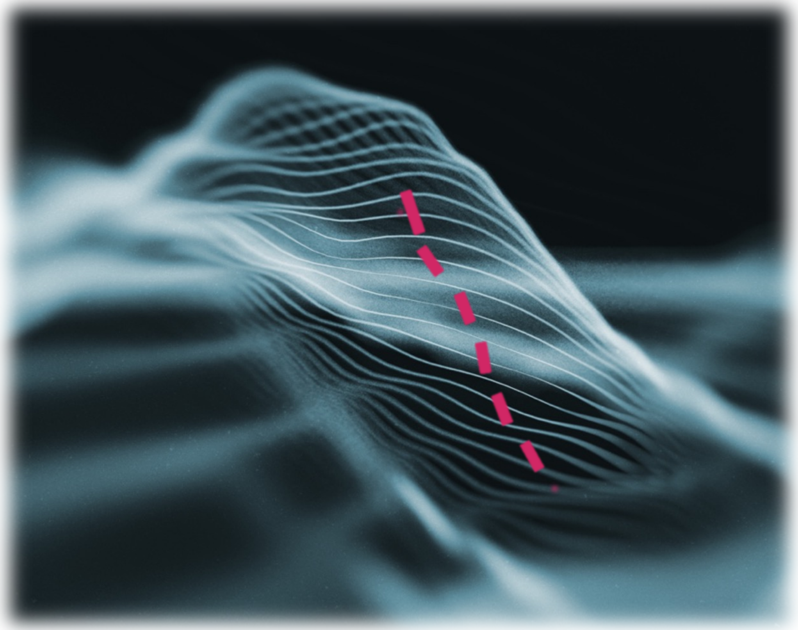
From Tight Gradient Bounds for Parameterized Quantum Circuits to the Absence of Barren Plateaus in QGANs

Christa Zoufal
Research Scientist
Quantum Computational Science
IBM Quantum, IBM Research Europe - Zurich

Quantum Techniques in Machine Learning 2023

Trainability

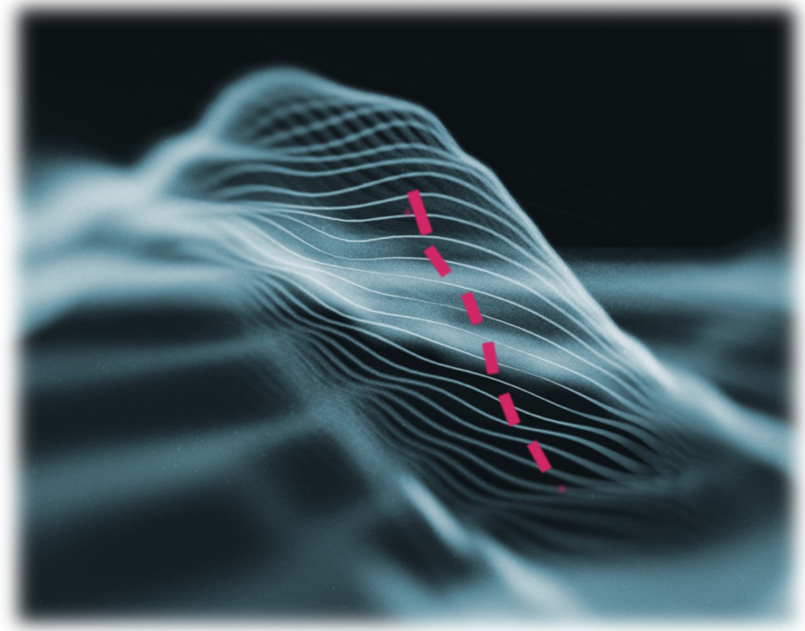
Efficient access to gradients
→ shot noise



