Barren plateaus, trainability issues and how to avoid them

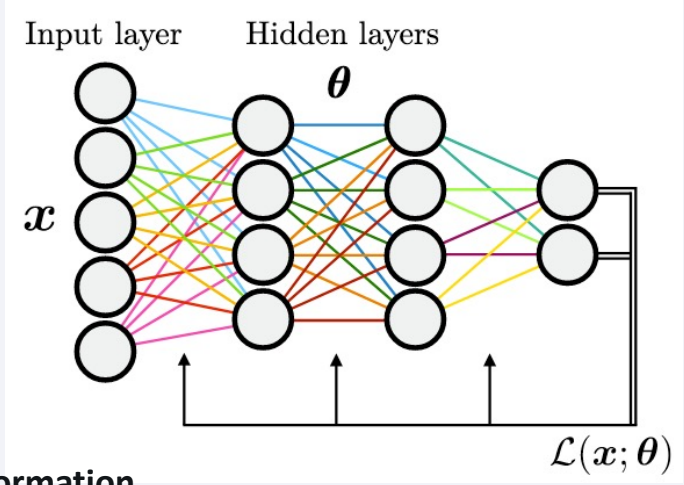
Francesco Tacchino
Quantum Applications Researcher
IBM Quantum, IBM Research – Zurich

Qiskit

Quantum neural networks

Qiskit

Classical feed-forward design



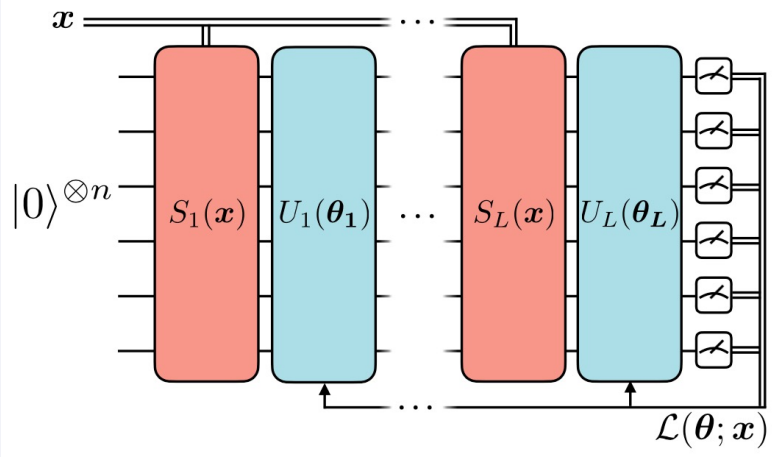
Information

Training

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(W_{\boldsymbol{\theta}}\boldsymbol{x} + b_{\boldsymbol{\theta}})$$

$$y_{NN}(\boldsymbol{x}) = f_{\boldsymbol{\theta}_L} \circ f_{\boldsymbol{\theta}_{L-1}} \dots \circ f_{\boldsymbol{\theta}_1}(\boldsymbol{x})$$

QNN ~ parametrized quantum circuits



$$y_{QNN}(\boldsymbol{x}) = \langle 0|\mathcal{U}^{\dagger}(\boldsymbol{x};\boldsymbol{\theta})M\mathcal{U}(\boldsymbol{x};\boldsymbol{\theta})|0\rangle$$

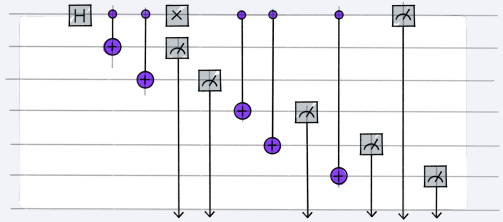
$$\mathcal{U}(\boldsymbol{x};\boldsymbol{\theta}) = \prod_i U_i(\boldsymbol{\theta}_i)S_i(\boldsymbol{x})$$

- Similar layered structure, but different “information flow”
- Quantum layers: **(linear)** unitary operations

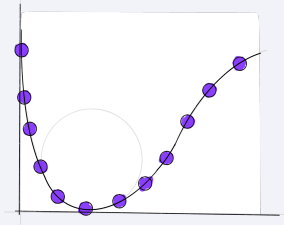
Solving problems with parametrized quantum circuits

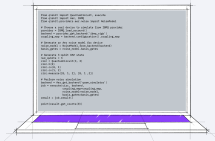
Qiskit

PROBLEM

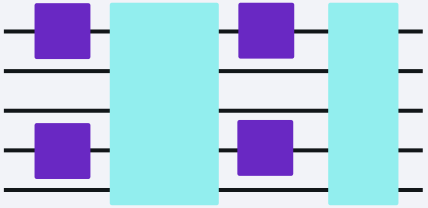


SOLUTION



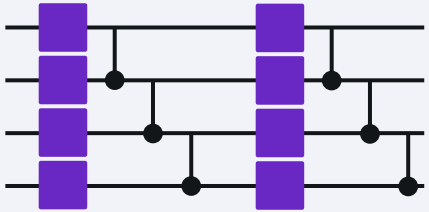


“Educated guess”

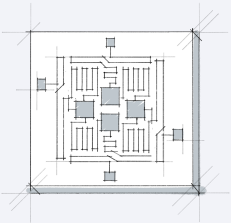


Example: UCC quantum circuits for chemistry

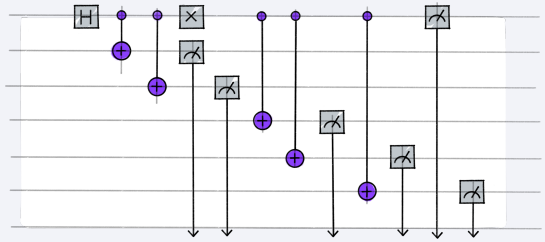
Heuristic/Hardware efficient



Example: R_y-$CNOT$ ansatz

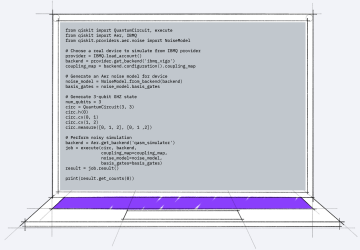


Vanishing gradients



$\boldsymbol{C(\theta)}$

$\boldsymbol{\theta'}$



$$C(\boldsymbol{\theta}) = \langle 0|U^{\dagger}(\boldsymbol{\theta})OU(\boldsymbol{\theta})|0\rangle$$

