Centre for Quantum Technologies × the entangled computing lab

FREIE UNIVERSITÄT BERLIN

Concept learning of parameterized quantum models from limited measurements

Po-Wei Huang

arXiv:2408.05116 [quant-ph]

[redacted]

Slides adapted from Beng Yee Gan

Beng Yee Gan

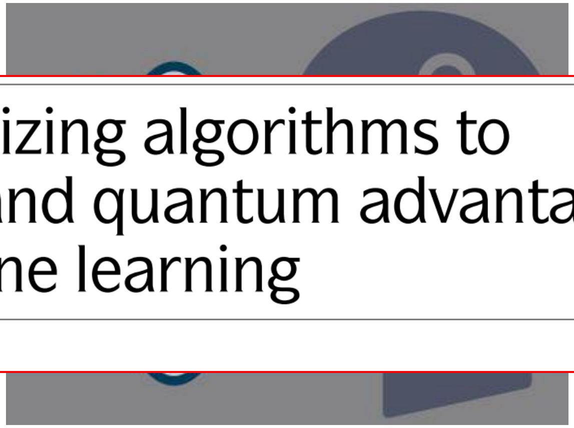
Po-Wei Huang

Elies Gil-Fuster

Patrick Rebentrost

Power of quantum computers

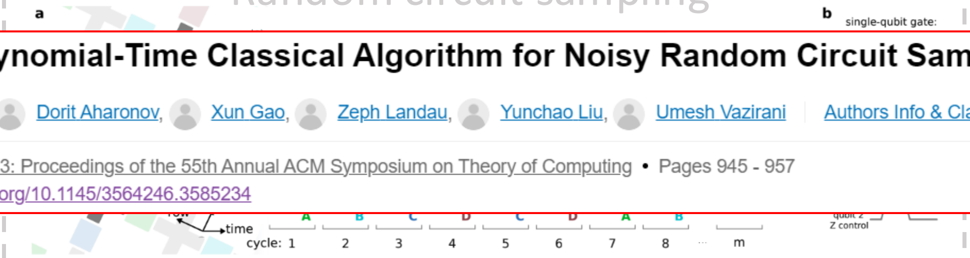
Quantum machine learning



Dequantizing algorithms to understand quantum advantage in machine learning

Ewin Tang

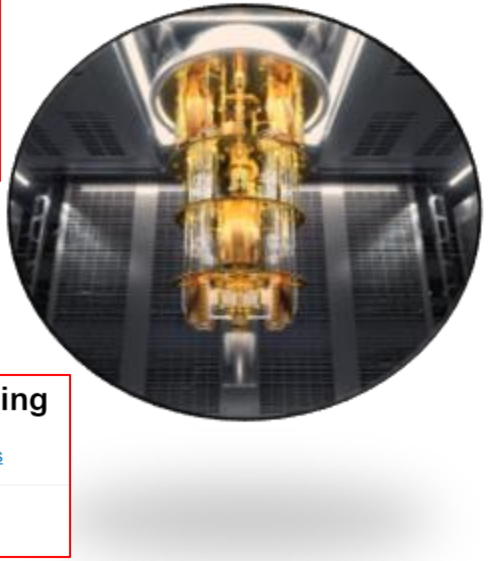
Random circuit sampling



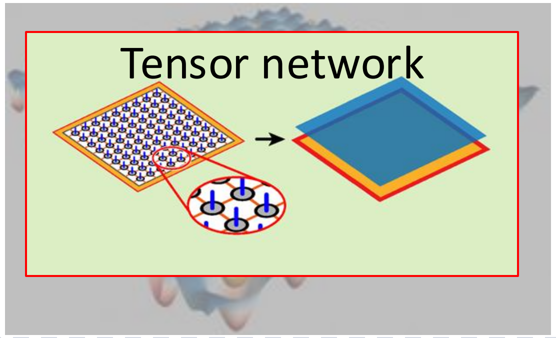
A Polynomial-Time Classical Algorithm for Noisy Random Circuit Sampling

Authors: Dorit Aharonov, Xun Gao, Zeph Landau, Yunchao Liu, Umesh Vazirani Authors Info & Claims

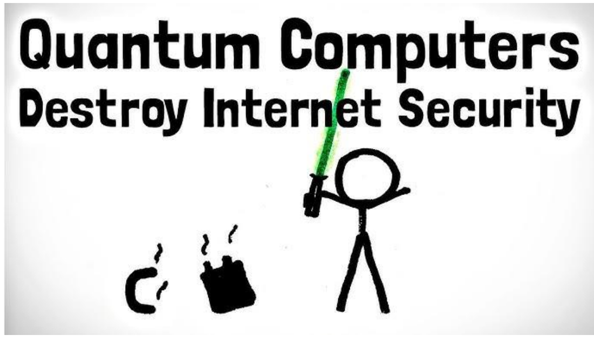
STOC 2023: Proceedings of the 55th Annual ACM Symposium on Theory of Computing • Pages 945 - 957
https://doi.org/10.1145/3564246.3585234



Quantum simulations



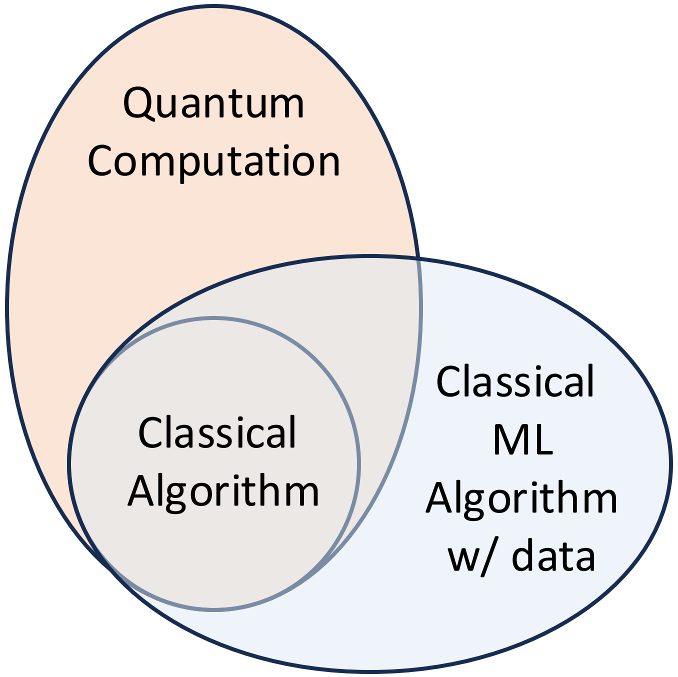
Shor's algorithm



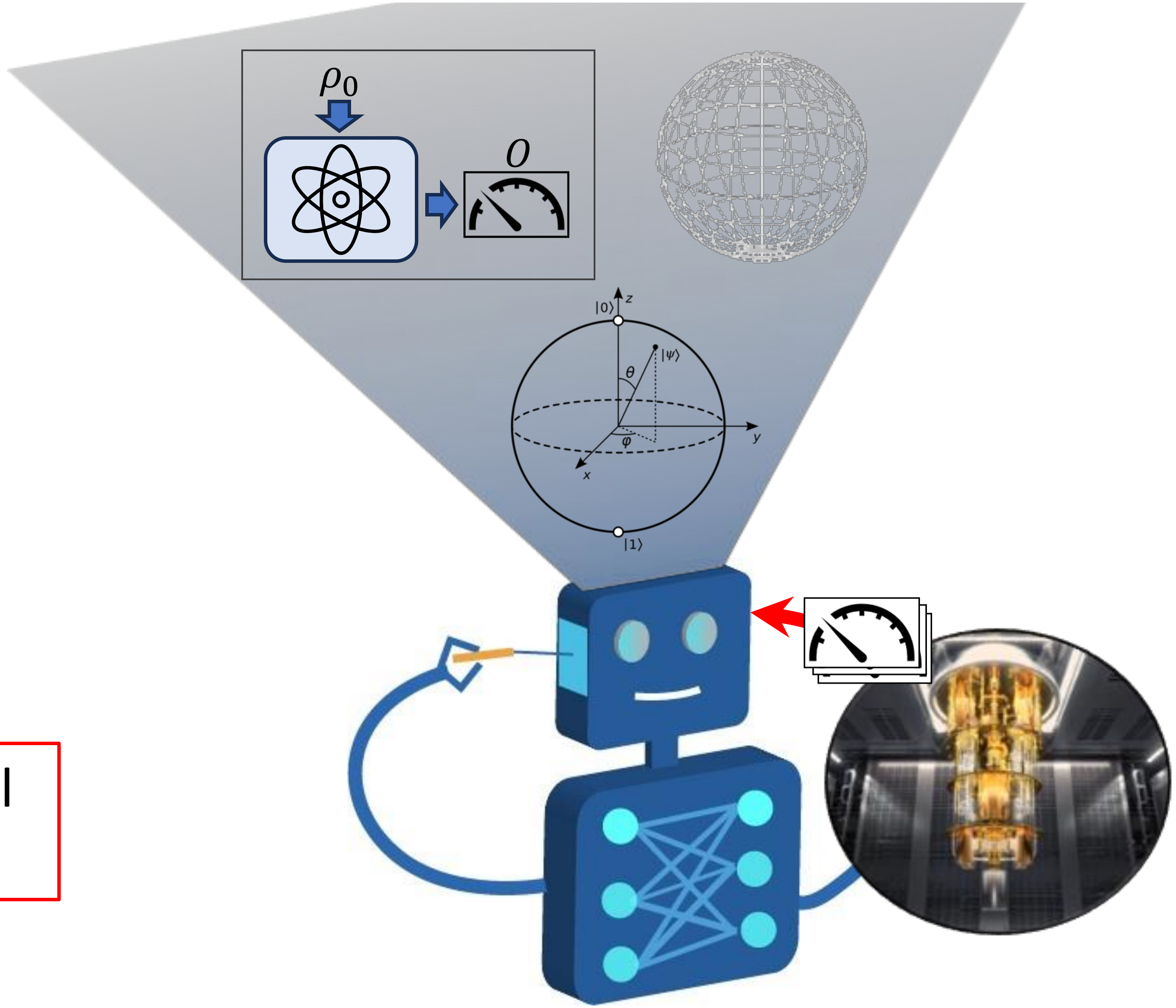
Performance is benchmarked against classical computers.

Centre for Quantum Technologies × the entangled computing lab

Replacing (some) quantum with classical

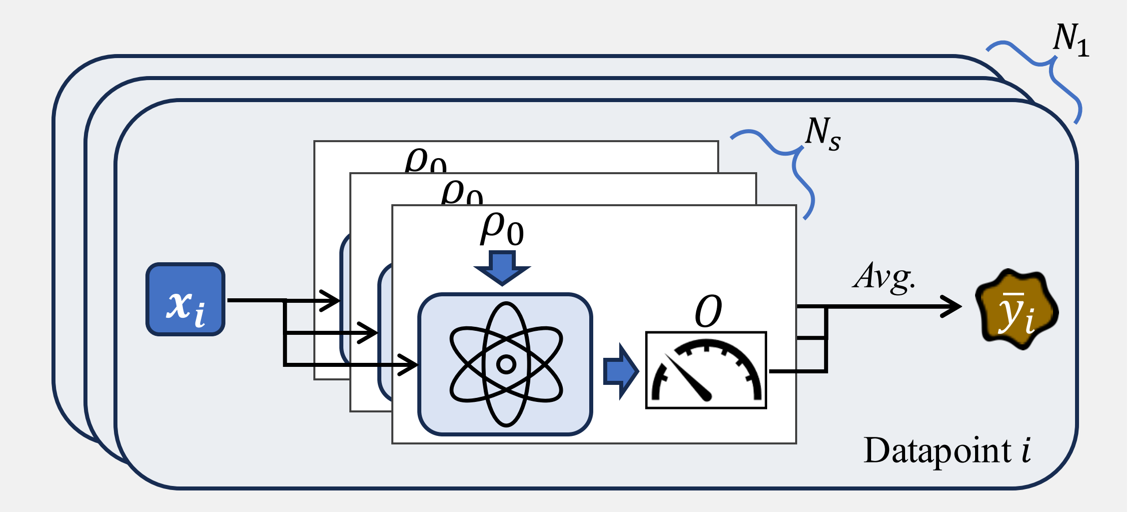


Nat. Commun., *12*(1), 2631 (2021).



