Problem-Dependent Power of Quantum Neural Networks on Multi-Class Classification

Yuxuan Du, Yibo Yang, Min-hsiu Hsieh, Dacheng Tao

Phys. Rev. Lett. **131**, 140601

Quantum neural networks in multi-class classification

Why exploring multi-class classification tasks are important?

Many fundamental problems in classical and quantum machine learning can be categorized to classifications.

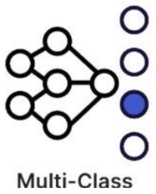
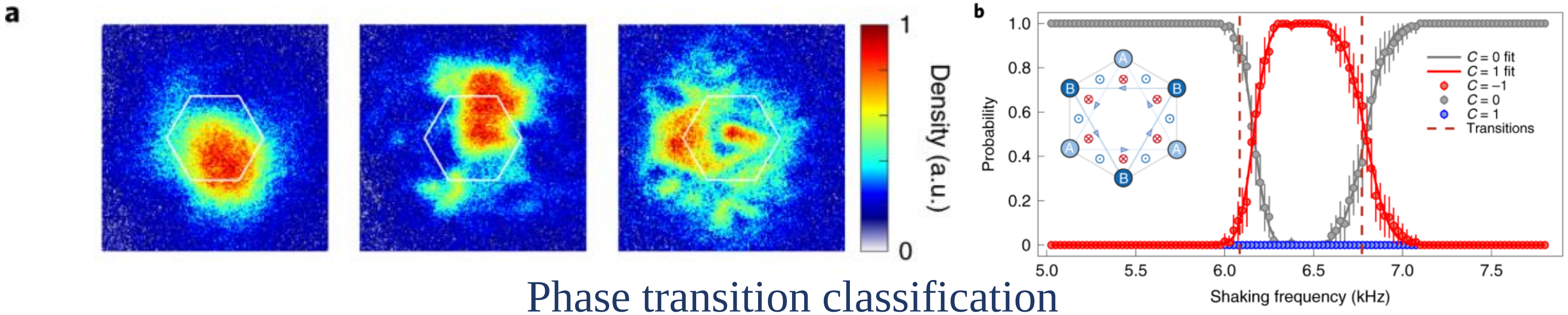


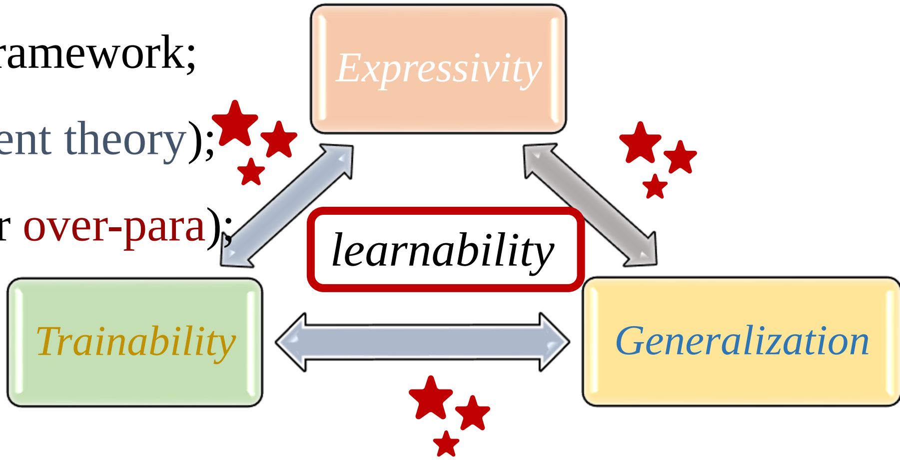
Image classification



Phase transition classification

Our main contributions:

- ✓ Unify the trainability, generalization, and expressivity into a single and general framework;
- ✓ Interpret the trainability of QCs from a geometric view (link quantum measurement theory);
- ✓ Derive a non-vacuous generalization error bound of quantum classifiers (even for over-para);
- ✓ Unravel trainability is more deterministic than generalization of QCs;
- ✓ Show disparate risk curves between classical classifiers and quantum classifiers;
- ✓ Devise an efficient method to examine the power of QCs.



Rem, B.S., Käming, N., Tarnowski, M. et al. Identifying quantum phase transitions using artificial neural networks on experimental data. Nat. Phys. ***15****, 917–920 (2019).*

Quantum neural networks in multi-class classification

Problem setup of K-class classification ($K \geq 2$) [applied to both classical and quantum classifiers]

Notations

Input data space: $\mathcal{X}$;

Label space: $\mathcal{Y} = \{1, 2, \dots, K\}$;

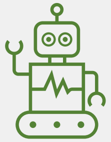
Train set: $\mathcal{D} = \cup_{k=1}^{K} \left\{\left(x^{(i,k)}, y^{(i,k)}\right)\right\}_{i=1}^{n_k}$, each example drawn i.i.d from an unknown distribution $\mathbb{D}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$;

The i-th example in the k-th class: $(x^{(i,k)}, y^{(i,k)})$;

Balanced dataset: $n_1 = n_2 = \cdots = n_k = \cdots n_K$, total number of training data points is $|\mathcal{D}| = n = K \cdot n_k$;

Intuition

An algorithm $\mathcal{A}$ aims to use $\mathcal{D}$ to infer a hypothesis (i.e., classifier) $h_{\mathcal{A}_\mathcal{D}}: \mathcal{X} \to \mathbb{R}^K$ from hypothesis space $\mathcal{H}$ to accurately separate examples sampling from $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ to the corresponding classes.