Trainability

Efficient access to gradients
→ shot noise

Non-exponentially vanishing gradients
→ barren plateaus

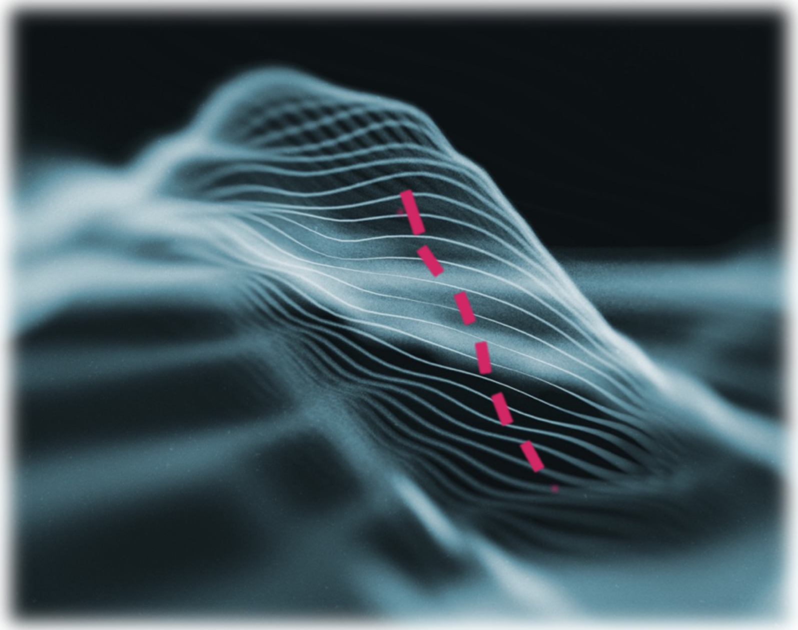


Trainability

Efficient access to gradients
→ shot noise

Non-exponentially vanishing gradients
→ barren plateaus

Functional ability to represent optimum
→ ansatz design

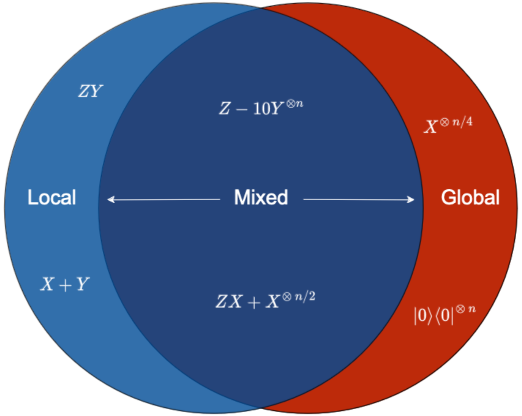


Observables, Loss Functions & Gradients

Ansatz: $U(\boldsymbol{\omega}) = \prod_{i=0}^{k-1} V_i U_i(\boldsymbol{\omega}_i)$

Initial state: $|\psi_{\text{in}}\rangle$

Observable: $H = \sum_j a_j h_j,\ a_j \in \mathbb{R}$



Loss Function: $L(\boldsymbol{\omega}) = \langle\psi_{\text{in}}|U^{\dagger}(\boldsymbol{\omega})\, HU(\boldsymbol{\omega})\,|\psi_{\text{in}}\rangle$

↑

QML, quantum chemistry, quantum time simulation, etc.

↓

Loss Function Gradient:

$$\partial_\omega L(\omega) = 2\mathrm{Re}\left[\langle\psi_{\text{in}}|\left(\partial_\omega U^{\dagger}(\omega)\right) HU(\omega)\,|\psi_{\text{in}}\rangle\right]$$

Barren Plateaus

Exponentially vanishing gradients ↔ exponentially flat loss landscape

$$\mathbb{E}_\omega\left[\partial_\omega L\left(\omega\right)\right] = 0$$

$$\mathrm{Var}_\omega\left[\partial_\omega L\left(\omega\right)\right] \in \mathcal{O}\left(\frac{1}{b^n}\right),\ b > 1$$

n representing the number of qubits

Known causes

- Ansatz close to a t-design [1-3]
- Global observable [4-5]
- Extensive entanglement paired with partial traces [6]
- Particular noise models [7]

[1] J. McClean, et al., Nature Communications 9 (2018).
[2] Z. Holmes, et al., PRX Quantum 3 (2022).
[3] M. Larocca, et al., Quantum 6, 824 (2022).
[4] M. Cerezo, et al., Nature Communications 12 (2021).
[5] S. Thanasilp, et al., Quantum Machine Intelligence 5, 21 (2023).
[6] C. Ortiz Marrero, et al., PRX Quantum 2, 040316 (2021).
[7] S. Wang, et al., Nature Communications 12 (2021).

Barren Plateaus

Exponentially vanishing gradients ↔ exponentially flat loss landscape

$$\mathbb{E}_\omega\left[\partial_\omega L\left(\omega\right)\right] = 0$$

$$\mathrm{Var}_\omega\left[\partial_\omega L\left(\omega\right)\right] \in \mathcal{O}\left(\frac{1}{b^n}\right), \ b > 1$$

n representing the number of qubits

Known causes

- Ansatz close to a t-design [1-3]
- Global observable [4-5]
- Extensive entanglement paired with partial traces [6]
- Particular noise models [7]

[1] J. McClean, et al., Nature Communications 9 (2018).
[2] Z. Holmes, et al., PRX Quantum 3 (2022).
[3] M. Larocca, et al., Quantum 6, 824 (2022).
[4] M. Cerezo, et al., Nature Communications 12 (2021).
[5] S. Thanasilp, et al., Quantum Machine Intelligence 5, 21 (2023).
[6] C. Ortiz Marrero, et al., PRX Quantum 2, 040316 (2021).
[7] S. Wang, et al., Nature Communications 12 (2021).

Motivation

Assume a parameterized ansatz $|\psi(\boldsymbol{\omega})\rangle$ and an observable

$$\hat{O} = \begin{pmatrix} x_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_k \end{pmatrix}$$

➔ Trainability issues due to **global** nature of the **observable** for $\langle\psi(\boldsymbol{\omega})|\hat{O}|\psi(\boldsymbol{\omega})\rangle$?

What if we rewrite the expectation value?

$$\langle\psi(\boldsymbol{\omega})|\hat{O}|\psi(\boldsymbol{\omega})\rangle = \sum_{i,j} e^{i\phi_i(\omega)} e^{-i\phi_j(\omega)} \sqrt{p_i(\boldsymbol{\omega}) p_j(\boldsymbol{\omega})} \langle i|\hat{O}|j\rangle$$

for

$$p_i(\boldsymbol{\omega}) = \langle\psi(\boldsymbol{\omega})|i\rangle\langle i|\psi(\boldsymbol{\omega})\rangle$$

➔ Again a **global observable**!!

Gradient Bounds

Previous Results

➔ T-design assumptions rarely sufficed in practice

Cerezo et al. [1]

Barren plateau for H a type of **global** observable, e.g., projector and U forming a local **2-design**

Uvarov & Biamonte [2]

Proof that Pauli strings make **independent contributions** to the gradient → lower bounds by Cerezo et al. extended to generic observables – still local 2-design

Napp [3]

Gradient bounds for **spatially** and **algebraically local** observable → tighter bounds than Uvarov and Biamonte – entangling gates chosen randomly according to any measure that forms a 2-design.

