# Understanding Generalization in Quantum Machine Learning with Margins

ICML 2025

**Tak Hur (w/ Daniel K. Park)**
Yonsei University
Department of Statistics and Data Science

# (Quantum) Supervised Learning: Classification

Data: $\rho \in \mathbb{C}^{2^n \times 2^n}$ n-qubit quantum states

Label: $y \in \mathbb{R}$ $y \in \{1,2,\ldots,k\}$ for $k$-class classification

Unknown probability distribution $(\rho, y) \sim \mathscr{D}$,

# (Quantum) Supervised Learning: Classification

Data: $\rho \in \mathbb{C}^{2^n \times 2^n}$ n-qubit quantum states

Label: $y \in \mathbb{R}$ $y \in \{1,2,\ldots,k\}$ for $k$-class classification

Unknown probability distribution $(\rho, y) \sim \mathscr{D}$,

Goal:

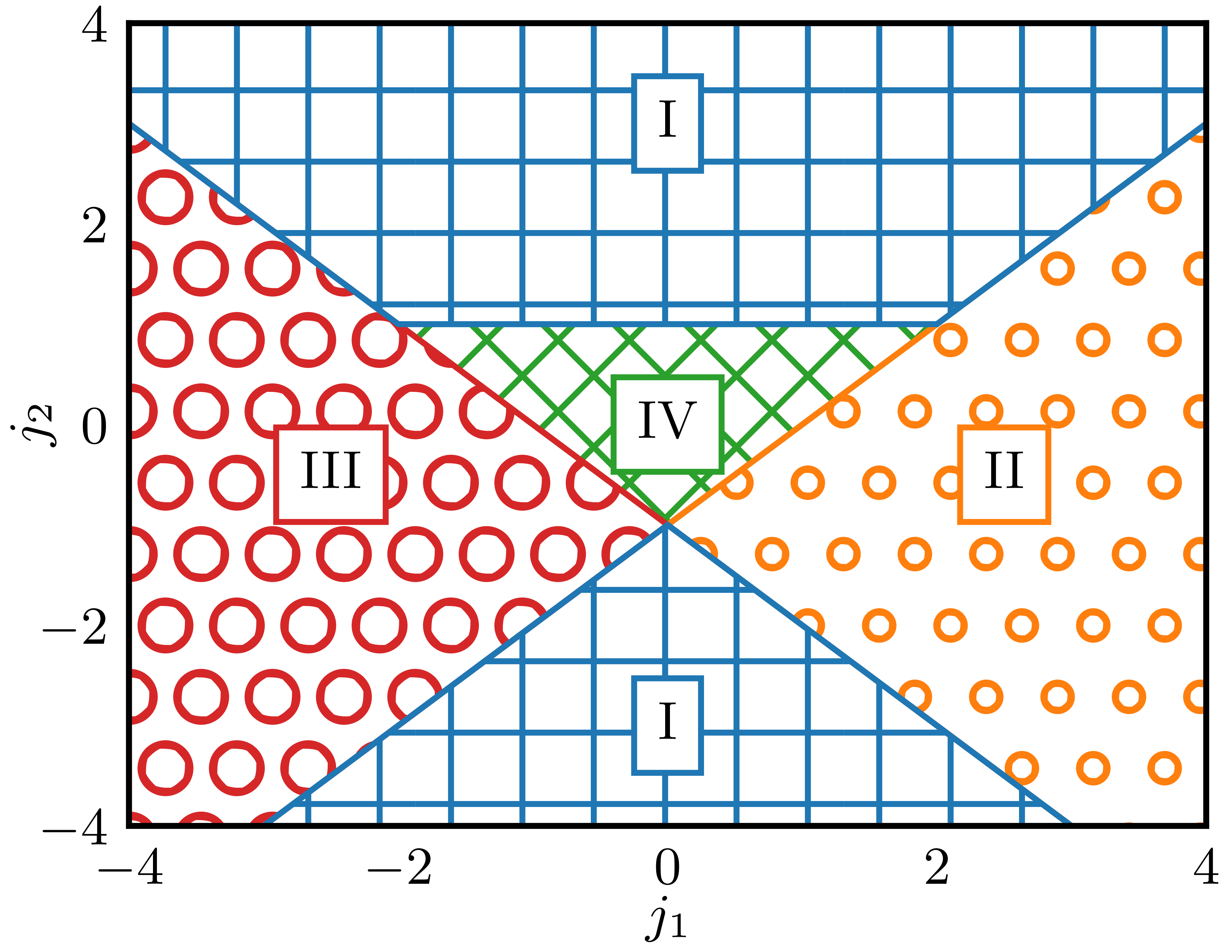
Find $h : \mathbb{C}^{2^n \times 2^n} \mapsto \mathbb{R}$ with small true error $R(h) = \Pr_{(\rho,y)\sim\mathscr{D}}\left[h(\rho) \neq y\right]$

with sample $S = \{(\rho_1, y_1), (\rho_2, y_2), \ldots, (\rho_m, y_m)\} \sim \mathscr{D}^m$

# Example: Quantum Phase Classification

Generalized Cluster Hamiltonian

$$H(j_1, j_2) = \sum_{j=1}^{N} \left( Z_j - j_1 X_j X_{j+1} - j_2 X_{j-1} Z_j X_{j+1} \right)$$




Data: $\rho(j_1, j_2)$ ground state of $H(j_1, j_2)$

Label: $y \in \{1,2,3,4\}$ quantum phases

1. Symmetry Protected Topological
2. Ferromagnetic
3. Anti-Ferromagnetic
4. Trivial

# (Quantum) Supervised Learning

How do we find a good hypothesis $h$?

# (Quantum) Supervised Learning

How do we find a good hypothesis $h$?

1. Choose a hypothesis class $\mathscr{H} = \{h_1, h_2, \ldots\}$
2. Empirical Risk Minimization: $h^* = \arg\min_{h \in \mathscr{H}} \hat{R}(h)$
3. Hope $\hat{R}(h^*) \approx R(h^*)$ 🙏

Note:

$R(h) = \Pr\left[h(x) \neq y\right] = \mathbb{E}_{(\rho,y)\sim\mathscr{D}}\left[1_{h(x)\neq y}\right]$

$\hat{R}(h) = \frac{1}{m} \sum_{(\rho,y)\in S} 1_{h(\rho)\neq y}$

# (Quantum) Supervised Learning

How do we find a good hypothesis $h$?

1. Choose a hypothesis class $\mathcal{H} = \{h_1, h_2, \ldots\}$
2. Empirical Risk Minimization: $h^* = \arg\min_{h \in \mathcal{H}} \hat{R}(h)$
3. Hope $\hat{R}(h^*) \approx R(h^*)$ 🙏

Note:

$R(h) = \text{Pr}\left[h(x) \neq y\right] = \mathbb{E}_{(\rho,y)\sim\mathcal{D}}\left[1_{h(x)\neq y}\right]$

$\hat{R}(h) = \frac{1}{m} \sum_{(\rho,y)\in S} 1_{h(\rho)\neq y}$

Generalization Gap $g(h) = |R(h) - \hat{R}(h)|$

Question: Do we have a rigorous guarantee for generalization?

Answer: Yes! With complexity measure of $\mathcal{H}$.

# Generalization: Finite Hypothesis Class

Consider a finite Hypothesis Class $\mathscr{H} = \{h_1, h_2, \ldots, h_N\}$. $|\mathscr{H}| = N$.

For any $\delta \geq 0$, with probability higher than $1 - \delta$, $\forall h \in \mathscr{H}$

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\log|\mathscr{H}| + \log 2/\delta}{2m}}$$

# Generalization: Finite Hypothesis Class

Consider a finite Hypothesis Class $\mathscr{H} = \{h_1, h_2, \ldots, h_N\}$. $|\mathscr{H}| = N$.