Cost function for the optimization problem

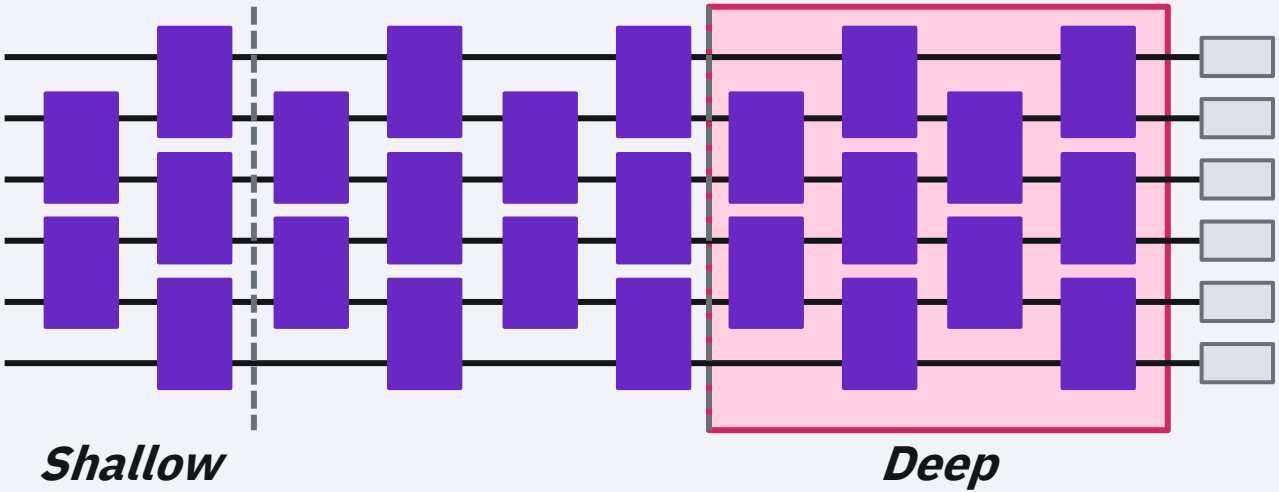
- "Deep" random parametrized circuits $d \sim O(\text{poly}(n))$
- Random initialization of parameters

$$\langle \partial_k C \rangle = 0$$

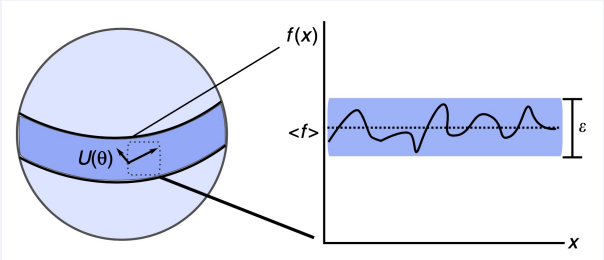
$$\text{Var}[\partial_k C] \approx 2^{-n}$$

- The circuit exhibits the characteristics of a **2-design**
- Gradients vanish **exponentially** with the number of qubits

Brief history of barren plateaus in QNNs (I)



J. McClean *et al.*, Nature Commun. **9**, 4812 (2018)



Random PQC exhibit barren plateaus for depth

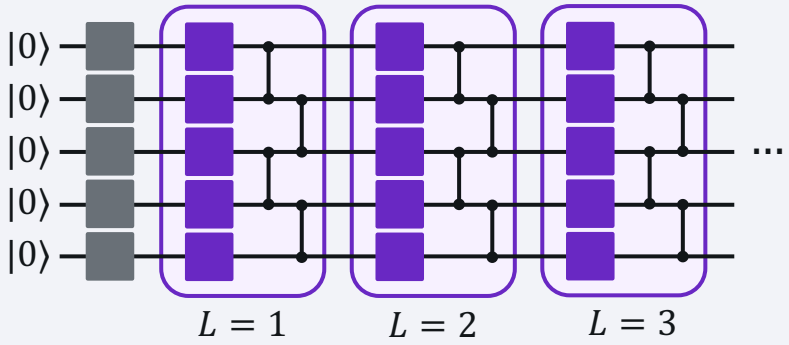
$$O(n^{1/L})$$

on L-dimensional arrays

$$C(\boldsymbol{\theta}) = \langle 0|U^{\dagger}(\boldsymbol{\theta})OU(\boldsymbol{\theta})|0\rangle$$

$$O = c_0\mathbb{I} + \sum_i \bigotimes_j c_i O_{ij}$$

Example: barren plateaus in deep QNNs



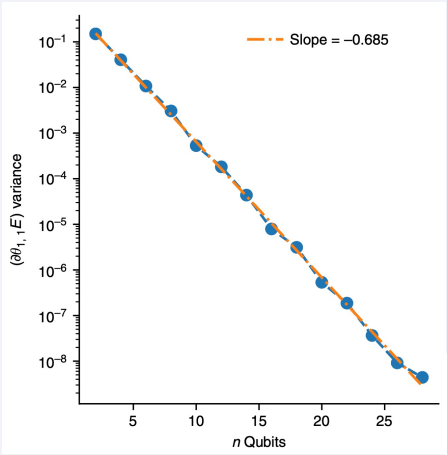
$= R_y\ (\pi/4)$

$= R_P\ (\theta_{ij})$

$= \mathrm{CZ}$

$$E(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|Z_1 Z_2|\psi(\boldsymbol{\theta})\rangle$$

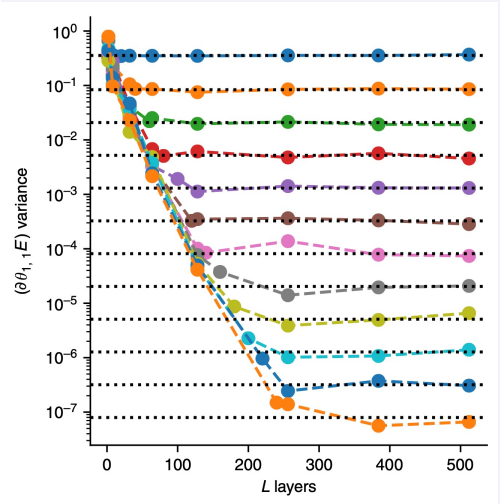
Similar results for $E(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|H|\psi(\boldsymbol{\theta})\rangle$ with $H = |0\rangle\langle 0|^{\otimes n}$



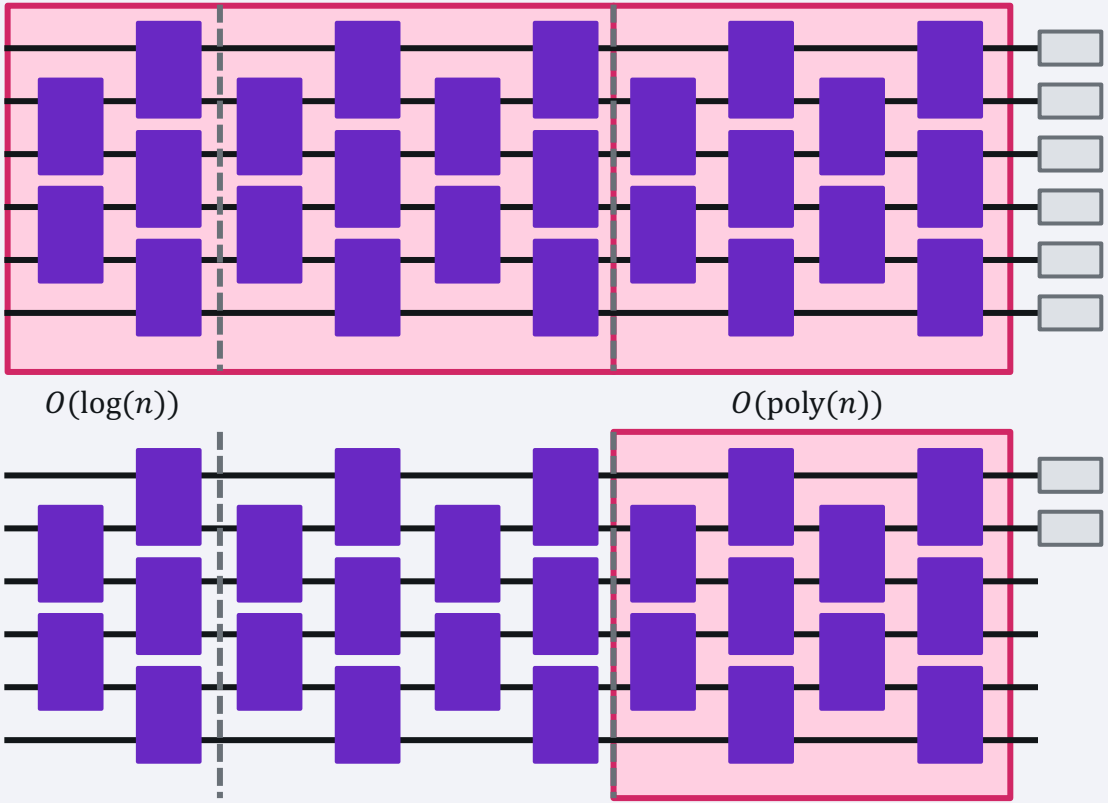
Convergence to the 2-design properties (the asymptotic value of the variance depends on n)