Quantum neural networks in multi-class classification

Problem setup of K-class classification ($K \geq 2$) [applied to both classical and quantum classifiers]

Definitions

The optimal hypothesis refers to $h^* = \min_{h \in \mathcal{H}} \mathfrak{R}(h)$, where $\mathfrak{R}(h) \coloneqq \mathbb{E}_{(x,y)\sim\mathbb{D}}[\ell\,(h(x), y)]$ is the *expected risk* of h;

$\ell(\cdot,\cdot)$ is the per-sample loss and we specify it to be mean square loss, i.e., $\ell(\mathrm{a}, \mathrm{b}) = ||\,\mathrm{a} - \mathrm{b}||_2^2$;

Since $\mathbb{D}$ is unknown, we approach $\mathfrak{R}(h)$ via empirical risk $\mathfrak{R}_{\mathrm{ERM}}$ by learning an empirical classifier $\hat{h} \in \mathcal{H}$ with

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \mathcal{L}(\boldsymbol{h}, \mathcal{D}) \coloneqq \frac{1}{n} \sum_{i=1,k=1}^{\mathrm{n}_c, K} l\left(y^{(i,k)}, \hat{y}^{(i,k)}\right) + \mathfrak{E}(\boldsymbol{h}).$$

Quantum neural networks in multi-class classification

Problem setup of K-class classification ($K \geq 2$) [applied to both classical and quantum classifiers]

The fundamental role of the expected risk → Use expected risk to explore quantum advantages:

When the algorithm $\mathcal{A}$ is *classical classifier* (CC), e.g., deep neural networks, the learned empirical classifier is denoted by $\hat{h}_C$;

When the algorithm $\mathcal{A}$ is *quantum classifier* (QC), e.g., QNNs, the learned empirical classifier is denoted by $\hat{h}_Q$;
The quantum advantage is verified when

$$\mathfrak{R}(\hat{h}_Q) < \mathfrak{R}(\hat{h}_C).$$

Reformulation of the expected risk

Recall the definition of expected risk is $\mathfrak{R}(h) \coloneqq \mathbb{E}_{(x,y)\sim\mathbb{D}}[\ell\,(h(x), y)]$, which is intractable. As such, we rewrite it as

$$\mathfrak{R}(h) = \mathfrak{R}_{\mathrm{ERM}}(h) + \mathfrak{R}_{\mathrm{Gene}}(h).$$

$\mathfrak{R}_{\mathrm{ERM}}$ → Trainability $\mathfrak{R}_{\mathrm{Gene}}$ → Generalization

Quantum neural networks in multi-class classification

Problem setup of K-class classification ($K \geq 2$) [applied to both classical and quantum classifiers]

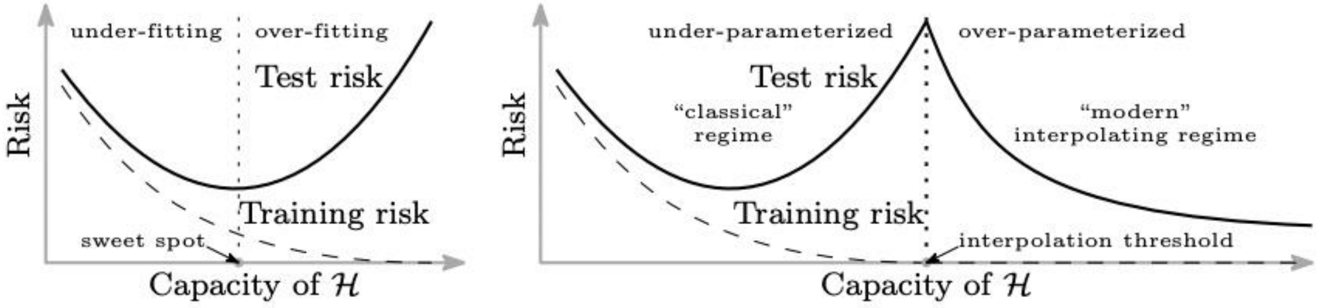
Reformulation of the expected risk

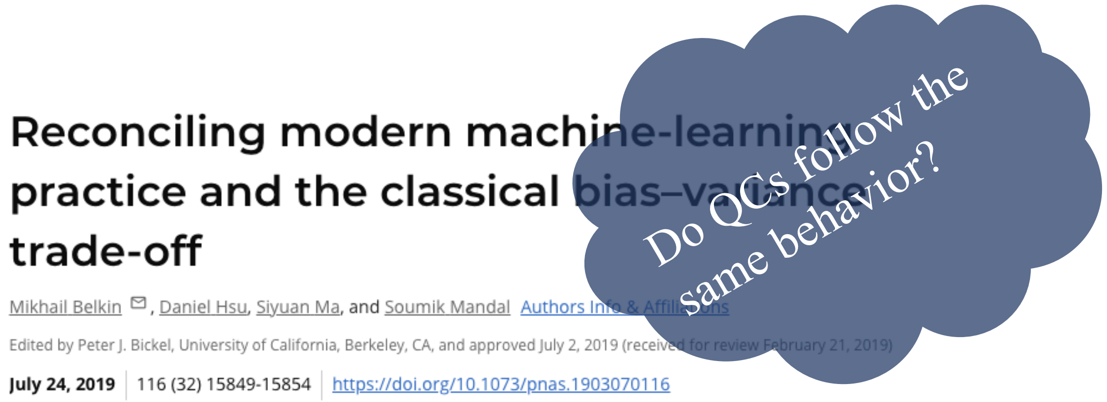
Recall the definition of expected risk is $\mathfrak{R}(h) \coloneqq \mathbb{E}_{(x,y)\sim\mathbb{D}}[\ell\,(h(x), y)]$, which is intractable. As such, we rewrite it as

$$\mathfrak{R}(h) = \mathfrak{R}_{\mathrm{ERM}}(h) + \mathfrak{R}_{\mathrm{Gene}}(h).$$

Trainability   Generalization

Risk curve of CCs is double-descent




Reconciling modern machine-learning practice and the classical bias–variance trade-off