For any $\delta \geq 0$, with probability higher than $1 - \delta$, $\forall h \in \mathscr{H}$

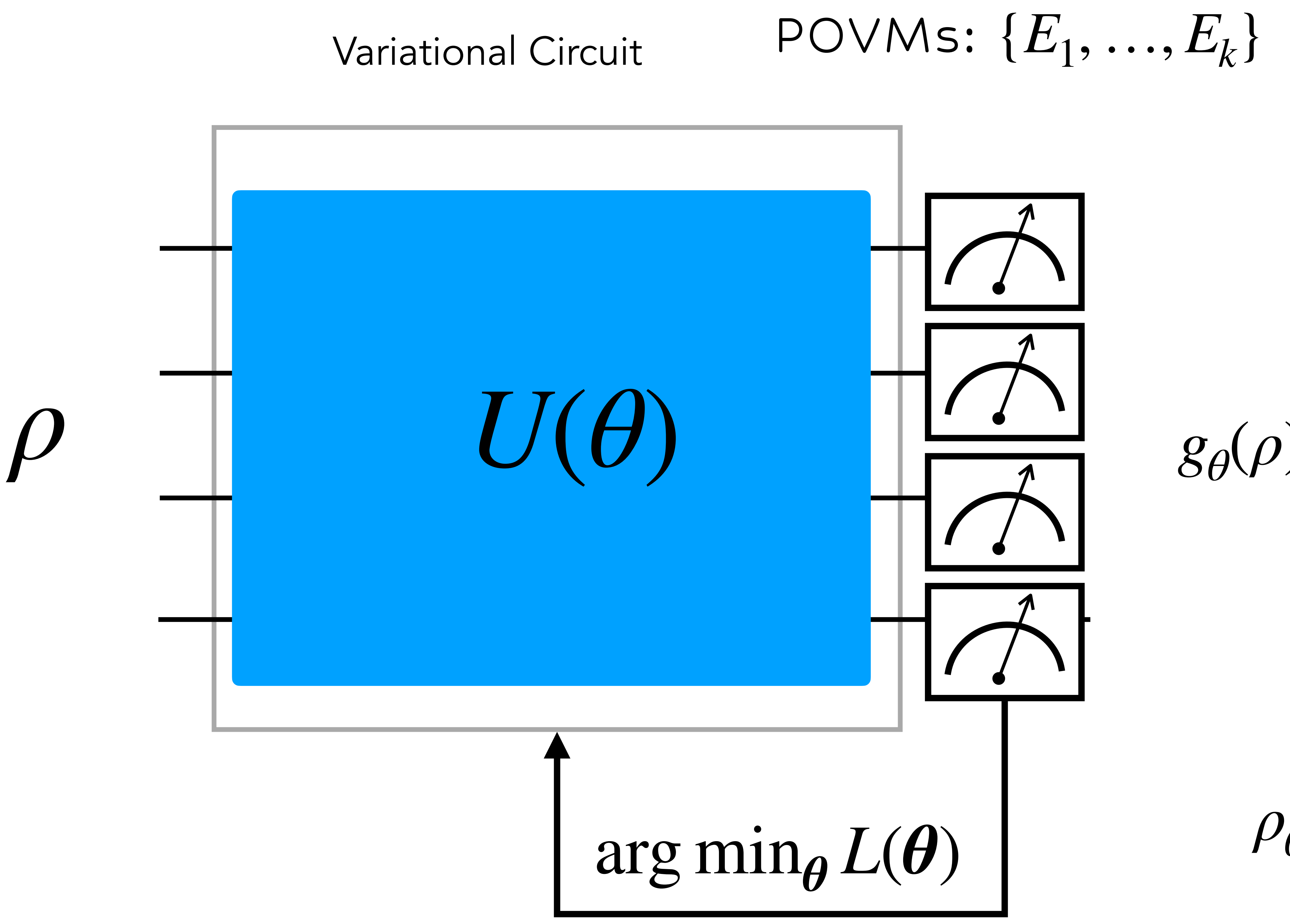
$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\log|\mathscr{H}| + \log 2/\delta}{2m}}$$

Proof Sketch:

$$\Pr\left[\max_{h \in \mathscr{H}} |R(h) - \hat{R}(h)| \geq \epsilon\right] \leq \sum_{h \in \mathscr{H}} \Pr\left[|R(h) - \hat{R}(h)| \geq \epsilon\right] \leq |\mathscr{H}| \times 2\exp\left(-2m\epsilon^2\right)$$

Here, Complexity Measure is simply $|\mathscr{H}|$

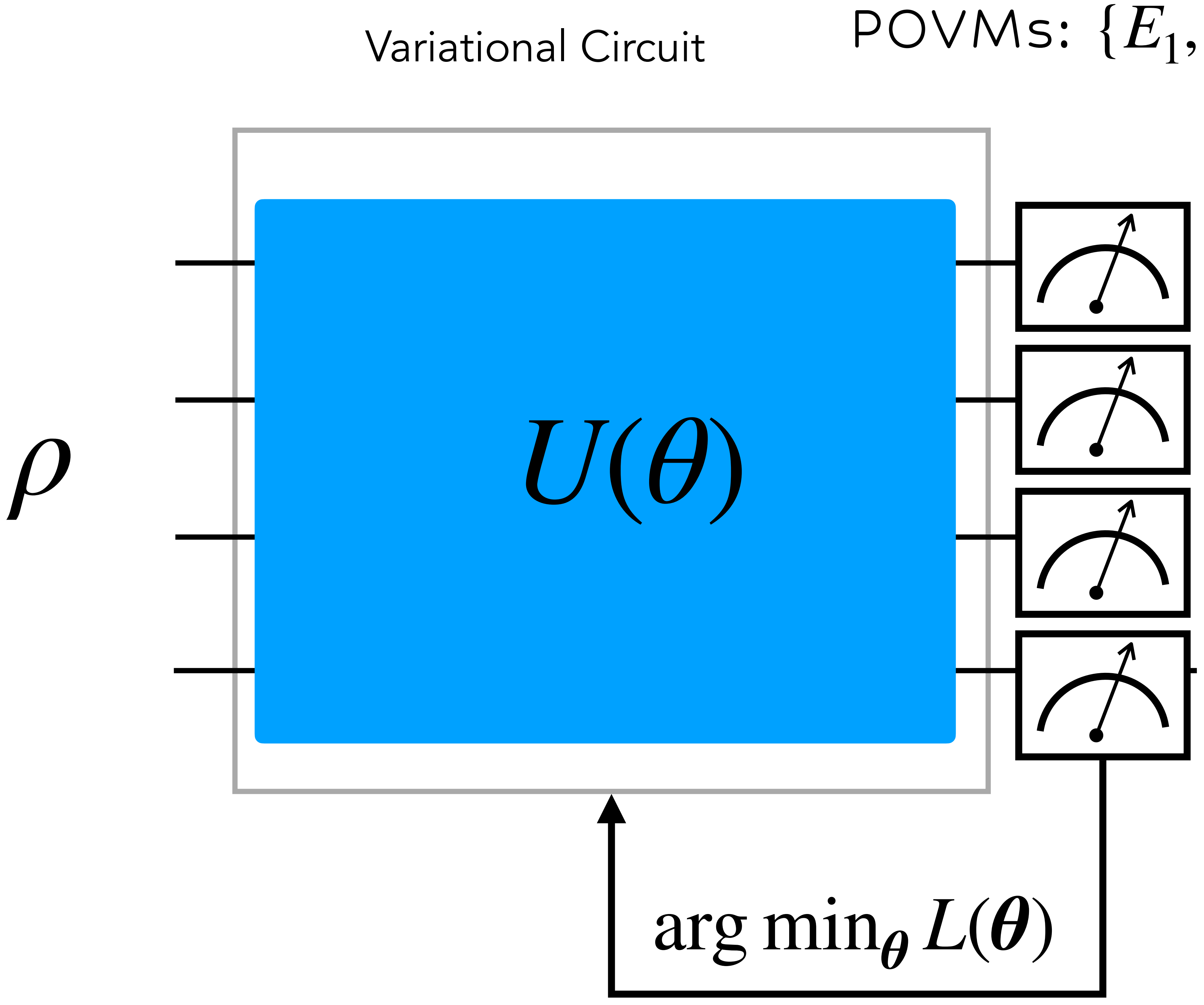
# Quantum Neural Networks (QNNs)




$$g_\theta(\rho) = \begin{pmatrix} \mathrm{Tr}\left(E_1 \rho_\theta\right) \\ \mathrm{Tr}\left(E_2 \rho_\theta\right) \\ \vdots \\ \mathrm{Tr}\left(E_k \rho_\theta\right) \end{pmatrix}$$

$$\rho_\theta = U(\theta) \rho U(\theta)^\dagger$$

# Quantum Neural Networks (QNNs)




POVMs: $\{E_1, \ldots, E_k\}$

$$g_\theta(\rho) = \begin{pmatrix} \mathrm{Tr}\left(E_1 \rho_\theta\right) \\ \mathrm{Tr}\left(E_2 \rho_\theta\right) \\ \vdots \\ \mathrm{Tr}\left(E_k \rho_\theta\right) \end{pmatrix}$$

$$\rho_\theta = U(\theta)\rho U(\theta)^\dagger$$

Prediction function (hypothesis) $h : \mathbb{C}^{2^n \times 2^n} \mapsto \{1,2,\ldots,k\}$:

$$h_\theta(\rho) = \arg\max_j g(\theta)_j$$

$$\mathscr{H}_{\mathrm{QNN}} = \{h_\theta : \theta \in \Theta\}$$

$$|\mathscr{H}_{\mathrm{QNN}}| = \infty$$

# Generalization: Rademacher Complexity

(Empirical) Rademacher Complexity $\hat{\mathfrak{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{z_i \in S} h(z_i)\sigma_i \right]$