[1] M. Cerezo, et al., Nature Communications 12 (2021).
[2] A. V. Uvarov and J. D. Biamonte, Journal of Physics A: Mathematical and Theoretical 54, 245301 (2021).
[3] J. Napp, arXiv preprint - arXiv:2203.06174 (2022).

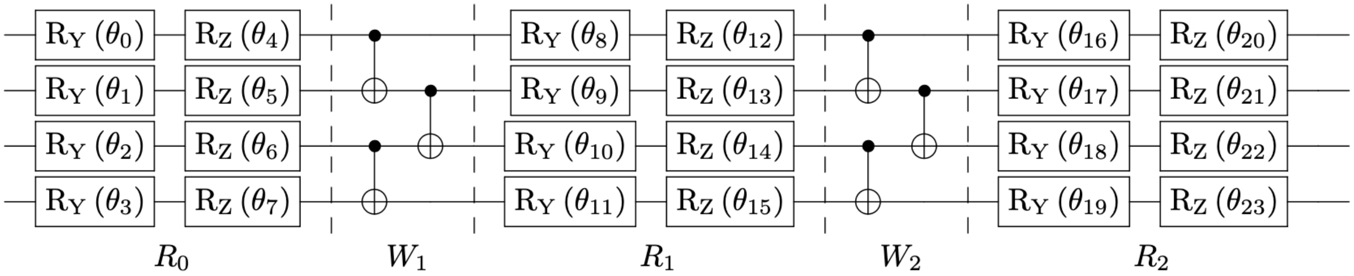
Circuits

Clifford gates

Product of rotation gates generated by Pauli strings

$$U(\boldsymbol{\omega}) = \sum_{k=0}^{K} W_k \, R_k(\boldsymbol{\omega})$$

- with the first layer two layers of orthogonal single-qubit rotations
- with $\boldsymbol{\omega} = \{\omega_0, \dots, \omega_{l-1}\}$ for all ω_i being independent and initialized over $[-\pi, \pi]$



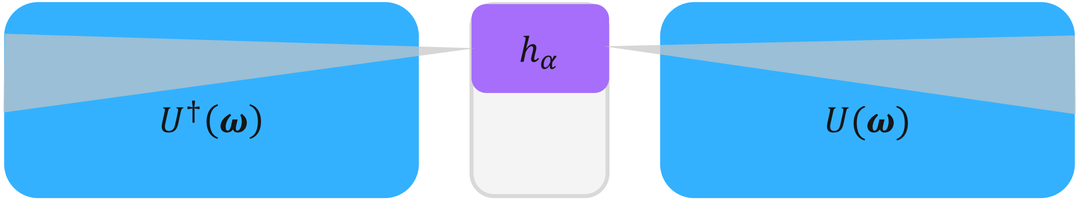
EfficientSU2 for depth 2 with pairwise entanglement

Initial State and Hamiltonian

Initial State $\rho = \otimes_j \rho_j$

Hamiltonian $H = \sum_\alpha a_\alpha h_\alpha, a_\alpha \in \mathbb{R}, h_\alpha \in \{I, X, Y, Z\}^{\otimes n}$

Mean and min. light-cones $\Delta_\alpha^{mean}, \Delta_\alpha^{min}$, i.e., mean and min number of qubits on which $U^\dagger(\boldsymbol{\omega}) h_\alpha U(\boldsymbol{\omega})$ acts non-trivially for $\omega_i \in [0, \frac{\pi}{2}]$



Gradient Bounds

Each Pauli term h_α makes an **independent contribution** to the loss and gradient concentrations!

$$\Omega\left(\rho\right)\left(\frac{1}{4}\right)^{\Delta_\alpha^{\mathrm{mean}}} \leq \mathrm{Var}_\omega\left[L_\alpha\left(\omega\right)\right] \leq \left(\frac{1}{2}\right)^{\Delta_\alpha^{\mathrm{min}}}$$

Measure of orthogonality between ρ and the eigenvectors of the first single-qubit rotation layer

At **least one** parameter sufficing the **lower** bound and **all** parameters sufficing the **upper** bound

Highlights

Drop t-design assumptions

Consider realistic circuit classes which are compatible with the realization of hardware efficient ansatz classes

Highlights

Drop t-design assumptions

Consider realistic circuit classes which are compatible with the realization of hardware efficient ansatz classes

Capture observable-circuit interaction in finer detail

Gradient concentration not determined by algebraic locality, circuit depth or entanglement per se, but the observable-circuit interaction → light-cones

Assumptions Justification

Orthogonal layers
→ commutation shield helps to avoid constant loss

Assumptions Justification

Orthogonal layers
→ commutation shield helps to avoid constant loss

Initial States
alignment between initial state and the first layer of rotations can lead to vanishing gradients

Assumptions Justification

Orthogonal layers
→ commutation shield helps to avoid constant loss

Initial States
alignment between initial state and the first layer of rotations can lead to vanishing gradients

Clifford gates
no restriction on circuit expressivity, necessary to ensure uncorrelated contributions of Pauli terms

Assumptions Justification

Orthogonal layers
→ commutation shield helps to avoid constant loss

Initial States
alignment between initial state and the first layer of rotations can lead to vanishing gradients