J. McClean *et al.*, Nature Commun. **9**, 4812 (2018)

The variance of the gradient decreases exponentially with the number of qubits $(L{\sim}10n)$



Brief history of barren plateaus in QNNs (II)



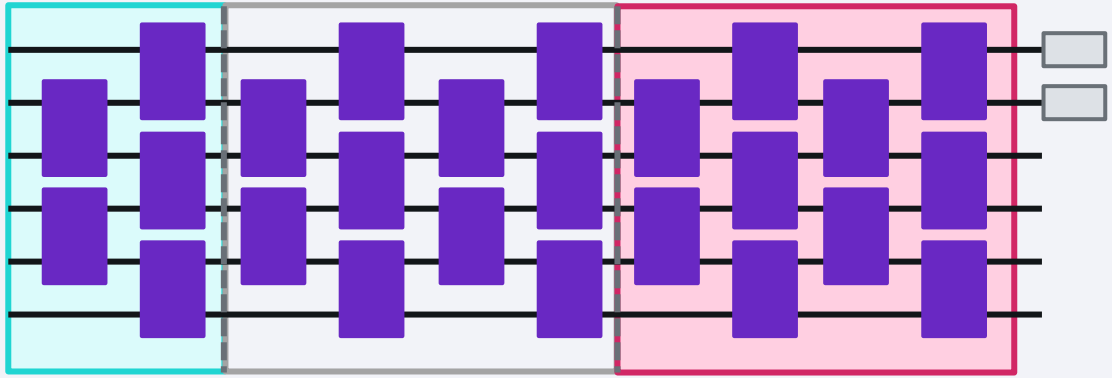
J. McClean *et al.*, Nature Commun. **9**, 4812 (2018)

$$O = c_0 \mathbb{I} + \sum_i \bigotimes_j c_i O_{ij}$$

M. Cerezo *et al.*, Nature Commun. **12**, 1791 (2021)

$$O = c_0 \mathbb{I} + \sum_i c_i O_i$$

A new hope: shallow circuits, local cost functions



M. Cerezo *et al.*, Nature Commun. **12**, 1791 (2021)

$$O = c_0 \mathbb{I} + \sum_i c_i O_i$$

where O_i acts on at most m qubits

For an alternating layered ansatz with a **local cost function**

$$G_n(L, l) \leq \text{Var}[\partial_v C]$$

If L is $O(\log(n))$ → $G_n(L, l) \in \Omega\left(\frac{1}{\text{poly}(n)}\right)$

If L is $O(\text{poly}(\log(n)))$ → $G_n(L, l) \in \Omega\left(\frac{1}{2^{\text{poly}(\log(n))}}\right)$

Example: the narrow gorge

Task: variational approx. of $\mathbb{I}$ on $|\mathbf{0}\rangle$

$$V(\boldsymbol{\theta})|\mathbf{0}\rangle = |\mathbf{0}\rangle$$

Ansatz

$$V(\boldsymbol{\theta}) = \bigotimes_j e^{-i\theta_j \sigma_x^{(j)}/2}$$

Cost function

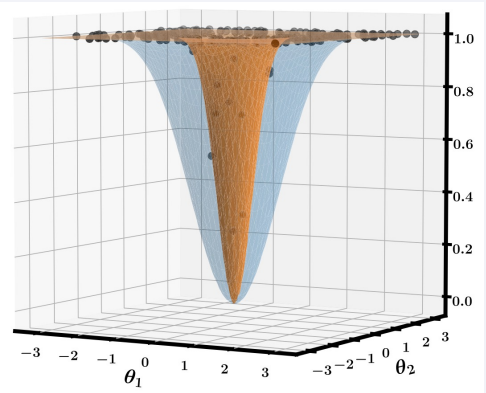
$$\mathrm{C_X}(\boldsymbol{\theta}) = \mathrm{Tr}[O_X V(\boldsymbol{\theta})|0\rangle\langle 0|V^\dagger(\boldsymbol{\theta})]$$

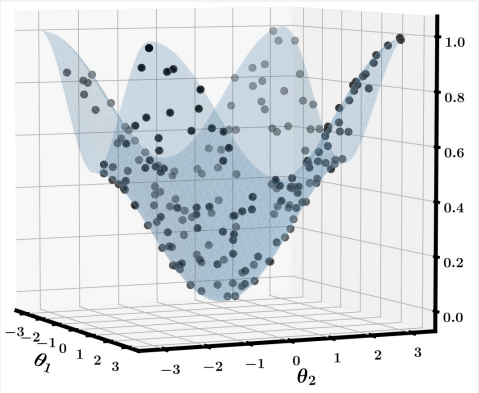
Global: $O_G = \mathbb{I} - |\mathbf{0}\rangle\langle\mathbf{0}|$

Local: $O_L = \mathbb{I} - \frac{1}{n}\sum_j |0\rangle\langle 0|_j \otimes \mathbb{I}_j$

Gradients

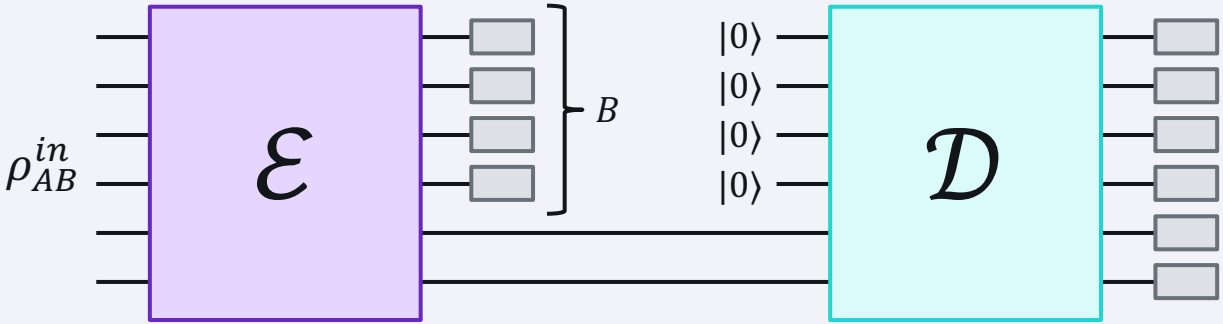
$$\mathrm{Var}[\partial_j C_G] = \frac{1}{8}\left(\frac{3}{8}\right)^{n-1} \qquad \mathrm{Var}[\partial_j C_L] = \frac{1}{8n^2}$$