Rademacher Random Variable $\sigma_i$

$\Pr(\sigma_i = +1) = +\frac{1}{2}$

$\Pr(\sigma_i = -1) = +\frac{1}{2}$

# Generalization: Rademacher Complexity

(Empirical) Rademacher Complexity $\hat{\mathfrak{R}}_S(\mathscr{H}) = \mathbb{E}_\sigma \left[ \sup_{h \in \mathscr{H}} \frac{1}{m} \sum_{z_i \in S} h(z_i)\sigma_i \right]$

Rademacher Random Variable $\sigma_i$

$\Pr(\sigma_i = +1) = +\frac{1}{2}$

$\Pr(\sigma_i = -1) = +\frac{1}{2}$

For any $\delta \geq 0$, with probability higher than $1 - \delta$, $\forall h \in \mathscr{H}$

$$R(h) \leq \hat{R}(h) + \hat{\mathfrak{R}}_S(\mathscr{H}) + 3\sqrt{\frac{\log 2/\delta}{2m}}$$

Previous result from finite Hypothesis class

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\log |\mathscr{H}| + \log 2/\delta}{2m}}$$

# Generalization: Rademacher Complexity

(Empirical) Rademacher Complexity $\hat{\mathfrak{R}}_S(\mathscr{H}) = \mathbb{E}_\sigma \left[ \sup_{h \in \mathscr{H}} \frac{1}{m} \sum_{z_i \in S} h(z_i)\sigma_i \right]$

Rademacher Random Variable $\sigma_i$

$\Pr(\sigma_i = +1) = +\frac{1}{2}$

$\Pr(\sigma_i = -1) = +\frac{1}{2}$

For any $\delta \geq 0$, with probability higher than $1 - \delta$, $\forall h \in \mathscr{H}$

$$R(h) \leq \hat{R}(h) + \hat{\mathfrak{R}}_S(\mathscr{H}) + 3\sqrt{\frac{\log 2/\delta}{2m}}$$

Previous result from finite Hypothesis class

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\log|\mathscr{H}| + \log 2/\delta}{2m}}$$

"Generalization in Quantum Machine Learning from few training data" (Nat. Comms. 2022)

$$\hat{\mathfrak{R}}_S(\mathscr{H}_{\text{QNN}}) \in \tilde{O}\left(\sqrt{\frac{T}{m}}\right)$$

T is number of parameters in QNN

nature communications

Article

https://doi.org/10.1038/s41467-022-32550-3

# Generalization in quantum machine learning from few training data

Received: 12 April 2022

Accepted: 4 August 2022

Published online: 22 August 2022

Check for updates

Matthias C. Caro[1,2], Hsin-Yuan Huang[3,4], M. Cerezo[5,6], Kunal Sharma[7], Andrew Sornborger[5,8], Lukasz Cincio[9] & Patrick J. Coles[9]

Modern quantum machine learning (QML) methods involve variationally optimizing a parameterized quantum circuit on a training data set, and subsequently making predictions on a testing data set (i.e., generalizing). In this work, we provide a comprehensive study of generalization performance in QML after training on a limited number $N$ of training data points. We show that the generalization error of a quantum machine learning model with $T$ trainable gates scales at worst as $\sqrt{T/N}$. When only $K \ll T$ gates have undergone substantial change in the optimization process, we prove that the generalization error improves to $\sqrt{K/N}$. Our results imply that the compiling of unitaries into a polynomial number of native gates, a crucial application for the quantum computing industry that typically uses exponential-size training data, can be sped up significantly. We also show that classification of quantum states across a phase transition with a quantum convolutional neural network requires only a very small training data set. Other potential applications include learning quantum error correcting codes or quantum dynamical simulation. Our work injects new hope into the field of QML, as good generalization is guaranteed from few training data.

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{T}{m}}\right)$$

$T$ : # of trainable parameters

nature communications

Article

https://doi.org/10.1038/s41467-022-32550-3

# Generalization in quantum machine learning from few training data

Received: 12 April 2022

Accepted: 4 August 2022

Published online: 22 August 2022

Check for updates

Matthias C. Caro[1,2], Hsin-Yuan Huang[3,4], M. Cerezo[5,6], Kunal Sharma[7], Andrew Sornborger[5,8], Lukasz Cincio[9] & Patrick J. Coles[9]

Modern quantum machine learning (QML) methods involve variationally optimizing a parameterized quantum circuit on a training data set, and subsequently making predictions on a testing data set (i.e., generalizing). In this work, we provide a comprehensive study of generalization performance in QML after training on a limited number $N$ of training data points. We show that the generalization error of a quantum machine learning model with $T$ trainable gates scales at worst as $\sqrt{T/N}$. When only $K \ll T$ gates have undergone substantial change in the optimization process, we prove that the generalization error improves to $\sqrt{K/N}$. Our results imply that the compiling of unitaries into a polynomial number of native gates, a crucial application for the quantum computing industry that typically uses exponential-size training data, can be sped up significantly. We also show that classification of quantum states across a phase transition with a quantum convolutional neural network requires only a very small training data set. Other potential applications include learning quantum error correcting codes or quantum dynamical simulation. Our work injects new hope into the field of QML, as good generalization is guaranteed from few training data.