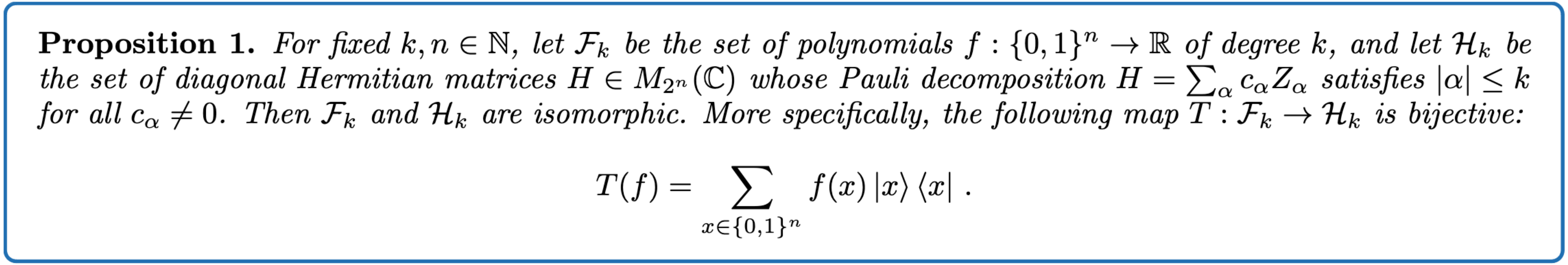
Clifford gates
no restriction on circuit expressivity, necessary to ensure uncorrelated contributions of Pauli terms

Independent parameters
dependent parameters for different Pauli terms with contribution in opposite directions can lead to uniformly zero gradients

Equivalence between k-Degree Polynomials and k-Local Diagonal Observables

IBM **Quantum**

Bijection between binary polynomials $f\colon \{0,1\}^n \rightarrow \mathbb{R}$ of degree k and diagonal Hermitian, k-algebraically local observables H



Proposition 1. *For fixed* $k, n \in \mathbb{N}$, *let* $\mathcal{F}_k$ *be the set of polynomials* $f : \{0,1\}^n \to \mathbb{R}$ *of degree* k, *and let* $\mathcal{H}_k$ *be the set of diagonal Hermitian matrices* $H \in M_{2^n}(\mathbb{C})$ *whose Pauli decomposition* $H = \sum_\alpha c_\alpha Z_\alpha$ *satisfies* $|\alpha| \leq k$ *for all* $c_\alpha \neq 0$. *Then* $\mathcal{F}_k$ *and* $\mathcal{H}_k$ *are isomorphic. More specifically, the following map* $T : \mathcal{F}_k \to \mathcal{H}_k$ *is bijective:*

$$T(f) = \sum_{x \in \{0,1\}^n} f(x) \left|x\right\rangle \left\langle x\right| .$$

Fixed polynomials with small degree monomials (such as QUBOs) do not induce barren plateaus

Application Areas of Relevance

Binary Optimization

Minimize a function $f:\{0,1\}^n \to \mathbb{R}$ in terms of a ground state [1] problem for

$$\mathrm{H} = \sum_x f(x)|x\rangle\langle x| = \sum_\alpha \left(\frac{1}{2^n}\sum_x (-1)^{\alpha x} f(x)\right) Z_\alpha \,.$$

Quantum Chemistry

VQAs for ground state and time simulation mixed observable, e.g., hydrogen chains [2], vibrational bosonic systems [3], and downfolded electronic Hamiltonians [4].

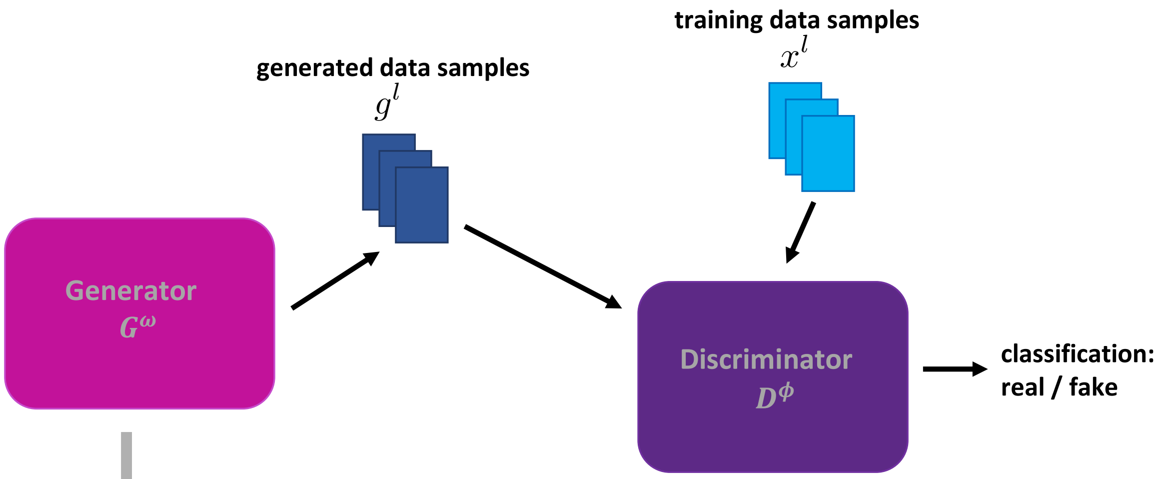
Generative QML

Often representable as ground state problem of a diagonal Hamiltonian, see e.g. hybrid qGANs [5-7] .