Access to data makes classical machines more powerful.

Learning from limited measurements

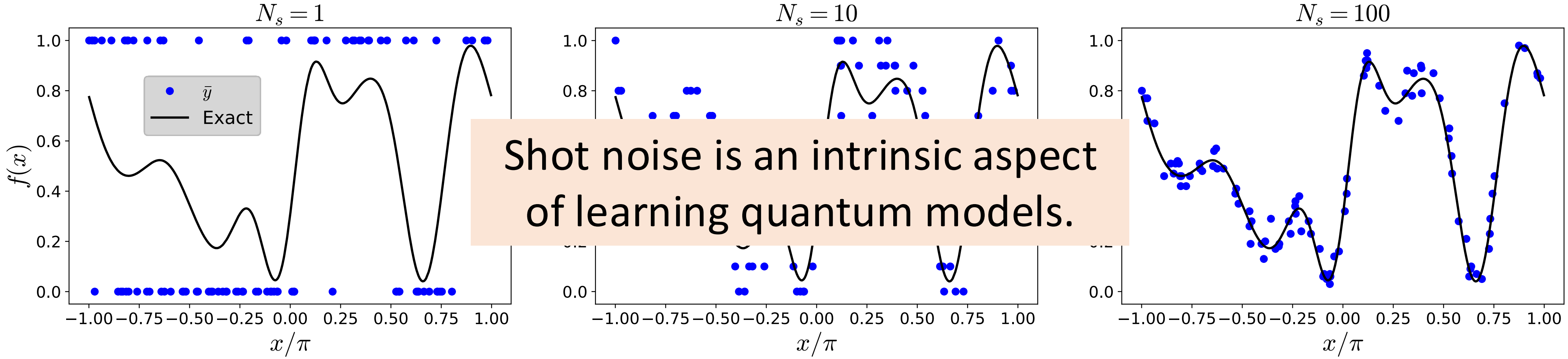


Quantum models: $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathrm{tr}(\rho_{\boldsymbol{\theta}}(\boldsymbol{x})O)$

$\mathbb{E}_{\bar{y}}[\bar{y}|\boldsymbol{x}]$

- Dataset: $(\boldsymbol{x}_i, \bar{y}_i)_{i=1}^{N_1}$

Estimated with N_s shots

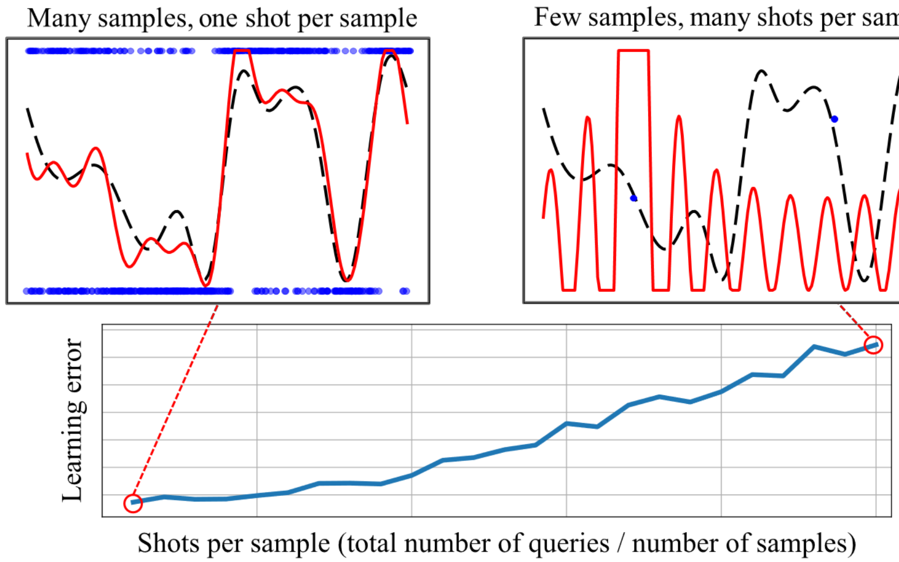


Shot noise is an intrinsic aspect of learning quantum models.

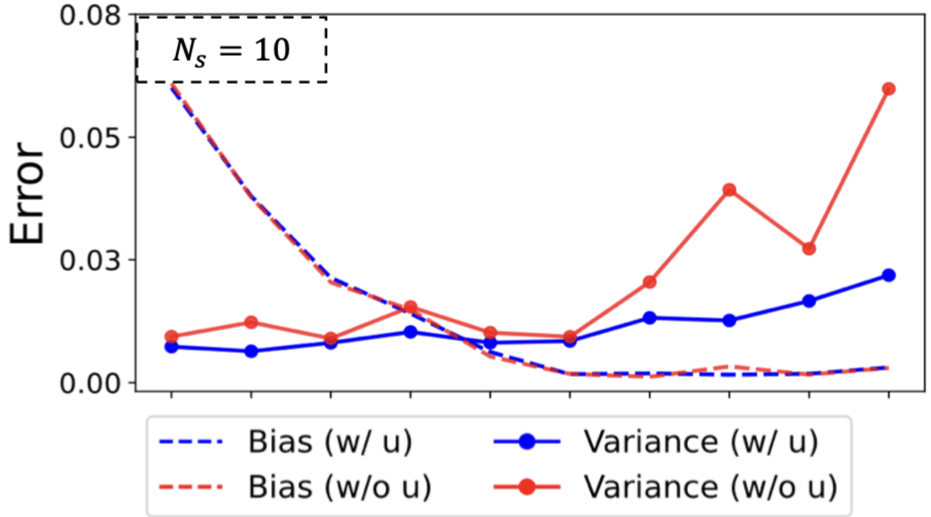
Results overview

Can we obtain provable guarantees of learning that exemplify the relationship between N_1 and N_s?

(1) Asymmetrical trade-offs between N_1 and N_s



(2) Gradient descent* can be made robust and provide tighter guarantees



*Technically a gradient-descent-esque algorithm instead of vanilla GD.

Centre for Quantum Technologies × the entangled computing lab

Probabilistic concept learning

Probabilistic concept class

$$\mathcal{F} = \{f(\boldsymbol{x}) = \mathbb{E}_y[y|\boldsymbol{x}]\}$$

$$\boxed{\mathcal{D} = p(\boldsymbol{x})p(y|\boldsymbol{x})}$$

Loss function: $\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), h_{\boldsymbol{w}}(\boldsymbol{x}))$

Hypothesis class

$$\mathcal{H} = \{h_{\boldsymbol{w}}(\boldsymbol{x}), \boldsymbol{w} \in \mathbb{R}^D\}$$

Learning task

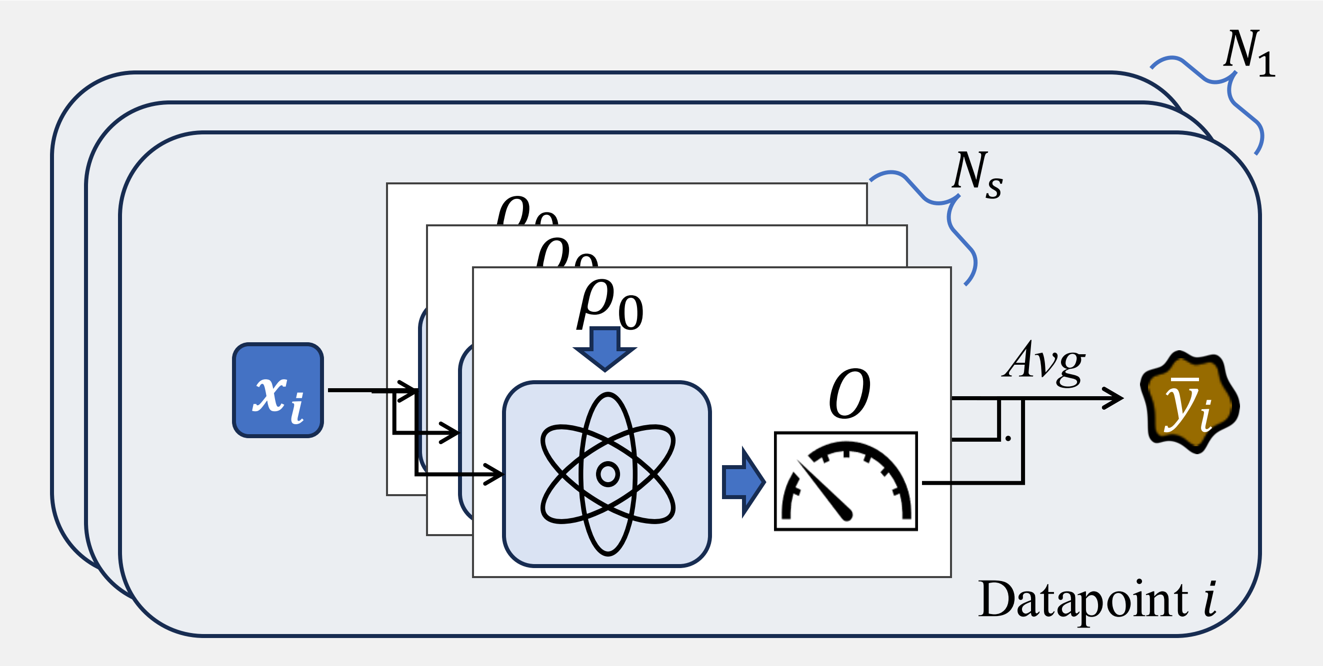
- Unknown $f \in \mathcal{F}$
- Goal: $\overbrace{\mathbb{E}_{\boldsymbol{x}}[\ell(f(\boldsymbol{x}), h_{\boldsymbol{w}}(\boldsymbol{x}))]}^{R(h_{\boldsymbol{w}})}$

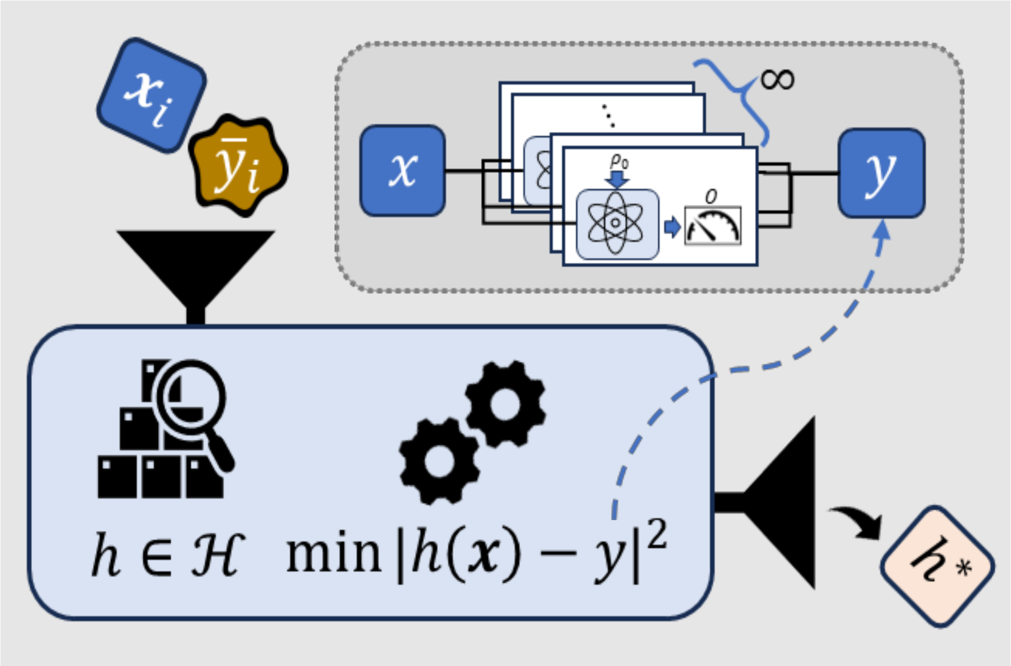
- Since $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_{N_1}} \sim \boldsymbol{\mathcal{D}}$
- We can only get: $\overbrace{\frac{1}{N_1}\sum_{i=1}^{N_1} \ell(\bar{y}, h_{\boldsymbol{w}}(\boldsymbol{x_i}))}^{\hat{R}(h_{\boldsymbol{w}})}$
- Task: $\hat{R}(h_{\boldsymbol{w}}) - R(h_{\boldsymbol{w}^*}) \leq \epsilon$

