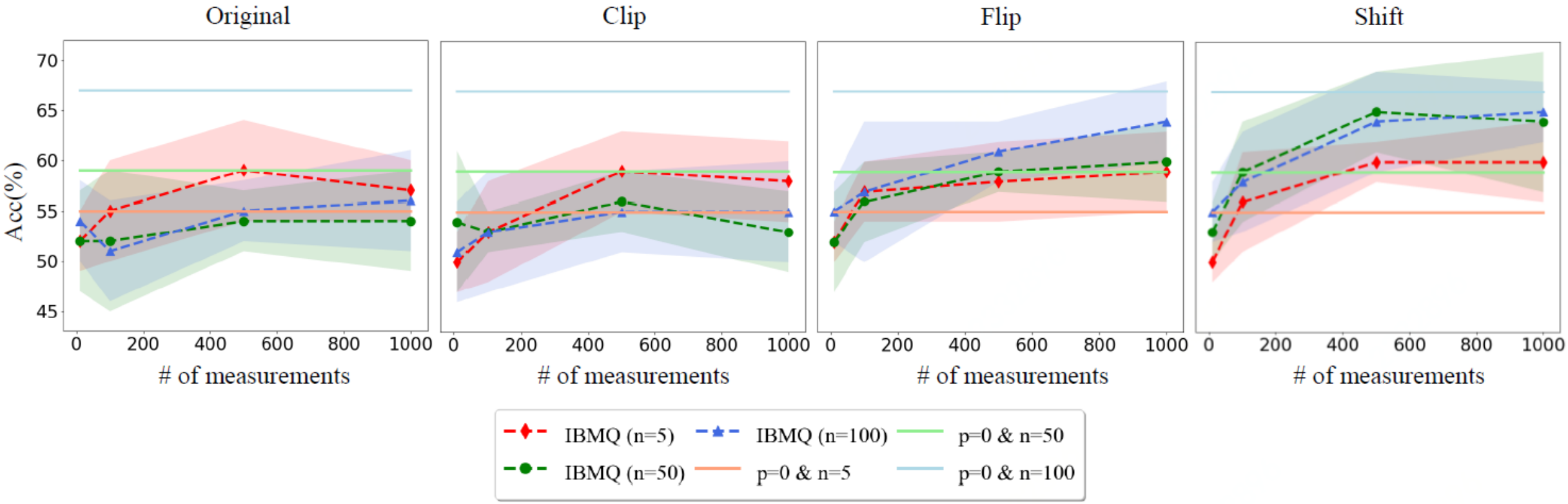
Numerical simulations achieved with the depolarization noise model



Enhance Performance of Noise Quantum Kernels

JDT 京东科技

Numerical simulations achieved on IBMQ-Melbourne's noisy settings.



JDT 京东科技

Thank You!