[1] C. Zoufal et al., Quantum 7 (2023).
[2] I. O. Sokolov, et al. The Journal of Chemical Physics 152 (2020).
[3] P. J. Ollitrault, et al. Chem. Sci. 11, 6842 (2020).
[4] K. Kowalski, Phys. Rev. A 104, 032804 (2021).
[5] C. Zoufal, Aet al., npj Quantum Information 5 (2019).
[6] P.-L. Dallaire-Demers and N. Killoran, Phys. Rev. A 98 (2018).
[7] J. Romero and A. Aspuru-Guzik, Advanced QuantumTechnologies 4 (2021).

Gradient Bounds for qGANs

Hybrid Quantum Generative Adversarial Networks



training data samples

x^l

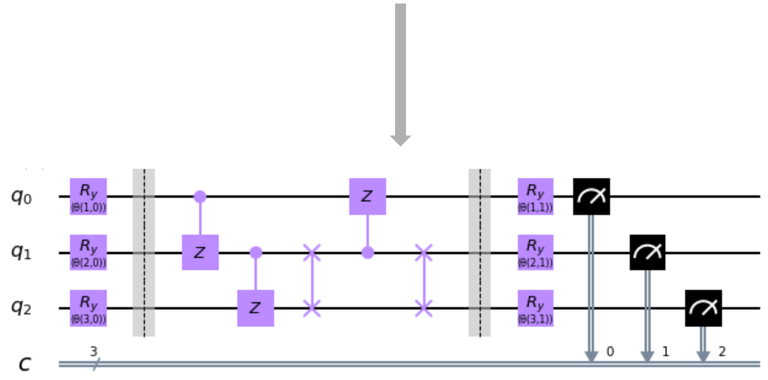
generated data samples

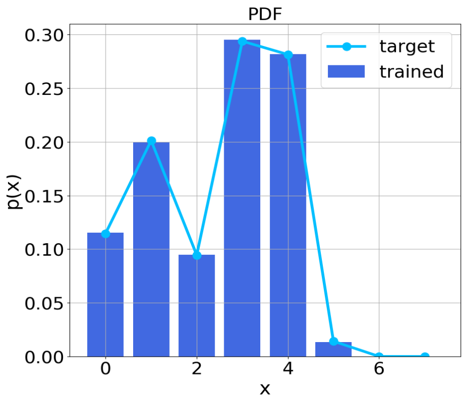
g^l

Generator
G^{ω}

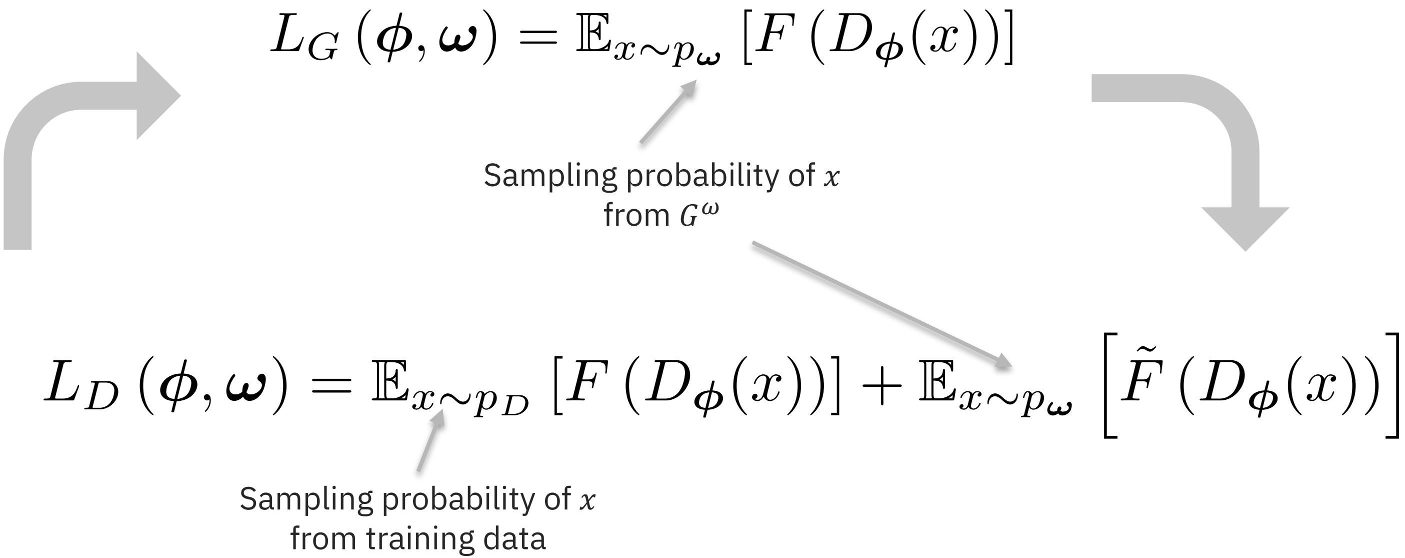
Discriminator
D^{ϕ}

classification:
real / fake





Loss Functions



$$L_G\left(\boldsymbol{\phi},\boldsymbol{\omega}\right) = \mathbb{E}_{x\sim p_{\boldsymbol{\omega}}}\left[F\left(D_{\boldsymbol{\phi}}(x)\right)\right]$$

Sampling probability of x
from G^{ω}

$$L_D\left(\boldsymbol{\phi},\boldsymbol{\omega}\right) = \mathbb{E}_{x\sim p_D}\left[F\left(D_{\boldsymbol{\phi}}(x)\right)\right] + \mathbb{E}_{x\sim p_{\boldsymbol{\omega}}}\left[\tilde{F}\left(D_{\boldsymbol{\phi}}(x)\right)\right]$$

Sampling probability of x
from training data

Compatible with min-max [1] and Wasserstein [2] GANs

[1] I. Goodfellow, et al., in Advances in Neural Information Processing Systems 27 (Curran Associates, Inc., 2014).