Provable guarantee

Getting data – the black box model

<u>Family of PQC models:</u> $\mathcal{F}_{U,O} = \left\{ f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \underbrace{\langle \mathbf{0} | U^\dagger(\boldsymbol{x}, \boldsymbol{\theta}) O U(\boldsymbol{x}, \boldsymbol{\theta}) | \mathbf{0} \rangle}_{\mathrm{tr}(\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) O)} \right\}$





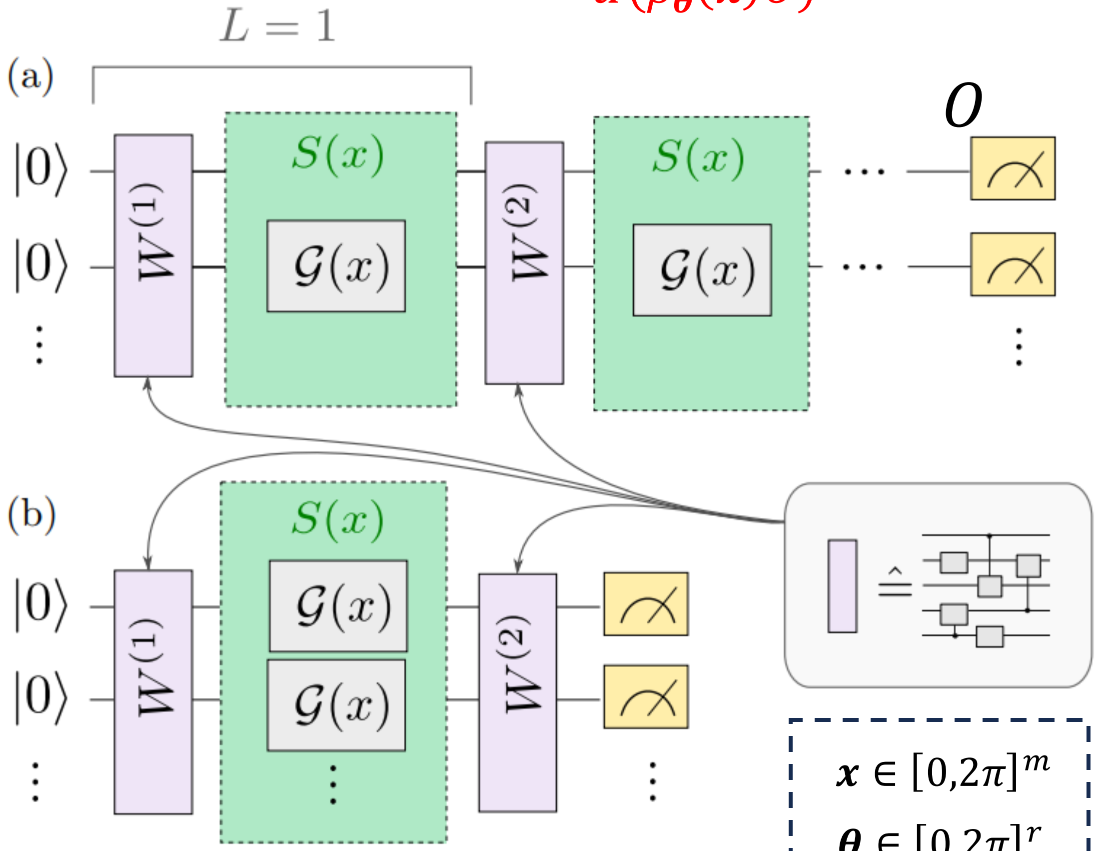
Which PQCs are learnable? How do we get the hypothesis class $\mathcal{H}$?

Getting data – the grey box model

<u>Family of PQC models:</u> $\mathcal{F}_{U,O} = \{f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \underbrace{\langle \mathbf{0}|U^{\dagger}(\boldsymbol{x},\boldsymbol{\theta})OU(\boldsymbol{x},\boldsymbol{\theta})|\mathbf{0}\rangle}_{\mathrm{tr}(\rho_{\boldsymbol{\theta}}(\boldsymbol{x})O)}\}$

Which PQCs are learnable?
No clear answer yet.

But we know
some PQCs are learnable,
and we know how to provide
their classical surrogates.



$\boldsymbol{x} \in [0,2\pi]^m$

$\boldsymbol{\theta} \in [0,2\pi]^r$

PRA 103, 032430 (2021).

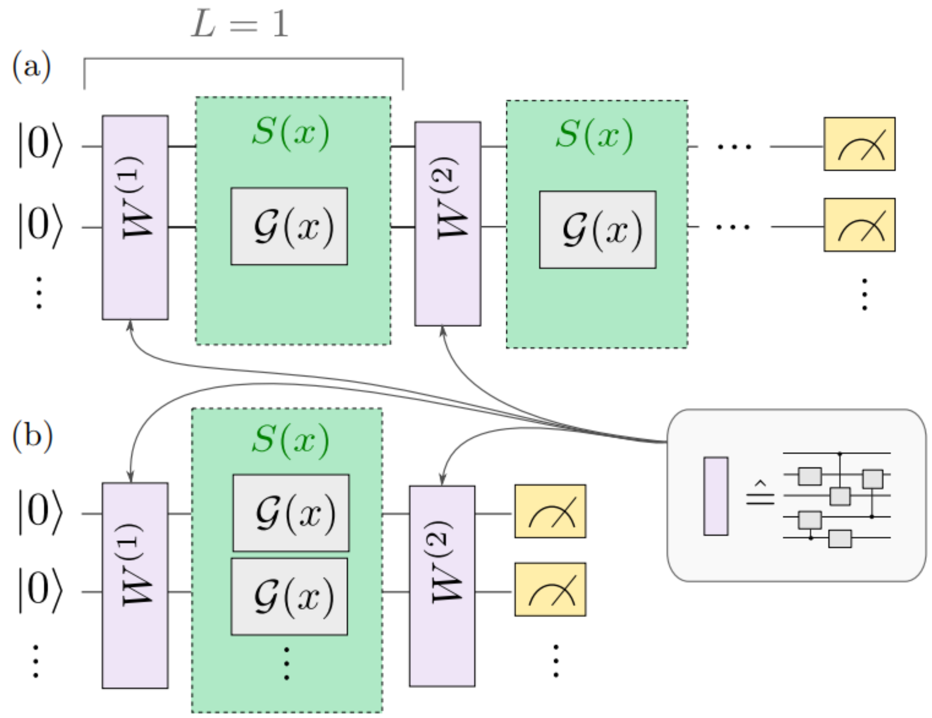
Fourier representation of PQC

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = c_{\omega_0}(\boldsymbol{\theta}) + \sum_{i=1}^{|\Omega|-1} a_{\omega_i}(\boldsymbol{\theta}) \cos(\langle \boldsymbol{\omega}_i, \boldsymbol{x} \rangle) + b_{\omega_i}(\boldsymbol{\theta}) \sin(\langle \boldsymbol{\omega}_i, \boldsymbol{x} \rangle)$$

$$= \langle \boldsymbol{w}_F(\boldsymbol{\theta}), \boldsymbol{\phi}_F(\boldsymbol{x}) \rangle$$

$$\sqrt{|\Omega|} \begin{pmatrix} c_{\omega_0}(\boldsymbol{\theta}) \\ a_{\omega_1}(\boldsymbol{\theta}) \\ b_{\omega_1}(\boldsymbol{\theta}) \\ \vdots \\ a_{\omega_{|\Omega|-1}}(\boldsymbol{\theta}) \\ b_{\omega_{|\Omega|-1}}(\boldsymbol{\theta}) \end{pmatrix} \qquad \frac{1}{\sqrt{|\Omega|}} \begin{pmatrix} 1 \\ \cos(\langle \boldsymbol{\omega}_1, \boldsymbol{x} \rangle) \\ \sin(\langle \boldsymbol{\omega}_1, \boldsymbol{x} \rangle) \\ \vdots \\ \cos(\langle \boldsymbol{\omega}_{|\Omega|-1}, \boldsymbol{x} \rangle) \\ \sin(\langle \boldsymbol{\omega}_{|\Omega|-1}, \boldsymbol{x} \rangle) \end{pmatrix}$$

$$\mathcal{F}_{U,O} = \{ f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \underbrace{\langle \boldsymbol{w}_F(\boldsymbol{\theta}), \boldsymbol{\phi}_F(\boldsymbol{x}) \rangle}_{\mathrm{tr}(\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) O)} \}$$



Classical machine learning models

PQC models: $\boxed{f_{\boldsymbol{\theta}}(\boldsymbol{x}) \in [0,1]}$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{w}_F, \boldsymbol{\phi}_F(\boldsymbol{x}) \rangle = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle + \boxed{\xi(\boldsymbol{x})}$$

$$\xi(\boldsymbol{x}) = \langle \boldsymbol{w}_F, \boldsymbol{\phi}_F(\boldsymbol{x}) \rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle$$
$$\xi(\boldsymbol{x}) \in [-M, M]$$
$$\mathbb{E}_{\boldsymbol{x}}[\xi(\boldsymbol{x})^2] \leq \epsilon_1$$

Hypothesis class: $\mathcal{H} = \{h(\boldsymbol{x}) = \langle \boldsymbol{w}, \underbrace{\boldsymbol{\phi}(\boldsymbol{x})} \rangle\}$

$\boxed{\|\boldsymbol{w}\|_2 \leq B, \|\boldsymbol{\phi}(\boldsymbol{x})\|_2 \leq 1}$

Simplify hypothesis class:
- Directly truncate $\Omega \rightarrow D$
- Random Fourier Features

$$\boldsymbol{\phi}(\boldsymbol{x}) = \frac{1}{\sqrt{D}} \begin{pmatrix} 1 \\ \cos(\langle \boldsymbol{\omega}_1, \boldsymbol{x} \rangle) \\ \sin(\langle \boldsymbol{\omega}_1, \boldsymbol{x} \rangle) \\ \vdots \\ \cos(\langle \boldsymbol{\omega}_D, \boldsymbol{x} \rangle) \\ \sin(\langle \boldsymbol{\omega}_D, \boldsymbol{x} \rangle) \end{pmatrix}$$

Empirical risk minimization

PQC models: $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \in [0,1]$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle + \xi(\boldsymbol{x})$$

$$\boldsymbol{\phi}(\boldsymbol{x}) = \frac{1}{\sqrt{D}} \begin{pmatrix} 1 \\ \cos(\langle \boldsymbol{\omega}_1, \boldsymbol{x} \rangle) \\ \sin(\langle \boldsymbol{\omega}_1, \boldsymbol{x} \rangle) \\ \vdots \\ \cos(\langle \boldsymbol{\omega}_D, \boldsymbol{x} \rangle) \\ \sin(\langle \boldsymbol{\omega}_D, \boldsymbol{x} \rangle) \end{pmatrix}$$

$\xi(\boldsymbol{x}) = \langle \boldsymbol{w}_F, \boldsymbol{\phi}_F(\boldsymbol{x}) \rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle$