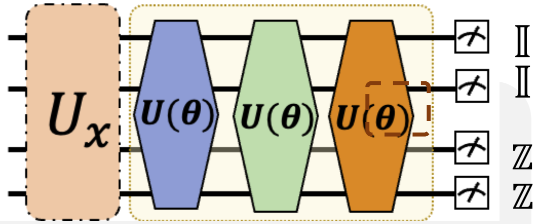
Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal Authors Info & Affiliations

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved July 2, 2019 (received for review February 21, 2019)

July 24, 2019 | 116 (32) 15849-15854 | https://doi.org/10.1073/pnas.1903070116

Quantum neural networks in multi-class classification

Problem setup of K-class classification ($K \geq 2$) [applied to both classical and quantum classifiers]



Mathematical formulation of QCs

Empirical loss to be minimized: $\hat{h}_Q = \arg\min_{h_Q \in \mathcal{H}_Q} \mathcal{L}(h_Q, \mathcal{D}) \coloneqq \frac{1}{n}\sum_{i=1,k=1}^{n_c,K} l\left(y^{(i,k)}, \hat{y}^{(i,k)}\right) + \mathfrak{E}\left(h_Q\right).$

The hypothesis space for an N-qubit QC is $\mathcal{H}_Q = \left\{\left[h_Q\left(\cdot, U(\theta), O^{(k')}\right]_{k'=1:K}\right| \theta \in \Theta\right\}$

- $\hat{y}^{(i,k)} = [h_Q\left(x^{(i,k)}, U(\theta), O^{(k')}\right]_{k'=1:K}, where\ [\cdot]_{k=1:K}$ is a K−dim vector;
- The k-th entry $h_Q\left(x, \mathrm{U}(\theta), \mathrm{O}^{(k)}\right) = \mathrm{Tr}(O^{(k)}U(\theta)\sigma(x)U(\theta)^{\top})$ refers to the prediction of QC for the k-th label;
- $\sigma(x)$ is the input state of x;
- $U(\theta)$ is the ansatz, i.e., $U(\theta) = \prod_{l=1}^{N_t} u_l(\theta) u_e \in \mathcal{U}(2^N)$ with $u_l \in \mathcal{U}(2^m)\ \forall l \in [N_t]$ operated on at most m-qubits;
- $\boldsymbol{O} = \left\{O^{(k)}\right\}_{k=1}^{K}$ is the set of measurement operators.

Quantum neural networks in multi-class classification

Problem setup of K-class classification ($K \geq 2$) [applied to both classical and quantum classifiers]

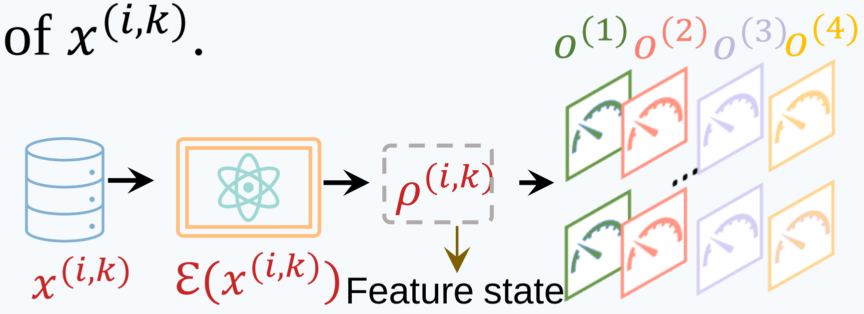
Unification of QCs

The diverse choice of ansatz, encodings, and measurement operators challenges the analysis. Fortunately, we can design a unified model to cover all these diversities.

$$h_Q\left(x^{(i,k)}, \mathrm{U}(\theta), \mathrm{O}^{(k)}\right) \to h_Q\left(\rho^{(i,k)}, o^{(k)}\right) \equiv \mathrm{Tr}\left(\rho^{(i,k)} o^{(k)}\right), \forall k \in [K]$$

$O^{(k)} = \mathbb{I}_{2^{N-D}} \otimes o^{(k)}$ with $o^{(k)}$ being non-trivial operator on 2^D-dim system;

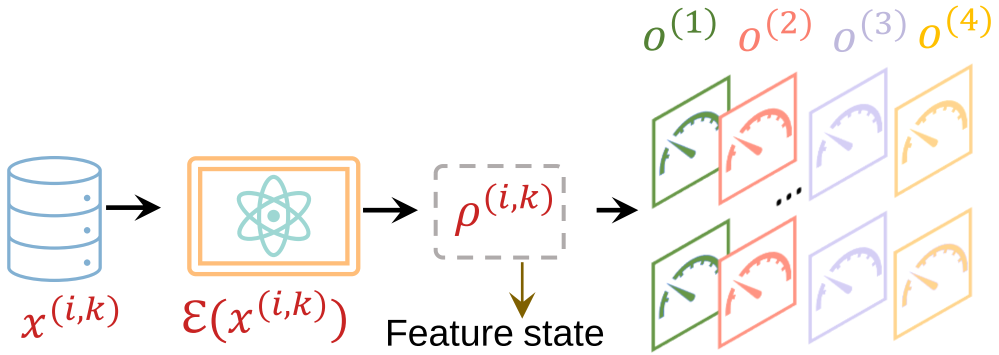
$\rho^{(i,k)} = \mathrm{Tr}_D\left(U(\theta)\sigma\left(x^{(i,k)}\right)U(\theta)^\top\right) \in \mathbb{C}^{2^D \times 2^D}$ is the *feature state* of $x^{(i,k)}$.