$$R(h) \le \hat{R}(h) + O\left(\sqrt{\frac{T}{m}}\right)$$

$T$ : # of trainable parameters

nature communications

Article

https://doi.org/10.1038/s41467-024-45882-z

# Understanding quantum machine learning also requires rethinking generalization

Received: 3 July 2023

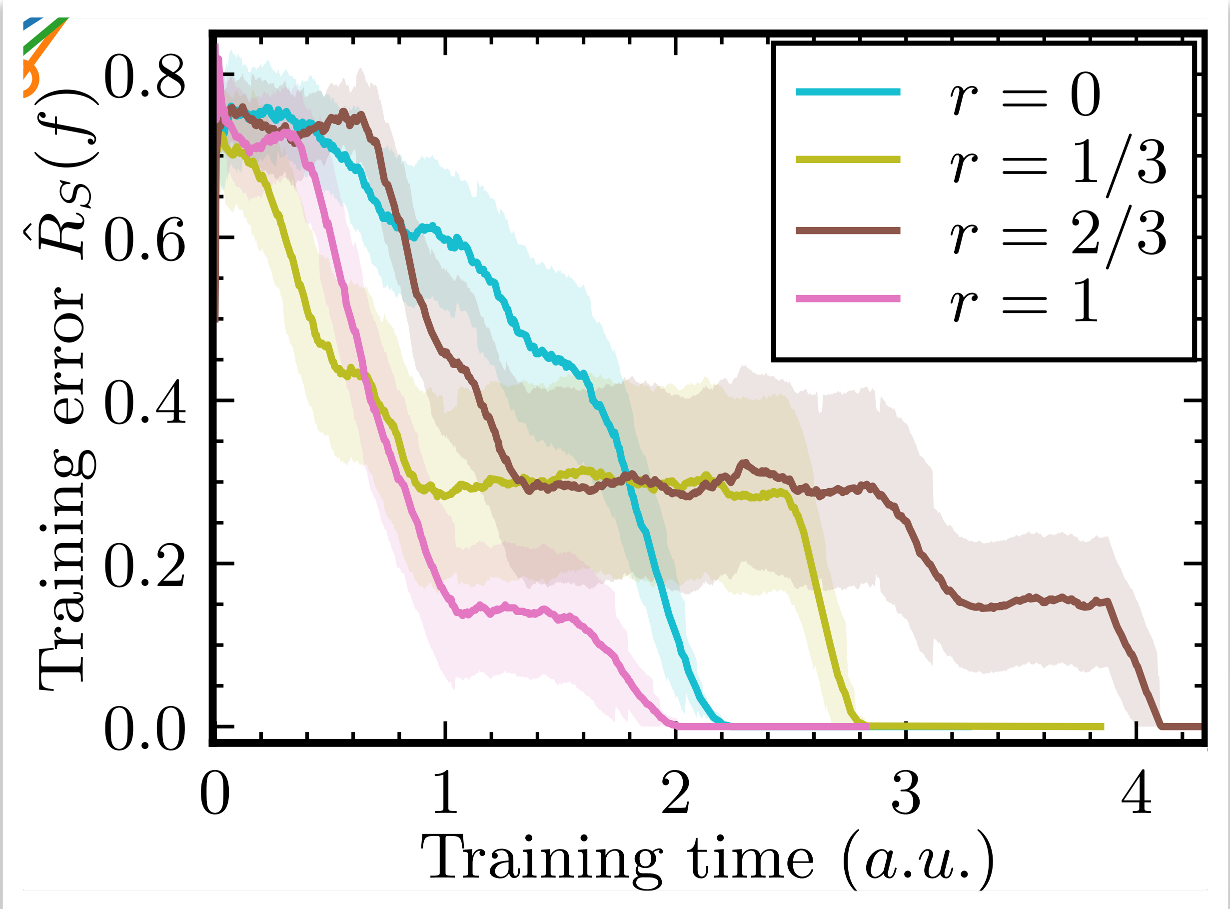
Accepted: 6 February 2024

Published online: 13 March 2024

Check for updates

Elies Gil-Fuster[1,2], Jens Eisert[1,2,3] & Carlos Bravo-Prieto[1]

Quantum machine learning models have shown successful generalization performance even when trained with few data. In this work, through systematic randomization experiments, we show that traditional approaches to understanding generalization fail to explain the behavior of such quantum




This is an "uniform bound".

Can be vacuous.

# Margin Generalization

Theorem

For any $\delta > 0$ and $\gamma > 0$, with probability at least $1 - \delta$ over the random draw of an i.i.d sample $S$ of size $m$, the following inequality holds for all $h \in \mathscr{H}$ :

$R(h)$ : True Error
$\hat{R}_\gamma(h)$ : Empirical Margin Error

$$R(h) \leq \hat{R}_\gamma(h) + \tilde{O}\left(\frac{nb}{\gamma}\sqrt{\frac{\sum_{i=1}^{k} \|E_i\|_\sigma^2}{m}} + \sqrt{\frac{\ln(1/\delta)}{m}}\right).$$

- $n$ : # of qubits
- $m$ : # of sample data
- $E_i$ : Measurement Operators
- $b$ : distance bound, $\|U - U_{\text{ref}}\|_{2,1} \leq b$

Theorem

For any $\delta > 0$ and $\gamma > 0$, with probability at least $1 - \delta$ over the random draw of an i.i.d sample $S$ of size $m$, the following inequality holds for all $h \in \mathscr{H}$ :

$R(h)$ : True Error

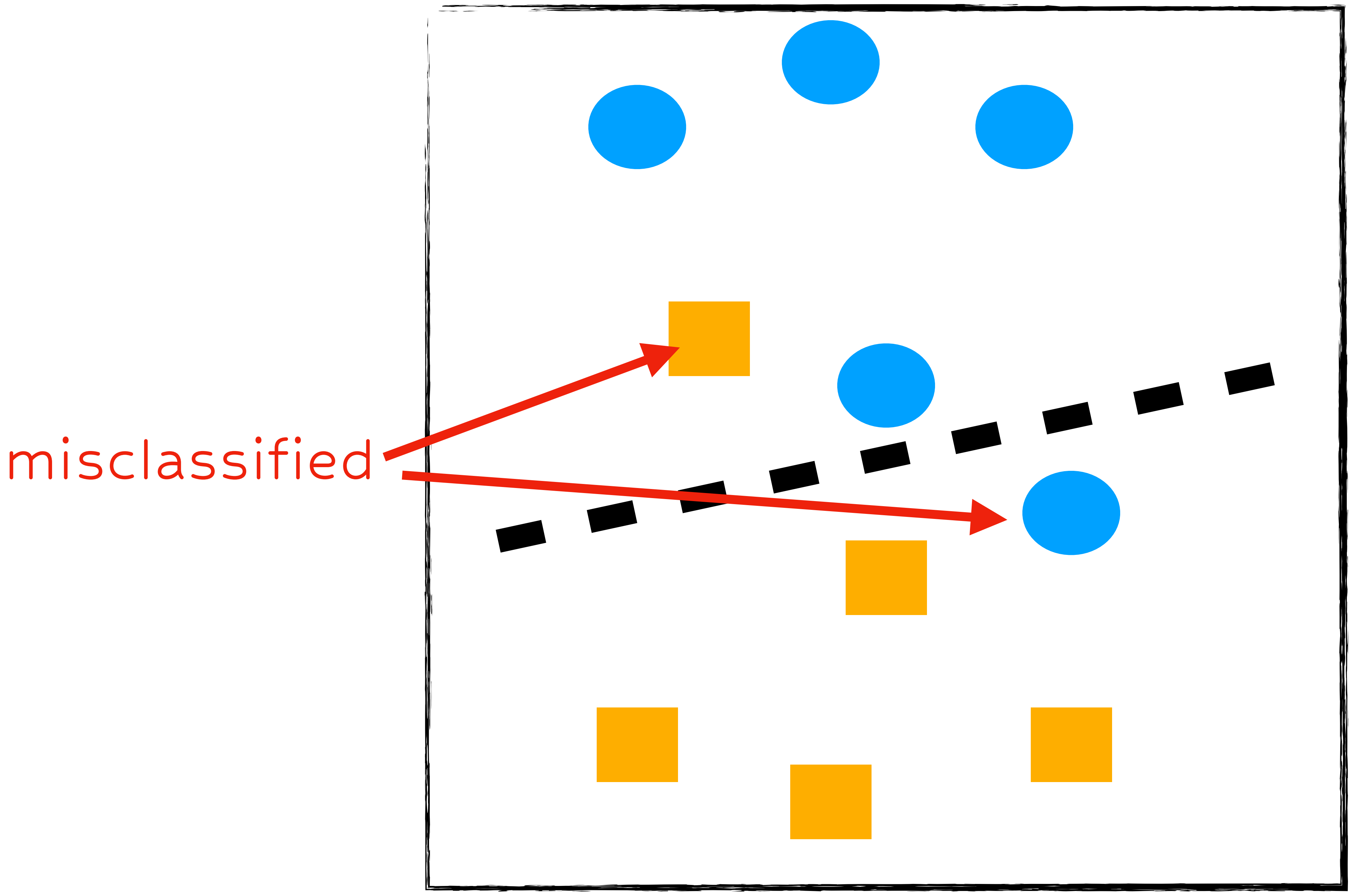
$\hat{R}_\gamma(h)$ : Empirical Margin Error

$$R(h) \le \hat{R}_\gamma(h) + \tilde{O}\left(\frac{nb}{\gamma}\sqrt{\frac{\sum_{i=1}^{k} \|E_i\|_\sigma^2}{m}} + \sqrt{\frac{\ln(1/\delta)}{m}}\right).$$