$\xi(\boldsymbol{x}) \in [-M, M]$

$\mathbb{E}_{\boldsymbol{x}}[\xi(\boldsymbol{x})^2] \leq \epsilon_1$

Minimize $\hat{R}(h)$ to achieve low $R(h)$

Hypothesis class: $\mathcal{H} = \{h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle\}$

$$\boldsymbol{w}^* = \underset{\|\boldsymbol{w}\|_2 < B}{\arg\min} \frac{1}{N_1} \sum_{i=1}^{N_1} (\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}_i) \rangle - y_i)^2$$

$$R(h_{\boldsymbol{w}^*}) = \mathbb{E}_{\boldsymbol{x}}[(h_{\boldsymbol{w}^*}(\boldsymbol{x}) - f_{\boldsymbol{\theta}}(\boldsymbol{x}))^2]$$

- Constrained convex optimization
- Kernel ridge regression with line search

Centre for Quantum Technologies × the entangled computing lab

Guarantees for empirical risk minimization

PQC models: $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \in [0,1]$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle + \xi(\boldsymbol{x})$$

$$\boldsymbol{\phi}(\boldsymbol{x}) = \frac{1}{\sqrt{D}} \begin{pmatrix} 1 \\ \cos(\langle \boldsymbol{\omega}_1, \boldsymbol{x} \rangle) \\ \sin(\langle \boldsymbol{\omega}_1, \boldsymbol{x} \rangle) \\ \vdots \\ \cos(\langle \boldsymbol{\omega}_D, \boldsymbol{x} \rangle) \\ \sin(\langle \boldsymbol{\omega}_D, \boldsymbol{x} \rangle) \end{pmatrix}$$

$\xi(\boldsymbol{x}) = \langle \boldsymbol{w}_F, \boldsymbol{\phi}_F(\boldsymbol{x}) \rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle$

$\xi(\boldsymbol{x}) \in [-M, M]$

$\mathbb{E}_{\boldsymbol{x}}[\xi(\boldsymbol{x})^2] \leq \epsilon_1$

Minimize $\hat{R}(h)$ to achieve low $R(h)$

Hypothesis class: $\mathcal{H} = \{h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle\}$

Lemma 1: $R(h_{\boldsymbol{w}^*}) \leq \epsilon_1 + \tilde{\mathcal{O}}\left(D^2 \sqrt{\frac{1}{N_1}} \right)$

Amount of data to learn: $N_1 \in \mathcal{O}(D^4)$

- Can we do better for the number of data?
- No indication on number of shots N_S
- Does a trade off between N_1 and N_S exist?

Classical machine learning models

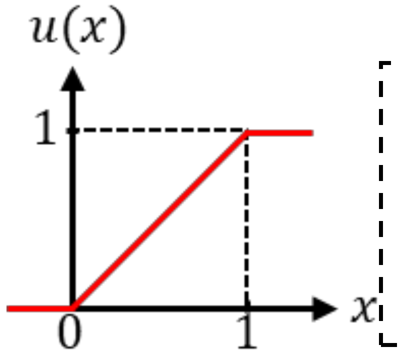
PQC models: $\boxed{f_{\boldsymbol{\theta}}(\boldsymbol{x}) \in [0,1]}$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{w}_F, \boldsymbol{\phi}_F(\boldsymbol{x})\rangle = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x})\rangle + \boxed{\xi(\boldsymbol{x})}$$
$$= u\big(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x})\rangle + \xi(\boldsymbol{x})\big)$$

$$\xi(\boldsymbol{x}) = \langle \boldsymbol{w}_F, \boldsymbol{\phi}_F(\boldsymbol{x})\rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x})\rangle$$
$$\xi(\boldsymbol{x}) \in [-M, M]$$
$$\mathbb{E}_{\boldsymbol{x}}[\xi(\boldsymbol{x})^2] \le \epsilon_1$$

Hypothesis class: $\mathcal{H} = \{h(\boldsymbol{x}) = u(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x})\rangle)\}$

$\boxed{\|\boldsymbol{w}\|_2 \le B, \|\boldsymbol{\phi}(\boldsymbol{x})\|_2 \le 1}$



$$u(x) = \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{if } 0 \le x \le 1, \\ 1, & \text{if } x > 1, \end{cases}$$

$$\boldsymbol{\phi}(\boldsymbol{x}) = \frac{1}{\sqrt{D}} \begin{pmatrix} 1 \\ \cos(\langle \boldsymbol{\omega}_1, \boldsymbol{x}\rangle) \\ \sin(\langle \boldsymbol{\omega}_1, \boldsymbol{x}\rangle) \\ \vdots \\ \cos(\langle \boldsymbol{\omega}_D, \boldsymbol{x}\rangle) \\ \sin(\langle \boldsymbol{\omega}_D, \boldsymbol{x}\rangle) \end{pmatrix}$$

The learning algorithm

Algorithm 1: The learning algorithm

Input: Training data size N_1, validation data size N_2, number of measurement shots N_s, parameter setting of quantum model $\boldsymbol{\theta}$, distribution of input $p(\boldsymbol{x})$, non-decreasing L-Lipschitz function $u : \mathbb{R} \to \mathcal{Y}$, kernel function k corresponding to feature map $\boldsymbol{\phi}$, learning rate $\lambda > 0$, number of iterations T

1 Sample N_1 training data inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_1} \sim p(\boldsymbol{x})$.

Just think of it as kernelized gradient descent with a validation dataset.

training dataset $(\boldsymbol{x}_i, y_i)_{i=1}$.

3 Repeat steps 1 and 2 to collect labelled validation data of size N_2, $(\boldsymbol{p}_j, \bar{q}_j)_{j=1}^{N_2}$.

4 Initialize $\alpha^i := 0 \in \mathbb{R}^{N_1}$.

5 **for** $t = 1, \ldots, T$ **do**

6 $\quad h^t(\boldsymbol{x}) := u\left(\sum_{i=1}^{N_1} \alpha_i^t k(\boldsymbol{x}, \boldsymbol{x}_i)\right)$

7 $\quad$ **for** $i = 1, 2, \ldots, N_1$ **do**

8 $\quad\quad \alpha_i^{t+1} := \alpha_i^t + \frac{\lambda}{N_1}(\bar{y}_i - h^t(\boldsymbol{x}_i))$

Output: h^r where

$$r = \arg\min_{t \in \{1,\ldots,T\}} \frac{1}{N_2} \sum_{j=1}^{N_2} (\bar{q}_j - h^t(\boldsymbol{p}_j))^2$$

Provable guarantee on concept learning

Apply gradient descent* on data to achieve low $R(h)$.

Theorem 1: $R(h) \leq \tilde{O}\left(\sqrt{\epsilon_1} + M\sqrt[4]{\frac{1}{N_1}} + D\sqrt{\frac{1}{N_1}} + D\sqrt{\frac{1}{N_1 N_s}}\right)$

Lemma 1: $R(h_{\boldsymbol{w}^*}) \leq \epsilon_1 + \tilde{O}\left(D^2\sqrt{\frac{1}{N_1}}\right)$

Provable guarantee on concept learning

Apply gradient descent* on data to achieve low $R(h)$.

Theorem 1: $R(h) \leq \tilde{O}\left(\sqrt{\epsilon_1} + M\sqrt[4]{\frac{1}{N_1}} + D\sqrt{\frac{1}{N_1}} + D\sqrt{\frac{1}{N_1 N_s}}\right)$

$\mathbb{E}_{\boldsymbol{x}}[\xi(\boldsymbol{x})^2] \leq \epsilon_1 \qquad \xi(\boldsymbol{x}) \in [-M, M]$

$\xi(\boldsymbol{x}) = \langle \boldsymbol{w}_F, \boldsymbol{\phi}_F(\boldsymbol{x}) \rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle$

$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{w}_F, \boldsymbol{\phi}_F(\boldsymbol{x}) \rangle \qquad h(\boldsymbol{x}) = u(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle)$

Quantum models cannot be learned without a (fairly) efficient and good classical approximation.

Bias from the model representation.

Centre for Quantum Technologies × the entangled computing lab

Asymmetrical effects of N_1 and N_s

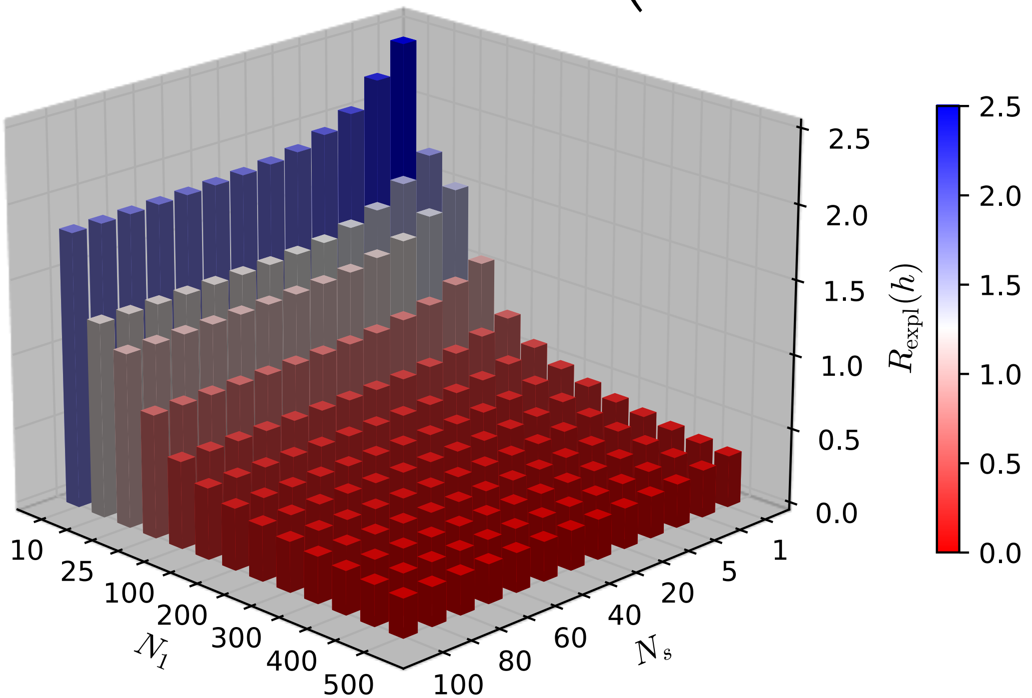
Corollary 1: $R(h) \leq \tilde{O}\left(D\sqrt{\frac{1}{N_1}} + D\sqrt{\frac{1}{N_1 N_s}}\right)$