M. Cerezo *et al.*, Nature Commun. **12**, 1791 (2021)

Example: quantum autoencoder

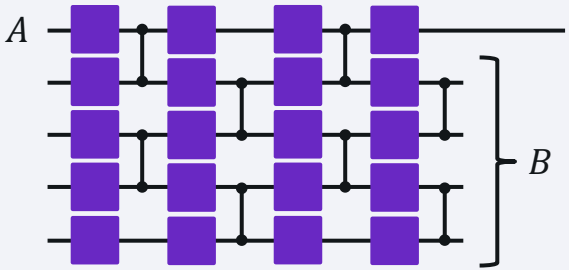


Goal

Compress the input state on subsystem A, recover it later with a decoder

J. Romero *et al.*, Quant. Sci. Technol. **2**, 045001 (2017)

M. Cerezo *et al.*, Nature Commun. **12**, 1791 (2021)

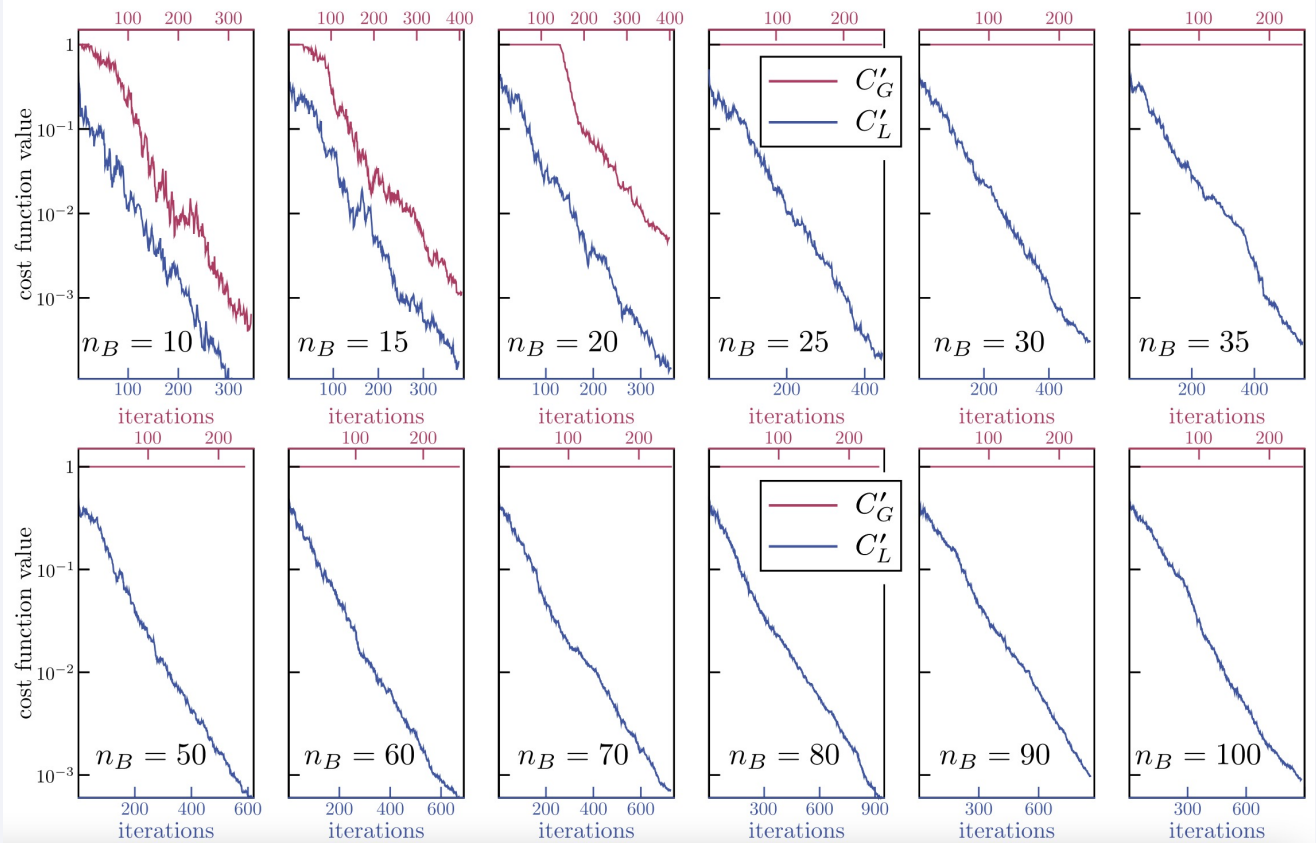


$$C_X(\theta) = \mathrm{Tr}[O_X V(\boldsymbol{\theta}) \rho_{AB}^{in} V^\dagger(\boldsymbol{\theta})]$$ **Degree of compression**

Global: $$O_G = \mathbb{I}_{AB} - \mathbb{I}_A \otimes |\mathbf{0}\rangle\langle\mathbf{0}|_B$$

Local: $$O_L = \mathbb{I}_{AB} - \frac{1}{n_B} \sum_j \mathbb{I}_A \otimes |0\rangle\langle 0|_j$$

Example: quantum autoencoder



Barren plateau for global cost function (untrainable for $n_B > 20$) and shallow ansatz.

The barren plateau is avoided with the local cost function (training is always possible in this example)

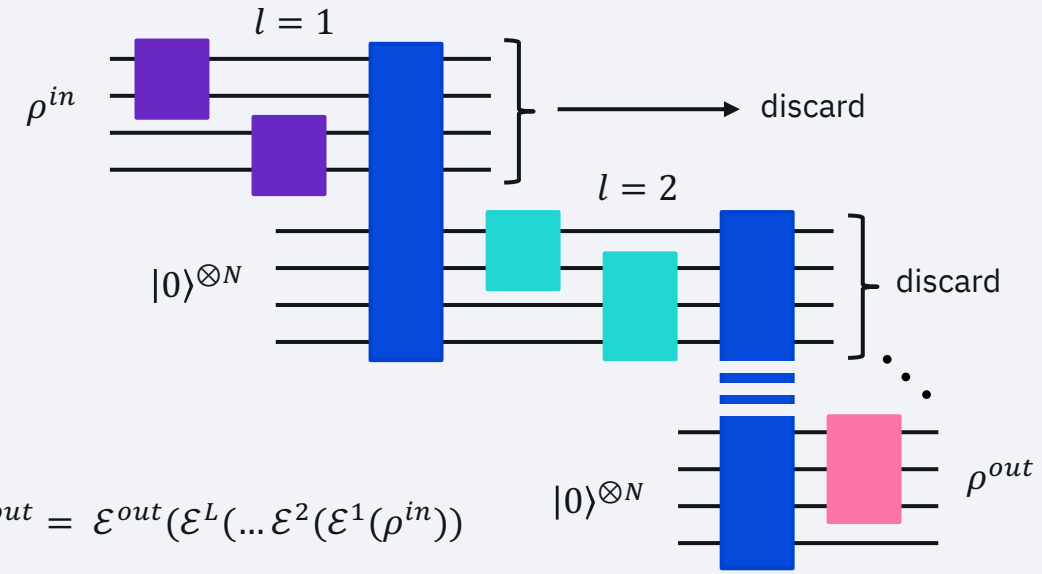
M. Cerezo *et al.*, Nature Commun. **12**, 1791 (2021)