[2] M. Arjovsky, et al., in Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, Vol. 70, (PMLR, 2017).

Loss Function Gradients

$\nabla_{\boldsymbol{\phi}} L_D\left(\boldsymbol{\phi}, \boldsymbol{\omega}\right)$ ➔ Standard automatic differentiation techniques

Loss Function Gradients

$\nabla_{\boldsymbol{\phi}} L_D\left(\boldsymbol{\phi}, \boldsymbol{\omega}\right)$ ➔ Standard automatic differentiation techniques

$\nabla_{\boldsymbol{\omega}} L_G\left(\boldsymbol{\phi}, \boldsymbol{\omega}\right) = \nabla_{\boldsymbol{\omega}} \langle 0|^{\otimes n} \left(G^{\boldsymbol{\omega}}\right)^\dagger H\left(\boldsymbol{\phi}\right) G^{\boldsymbol{\omega}} |0\rangle^{\otimes n}$ ➔ Quantum gradient

for $H\left(\boldsymbol{\phi}\right) = \sum_x F(D_\phi(x)) \left|x\right\rangle \left\langle x\right| = \sum_\alpha c_\alpha(\phi) Z_\alpha$

Enables reformulation in terms of Pauli Basis → compatibility with our **gradient bounds**

$L_G\left(\boldsymbol{\phi}, \boldsymbol{\omega}\right) = \sum_\alpha c_\alpha(\boldsymbol{\phi}) \langle 0|^{\otimes n} \left(G^{\boldsymbol{\omega}}\right)^\dagger Z_\alpha G^{\boldsymbol{\omega}} |0\rangle^{\otimes n}$

No barren plateaus for sufficiently large $c_\alpha(\boldsymbol{\phi})$ for local contribution terms Z_α

Discriminator

Definition

$$D^{\phi}: \{0,1\}^n \to \mathbb{R}$$

- Fully-connected neural network with L hidden layers
- Number of nodes for each layer l denoted by m_l
- Standard deviation of initial weights σ_l for layer l
- Weights and biases initialized with i.i.d. symmetric distributions
- Leaky-ReLu hidden activation functions with parameter γ_l

Gradient Bounds for qGANs

Let D_ϕ be a discriminator of depth L. For any 1-local weight α, we have

$$\mathbb{E}_\phi\left[c_\alpha(\phi)^2\right] \geq \frac{\sigma_{L+1}^2}{16} \prod_{l=1}^{L} \frac{m_l \sigma_l^2 (1+\gamma_l)^2}{4} .$$

In particular, initialising parameters such that $m_l \sigma_l^2 \geq 4$ for each l, the bound reduces to $\mathbb{E}_\phi\left[c_\alpha(\phi)^2\right] \geq \sigma_{L+1}^2/16$, which is **constant** both in the number of qubits n and the discriminator depth L.

➔ Guarantees that $H(\boldsymbol{\phi})$ introduces 1-local weights that are constant in the number qubits!

Consequences for Quantum Generative Adversarial Networks

IBM **Quantum**

For any discriminator $D_{\boldsymbol{\phi}}$ satisfying $m_l \sigma_l^2 \geq 4$ for each layer l, and any EfficientSU2 generator $G_{\boldsymbol{\omega}}$ with pairwise entanglement and logarithmic depth $K \in O(\log n)$, there exists a parameter ω_k such that

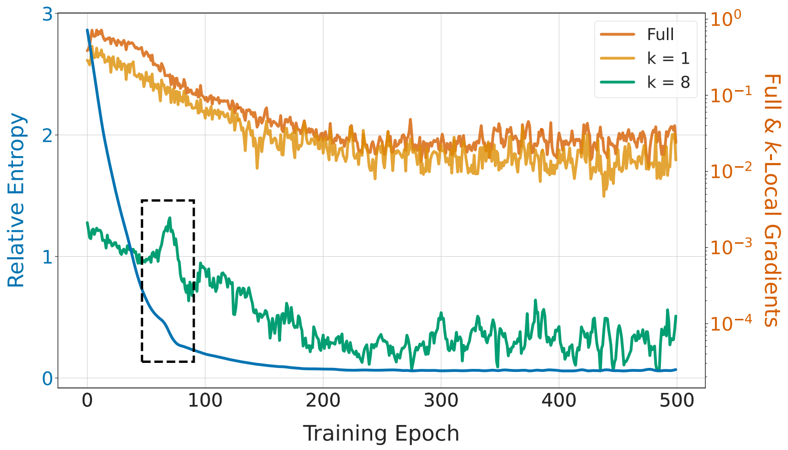
$$\mathrm{Var}_{\omega,\phi}\left[\partial_{\omega_k} L_G\right] \in \Omega\left(\frac{1}{\mathrm{poly}(n)}\right)$$