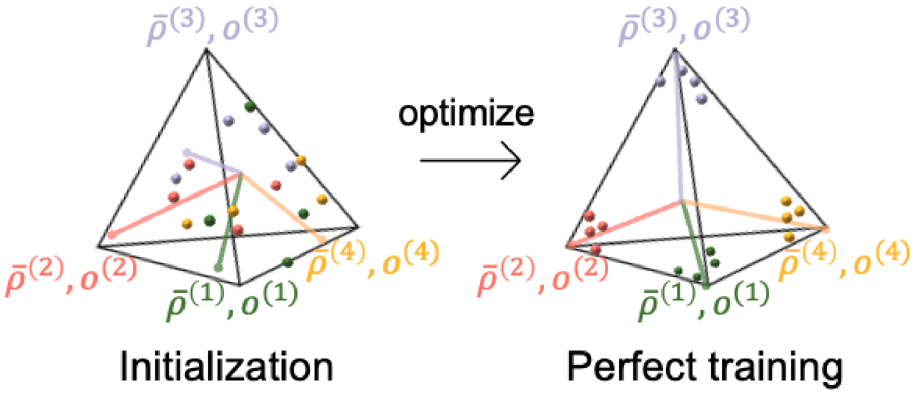


Quantum neural networks in multi-class classification

Results of QCs for K-class classification ($K \geq 2$) [applied to both classical and quantum classifiers]

Theorem 1 (informal). *Following the above notations, when the train data size is* $n \sim O(KN_{ge}\log\frac{KN_{ge}}{\epsilon\delta})$ *with* ϵ *being the tolerable error, and the optimal sets of* $\boldsymbol{\rho}^*$ *and* $\boldsymbol{o}^*$ *satisfy three conditions: (i) the feature states have the vanished variability in the same class; (ii) all feature states are equal length and are orthogonal in the varied classes; (iii) any feature state is alignment with the measure operator in the same class, with probability 1−*δ*, the expected risk of QC tends to be zero, i.e.,* $\mathfrak{R}(\hat{h}_Q) \to 0$.

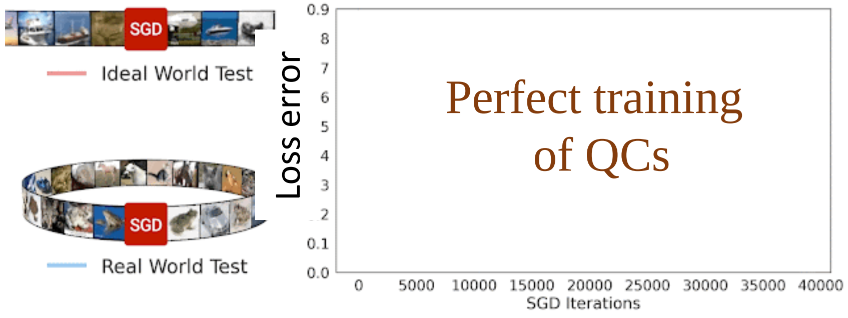


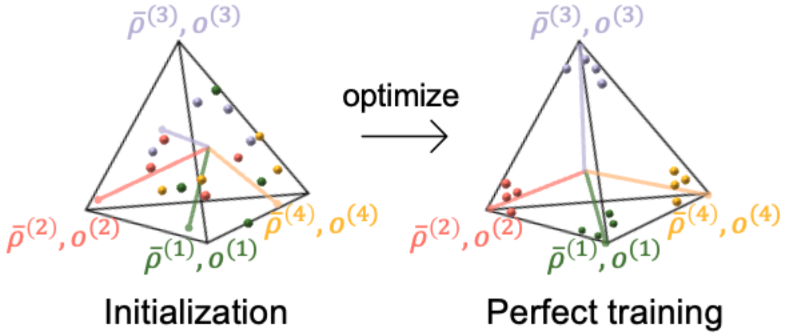


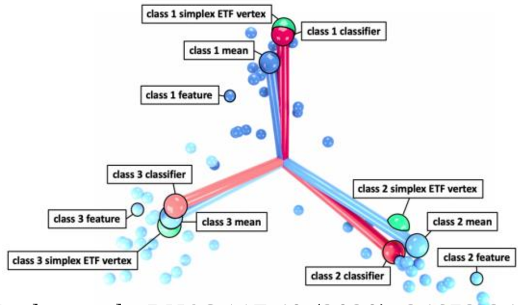
Quantum neural networks in multi-class classification

Results of QCs for K-class classification ($K \geq 2$) [applied to both classical and quantum classifiers]

Theorem 1 (informal). *Following the above notations, when the train data size is $n \sim O(KN_{ge}\log\frac{KN_{ge}}{\epsilon\delta})$ with ϵ being the tolerable error, and the optimal sets of $\boldsymbol{\rho}^*$ and $\mathbf{o}^*$ satisfy three conditions: (i) the feature states have the vanished variability in the same class; (ii) all feature states are equal length and are orthogonal in the varied classes; (iii) any feature state is alignment with the measure operator in the same class, with probability $1-\delta$, the expected risk of QC tends to be zero, i.e., $\Re(\hat{h}_Q) \rightarrow 0$.*







Papyan, Vardan et.al . *PNAS* 117.40 (2020): 24652-24663

Implications:

1. The scaling $n \sim O(KN_{ge}\log\frac{KN_{ge}}{\epsilon\delta})$ shows that a low generalization error $\Re_{\text{Gene}}(\hat{\text{h}}_{\text{Q}})$ can be achieved using few training data;
2. Conditions (i)-(iii) sculpt the geometric interpretations of the set of feature states $\boldsymbol{\rho}^*$ and the set of local observables $\mathbf{o}^*$ to achieve $\Re(\hat{h}_Q) \rightarrow 0$, i.e., $\Re_{\text{ERM}}(\hat{h}_Q) \rightarrow 0$ (when *perfect training* happens):
   - Classical view: link with *neural collapse* (Condition I + II);
   - Quantum view: connect with quantum state discrimination, i.e., for any two varied classes, $\mathbf{o}^{*(k)}$ and $\mathbf{o}^{*(k')}$ classifies $\bar{\rho}^{(k)}$ and $\bar{\rho}^{(k')}$ with prob 1 [maximize the Helstrom bound] (Condition I + II + III).