$N_1 \to \infty$

$R(h) \leq 0$

$N_s \to \infty$

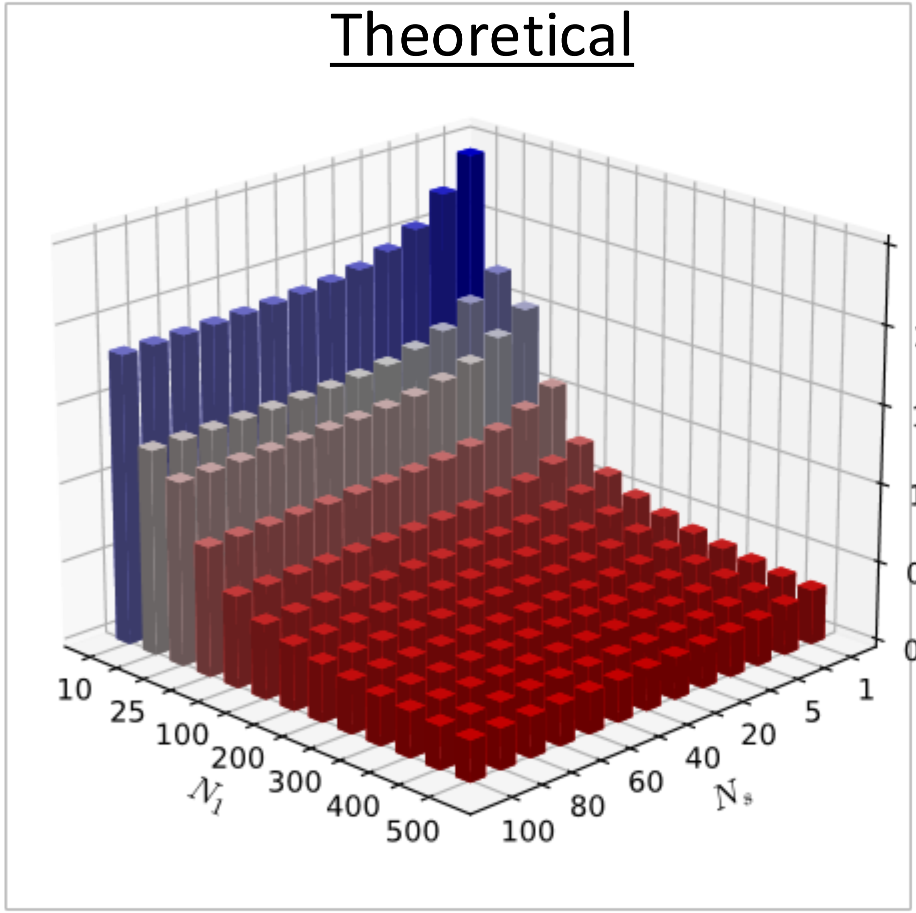
$R(h) \leq \tilde{O}\left(D\sqrt{\frac{1}{N_1}}\right)$

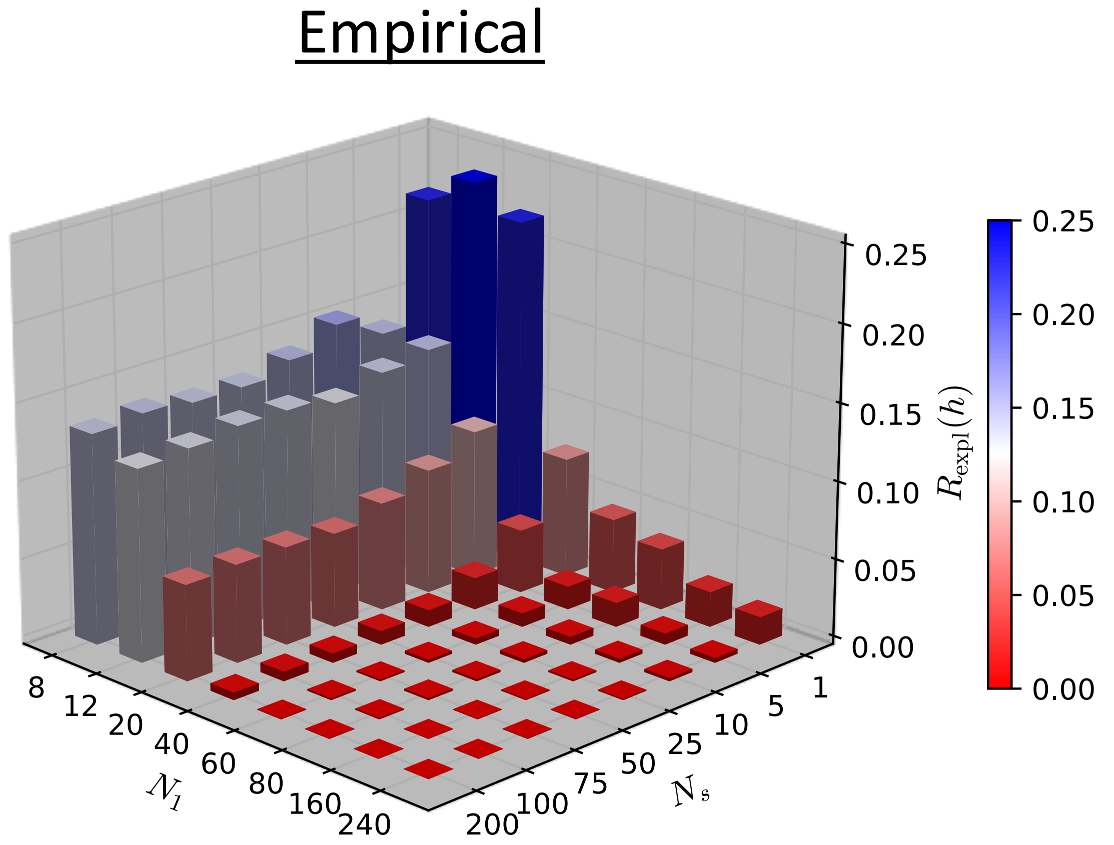


N_1 and N_s have asymmetrical effects on the model's learning performance.

Centre for Quantum Technologies × the entangled computing lab

Asymmetrical effects of N_1 and N_s



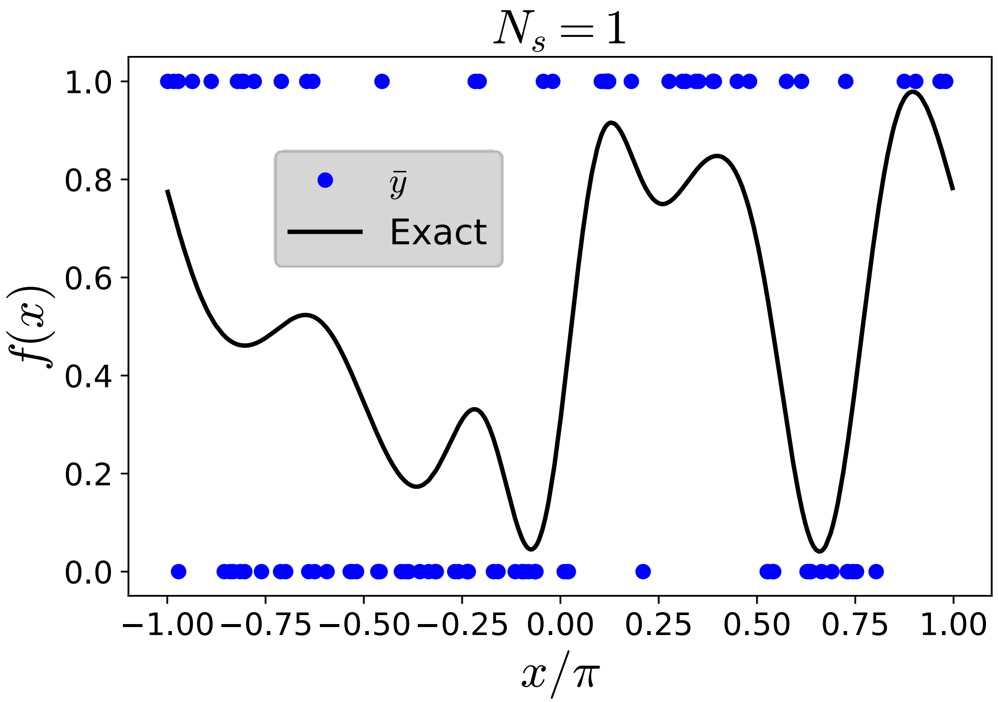


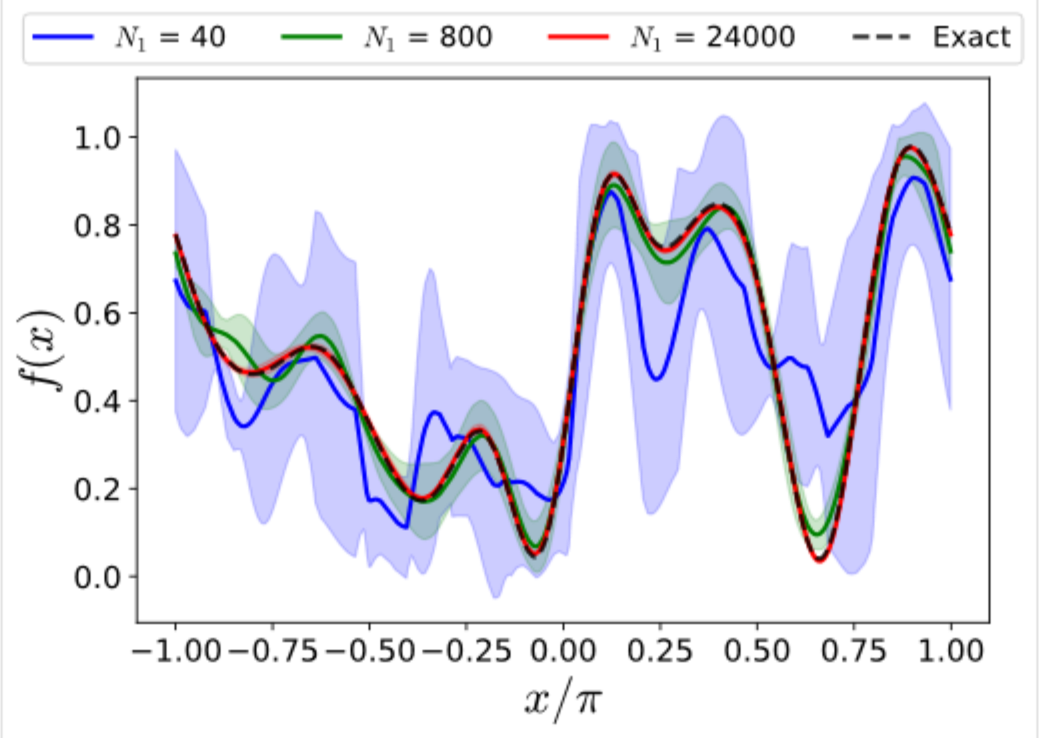
Learning PQC models with $N_s = 1$

Corollary 1: $R(h) \leq \tilde{O}\left(D\sqrt{\frac{1}{N_1}} + D\sqrt{\frac{1}{N_1 N_s}}\right)$

$N_1 \rightarrow \infty$

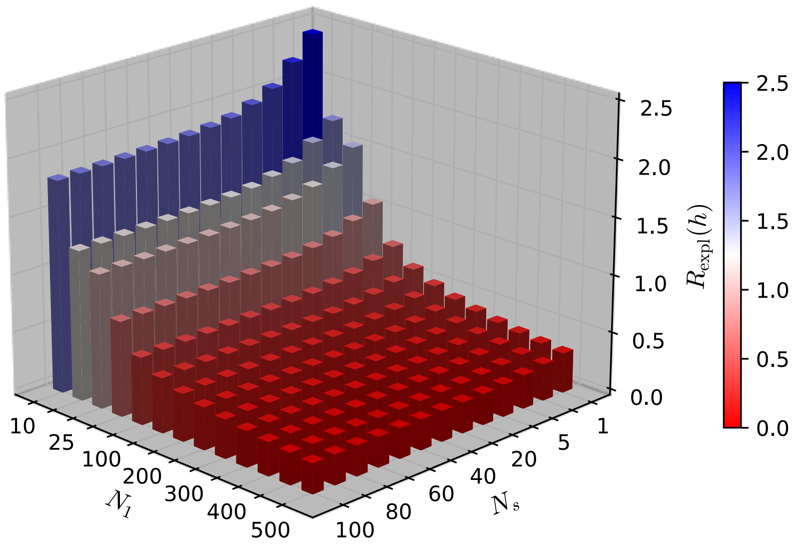
$R(h) \leq 0$





Centre for Quantum Technologies × the entangled computing lab

Trade-off between N_1 and N_s



Assumed we have unlimited queries to quantum systems.

Access to quantum computers is expensive.

Hardware Provider	QPU family	Per-task price	Per-shot price
IonQ	Harmony	\$0.30000	\$0.01000
IonQ	Aria	\$0.30000	\$0.03000
IQM	Garnet	\$0.30000	\$0.00145
QuEra	Aquila	\$0.30000	\$0.01000
Rigetti	Aspen-M	\$0.30000	\$0.00035

Some hardware is more expensive than others.