Dissipative quantum neural networks

Qiskit

Key Features

- It formulates in QML terms the task of learning an unknown quantum transformation
- The feedforward architecture is represented by the layer-wise composition of quantum CP maps
- Backpropagation-like training can be performed by acting on one layer at a time (depth of the network is not a bottleneck)



$$\rho^{out} = \mathcal{E}^{out}(\mathcal{E}^{L}(\dots\mathcal{E}^{2}(\mathcal{E}^{1}(\rho^{in}))$$

$$\rho^{l+1} = \mathrm{Tr}_{l}\left[\prod_{j} U_{j}^{l}\left(\rho^{l}\otimes|0\dots0\rangle_{l+1}\langle 0\dots 0|\right)\prod_{j} U_{j}^{l,\dagger}\right]$$

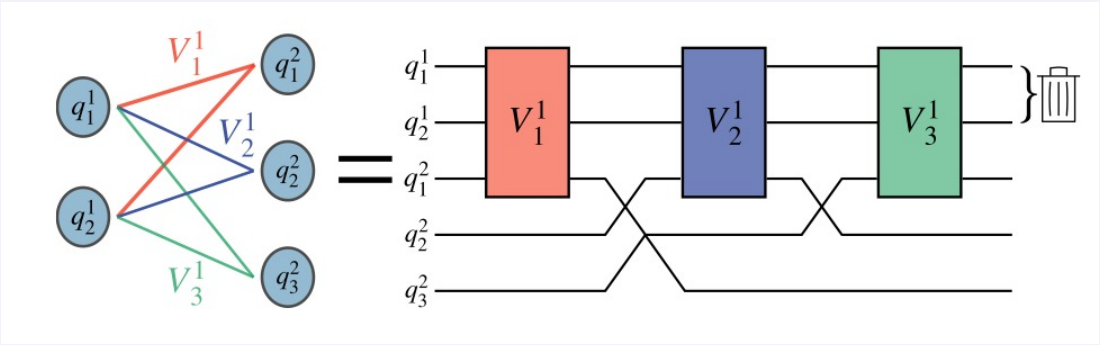
$$C = \frac{1}{|S|}\sum_{x\in S}\langle\phi_{x}^{out}|\rho_{x}^{out}|\phi_{x}^{out}\rangle$$

Target transformation

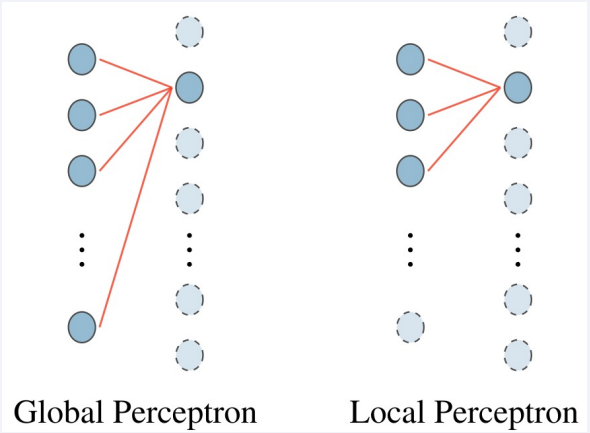
$$|\phi_{x}^{in}\rangle \rightarrow |\phi_{x}^{out}\rangle$$

K. Beer *et al.*, Nature Communications **11**, 808 (2020)
K. Sharma *et al.*, arXiv:2005.12458 (2020)

Barren plateaus in dissipative QNNs



Architecture	Cost	Trainable
Deep global perceptrons	Global	✗
Shallow local perceptrons	Global	✗
	Local	✓



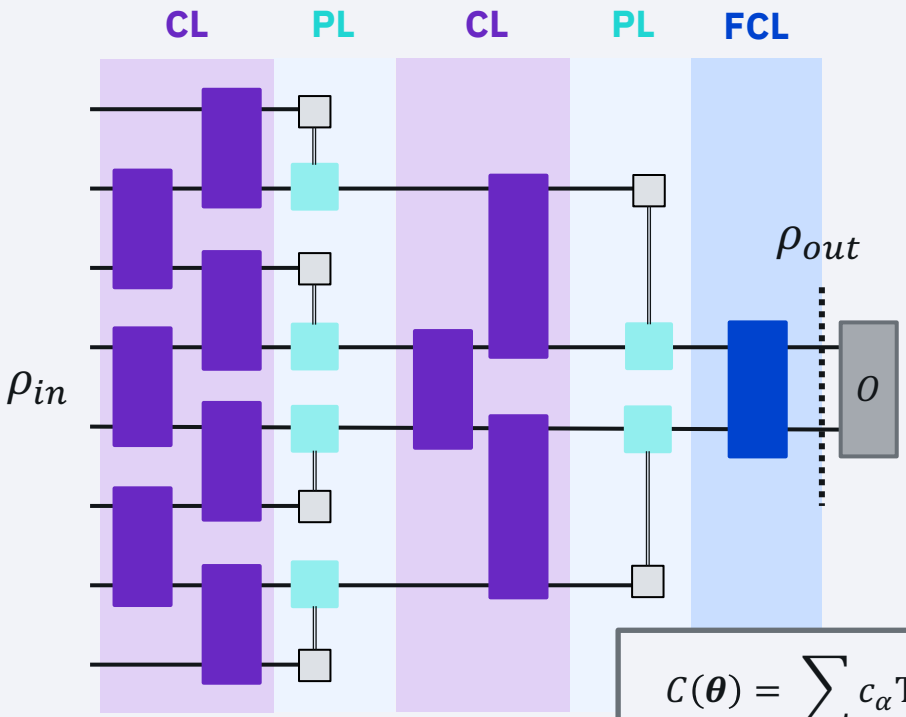
$$O_x^G = \mathbb{I} - |\phi_x^{out}\rangle\langle\phi_x^{out}|$$

(Assuming $|\phi_x^{out}\rangle = |\psi_{x,1}^{out}\rangle \otimes \cdots \otimes |\psi_{x,n_{out}}^{out}\rangle$)

$$O_x^L = \mathbb{I} - \frac{1}{n_{out}} \sum_{j=1}^{n_{out}} |\psi_{x,j}^{out}\rangle\langle\psi_{x,j}^{out}|$$

K. Sharma et al., arxiv:2005.12458

Quantum convolutional networks



$$C(\boldsymbol{\theta}) = \sum_{\alpha \in S} c_\alpha \mathrm{Tr}[\rho_{out}^{\alpha}(\boldsymbol{\theta}) O]$$

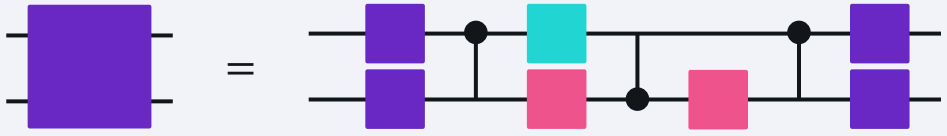
Key Features

- Can have as few as $O(\log N)$ parameters
- Related to hierarchical quantum circuits like tree-tensor networks and multi-scale entanglement renormalization ansatz
- Can be used to analyze classical data or quantum states
  - Quantum phase recognition, quantum error correction, entanglement detection, ...