Quantum neural networks in multi-class classification

Results of QCs for K-class classification ($K \geq 2$) [applied to both classical and quantum classifiers]

Theorem 1 also suggests the empirical risk dominates the expected risk, i.e., $\mathfrak{R}(h) = \mathfrak{R}_{\mathrm{ERM}}(h) + \mathfrak{R}_{\mathrm{Gene}}(h)$, since satisfying the number of training data is easy but satisfying Conditions (i)-(iii) is *challenging* and problem-dependent.

Given a QC, its abilities and limitations can be quantified by examining whether the three conditions can be fulfilled. The following Corollary shows the fundamental limitations of over-parameterized QCs.

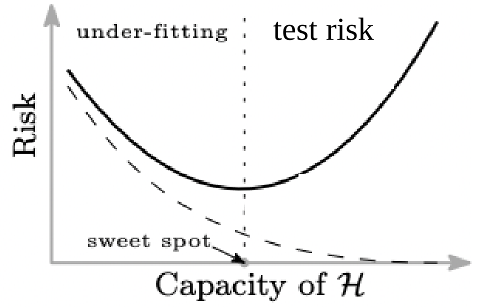
Corollary 1. *Given a QC, when its encoding unitary* $\{U_E(x)|x \in \mathcal{X}\}$ *follows 2-design, with probability* $1-\delta$*, the empirical QC follows* $\left|\mathrm{Tr}\left(\sigma\left(x^{(i,k)}\right) - \sigma(x)\right) - \frac{1}{2^N}\right| \leq \sqrt{3/(2^{2N}\delta)}$. *When its ansatz* $\{U(\theta)|\theta \in \Theta\}$ *follows 2-design, with probability* $1-\delta$*, the empirical QC follows* $\left|\mathrm{Tr}\left(o^{(k)}o^{(k')}\right) - \frac{\mathrm{Tr}(o^{(k')})}{2^D}\right| \leq \sqrt{\frac{\mathrm{Tr}\left(o^{(k')}\right)^2 + 2\mathrm{Tr}\left(\left(o^{(k')}\right)^2\right)}{4^D\delta}}$.

Proof relies on the results of concentration of measure and unitary t-design.

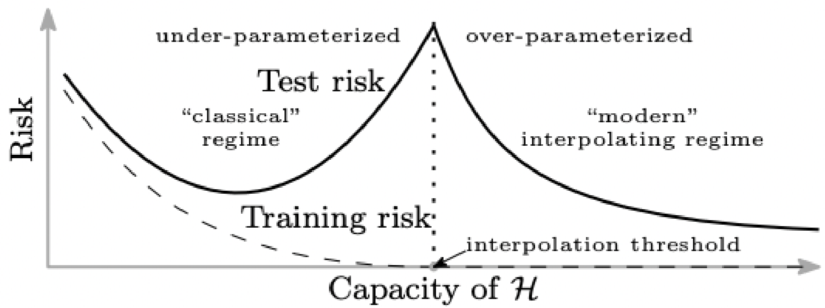
Over-parameterized $U(\theta)$ collapses Condition (iii)

No-perfect training for over-QCs & $\mathfrak{R}(h_Q) > 0$.

QCs: U-shape risk curve



CCs: double-descent risk curve

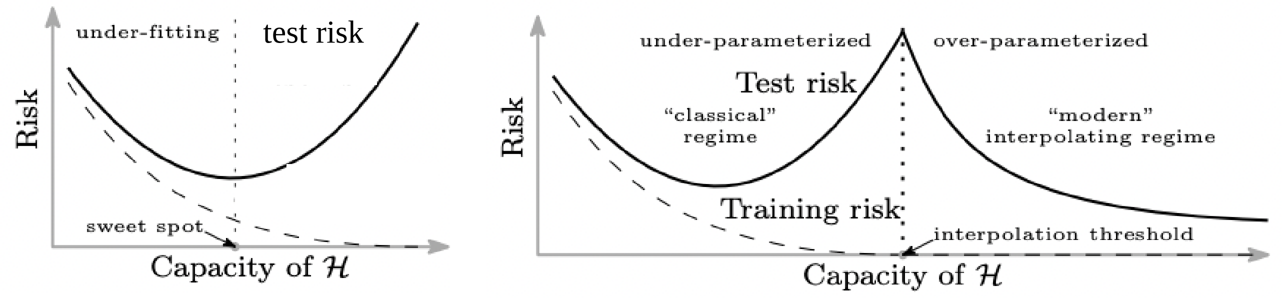


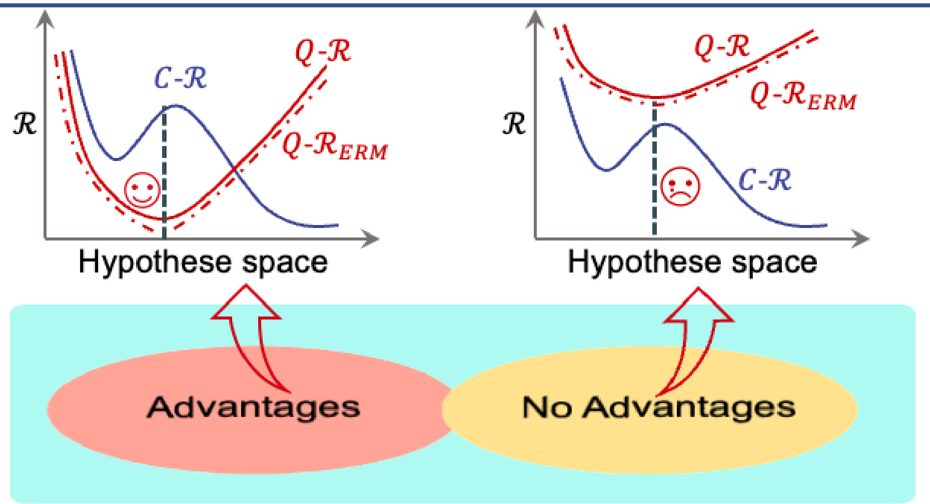
Quantum neural networks in multi-class classification

Results of QCs for K-class classification ($K \geq 2$) [applied to both classical and quantum data]

No-perfect training for over-QCs vs perfect training of over-CCs.