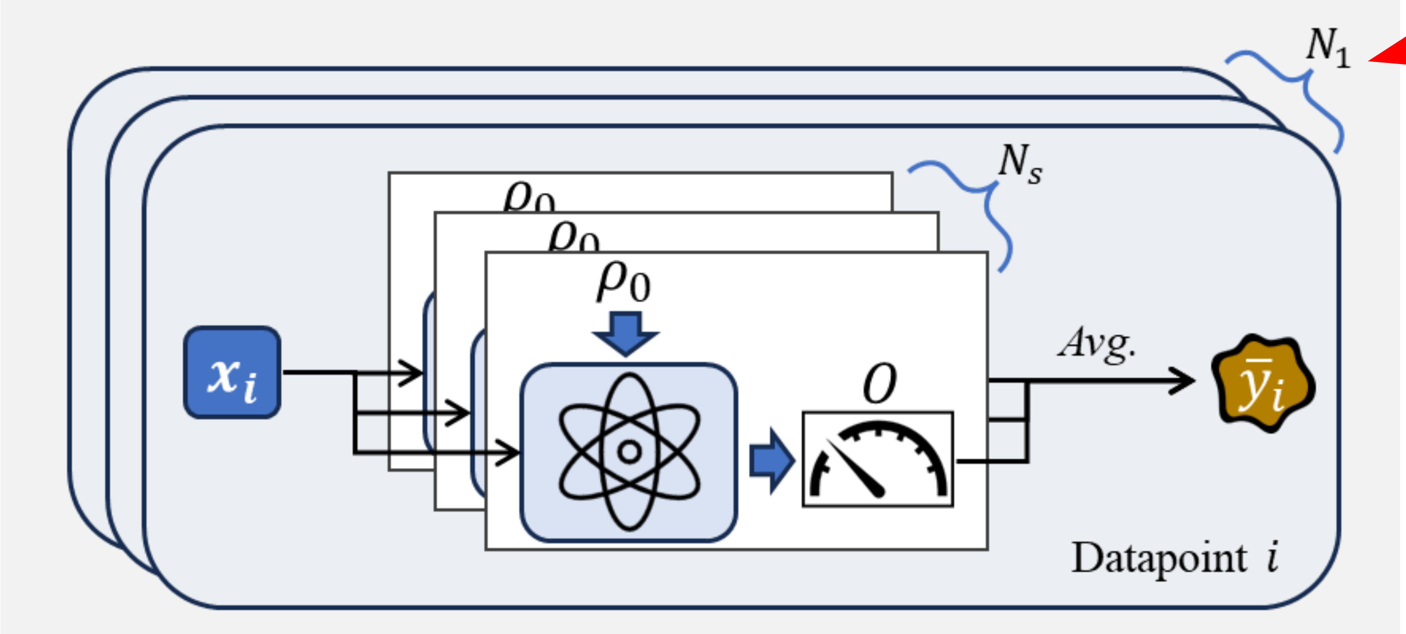
Centre for Quantum Technologies × the entangled computing lab

Trade-off between N_1 and N_s

Limit the total number of queries to quantum systems.

Total measurement budget: $N_{\text{tot}} = N_1 \cdot (N_s + \gamma)$

Extra $\gamma \in \mathbb{R}^+$ for changing parameter settings



- $\gamma = 0$: $N_{\text{tot}} = N_1 \cdot N_s$
- $\gamma_{\text{Trapped Ions}} > \gamma_{\text{Superconducting}}$

Trade-off between N_1 and N_s

Corollary 1:

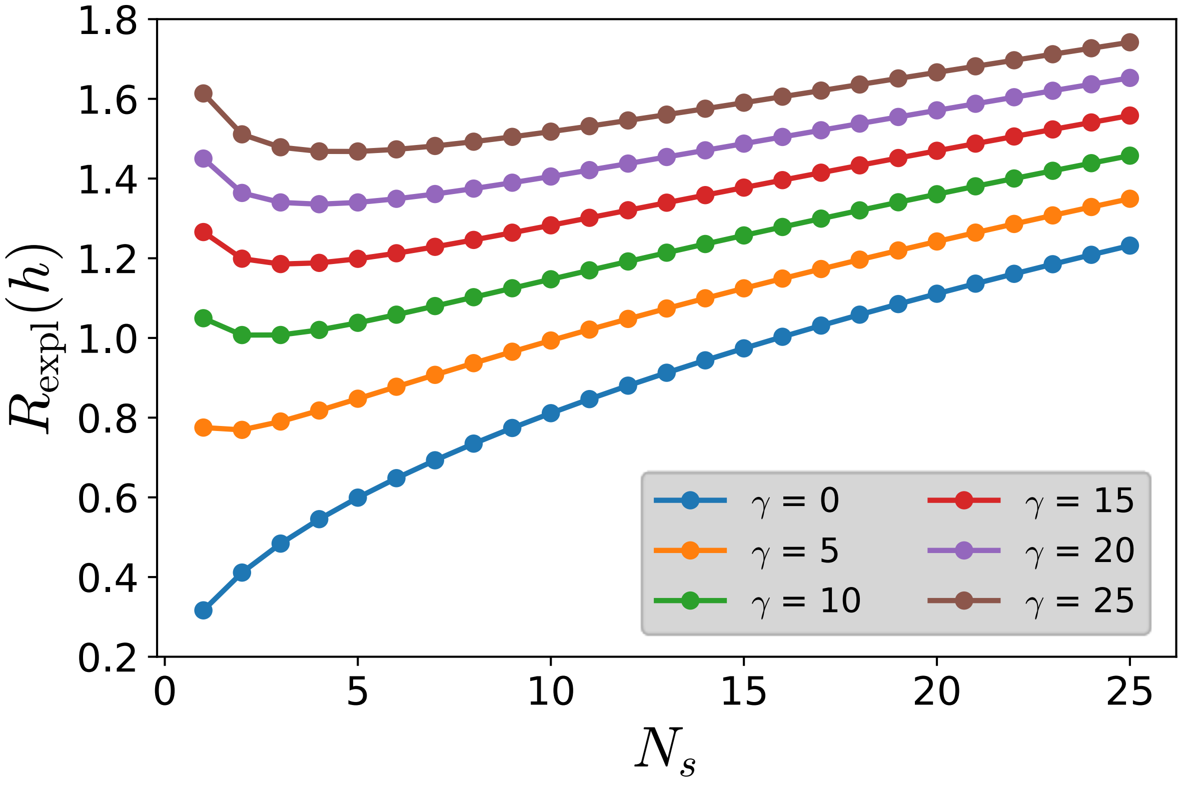
$$R(h) \le \tilde{\mathcal{O}}\left(D\sqrt{\frac{1}{N_1}} + D\sqrt{\frac{1}{N_1 N_s}} \right)$$

$$N_{\text{tot}} = N_1 \cdot (N_s + \gamma)$$

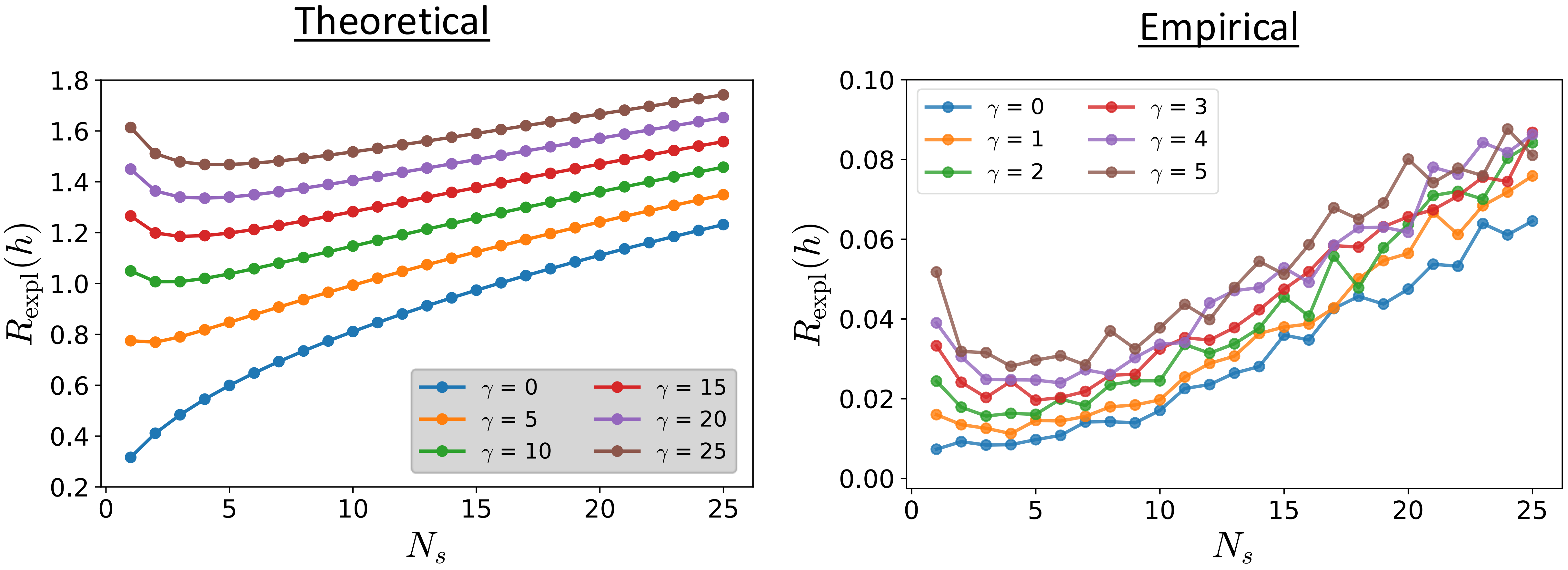
$$\frac{1}{N_1} = \frac{N_s + \gamma}{N_{\text{tot}}}$$

Corollary 2:

$$R(h) \le \tilde{\mathcal{O}}\left(D\sqrt{\frac{N_s + \gamma}{N_{\text{tot}}}} + D\sqrt{\frac{N_s + \gamma}{N_{\text{tot}} N_s}} \right)$$



Trade-off between N_1 and N_s



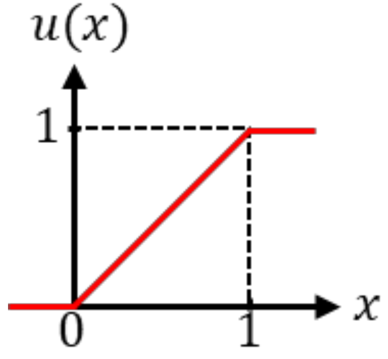
Centre for Quantum Technologies × the entangled computing lab