Grant *et al.*, npj Quant. Inf. **4**, 65 (2018)
I. Cong *et al.*, Nature Physics **15**, 1273 (2019)
Pesah *et al.*, arXiv:2011.02966 (2020)

No barren plateaus in Quantum CNNs

$$\mathrm{Var}[\partial_v C] \geq F_n(L, \ell)$$

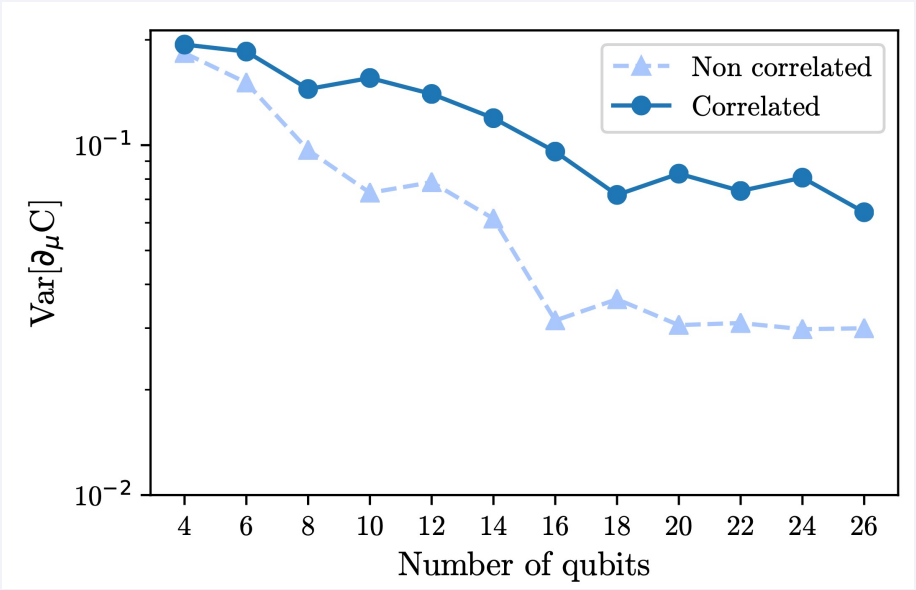


For L at most $O(\log n)$

$$F_n(L, \ell) \in \Omega\left(\frac{1}{\mathrm{poly}(n)}\right)$$

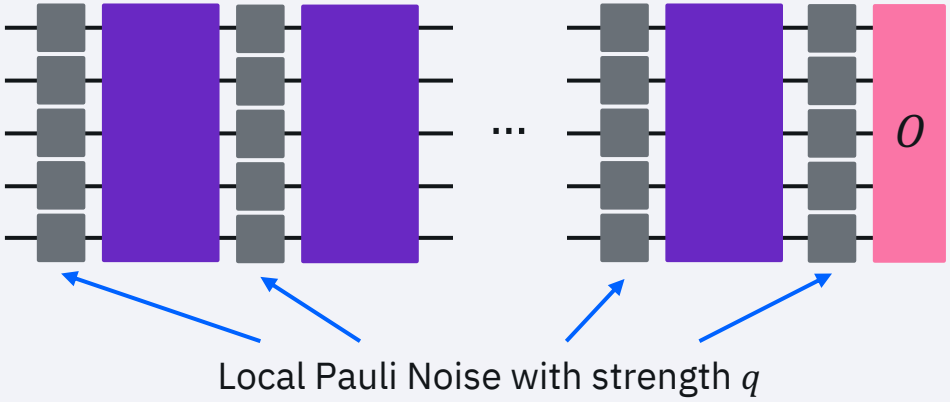
[QCNNs are "naturally" shallow]

- Variance of the gradient has sub-exponential scaling with n
- Correlations between variational blocks lead to larger variances
- Pooling-based QCNN is also trainable



Pesah *et al.*, arxiv:2011.02966

Noise-induced barren plateaus



- The cost function values concentrate around the one for maximally mixed states exponentially in L (number of layers)

$$\left|\tilde{C} - \frac{1}{2^n}\mathrm{Tr}[O]\right| \leq G(n)\left\|\rho - \frac{\mathbb{I}}{2^n}\right\|_1$$

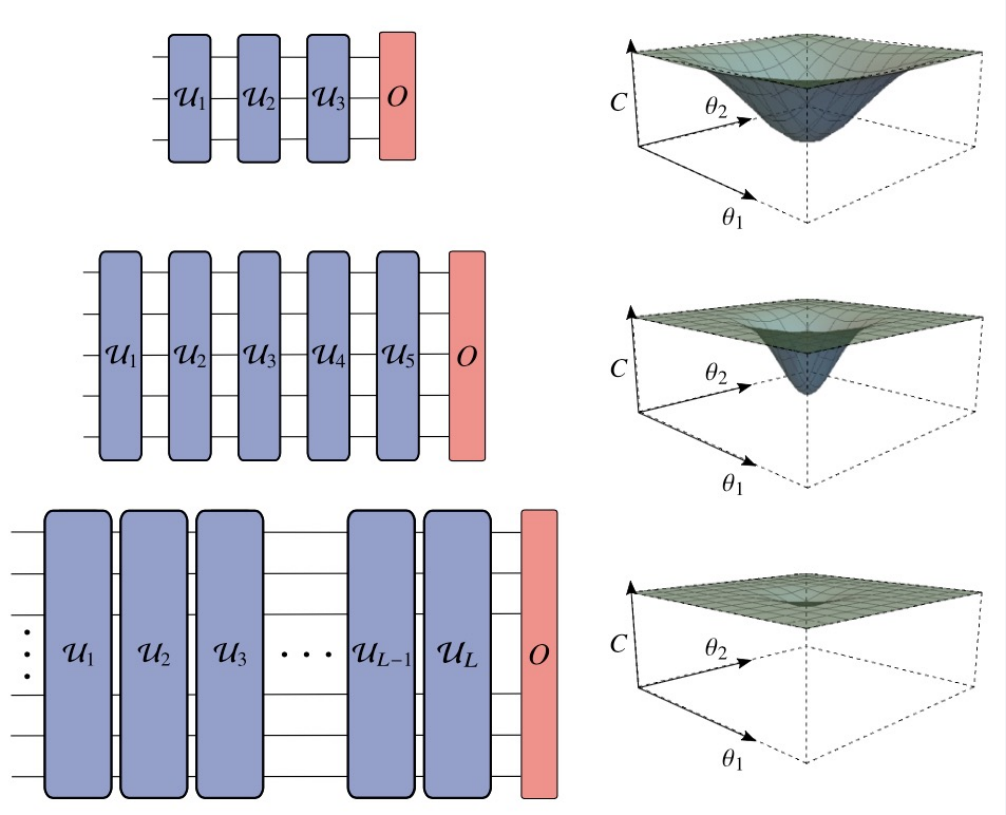
where $G(n) \propto q^{L+1}$.