QCs: U-shape risk curve   CCs: double-descent risk curve





Separation of VQAs in optimization and learning

Over-parameterization is the key of using VQAs to estimate the target result of optimization tasks (e.g., VQEs), but it forbids the optimility in learning (e.g., QCs);

A general method to achieve perfect training in learning is unknown.

For *learning*, the potential quantum advantages of QCs may posit in the regime with *the modest hypothesis space* $\mathcal{H}_Q$.

Q: how to recognize potential quantum advantages?

A: The observation in which the empirical risk dominates the expected risk of QCs allows an efficient method to probe power of QCs by *fitting loss dynamics* [Alg. 1, Page 20].

Quantum neural networks in multi-class classification

Numerical simulation results (binary classification for parity dataset)

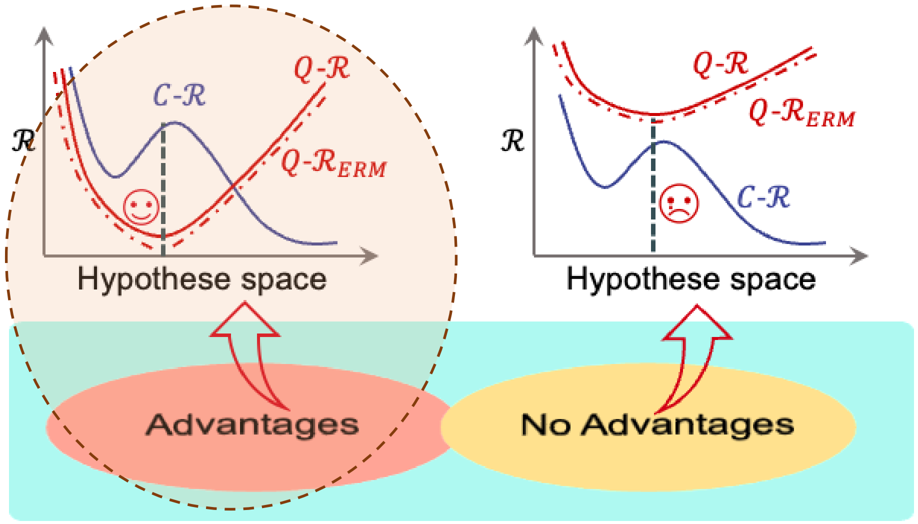
Dataset: $\mathcal{X} = \{0,1\}^6,\ \mathcal{Y} = \{0,1\},\ \mathcal{D} = \{x^{(i,k)}, y^{(i,k)}\},\ |\mathcal{D}| = n = 48$;

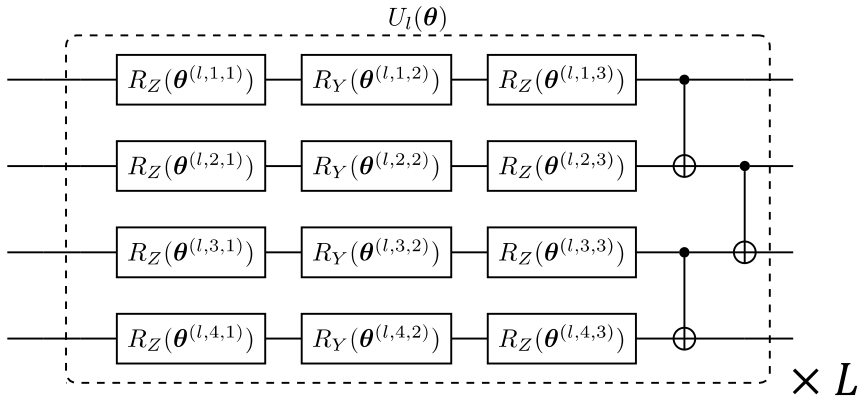
Labels: if the number of `0' in $x^{(i,k)}$ is even, $y^{(i,k)}$=0; otherwise, $y^{(i,k)}$=1.

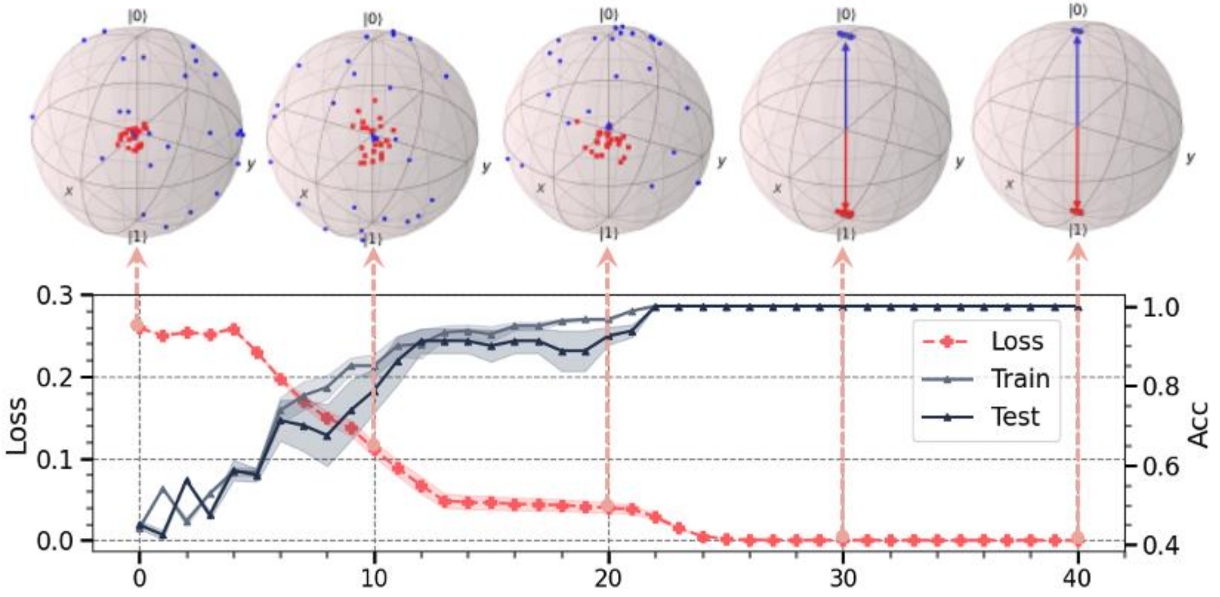
Number of qubits: N = 6 (basis encoding)