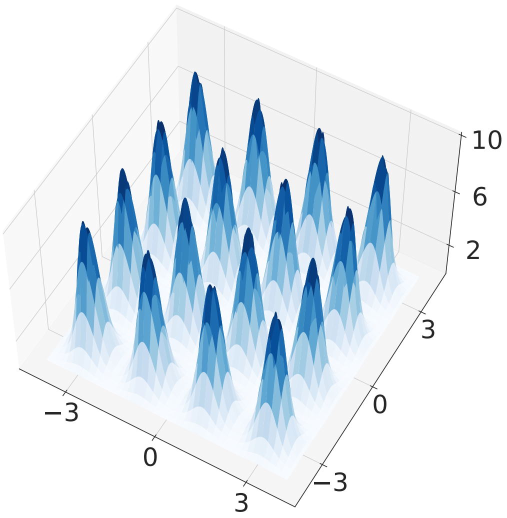
Unlike most other generative QML algorithms, QGANs do not suffer from barren plateaus for a large class of classical discriminators of arbitrary depth, and quantum generators of logarithmic depth.

Experiments

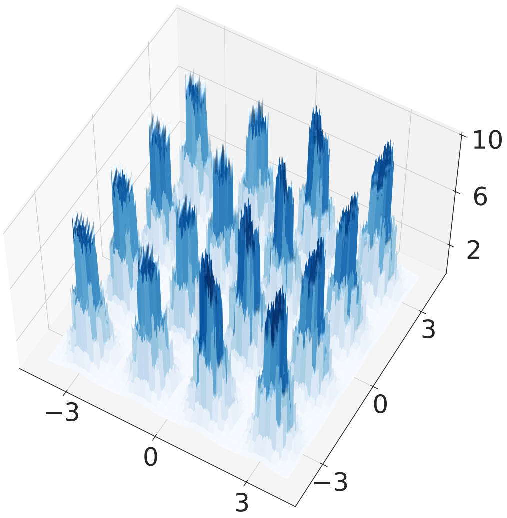
Model

Size: 16 qubits
Generator: EfficientSU2 (pairwise entanglement, depth 1), parameters uniformly initialized in $[-\pi, \pi]$
Discriminator: fully-connected, 192 nodes, leaky-ReLU hidden activation, parameter initialization Kaiming uniform $2\times\sigma$