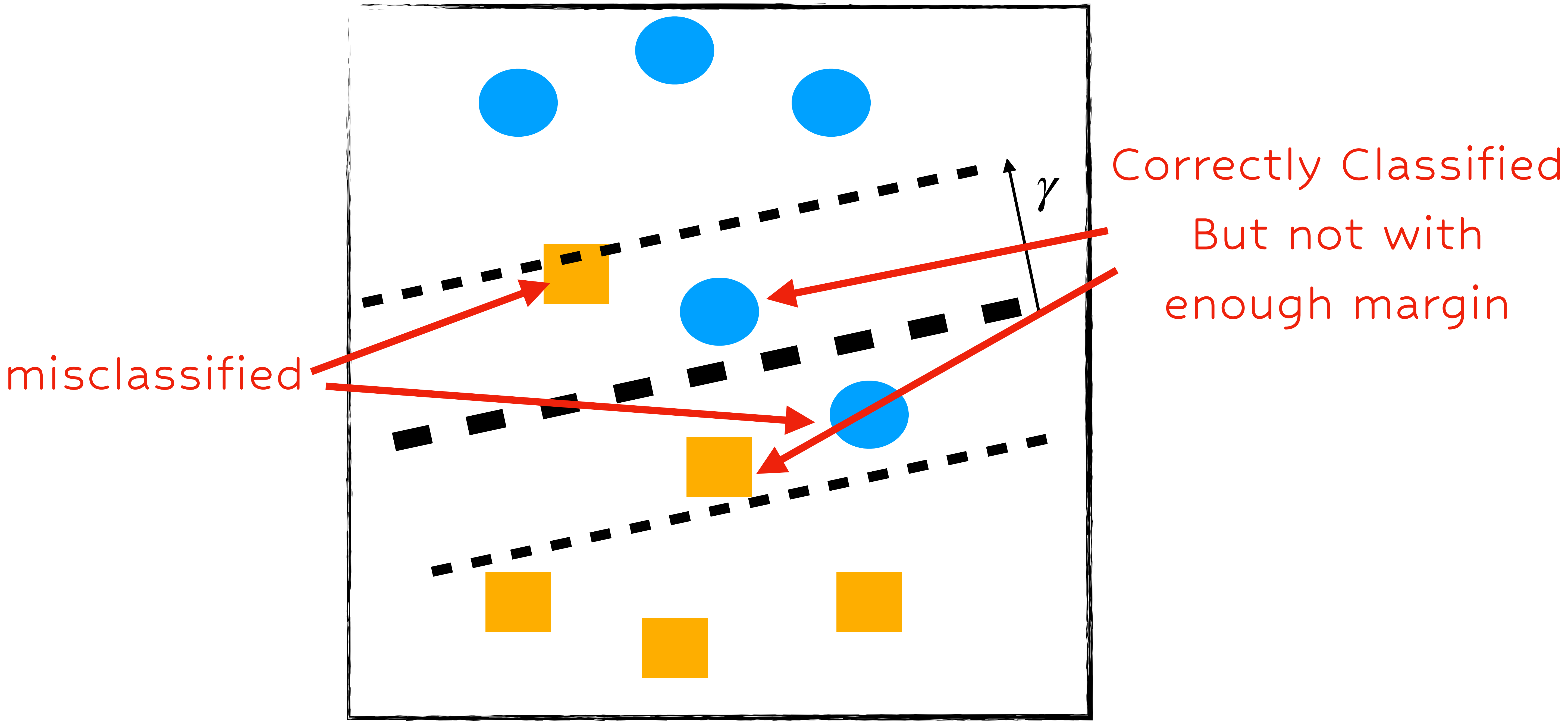
$\gamma$ { ↑ ↓ ➡ $\hat{R}_\gamma(h)$ ↑ ↓ vs. $\frac{nb}{\gamma}$ ↓ ↑

- $n$ : # of qubits
- $m$ : # of sample data
- $E_i$ : Measurement Operators
- $b$ : distance bound, $\|U - U_{\mathrm{ref}}\|_{2,1} \le b$

$\hat{R}(h) = 0.2$




$\hat{R}_\gamma(h) = 0.4$



# Margin Generalization

Theorem

For any $\delta > 0$ and $\gamma > 0$, with probability at least $1-\delta$ over the random draw of an i.i.d sample $S$ of size $m$, the following inequality holds for all $h \in \mathscr{H}$ :

$R(h)$ : True Error
$\hat{R}_\gamma(h)$ : Empirical Margin Error

$$R(h) \leq \hat{R}_\gamma(h) + \tilde{O}\left(\frac{nb}{\gamma}\sqrt{\frac{\sum_{i=1}^{k} \|E_i\|_\sigma^2}{m}} + \sqrt{\frac{\ln(1/\delta)}{m}}\right).$$

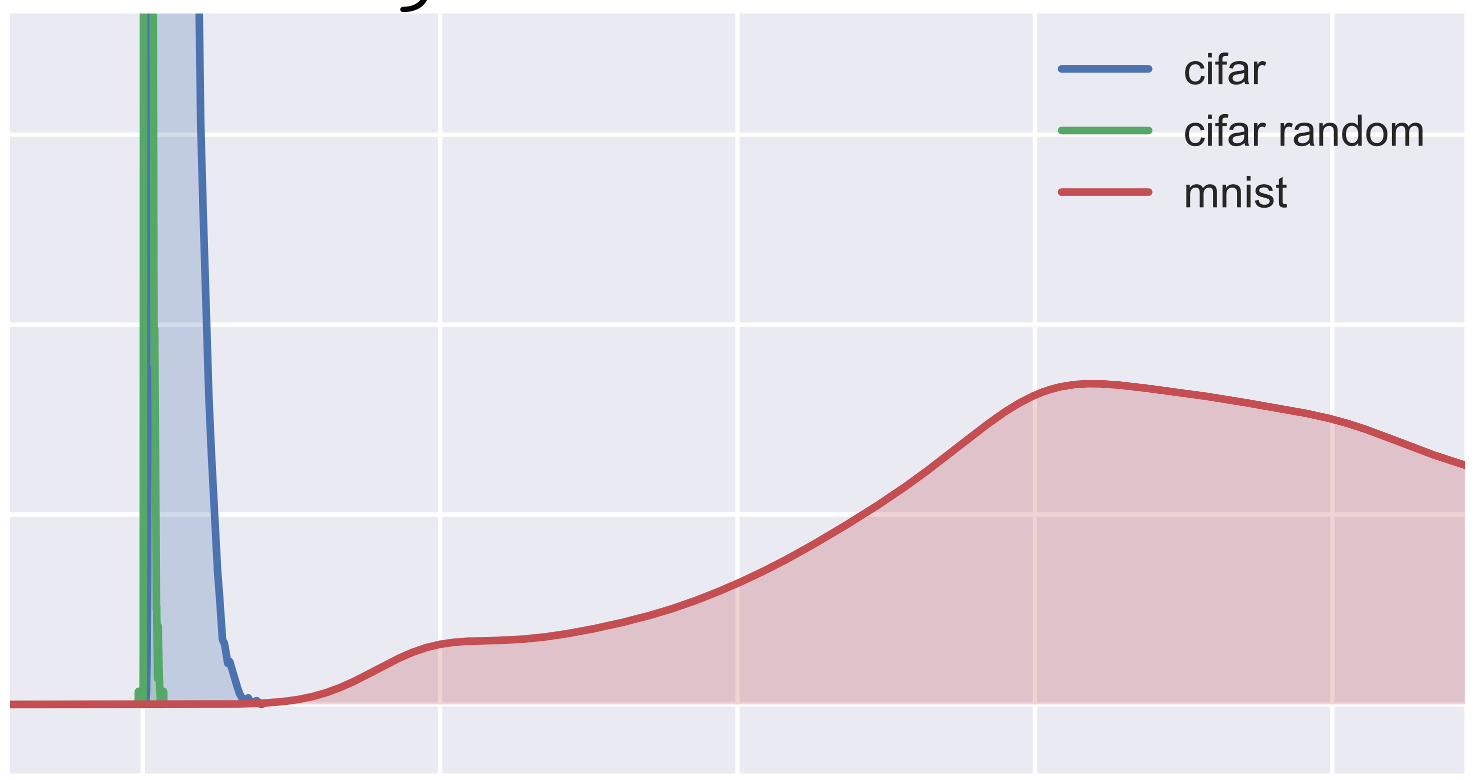
- $n$ : # of qubits
- $m$ : # of sample data
- $E_i$ : Measurement Operators
- $b$ : distance bound, $\|U - U_{\text{ref}}\|_{2,1} \leq b$

Consequences

Margin: $h(x)_y - \max_{i \neq y} h(x)_i$

Margin distribution is important to understand generalization

Margin Distribution Plot




Left skewed margin dist. ↦ Large generalization Upper bound
Right skewed margin dist. ↦ Small generalization Upper bound

# Margin Generalization

## Proof Sketch

$\mathfrak{R}$ : Rademacher Complexity

$\mathcal{N}$ : Covering Number

$\mathcal{F}$ : Hypothesis Class of QML model

1. Rademacher Complexity

$$R(h) \leq \hat{R}_\gamma(h) + 2\mathfrak{R}((\mathcal{F}_\gamma)_{|S}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}},$$

2. Dudley's Entropy Integral

$$\mathfrak{R}(U) \leq \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{\sqrt{m}} \sqrt{\ln \mathcal{N}(U, \beta, \|\cdot\|_2)d\beta}\right).$$