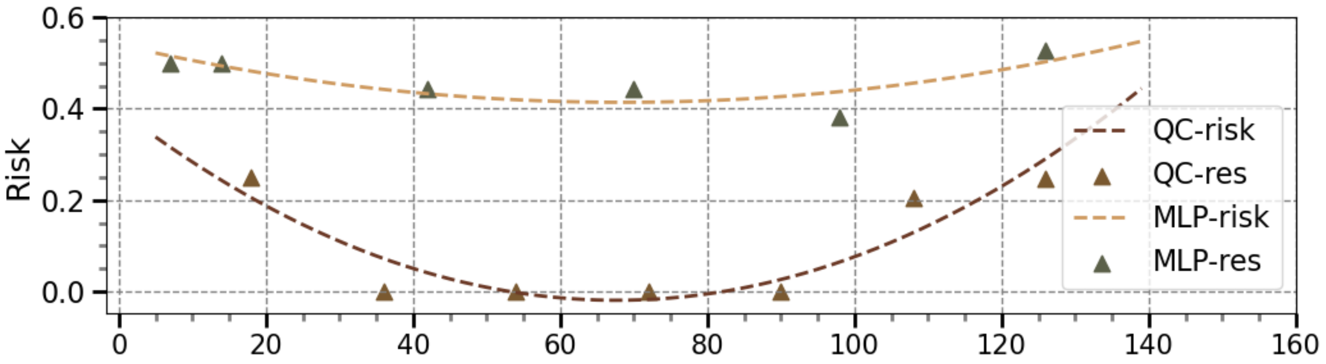
Ansatz $U(\theta)$: Hardware-efficient ansatz $U(\theta) = \prod_{l=1}^{L} U_l(\theta)$ with the varied number of L (varied hypothesis space);

Measurements: $o^{(k=1)} = |0\rangle\langle 0|$ and $o^{(k=2)} = |1\rangle\langle 1|$.









Quantum neural networks in multi-class classification

Numerical simulation results (9-class classification for Image dataset)

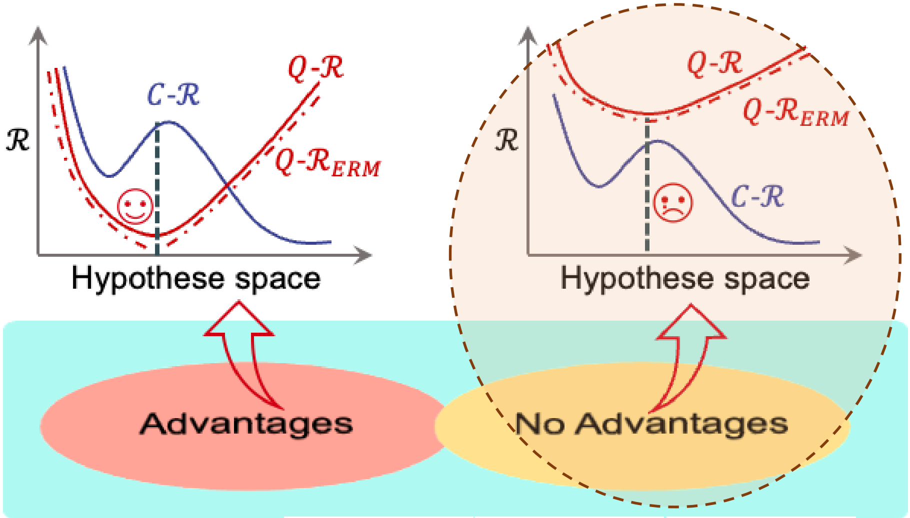
Dataset: Fashion-MNIST with the first 9 classes ($\mathcal{X} = \mathbb{R}^{28\times 28}$, $\mathcal{Y} = \{1, \ldots, 9\}$, $\mathcal{D} = \{x^{(i,k)}, y^{(i,k)}\}$, $|\mathcal{D}| = n = 180$);

Number of qubits: N = 10 (amplitude encoding with padding)

Ansatz $U(\theta)$: Hardware-efficient ansatz $U(\theta) = \prod_{l=1}^{L} U_l(\theta)$ with the varied number of L (varied hypothesis space);

Measurements: Pauli-based measurements on three qubits

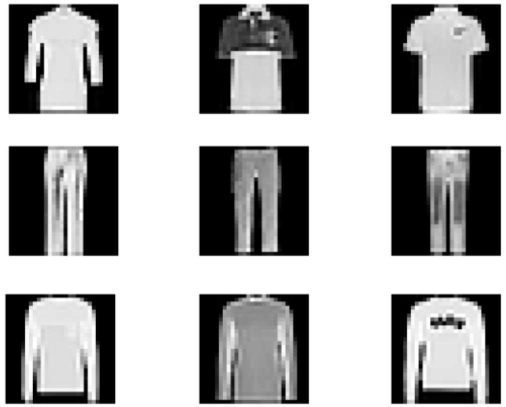
Epochs: 50; Optimizer: SGD; $L \in [25, 100]$;