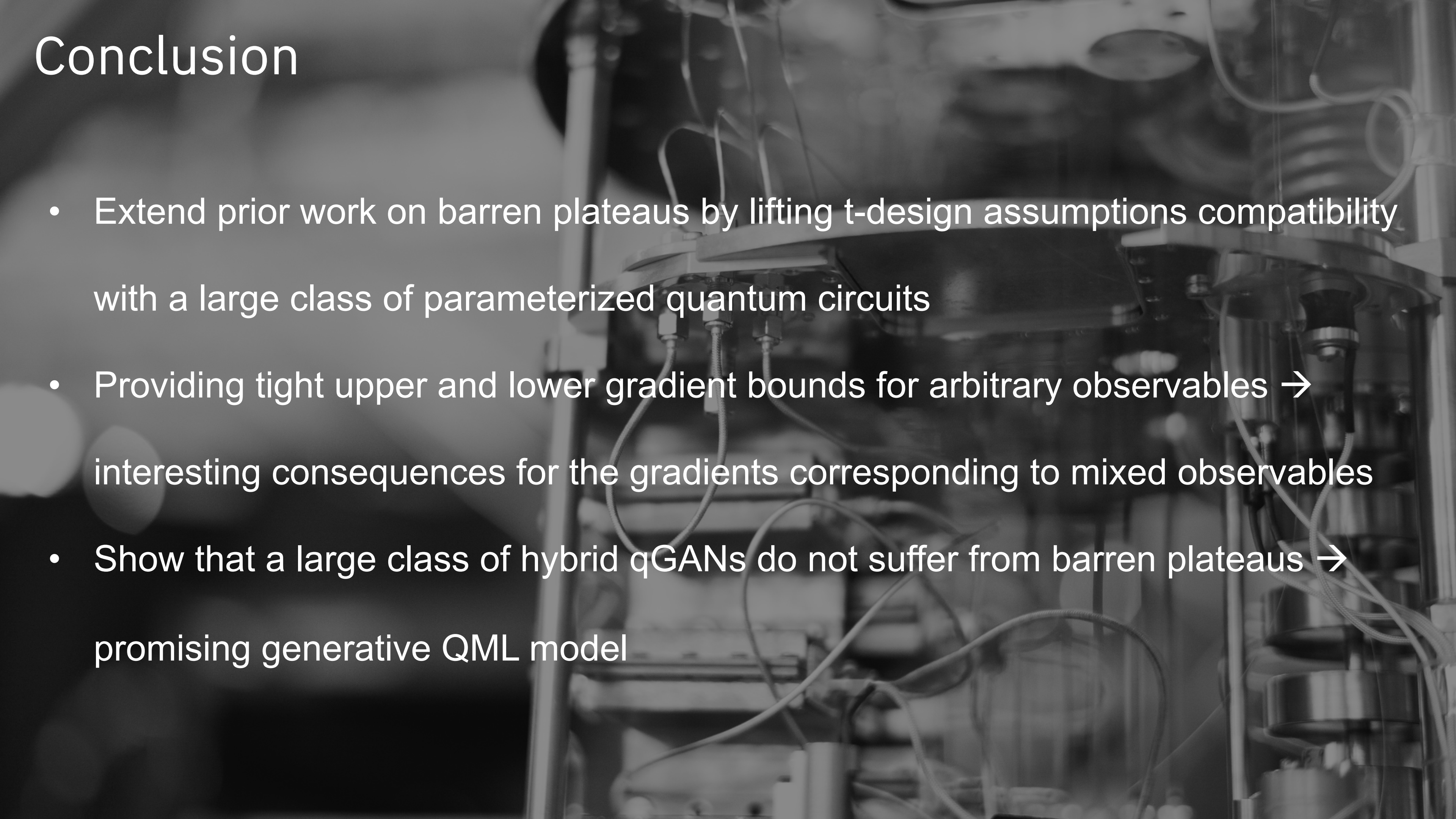


Target PDF



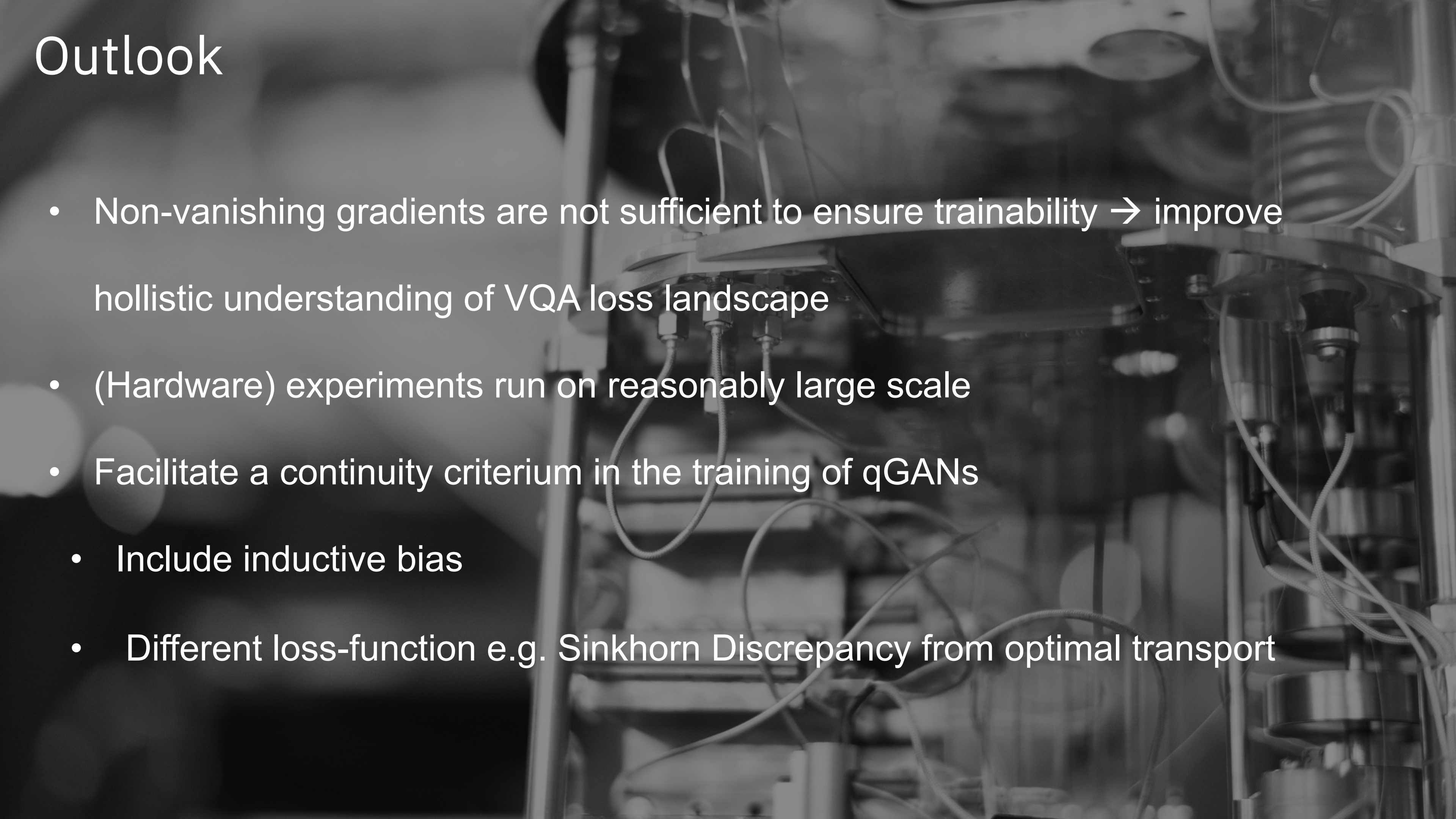
Trained PDF

Conclusion



- Extend prior work on barren plateaus by lifting t-design assumptions compatibility with a large class of parameterized quantum circuits
- Providing tight upper and lower gradient bounds for arbitrary observables → interesting consequences for the gradients corresponding to mixed observables
- Show that a large class of hybrid qGANs do not suffer from barren plateaus → promising generative QML model

Outlook



- Non-vanishing gradients are not sufficient to ensure trainability → improve hollistic understanding of VQA loss landscape
- (Hardware) experiments run on reasonably large scale
- Facilitate a continuity criterium in the training of qGANs
  - Include inductive bias
  - Different loss-function e.g. Sinkhorn Discrepancy from optimal transport

Thank you!


Alistair Letcher


Stefan Woerner

… and you for your attention! Questions?

arXiv:2309.12681

IBM **Quantum**

ouf@zurich.ibm.com