Role of the link function u

With link function

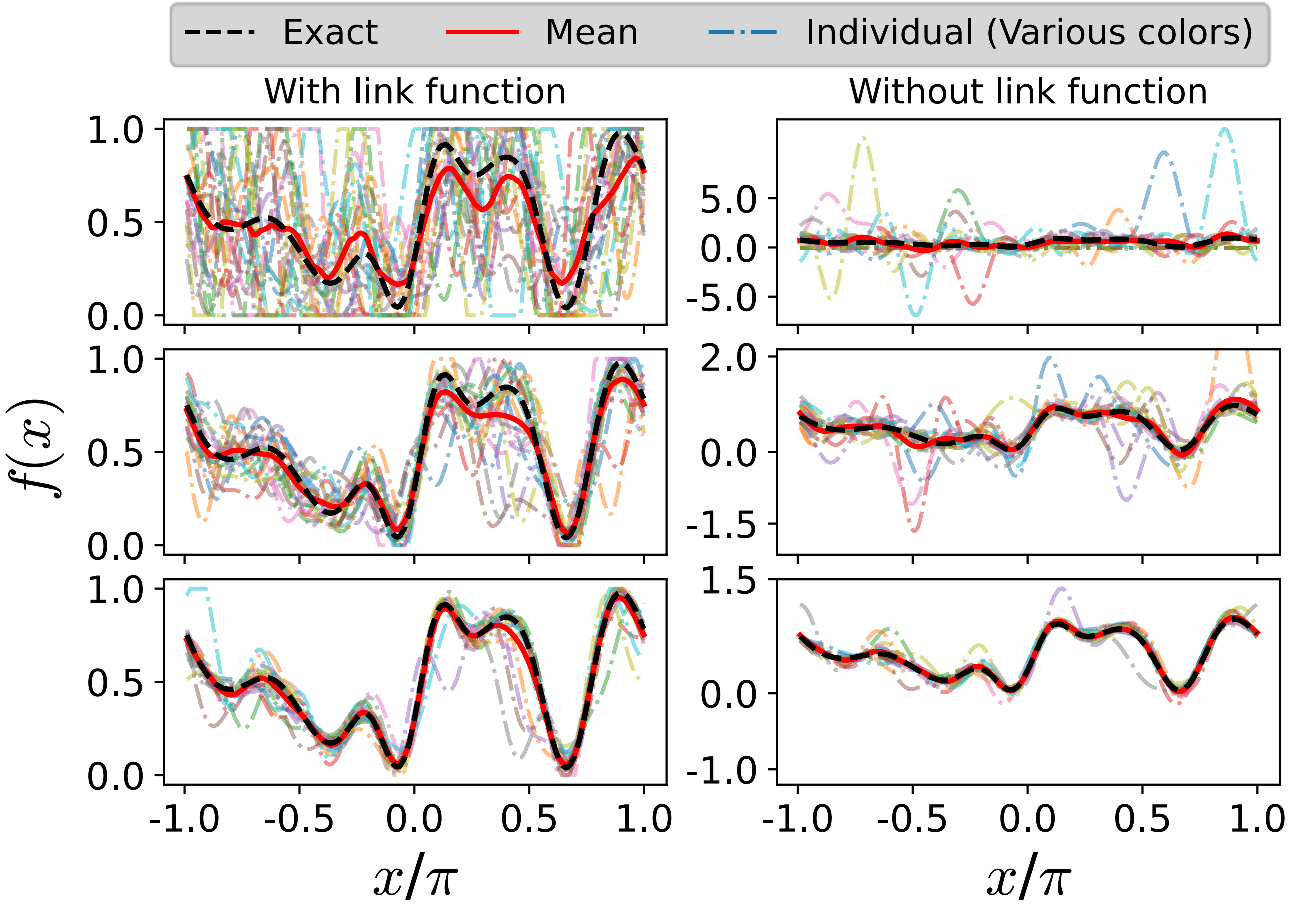
$$h(\boldsymbol{x}) = \underbrace{u}(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle)$$

$$u(x) = \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{if } 0 \le x \le 1, \\ 1, & \text{if } x > 1, \end{cases}$$



Without link function

$$g(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle$$

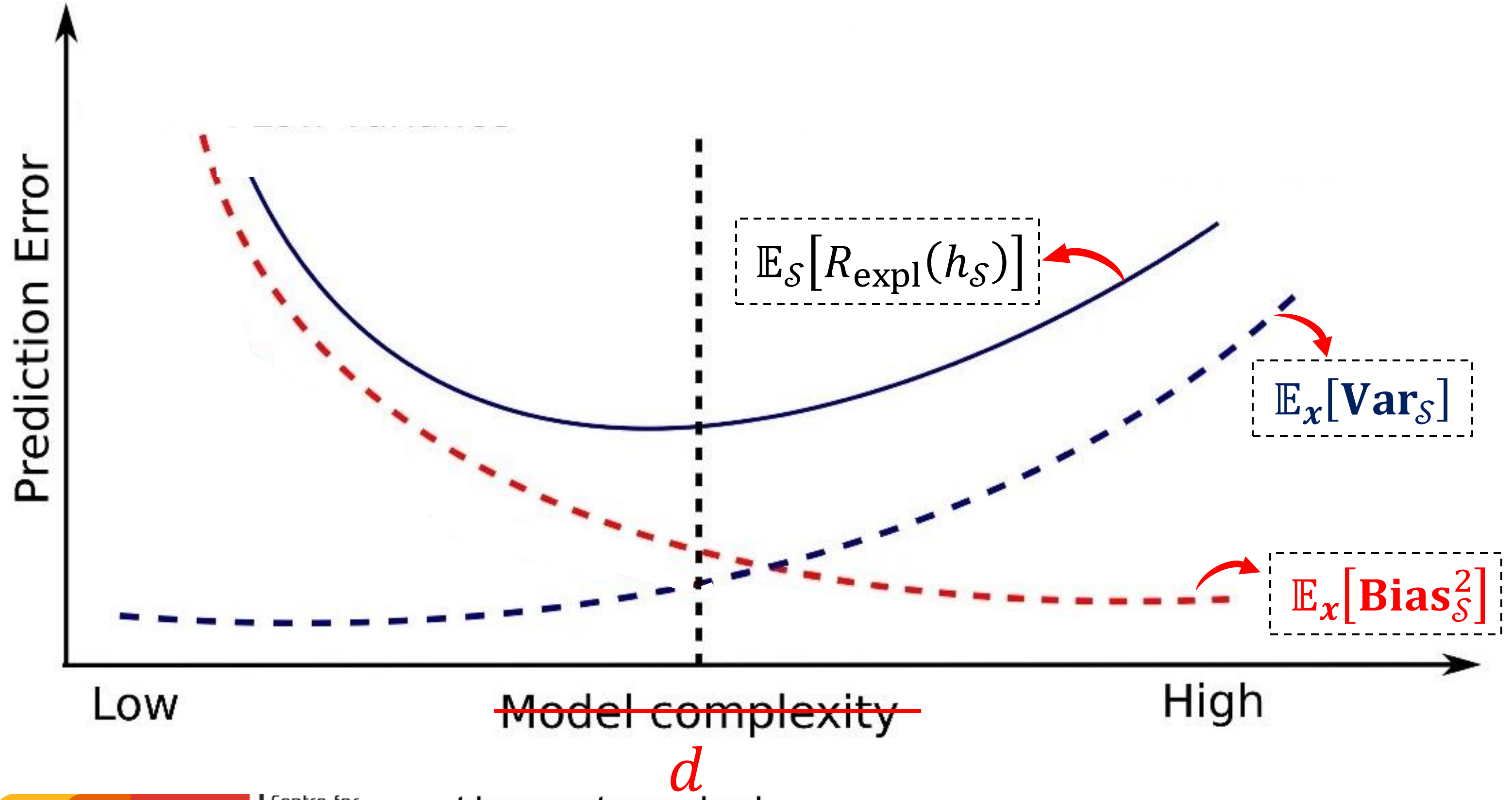


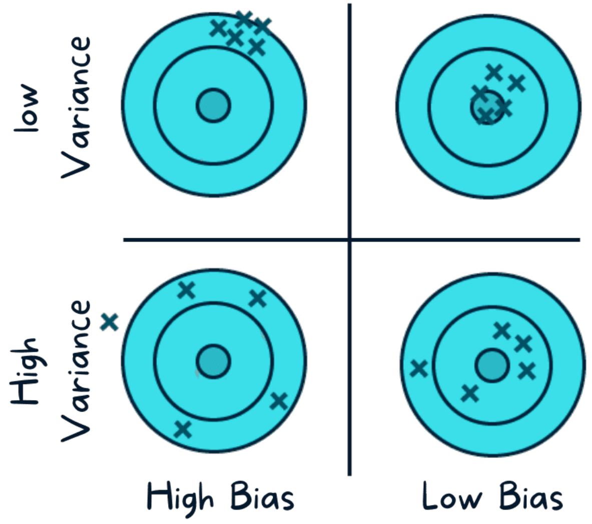
$$R(h) = \mathbb{E}_{(\boldsymbol{x},\bar{y})\sim\bar{\mathcal{D}}}\left[\left(h(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right]$$

Bias-variance trade-off

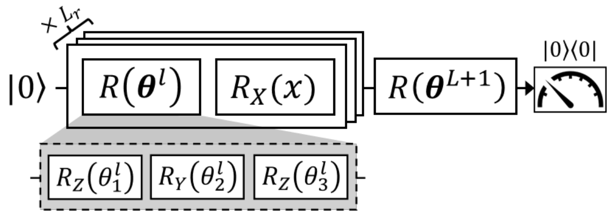
Analyze the average $R(h)$ over all possible training datasets

$$\mathbb{E}_{\mathcal{S}}[R(h_{\mathcal{S}})] = \mathbb{E}_{\boldsymbol{x}}\left[\mathbf{Bias}_{\mathcal{S}}^2\right] + \mathbb{E}_{\boldsymbol{x}}[\mathbf{Var}_{\mathcal{S}}]$$





Data re-uploading models:



Concept class:

$$\mathcal{F}_{L_r} = \left\{ f_{\boldsymbol{\theta}}(\boldsymbol{x}) = c_0(\boldsymbol{\theta}) + \sum_{\omega=1}^{L_r} a_\omega(\boldsymbol{\theta})\cos(\omega x) + b_\omega(\boldsymbol{\theta})\sin(\omega x) \right\}$$

Hypothesis class:

$$\mathcal{H}_d = \left\{ h_d(x) = u\left(v_0 + \sum_{\omega=1}^{d} \alpha_\omega \cos(\omega x) + \beta_\omega \sin(\omega x) \right) \right\}$$

Centre for Quantum Technologies × the entangled computing lab

Provable guarantee on concept learning

Apply gradient descent* on data to achieve low $R(h)$.

Theorem 1: $R(h) \leq \tilde{O}\left(\sqrt{\epsilon_1} + M\sqrt[4]{\frac{1}{N_1}} + D\sqrt{\frac{1}{N_1}} + D\sqrt{\frac{1}{N_1 N_s}}\right)$

Bias

Variance

Lemma 1: $R(h_{\boldsymbol{w}^*}) \leq \epsilon_1 + \tilde{O}\left(D^2\sqrt{\frac{1}{N_1}}\right)$

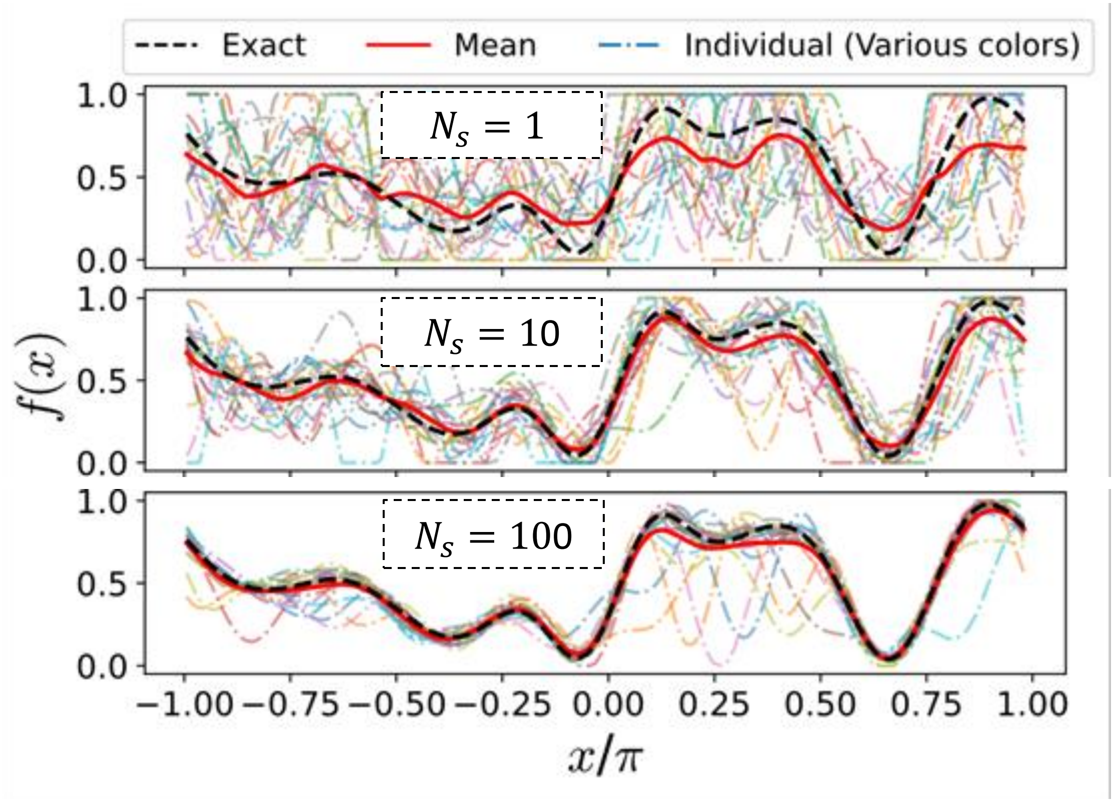
The link function limits the model complexity and forcibly compresses the variance upper bound.