- Bound on the gradient of the cost function → $\left|\partial_v \tilde{C}\right| \leq F(n)$ with $F(n) \propto n^{1/2} q^{L+1}$
- Noise-Induced barren plateau → $L \in \Omega(n) \Longrightarrow F(n) \in O(2^{-\alpha n})$
- Adding measurement noise (local bit-flip channels) with stength q_M → $\left|\partial_v \tilde{C}\right| \leq q_M^w F(n)$

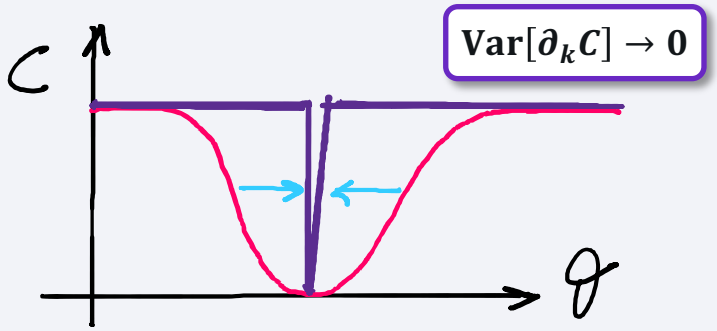
w: minimum Pauli weight in the expansion of O over Pauli strings

Wang et al., arXiv:2007.14384

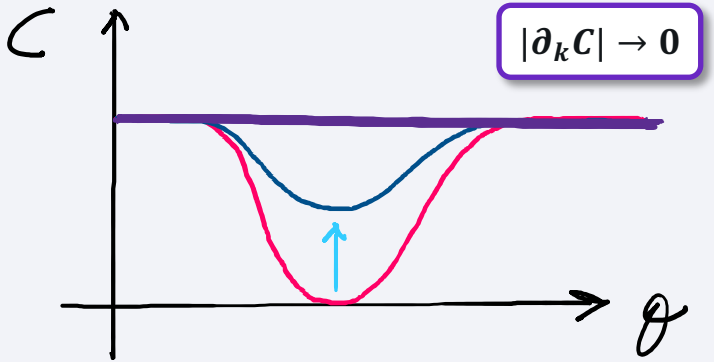
Noise-induced barren plateaus



Noise free barren plateaus



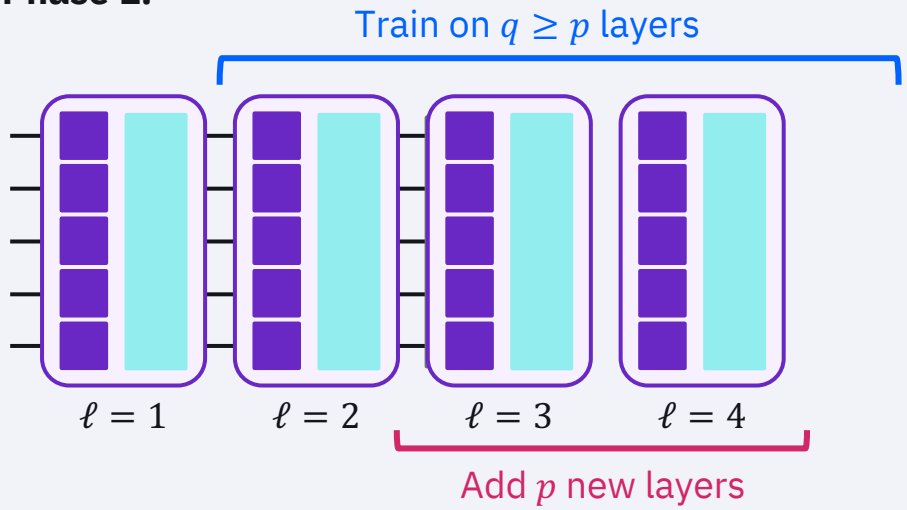
Noise-induced barren plateaus



Wang et al., arXiv:2007.14384

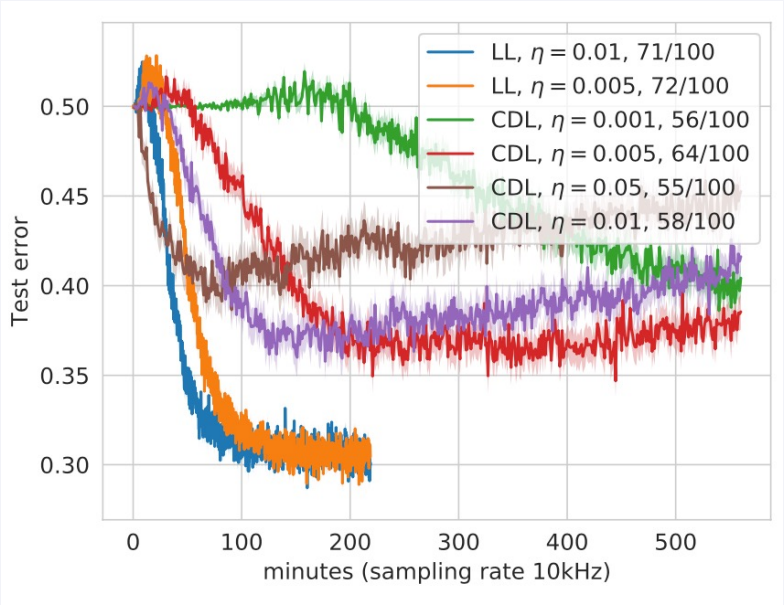
Mitigating barren plateaus: layerwise learning