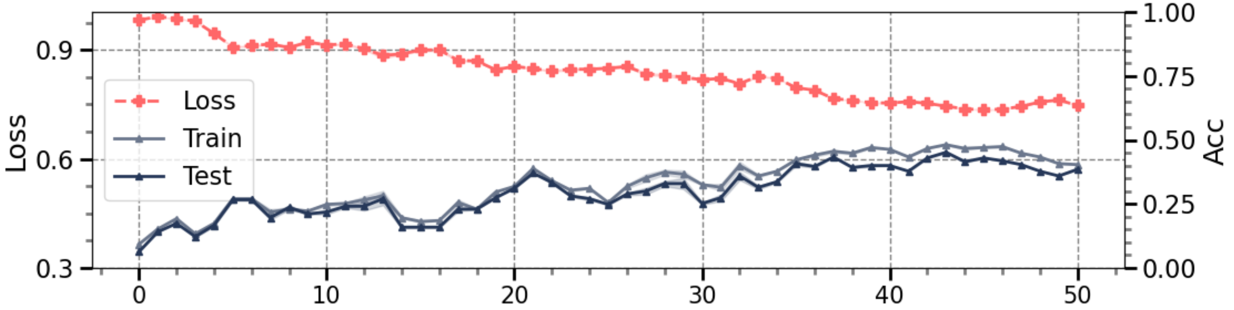


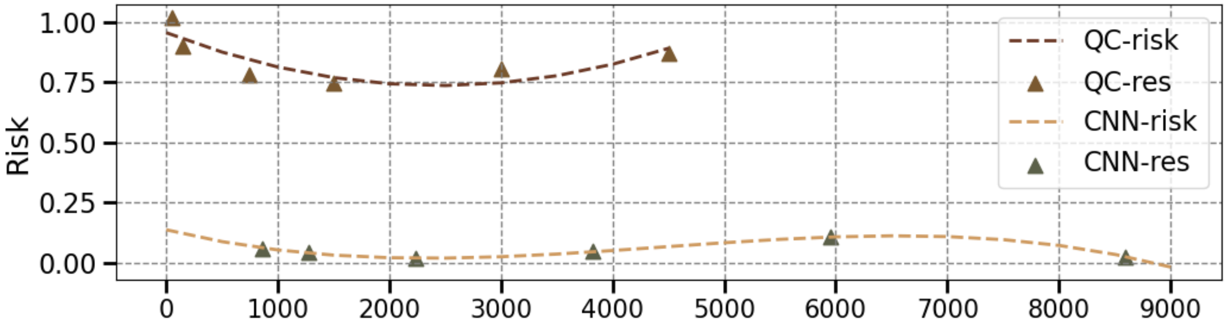
Class 1:

Class 2:

Class 3:







Outlook

- The demystified U-shape risk curve of current QCs pushes in the stage of creating new quantum learning models that can effectively learn classical data while providing proven benefits, especially the ability of *perfect training*.

- Is nonlinearity necessary for QCs? Moreover, is double-descent risk curve necessary for QCs to gain computational advantages in learning classical data? If necessary, how to design these nonlinear QCs?

- Are current QCs sufficient to gain computational advantages in learning *quantum data*? If so, how to prove these advantages theoretically?

Reference list:

- *Du, Yuxuan, et al. "**Demystify Problem-Dependent Power of Quantum Neural Networks on Multi-Class Classification**." arXiv:2301.01597 (2022).*
- *Du, Yuxuan, et al. "Learnability of quantum neural networks." PRX Quantum 2.4 (2021): 040337.*
- *Du, Yuxuan, et al. "Efficient Measure for the Expressivity of Variational Quantum Algorithms." Physical Review Letters 128.8 (2022): 080506.*
- *Wang, X., Liu, J., Liu, T., Luo, Y., **Du, Y.**, & Tao, D. (2022). Symmetric Pruning in Quantum Neural Networks. International conference learning representation (ICLR) 2023 (Spotlight) [arXiv preprint arXiv:2208.14057];*
- *Qian, Y., Wang, X., **Du, Y.**, Wu, X., & Tao, D. (2022). The dilemma of quantum neural networks. IEEE Transactions on Neural Networks and Learning Systems.*

- *Arrasmith, Andrew, et al. "Effect of barren plateaus on gradient-free optimization." Quantum 5 (2021): 558.*
- *Cerezo, Marco, et al. "Cost function dependent barren plateaus in shallow parametrized quantum circuits." Nature communications 12.1 (2021): 1-12.*
- *Caro, Matthias C., et al. "Encoding-dependent generalization bounds for parametrized quantum circuits." Quantum 5 (2021): 582.*
- *Caro, Matthias C., Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles. "Generalization in quantum machine learning from few training data." arXiv preprint arXiv:2111.05292 (2021).*
- *Cerezo M, Arrasmith A, Babbush R, et al. Variational quantum algorithms[J]. Nature Reviews Physics, 2021, 3(9): 625-644.*
- *Schreiber F J, Eisert J, Meyer J J. Classical surrogates for quantum learning models[J]. arXiv preprint arXiv:2206.11740, 2022.*
- *Shirai, N., Kubo, K., Mitarai, K., & Fujii, K. (2021). Quantum tangent kernel. arXiv preprint arXiv:2111.02951.*
- *Liu, J., Najafi, K., Sharma, K., Tacchino, F., Jiang, L., & Mezzacapo, A. (2022). An analytic theory for the dynamics of wide quantum neural networks. arXiv preprint arXiv:2203.16711.*

Thank You!

Contact: duyuxuan123@gmail.com

https://yuxuan-du.github.io/