3. Covering Number Bound for QML model

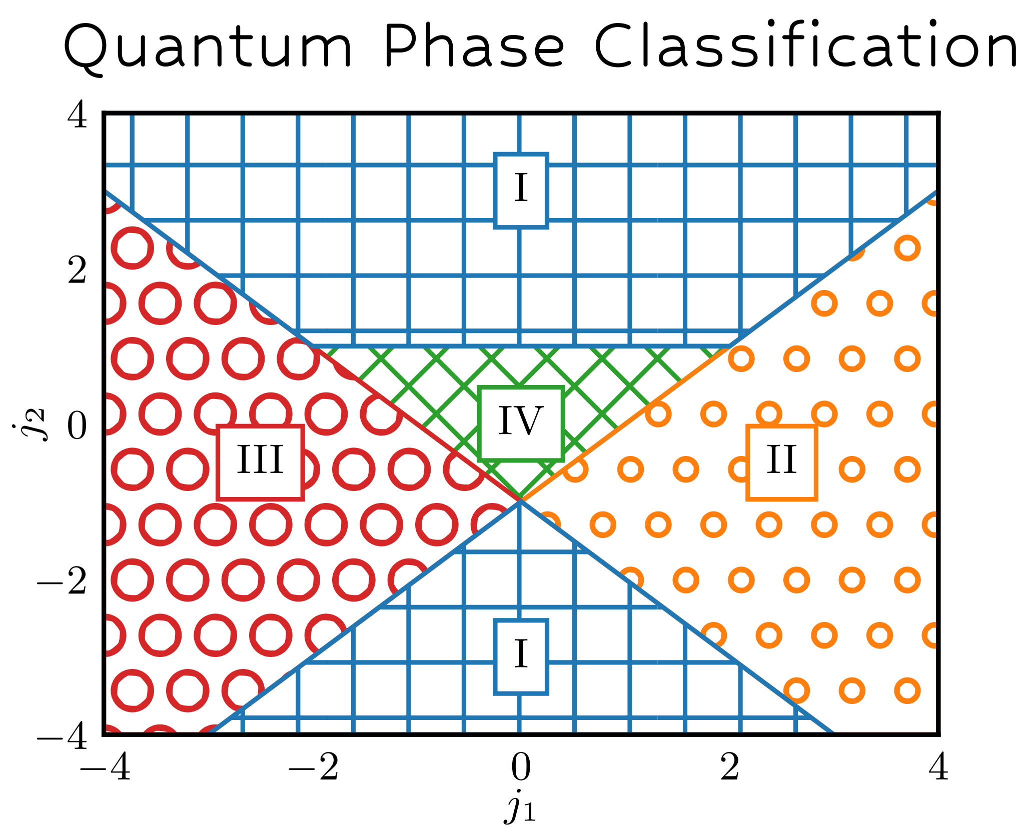
$$\ln \mathcal{N}\left((\mathcal{F}_\gamma)_{|S}, \epsilon, ||\cdot||_2\right) \leq \ln \mathcal{N}\left(\{UX : U \in \mathbb{U}_{\mathrm{QNN}}\}, \frac{\epsilon\gamma}{4E}, ||\cdot||_2\right) \leq \left\lceil \frac{32mb^2E^2}{\epsilon^2\gamma^2} \right\rceil \ln 4N^2,$$