- **Phase 1:**



All new layers initialized at $\vec{\theta} = 0$ for a smooth transition

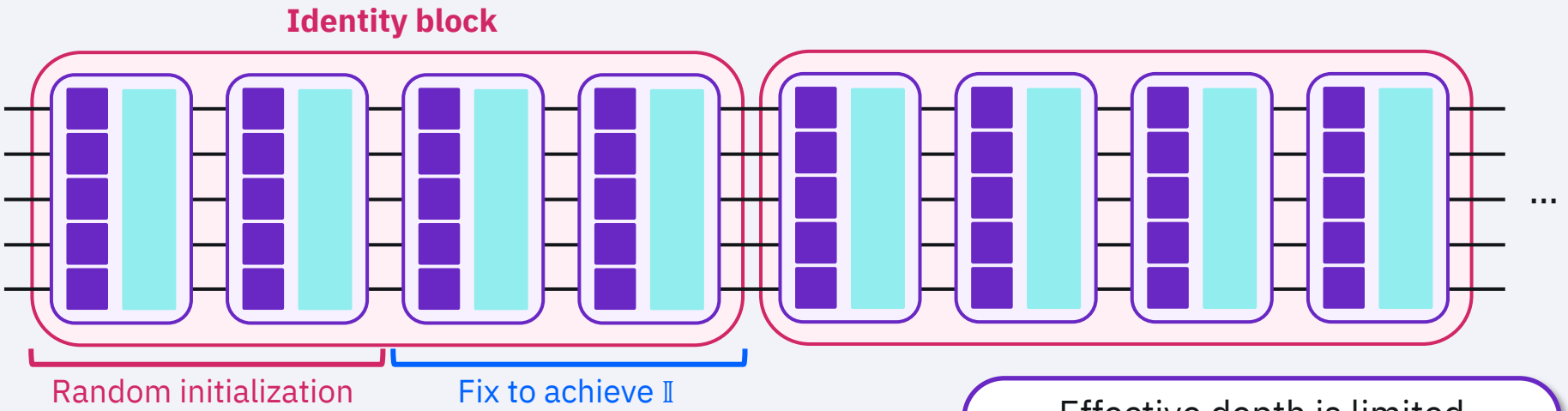
- **Phase 2:** Take pre-trained circuit from phase 1, train larger contiguous partitions of layers (the fraction of layers involved can be treated as a hyperparameter)



Layerwise learning (**LL**) vs complete depth learning (**CDL**) for a MNIST classification experiment with a 8-qubit, 21-layer circuit

A. Skolik *et al.*, Quantum Machine Intelligence 3, 5 (2021)

Mitigating barren plateaus: initialization



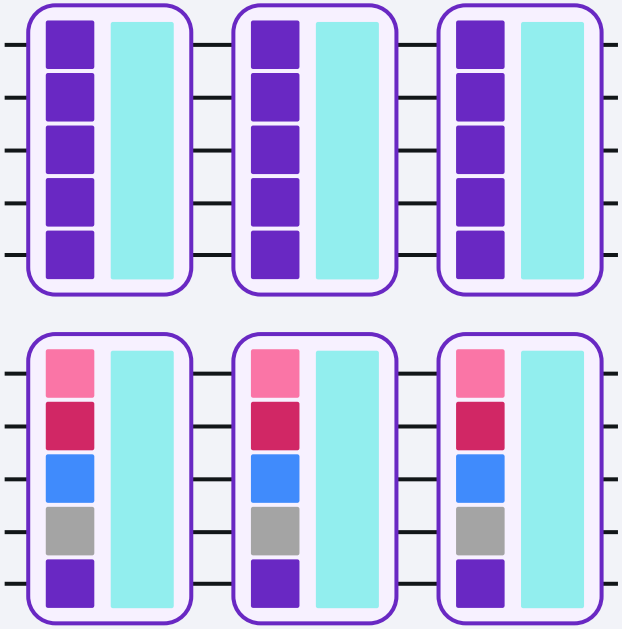
- Partition the circuit in M (shallow) blocks of depth L
- Initialize some parameters with random values
- Fix the remaining ones such that each block implements the identity operation

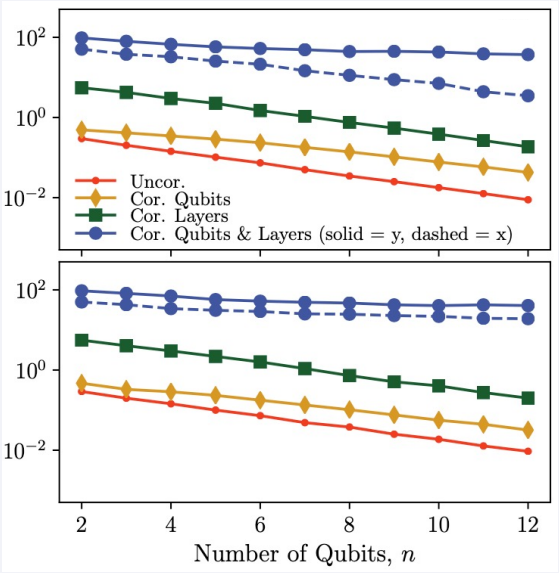
Effective depth is limited

Gradients w.r.t. most parameters will not be exponentially small **for the first training step**

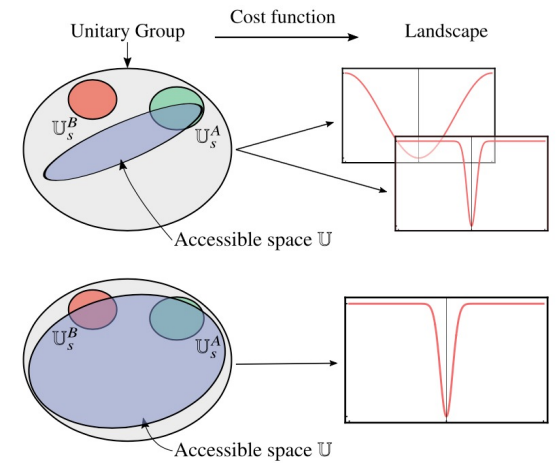
Grant *et al.*, Quantum 3, 214 (2019)

Mitigating barren plateaus: correlations





Part of a larger study on the relationhip between expressibility, gradients and barren plateaus



Z. Holmes *et al.*, arXiv:2101.02138 (2021)

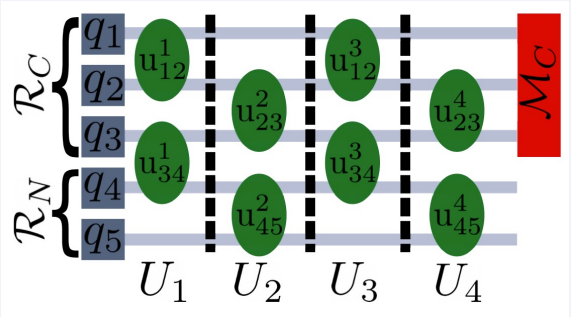
T. Volkoff and P. J. Coles, arXiv:2005.12200 (2020)

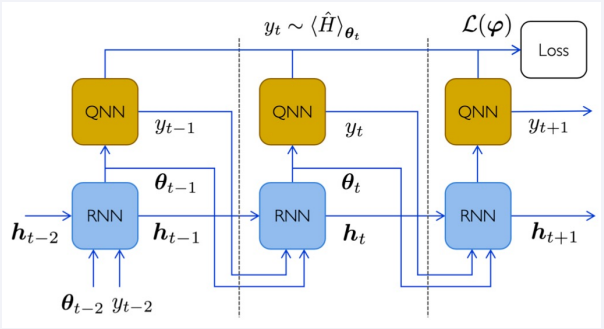
Additional results and training strategies

Entanglement induced barren plateaus and entanglement-aware barren plateau mitigation

C. O. Marrero *et al.*, arXiv:2010.15968 (2020)

T. Patti *et al.*, arXiv:2012.12658 (2020)





Meta-learning ("learning to learn") of initialization heuristics via classical Recurrent Neural Networks

G. Verdon *et al.*, arXiv:1907.05415 (2019)

Variational quantum unsampling

J. Carolan *et al.*, Nature Physics **16**, 322 (2020)

F. Tacchino *et al.*, IEEE Trans. Quantum Eng. **2**, 3101110 (2021)