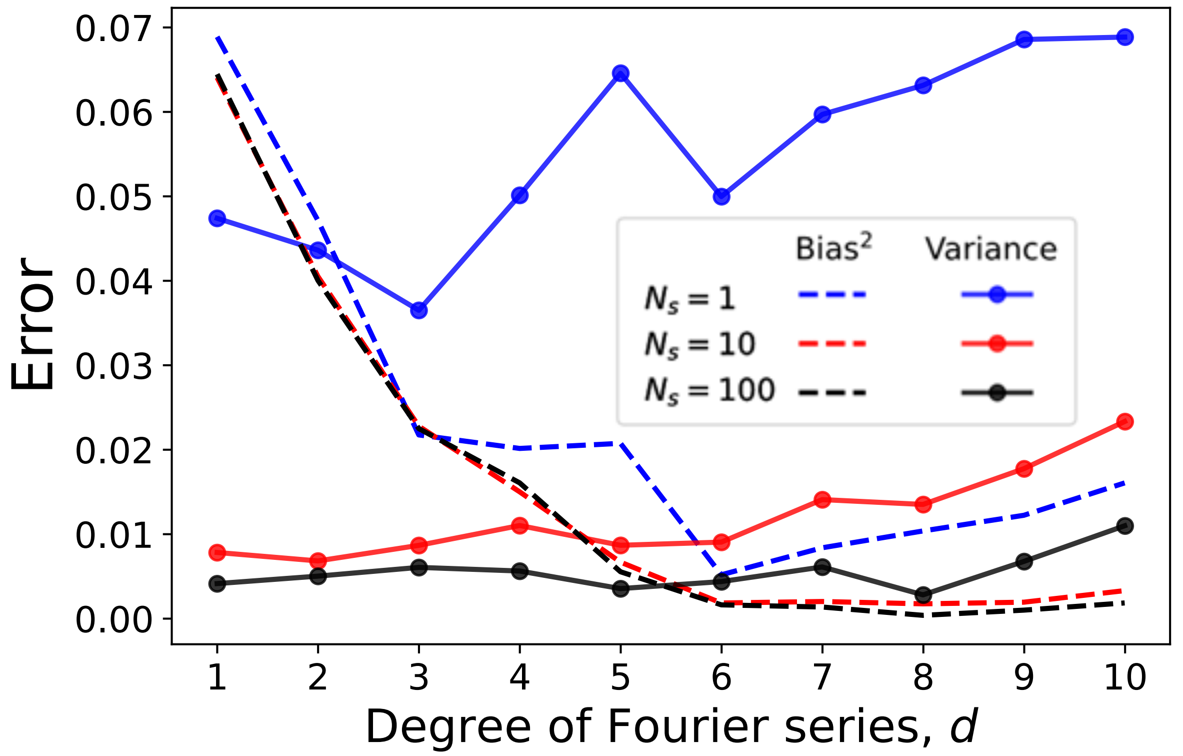
$R_{\mathrm{expl}}(h) = \mathbb{E}_{(\boldsymbol{x},\bar{y})\sim\bar{\mathcal{D}}}[(h(\boldsymbol{x}) - f(\boldsymbol{x}))^2]$

Effects of shot noise on bias and variance

Analyze the average $R_{\mathrm{expl}}(h)$ over all possible training datasets

$$\mathbb{E}_S\left[R_{\mathrm{expl}}(h_S)\right] = \mathbb{E}_{\boldsymbol{x}}\left[\mathbf{Bias}_S^2\right] + \mathbb{E}_{\boldsymbol{x}}[\mathbf{Var}_S]$$





Centre for Quantum Technologies × the entangled computing lab

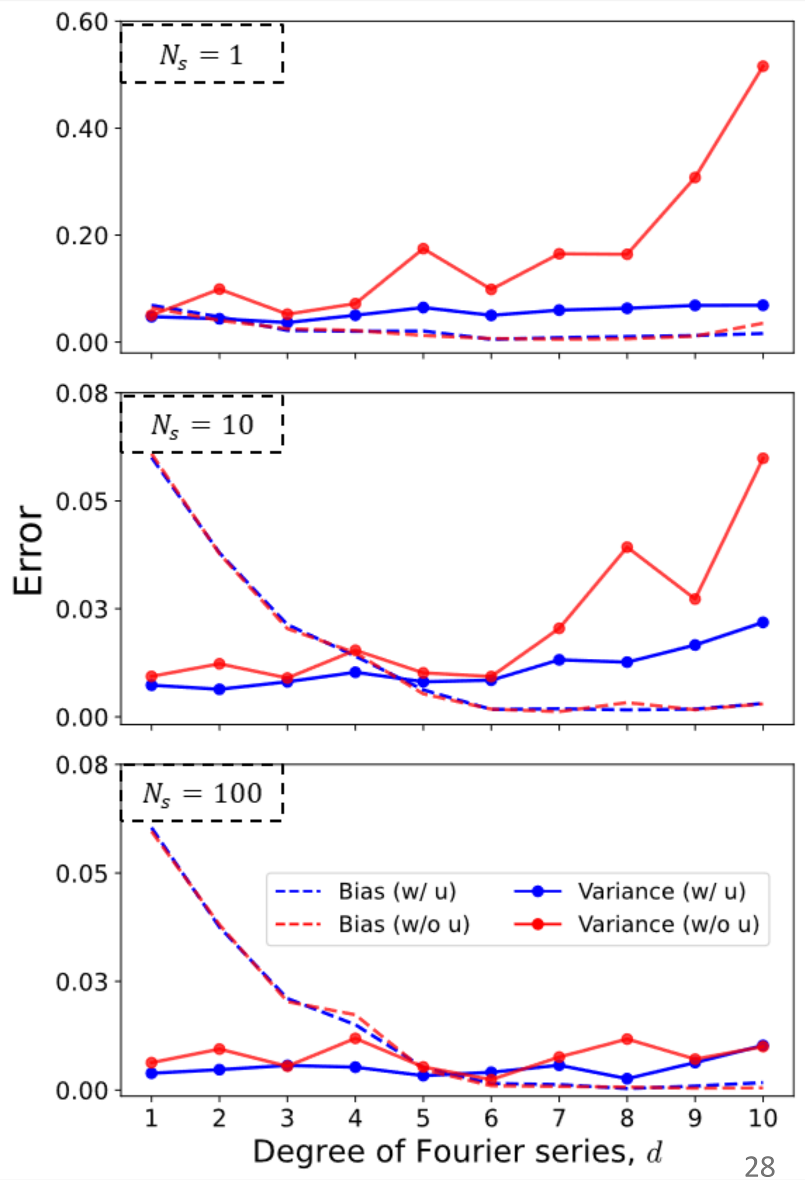
Learning with and without the link function

<u>Hypothesis class (with link function):</u>

$$\mathcal{H}_d = \left\{ h_d(x) = u\left(v_0 + \sum_{\omega=1}^{d} \alpha_\omega \cos(\omega x) + \beta_\omega \sin(\omega x) \right) \right\}$$

<u>Hypothesis class (without link function):</u>

$$G_d = \left\{ g_d(x) = v_0 + \sum_{\omega=1}^{d} \alpha_\omega \cos(\omega x) + \beta_\omega \sin(\omega x) \right\}$$



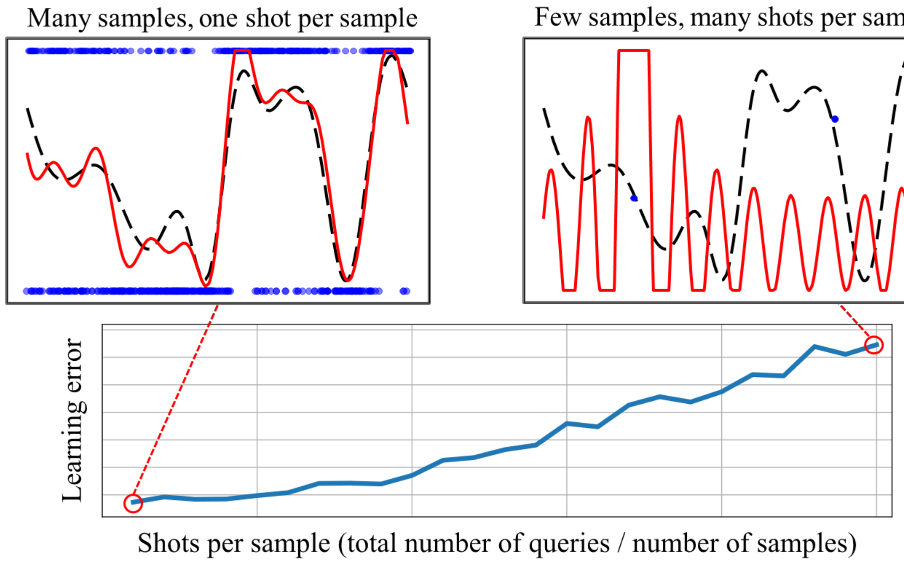
Centre for Quantum Technologies × the entangled computing lab

Conclusion

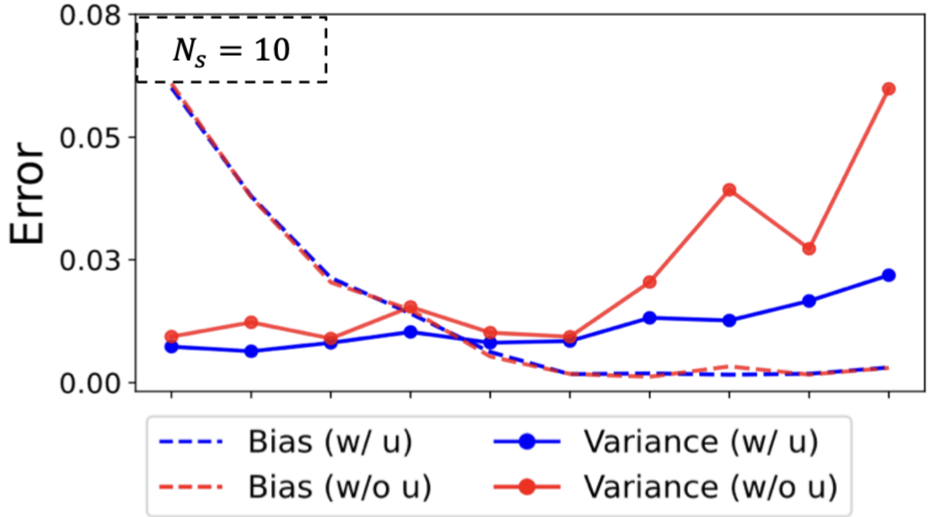
Can we obtain provable guarantees of learning that exemplify the relationship between N_1 and N_s?

arXiv:2408.05116 [quant-ph]

(1) Asymmetrical trade-offs between N_1 and N_s



(2) Gradient descent* can be made robust and provide tighter guarantees



Centre for Quantum Technologies × the entangled computing lab