Lipschitz property of quantum measurement function $g(x)$

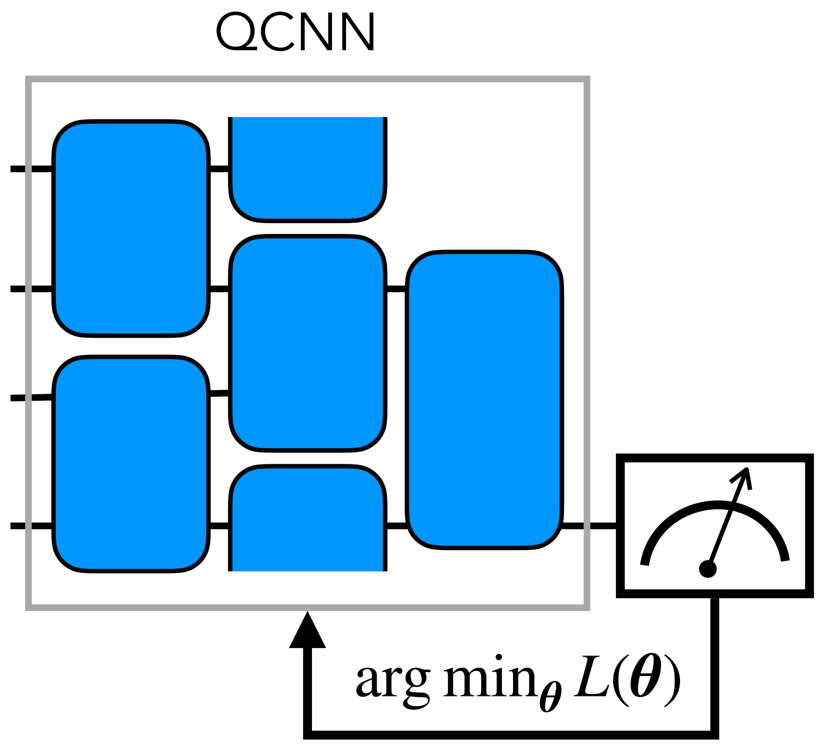
Maurey's Sparsification Lemma

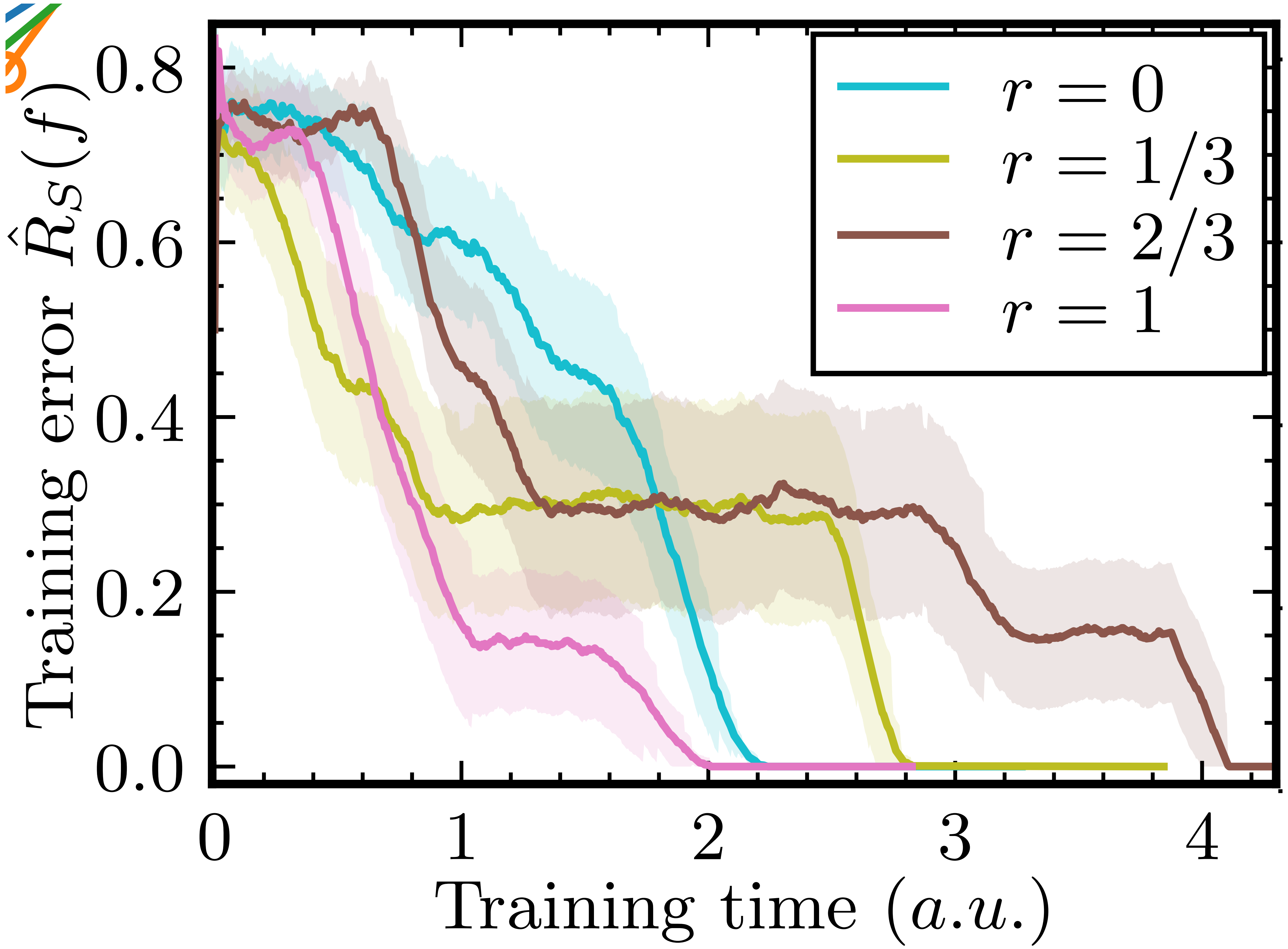
# Experimental Results: Margin Boxplot

Solve




with



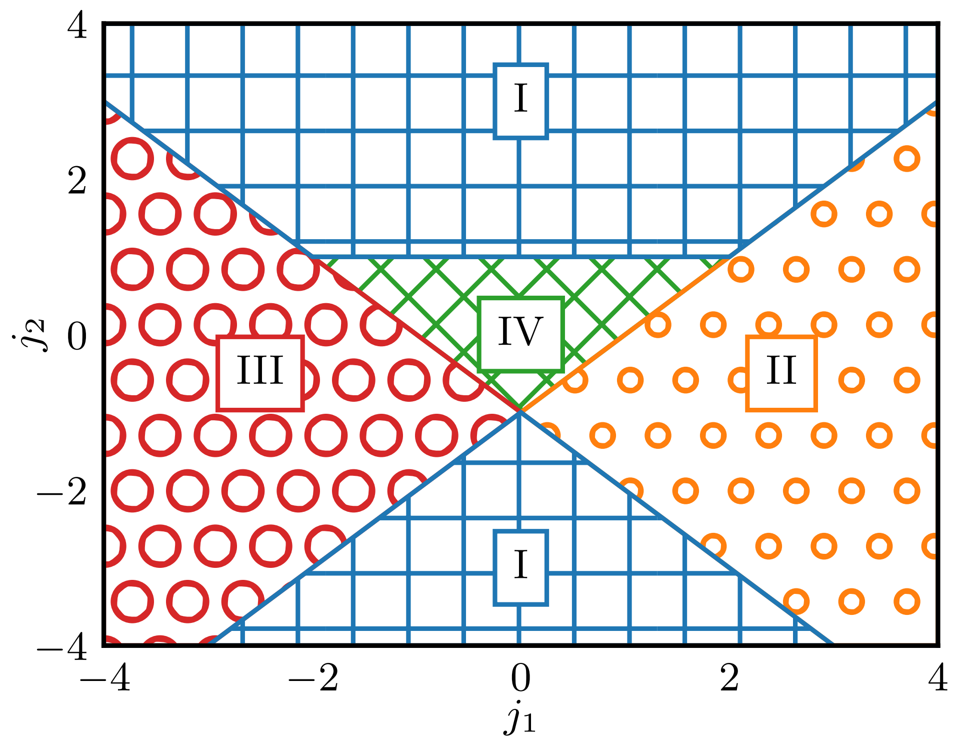




"Understanding Quantum Machine Learning Also Requires Rethinking Generalization" (Nat. Comm. 2024)

# Experimental Results: Margin Boxplot

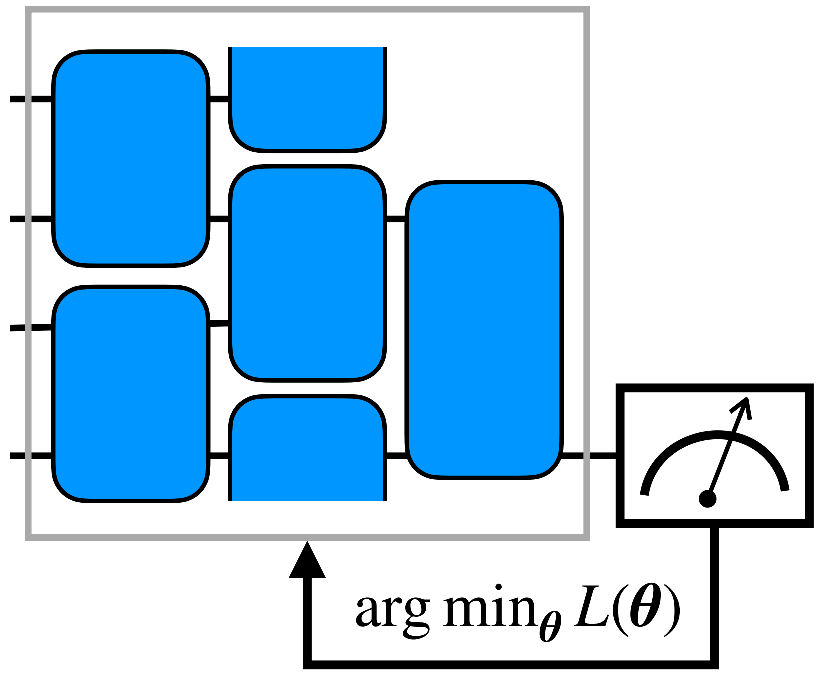
Solve

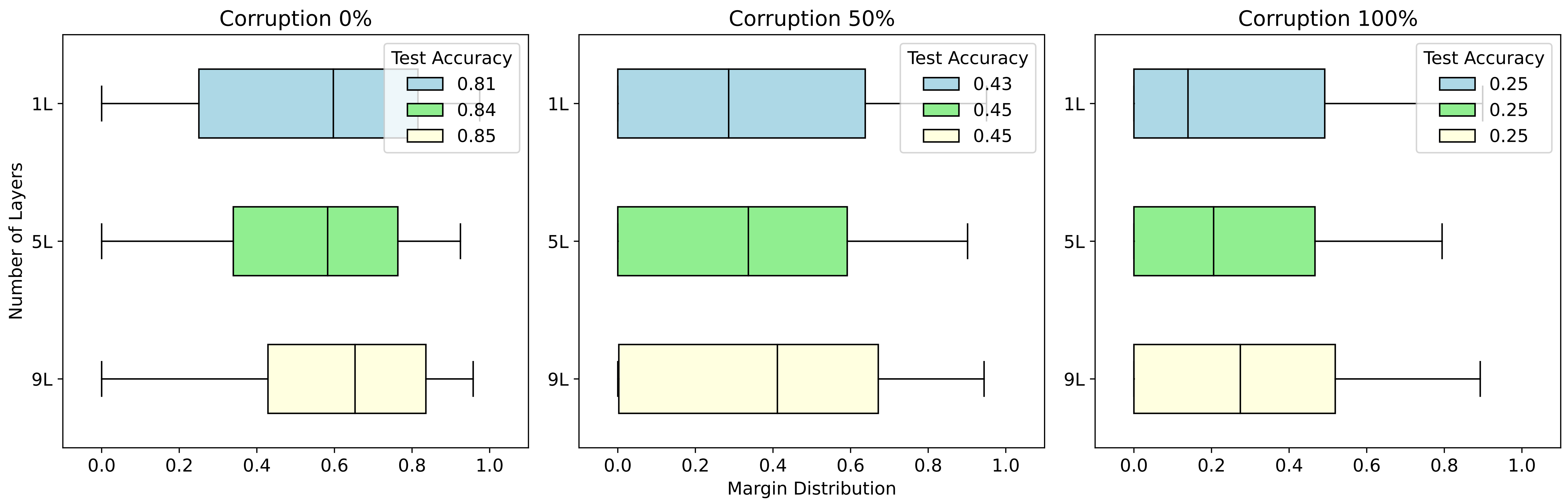
Quantum Phase Classification




with

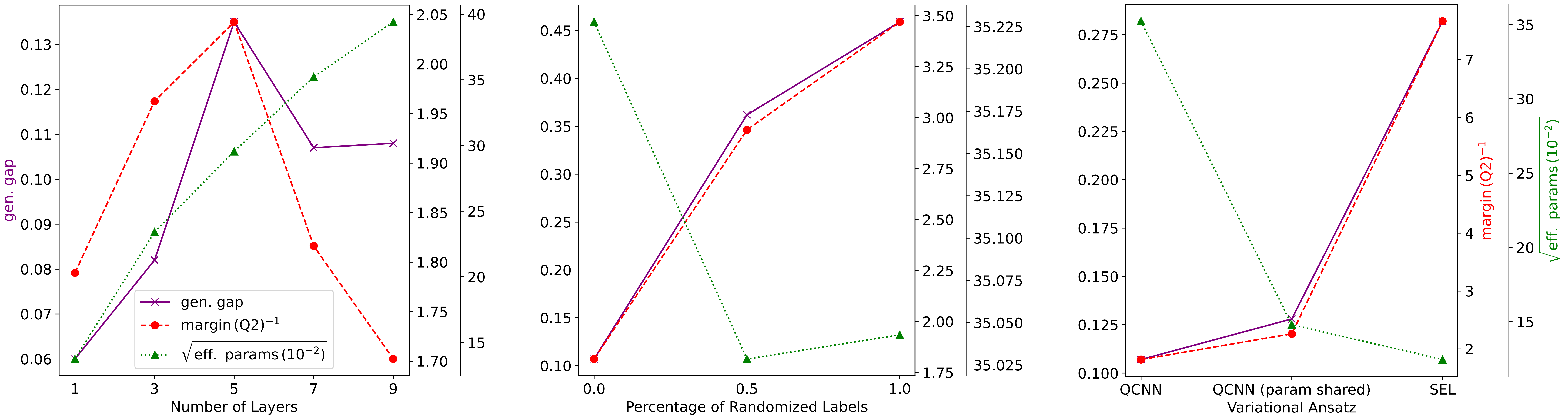
QCNN




Unlike previous uniform bounds, margin bound captures generalization behaviour of QML models under label corruption.

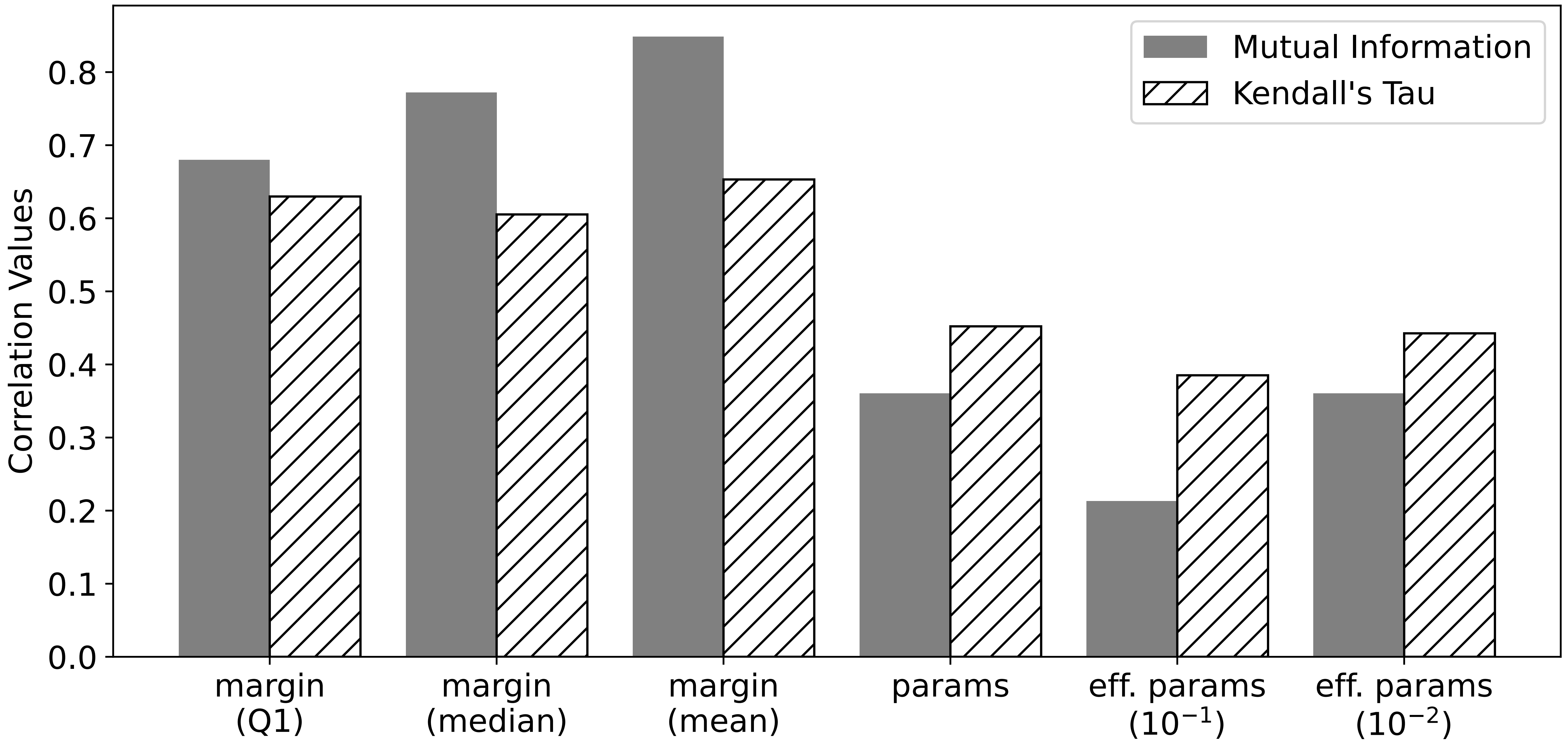
# Experiment 2: Parmeter vs. Margin




(a) (b) (c)

Margin effectively captures generalization behaviours (Better than the # of parameters).

# Experiment 2: Parmeter vs. Margin




1. Mutual Information

$$I(g;\mu) = H(g) - H(g\,|\,\mu)$$

How much information does $\mu$ provide about $g$?

2. Kendall's Rank Correlation

$$\tau(G,M) = \frac{1}{2n(n-1)} \sum_{i<j} \left[1 + \mathrm{sgn}(g_i - g_j)\,\mathrm{sgn}(\mu_i - \mu_j)\right].$$

Are orderings of $g$ and $\mu$ aligned?

# Experiment 3

Neural Quantum Embedding: Pushing the Limits of Quantum Supervised Learning. Phys. Rev. A (2024)

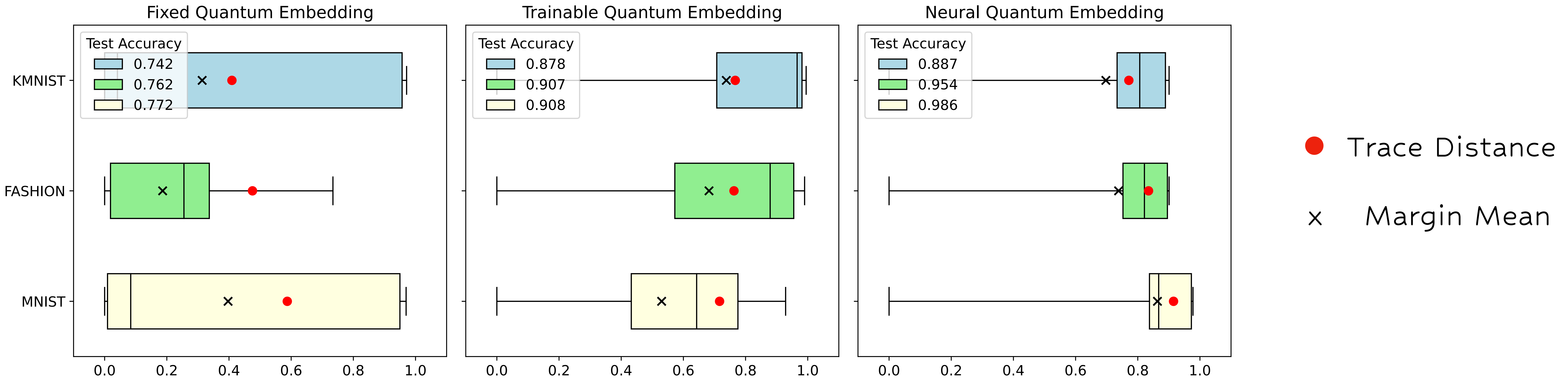
- Large "Trace distance" ↦ Small Training Error
- Open question: Why does it generalizes well?

Answer: Margin Generalization Bounds!

- "Trace Distance" upper bounds Margin mean

$$\mu_{\text{mean}} \leq D_{\text{tr}}(p^+\rho^+, p^-\rho^-)$$

- Large Trace Distance ↦ Right Skewed Margin Dist. ↦ Better Generalization



# Summary

- Established margin-based generalization bound for QML models.
- Experimentally demonstrated strong correlation between generalization and margin.
- Established connection between margins and quantum state discrimination.

# Thank You